XI-th International Conference

Knowledge-Dialogue-Solution

June 20-30, 2005, Varna (Bulgaria)

P R O C E E D I N G S

VOLUME 2

FOI-COMMERCE

SOFIA, 2005

Gladun V.P., Kr.K. Markov, A.F. Voloshin, Kr.M. Ivanova (editors)

Proceedings of the XI-th International Conference "Knowledge-Dialogue-Solution" – Varna, 2005
Volume 2

Sofia, FOI-COMMERCE – 2005

Volume 1 ISBN: 954-16-0032-8
Volume 2 ISBN: 954-16-0033-6
First Edition

The XI-th International Conference "Knowledge-Dialogue-Solution" (KDS 2005) continues the series of annual international KDS events organized by Association of Developers and Users of Intelligent Systems (ADUIS).

The conference is traditionally devoted to discussion of current research and applications regarding three basic directions of intelligent systems development: knowledge processing, natural language interface, and decision making.

Edited by :
Association of Developers and Users of Intelligent Systems, Ukraine
Institute of Information Theories and Applications FOI ITHEA, Bulgaria
Printed in Bulgaria by FOI ITHEA
Sofia-1090, P.O.Box 775, Bulgaria
e-mail: foi@nlcv.net
www.foibg.com

All Rights Reserved

© 2005 Viktor P. Gladun, Krassimir K. Markov, Alexander F. Voloshin, Krassimira M. Ivanova - Editors
© 2005 Krassimira Ivanova - Technical editor
© 2005 Association of Developers and Users of Intelligent Systems, Ukraine - Co-edition
© 2005 Institute of Information Theories and Applications FOI ITHEA, Bulgaria - Co-edition
© 2005 FOI-COMMERCE, Bulgaria - Publisher
© 2005 For all authors in the issue

Volume 1 ISBN: 954-16-0032-8
Volume 2 ISBN: 954-16-0033-6
C\o Jusautor, Sofia, 2005

PREFACE

The scientific Eleventh International Conference “Knowledge-Dialogue-Solution” took place in June, 20-30, 2005 in Varna, Bulgaria. These two volumes include the papers presented at this conference. Reports contained in the Proceedings correspond to the scientific trends, which are reflected in the Conference name.

The Conference continues the series of international scientific meetings, which were initiated more than fifteen years ago. It is organized owing to initiative of ADUIS - Association of Developers and Users of Intelligent Systems (Ukraine), Institute of Information Theories and Applications FOI ITHEA, (Bulgaria), and IJ ITA - International Journal on Information Theories and Applications, which have long-term experience of collaboration.

Now we can affirm that the international conferences “Knowledge-Dialogue-Solution” in a great degree contributed to preservation and development of the scientific potential in the East Europe.

The conference is traditionally devoted to discussion of current research and applications regarding three basic directions of intelligent systems development: knowledge processing, natural language interface, and decision making.

The basic approach, which characterizes presented investigations, consists in the preferential use of logical and linguistic models. This is one of the main approaches uniting investigations in Artificial Intelligence.

KDS 2005 topics of interest include, but are not limited to:

Cognitive Modelling	Knowledge Engineering
Data Mining and Knowledge Discovery	Logical Inference
Decision Making	Machine Learning
Informatization of Scientific Research	Multi-agent Structures and Systems
Intelligent NL Text Processing	Neural and Growing Networks
Intelligent Robots	Philosophy and Methodology of Informatics
Intelligent Technologies in Control and Design	Planning and Scheduling
Knowledge-based Society	Problems of Computer Intellectualization

The organization of the papers in KDS-2005 is based on specialized sessions. They are

1. Cognitive Modelling
2. Data Mining and Knowledge Discovery
3. Decision Making
4. Intelligent Technologies in Control, Design and Scientific Research
5. Mathematical Foundations of AI
6. Neural and Growing Networks
7. Philosophy and Methodology of Informatics

The official languages of the Conference are English and Russian. Sections are in alphabetical order. The sequence of the papers in the sections has been proposed by the corresponded chairs and is thematically based. The Program Committee recommends the accepted papers for free publishing in English in the International Journal on Information Theories and Applications (IJ ITA).

The Conference is sponsored by FOI Bulgaria (www.foibg.com).

We appreciate the contribution of the members of the KDS 2005 Program Committee.

On behalf of all the conference participants we would like to express our sincere thanks to everybody who helped to make conference success and especially to Kr.Ivanova, I.Mitov, N.Fesenko and V.Velichko.

V.P. Gladun, A.F. Voloshin, Kr.K. Markov

CONFERENCE ORGANIZERS

National Academy of Sciences of Ukraine
Association of Developers and Users of Intelligent Systems (Ukraine)
International Journal “Information Theories and Applications”
V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
Institute of Information Theories and Applications FOI ITHEA (Bulgaria)
Institute of Mathematics and Informatics, BAS (Bulgaria)
Institute of Mathematics of SD RAN (Russia)
New Technik Publishing Ltd. (Bulgaria)

PROGRAM COMMITTEE

Victor Gladun (Ukraine) – chair
Alexey Voloshin (Ukraine) - co-chair
Krassimir Markov (Bulgaria) - co-chair

Igor Arefiev (Russia)
Frank Brown (USA)
Alexander Eremeev (Russia)
Natalia Filatova (Russia)
Konstantin Gaindrik (Moldova)
Tatyana Gavrilova (Russia)
Vladimir Donskoy (Ukraine)
Krassimira Ivanova (Bulgaria)
Natalia Ivanova (Russia)
Vladimir Jotsov (Bulgaria)
Julia Kapitonova (Ukraine)
Vladimir Khoroshevsky (Russia)
Rumyana Kirkova (Bulgaria)
Nadezhda Kiselyova (Russia)
Alexander Kleshchev (Russia)
Valery Koval (Ukraine)
Oleg Kuznetsov (Russia)
Vladimir Lovitskii (GB)
Vitaliy Lozovskiy (Ukraine)
Ilia Mitov (Bulgaria)
Nadezhda Mishchenko (Ukraine)
Xenia Naidenova (Russia)
Olga Nevzorova (Russia)
Genady Osipov (Russia)
Alexander Palagin (Ukraine)
Vladimir Pasechnik (Ukraine)
Zinoviy Rabinovich (Ukraine)
Alexander Reznik (Ukraine)
Galina Rybina (Russia)
Vladimir Ryazanov (Russia)
Vasil Sgurev (Bulgaria)
Vladislav Shelepov (Ukraine)
Anatoly Shevchenko (Ukraine)
Ekaterina Solovyova (Ukraine)
Vadim Stefanuk (Russia)
Tatyana Taran (Ukraine)
Valery Tarasov (Russia)
Adil Timofeev (Russia)
Vadim Vagin (Russia)
Jury Valkman (Ukraine)
Neonila Vashchenko (Ukraine)
Stanislav Wrycza (Poland)
Nikolay Zagoruiko (Russia)
Larissa Zainutdinova (Russia)
Jury Zaichenko (Ukraine)
Arkady Zakrevskij (Belarus)

TABLE OF CONTENTS

VOLUME 1

Section 1. Cognitive Modelling

1.1. Conceptual Modelling of Thinking as Knowledge Processing during the Recognition and Solving the Problems

Концептуальное представление об опознании образов и решении проблем в памяти человека и возможностях его использования в искусственном интеллекте
З.Л. Рабинович 1

Новое содержание в старых понятиях: К пониманию механизмов мышления и сознания
Геннадий С. Воронков 9

Формирование нейронных элементов в обонятельной коре: обучение путем прорастания
Геннадий С. Воронков, Владимир А. Изотов 17

Mathematical and Computer Modelling and Research of Cognitive Processes in Human Brain. Part I. System Compositional Approach to Modelling and Research of Natural Hierarchical Neuron Networks. Development of Computer Tools
Yuriy A. Byelov, Sergiy V. Tkachuk, Roman V. Iamborak 23

Mathematical and Computer Modelling and Research of Cognitive Processes in Human Brain. Part II. Applying of Computer Toolbox to Modelling of Perception and Recognition of Mental Pattern by the Example of Odor Information Processing
Yuriy A. Byelov, Sergiy V. Tkachuk, Roman V. Iamborak 32

О моделировании образного мышления в компьютерных технологиях: общие закономерности мышления
Юрий Валькман, Вячеслав Быков 37

Модели биоритмов взаимодействия
Степан Г. Золкин 45

Section 2. Data Mining and Knowledge Discovery

2.1. Actual Problems of Data Mining

Автоматизация процессов построения онтологий
Николай Г. Загоруйко, Владимир Д. Гусев, Александр В. Завертайлов, Сергей П. Ковалёв, Андрей М. Налётов, Наталия В. Саломатина 53

Application of the Multivariate Prediction Method to Time Series
Tatyana Stupina, Gennady Lbov 60

К определению интеллектуального анализа данных
Ксения А. Найденова 67

The Development of the Generalization Algorithm based on the Rough Set Theory
M. Fomina, A. Kulikov, V. Vagin 76

Extreme Situations Prediction by Multidimensional Heterogeneous Time Series Using Logical Decision Functions
Svetlana Nedel'ko 84

Co-ordination of Probabilistic Expert's Statements and Sample Analysis in Recognition Problems
Tatyana Luchsheva 88

Evaluating Misclassification Probability Using Empirical Risk
Victor Nedel'ko 92

2.2. Structural-Predicate Models of Knowledge

SCIT — Ukrainian Supercomputer Project
Valeriy Koval, Sergey Ryabchun, Volodymyr Savyak, Ivan Sergienko, Anatoliy Yakuba 98

Discovery of New Knowledge in Structural-predicate Models of Knowledge
Valeriy N. Koval, Yuriy V. Kuk 104

Cluster Management Processes Organization and Handling
Valeriy Koval, Sergey Ryabchun, Volodymyr Savyak, Anatoliy Yakuba 112

Multi-agent User Behavior Monitoring System Based on Aglets SDK
Alexander Lobunets 119

2.3. Ontologies

Development of Educational Ontology for C-Programming
Sergey Sosnovsky, Tatiana Gavrilova 127

How Can Domain Ontologies Relate to One Another?
Alexander S. Kleshchev, Irene L. Artemjeva 132

Development of Procedures of Recognition of Objects with Usage Multisensor Ontology Controlled Instrumental Complex
Alexander Palagin, Victor Peretyatko 140

A concept of the Knowledge Bank on Computer Program Transformations
Margarita A. Knyazeva, Alexander S. Kleshchev 147

Implementation of Various Dialog Types Using an Ontology-based Approach to User Interface Development
Valeriya Gribova 153

Онтологии как перспективное направление интеллектуализации поиска информации в мульти-агентных системах е-коммерции
Анатолий Я. Гладун, Юлия В. Рогушина 158

Implementing Simulation Modules as Generic Components
Anton Kolotaev 165

Использование Semantic Web технологий при аннотировании программных компонентов
Михаил Рощин, Алла Заболеева-Зотова, Валерий Камаев 171

2.4. Computer Models of Common Sense Reasoning

DIAGaRa: An Incremental Algorithm for Inferring Implicative Rules from Examples (Part 1)
Xenia Naidenova 174

DIAGaRa: An Incremental Algorithm for Inferring Implicative Rules from Examples (Part 2)
Xenia Naidenova 182

Программные системы и технологии для интеллектуального анализа данных
Александр Е. Ермаков, Ксения А. Найденова 190

Модуль формирования таблиц соответствия измерительных шкал в подсистеме индуктивного вывода знаний проблемно-ориентированного инструментального средства
Александр Е. Ермаков, Вадим А. Ниткин 199

Section 3. Decision Making

3.1. Actual Problems of Decision Making

О проблемах принятия решений в социально-экономических системах
Алексей Ф. Волошин 205

Оптимальная траектория модели динамического межотраслевого баланса открытой экономики
Игорь Ляшенко, Елена Ляшенко 212

Нечеткие множества: Аксиома абстракции, статистическая интерпретация, наблюдения нечетких множеств
Владимир С. Донченко 218

Технология классификации электронных документов с использованием теории возмущения псевдообратных матриц
Владимир С. Донченко, Виктория Н. Омардибирова ... 223
Векторные равновесия во многокритериальных играх
Сергей Мащенко ... 226
Эволюционная кластеризация сложных объектов и процессов
Виталий Снитюк ... 232
Система качественного прогнозирования на основе нечетких данных и психографии экспертов
А.Ф. Волошин, В.М. Головня, М.В. Панченко ... 237
Процедуры локализации вектора весовых коэффициентов за обучающими выборками в задаче потребления
Елена В. Дробот ... 243
Нечеткие модели многокритериального коллективного выбора
Алексей Ф. Волошин, Николай Н. Маляр ... 247
Алгоритм последовательного анализа и отсеивания елементов в задаче определения медианы строгих ранжирований объектов
Павел П. Антосяк, Григорий Н. Гнатиенко ... 250
Один подход к модели теории инвестиционного анализа с учетом фактора нечеткости
Ольга В. Дьякова ... 253
Model of Active Monitoring
Sergey Mostovoi, Vasiliy Mostovoi ... 256
Towards the Problems of an Evaluation of Data Uncertainty in Decision Support Systems
Victor Krissilov, Daria Shabadash ... 262

3.2. Decision Support Systems

Применение квалиметрических моделей при решении социально-экономических задач
А. Крисилов, В. Степанов, И. Голяева, Б. Блюхер ... 265
Analogous Reasoning for Intelligent Decision Support Systems
A.P. Eremeev, P.R. Varshavsky ... 272
A MultIcriteria Decision Support System *MultiDecision-1*
Vassil Vassilev, Krasimira Genova, Mariyana Vassileva ... 279
Recognition on Finite Set of Events: Bayesian Analysis of Statistical Regularity and Classification Tree Pruning
Vladimir B. Berikov ... 286
Decision Forest versus Decision Tree
Vladimir Donskoy, Yuliya Dyulicheva ... 289
Generalized Scalarizing Problems *GENS* and *GENSLex* of Multicriteria Optimization
Mariyana Vassileva ... 297
Information System for Situational Design
T. Goyvaerts, A. Kuzemin, V. Levikin ... 305
Implementation of the System Approach in Designing Information System for Ensuring Ecological Safety of Mudflow and Creep Phenomenae
E. Petrov, A. Kuzemin, N.Gusar, D. Fastova, I. Starikova, O. Dytshenko ... 307
A Method of the Speaker Identification on Basis of the Individual Speech Code
M.F. Bondarenko, A.V. Rabotyagov, M.I. Sliptshenko ... 312
Mathematical Model for Situational Center Development Technology
V.M. Levykin ... 318

Index of Authors ... 319

VOLUME 2

Section 4. Intelligent Technologies in Control, Design and Scientific Research

4.1. Intelligent NL Processing

A Workbench for Document Processing
Karola Witschurke ... 321

Experiments in Detection and Correction of Russian Malapropisms by Means of the WEB
Elena I. Bolshakova, Igor A. Bolshakov, Aleksey P. Kotlyarov ... 328

Verbal Dialogue versus Written Dialogue
David Burns, Richard Fallon, Phil Lewis, Vladimir Lovitskii, Stuart Owen ... 336

Конспектирование естественноязыковых текстов
Виктор П. Гладун, Виталий Ю. Величко ... 344

О задаче семантического индексирования тематических текстов
Надежда Мищенко, Наталья Щеголева ... 347

Resolution of Functional Homonymy on the Basis of Contextual Rules for Russian Language
Olga Nevzorova, Julia Zin'kina, Nicolaj Pjatkin ... 351

Information Processing in a Cognitive Model of NLP
Velina Slavova, Alona Soschen, Luke Immes ... 355

4.2. Application of AI Methods for Prediction and Diagnostics

Application of Artificial Intelligence Methods to Computer Design of Inorganic Compounds
Nadezhda N. Kiselyova ... 364

К вопросу о развитии интерфейса «разработчик-заказчик»
Леонид Святогор ... 371

4.3. Planning and Sheduling

Two-machine Minimum-length Shop-Scheduling Problems with Uncertain Processing Times
Natalja Leshchenko, Yuri Sotskov ... 375

Learning Technology in the Scheduling Algorithm Based on the Mixed Graph Model
Yuri Sotskov, Nadezhda Sotskova, Leonid V. Rudoi ... 381

4.4. Intelligent Technologies in Control

Автоматный метод решения систем линейных ограничений в области {0,1}
Сергий Крывый, Людмила Матвеева, Виолета Гжывач ... 389

Logical Models of Composite Dynamic Objects Control
Vitaly J. Velichko, Victor P. Gladun, Gleb S. Gladun, Anastasiya V. Godunova, Yurii L. Ivaskiv, Elina V. Postol, Grigorii V. Jakemenko ... 395

The Information-analytical System for Diagnostics of Aircraft Navigation Units
Ilya Prokoshev, Vyacheslav Suminov ... 400

Динамические системы в описании нелинейных рекурсивных регрессионных преобразователей
Микола Ф. Кириченко, Владимир С. Донченко, Денис П. Сербаев ... 404

The Matrix Method of Determining the Fault Tolerance Degree of a Computer Network Topology
Sergey Krivoi, Miroslaw Hajder, Pawel Dymora, Miroslaw Mazurek ... 412

Robot Control Using Inductive, Deductive and Case Based Reasoning
Agris Nikitenko ... 418

Information Models for Robotics System with Intellectual Sensor and Self-organization
Valery Pisarenko, Ivan Varava, Julia Pisarenko, Viktoriya Prokopchuk ... 427

4.5. Intelligent Systems

Static and Dynamic Integrated Expert Systems: State of the Art, Problems and Trends
Galina Rybina, Victor Rybin 433

Adaptive Routing and Multi-Agent Control for Information Flows in IP-Networks
Adil Timofeev 442

The on-board Operative Advisory expert Systems for Anthropocentric Object
Boris E. Fedunov 446

Оптимизация телекоммуникационных сетей с технологией ATM
Леонид Л. Гуляницкий, Андрей А. Баклан 454

Testing AI in One Artificial World
Dimiter Dobrev 461

Concurrent Algorithm for Filtering Impulse Noise on Satellite Images
Nguyen Thanh Phuong 465

4.6. Macro-economical Modelling

Сравнительный анализ четкого и нечеткого методов индуктивного моделирования (МГУА) в задачах макроэкономического прогнозирования
Юрий П. Зайченко 473

Исследование нечеткой нейронной сети ANFIS в задачах макроэкономического прогнозирования
Юрий П. Зайченко, Фатма Севаее 479

Математическая модель реструктуризации сложных технико-экономических структур
Май Корнийчук, Инна Совтус, Евгений Цареградский 486

Section 5. Mathematical Foundations of AI

5.1. Algorithms

Raising Efficiency of Combinatorial Algorithms by Randomized Parallelization
Arkadij D. Zakrevskij 491

Specifying Agent Interaction Protocols with Parallel Control Algorithms
Dmitry Cheremisinov, Liudmila Cheremisinova 496

Об одной модификации TSS-алгоритма
Руслан А. Багрий 504

The Development of Parallel Resolution Algorithms Using the Graph Representation
Andrey Averin, Vadim Vagin 509

Магнитная гидродинамика жидкости и динамика упругих тел: моделирование в среде Mathematica
Ю.Г. Лега, В.В. Мельник, Т.И. Бурцева, А.Н. Папуша 517

Some Approaches to Distributed Encoding of Sequences
Artem Sokolov, Dmitri Rachkovskij 522

5.2. Modal Logic

Representing the Closed World Assumption in Modal Logic
Frank M. Brown 529

Representing Skeptical Logics in Modal Logic
Frank M. Brown 537

Automatic Fixed-point Deduction Systems for Five Different Propositional NonMonotonic Logics
Frank M. Brown 545

Nonmonotonic Systems Based on Smallest and Minimal Worlds Represented in World Logic, Modal Logic, and Second Order Logic
Frank M. Brown 553

Z Priorian Modal Second Order Logic
Frank M. Brown 560

Section 6. Neural and Growing Networks

6.1. Neural Network Applications

Parallel Markovian Approach to the Problem of Cloud Mask Extraction
Natalia Kussul, Andriy Shelestov, Nguyen Thanh Phuong, Michael Korbakov, Alexey Kravchenko 567

Идентификация нейросетевой модели поведения пользователей компьютерных систем
Н. Куссуль, С. Скакун 570

Jamming Cancellation Based on a Stable LSP Solution
Elena Revunova, Dmitri Rachkovskij 578

Graph Representation of Modular Neural Networks
Michael Kussul, Alla Galinskaya 584

Гетерогенные полиномиальные нейронные сети для распознавания образов и диагностики состояний
Адиль В.Тимофеев 591

Neuronal Networks for Modelling of Large Social Systems. Approaches for Mentality, Anticipating and Multivaluedness Accounting.
Alexander Makarenko 600

6.2. Neural Network Models

Представление нейронных сетей динамическими системами
Владимир С. Донченко, Денис П. Сербаев 605

Generalization by Computation Through Memory
Petro Gopych 608

Neural Network Based Approach for Developing the Enterprise Strategy
Todorka Kovacheva, Daniela Toshkova 616

Neuro-Fuzzy Kolmogorov's Network with a Hybrid Learning Algorithm
Yevgeniy Bodyanskiy, Yevgen Gorshkov, Vitaliy Kolodyazhniy 622

Нейросетевая классификация земного покрова на основании спектральных измерений
Алла Лавренюк, Лилия Гнибеда, Екатерина Яровая 627

Section 7. Philosophy and Methodology of Informatics

7.1. Knowledge Market

The Staple Commodities of the Knowledge Market
Krassimir Markov, Krassimira Ivanova, Ilia Mitov 631

Basic Interactions between Members of the Knowledge Market
Krassimira Ivanova, Natalia Ivanova, Andrey Danilov, Ilia Mitov, Krassimir Markov 638

7.2. Information Theories

Ценность информации
Андрей Данилов 649

The Main Question of the Informatics, 100 Years after its Poseing
Stoyan Poryazov 655

Objects, Functions and Signs
Stoyan Poryazov 656

7.3. The Intangible World

Approaching the Noosphere of Intangible – Esoteric from Materialistic Viewpoint
Vitaliy Lozovskiy 657

Informatics, Psychology, Spiritual Life
Larissa A. Kuzemina 669

Information Support of Passionaries
Alexander Ya. Kuzemin 670

Index of Authors 675

Section 4. Intelligent Technologies in Control, Design and Scientific Research

4.1. Intelligent NL Processing

A WORKBENCH FOR DOCUMENT PROCESSING

Karola Witschurke

***Abstract**: During the MEMORIAL project an international consortium has developed a software solution called DDW (Digital Document Workbench). It provides a set of tools to support the process of digitization of documents from the scanning up to the retrievable presentation of the content. The attention is focused to machine typed archival documents. One of the important features is the evaluation of quality in each step of the process. The workbench consists of automatic parts as well as of parts which request human activity. The measurable improvement of 20% shows the approach is successful.*

***Keywords**: Document Management, Digital Document Workbench, Image Processing, OCR, Machine Typed Document.*

Introduction

Several organizations have joined forces to solve a task of common interest.

A strategic goal of the international consortium undertaking this project was to support the creation of virtual archives based on documents which exist in libraries, archives, museums, memorials and public record offices, in order to allow computer aided information retrieval.

Machine type written documents have proved as especially hard to process.

Such documents may constitute less or more complex printed forms mixing printed and typed text, graphics, as well as hand written annotations, signatures, rubber stamps and photographs. Moreover, the color of typed characters may vary; characters may be overstricken, shifted up or down, or due to torn out or dried ribbon only partially typed. A special challenge are documents which represent a carbon copy instead of an original one, Finally, due to physical conditions a document may contain stained or damaged parts.

1. Approach to the Project

OCR systems may be used to extract the textual information from a document image.

OCR is used to working on binary images and assigns groups of black pixels to patterns of characters. The results of state-of-the-art OCR systems are satisfying if fresh printed office documents are processed. Historical documents in general look bad for different reasons. The problem is to get good looking binary images from bad looking colored original ones by filtering noise (stains, wrinklings, torns) out of the image and improving the shapes of characters. Due to the unsteady quality of a page it is not satisfactory to use the same threshold of binarisation (a parameter between 1 and 256 used for binarisation of a gray scale image and causes the assignment of pixel to be “white” or “black”) for the whole page. For those reason in the MEMORIAL-project a semantic driven approach was preferred. A special editor supports the user to draw regions of interest and mark

other regions with damages or illegible parts. After that a partial image improvement and background clearing using the color information is possible. Thus, the binarisation with the best possible threshold is processed adaptively for the different regions down to single characters. The subsequent figures are illustrating the advantage of this approach.

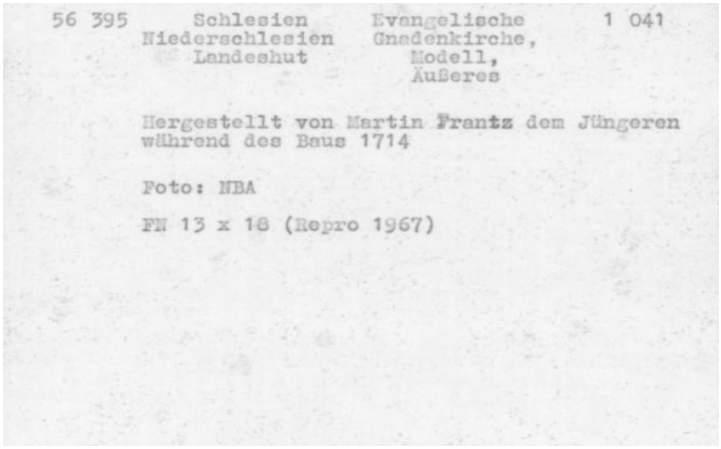


Figure 1 – inventory card from Herder-Institute Marburg (Germany)

The often automatically chosen binarsation threshold of 128 in this case provides a nearly blank image caused by the weakness of the typed characters. Figure 2 compares three manually chosen thresholds to the adaptively applied threshold of DDW. The marked region in figure 3 shows the advance of DDW vs. the best manual binarisation. In this case, the OCR is getting the best possible input.

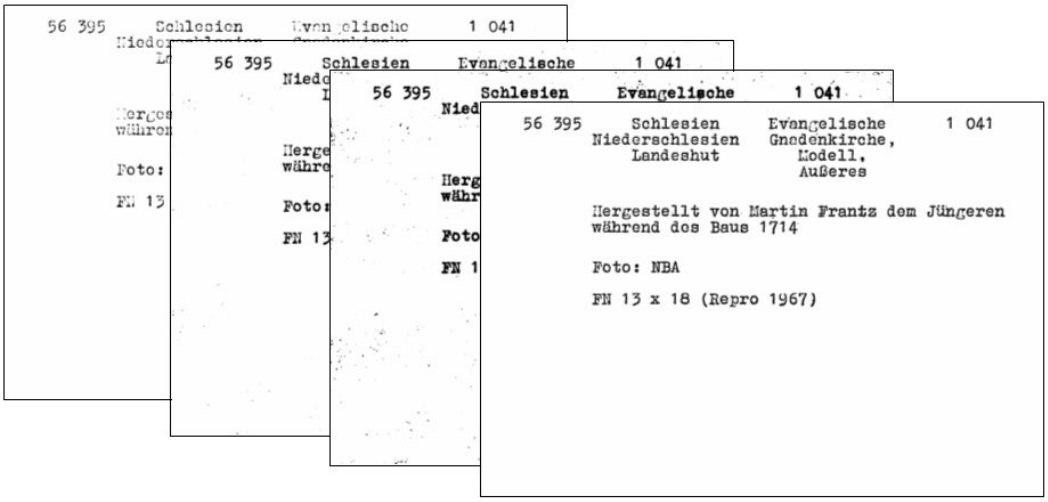


Figure 2 – comparisation of threshold (from left) 180, 200, 210 with adaptive threshold used by DDW (right)

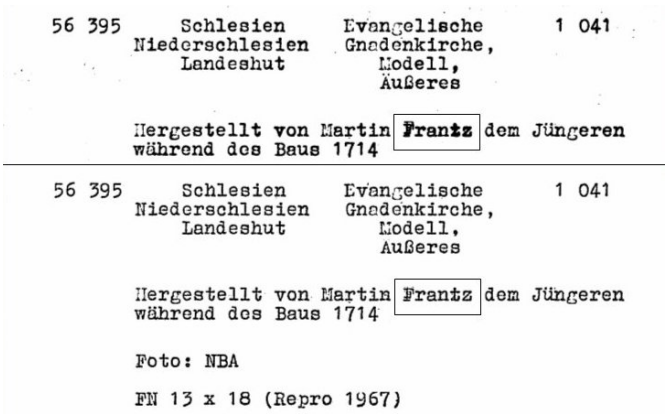


Figure 3 – best manual binarisation (above) versus adaptive thresholding

2. The DDLC Model

The transformation of a historical paper document into its electronic parallel is a complex multi-phased engineering process interpreted as a document life cycle. In this effect, the project consortium has developed a *Digital Document Workbench (DDW)* toolkit supporting a *Digital Document Life-Cycle Development (DDLC)* model.

The first phase of the DDLC model is *digitization*, which provides a raw digital image of a paper original. Depending on the constitution of the documents this process may be done manually or semiautomatically. The quality in this phase immediately influences the quality of the following phases. Furthermore the naming of the image files happens in this phase (mostly by the scanner) – each generated file must have a unique name to avoid processing duplicates or overwriting files during their processing later. A document *Repository Management Tool (RMT)* of the DDW toolkit has been developed to help the archivist in the namespace management and to store images in a database.

The second phase of DDLC is *qualification.* An expert user should attentively classify documents by building groups of similar in structure and meaning documents. Documents within the same semantic class can be processed together throughout the rest of the cycle. The output of this phase is a XML structure called *document template*, which contains the formal description of a semantic class.

The third phase of DDLC is *segmentation*, where the identification of major components (regions) of a document page image happens. The document template is used to control the segmentation phase, where the raw document image is cleaned and improved by the *Image Processing Tool (IPT)* which transforms original (colored) TIFF files into binarised clean images. The output of this phase is a *document content* XML file, which represents an interface for the next steps.

The fourth phase of DDLC is *extraction*, a key phase of the DDLC. Here the clean document image is processed by OCR; i.e. the textual information of an image is transformed into computer text. In this phase, the *document content* XML file is filled with the results of OCR.

The following *acceptance* phase allows the user to decide whether the recognized text is suitable to be introduced into the target database (digital archive). Corrections should be done to deliver the essential quality for the subsequent exploitation phase. This effort is supported by the content editor *Generator of Electronic Documents (GED)*. The multivalent browser *Viewer of Electronic Documents (VED)* however enables addition of notifications. This might be required to improve the quality of the results of queries to the target database.

The graphical representation of the DDLC model is shown in figure 4.

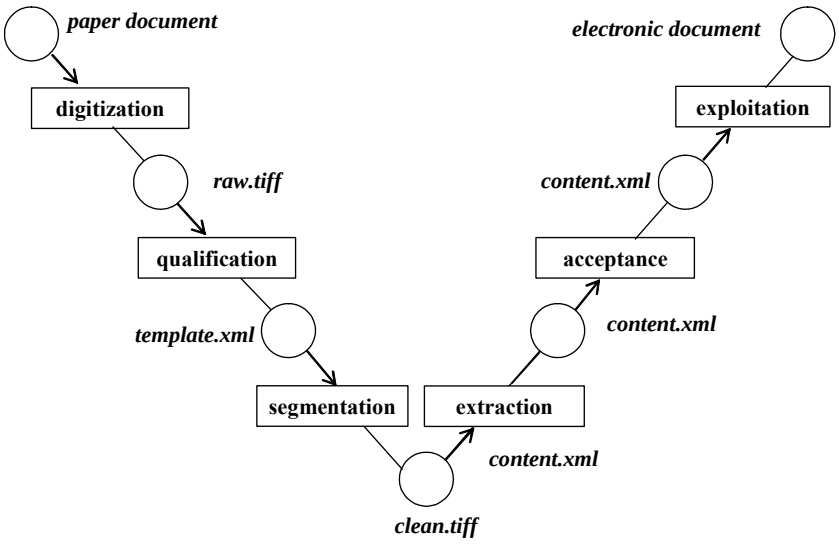


Figure 4 - DDLC model

The similarity of the DDLC model of document engineering to the well-known V model of software engineering is intended. The left arm of the DDLC model like in the V model, represents analysis of information aided by the user, whose domain knowledge is gradually being transformed into a control structure of processes

for engineering the final product, represented by the right arm. Similarly, in either model, verification of partial products of respective phases of the cycle plays a key role in assuring the quality of the final product. DDW tools enable a great deal of flexibility in extracting content of scanned paper documents into electronic documents. A sample representative subset of documents constituting a class can be processed in a *semiautomatic mode* with DDW tools to find the best settings of parameters for each DDLC phase, and next the remaining documents of the same class can be processed automatically in a batch mode, with the same settings. This approach introduced by the consortium enables fine-tuning of a document content extraction processes and quality management throughout the entire cycle.

3. Quality Management

In order to assure the best possible quality in each step of the DDLC model a quality management has been established. Therefore, human expertise is requested. The QED Tool of the DDW supports the fine-tuning of the process displaying the parameters and metrics on the one hand and setting up weights on the other hand. The interface for quality data exchange is the *qed.xml* file which is stored in the working repository (see figure 6 and table 1).

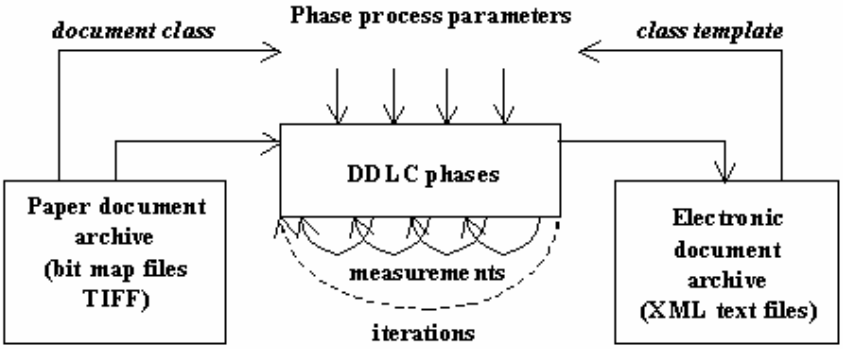


Figure 5 – Quality assessment during DDLC

The figure indicates the relationship between parameters influencing the phase processing and the measurements influencing the acceptance of the result. In dependence on the acceptance, a phase may rerun with tuned parameters.

When the quality of the paper document is *Q(PD)* and quality of the electronic counterpart is *Q(ED)* then three relations are possible.

Q(PD) > *Q(ED)*, the final product quality has deteriorated during processing along DDLC phases;

Q(PD) ≅ *Q(ED)*, the final document quality has not significantly changed compared to the original;

Q(PD) < *Q(ED)*, the final document quality has been improved during processing.

The first case indicates incorrect parameter settings. The achieved electronic document is unacceptable. The second case indicates correct but not optimal parameter settings. The third case is the desirable one - indicating an increasing quality during processing along DDLC. It is possible only when an expert user has been able to successfully contribute to the document engineering processes.

A typical situation observed during the project is the quality improvement combined with a significant reduction of time and effort in editing a document during acceptance phase, compared to manual reproduction of a document from scratch. Document quality assessment in any DDLC phase uses a specially developed Visual GQM (VGQM) method [9].

The VGQM method distinguishes between parameters and metrics. Parameters characterize processes of each phase, and their values may be used to control component DDW tools. The values of metrics, specific to each phase, are measured to characterize the respective input and output data. The value of Q is calculated based on a quality tree and normalized to a five-grade scale, from very low, through low and medium, up to high, and very high quality. Once all optimal settings for each phase are established by the quality expert, the processing of

the remaining documents of the class can be performed automatically in a batch by an archivist. Any document that cannot pass the quality threshold set up by an expert may now be rejected. Thorough selection of acceptance criteria for each respective class implies that either document processing progresses to the next phase, or is of such a poor quality that it must be processed manually (retyped).

4. Digital Document Workbench

DDW is a set of tools which aids the user in transferring typed content of paper documents into a digital archive. .

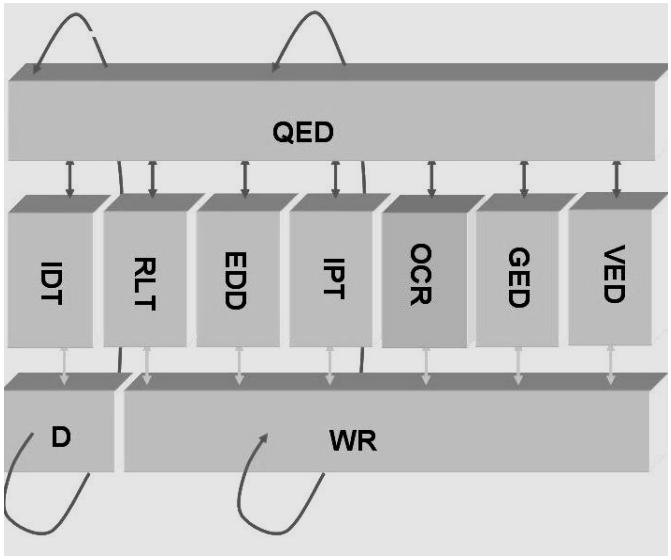


Figure 6 – Architecture of the DDW Toolset

Table 1 – DDW components

Acronym	*Full name*	*Functionality*
IDT	**I**n**D**exing **T**ool	Supportes name management, storage management and meta data description
RLT	**R**epository **L**oading **T**ool	Stores metadata and links to the image files as well as automatically generated jpeg-files and thumbnails in the working repository.
EDD	**E**lectronic **D**ocument class template e**D**itor	Creates and edits a template.xml file to describe a document class, allows to add (similar) documents to the class
IPT	**I**mage **P**rocessing **T**ool	Performs background cleaning and character improvement, creates a content.xml file
OCR	**O**ptical **C**haracter **R**ecognition tool	Separates the information relevant for OCR from the content.xml, runs the OCR (of-the-shelf product) and returns the results of OCR to the content.xml
GED	**G**enerator of **E**lectronic **D**ocuments	Enables editing badly recognized text or (in a extreme case) retyping a document
VED	**V**iewer of **E**lectronic **D**ocuments	Allows the archivist to browse layers of the electronic document with lenses.
QED	**D**ocument **Q**uality **E**valuation Tool	Supports measurement and evaluation of the quality of each respective DDLC phase. If a quality level is not satisfactory, the whole process can be repeated with changed settings.
WR	**W**orking **R**epository	Is an internal DDW database for storing project data

Any realistic use of DDW is possible, if the following assumptions are valid:
- all paper documents are type written;
- originals may be yellowed, dirty and otherwise bad looking;
- many documents which match the same layout may be found in the paper archive;
- a target digital archive should contain machine readable text of the documents.

The "backbone" of the DDW is a MS-SQL database (of-the-shelf product) called a Working Repository (**WR**). It contains information on the entire lifecycle of each document. Component tools post their results into the working repository, this way information exchange is guaranteed between all tools. Figure 6 gives summary of the several tools belonging to the DDW and the ways of information exchangeDDW toolkit can be used in two possible configurations:

- stand alone, without any connection to the working repository, with all relevant files stored directly in a common file system, useful when processing just a few documents in manual mode. In this case, the user has to control the file system.
- connected to the working repository (DDW database), providing a better control on intermediary document forms in between DDLC phases, in particular when operating DDW in a batch mode.

5. Final Remarks

One of the key advances of DDW is the semantic driven image processing. The success of the approach to handle color images of documents by developed image processing tools (IPT) is shown in the following impressive graphic. It maps the confidence rate of 50 documents (register cards of Herder - Institute Marburg) depending on the threshold of binarisation. In general, OCR systems choose t =128 as threshold automatically. It can also be chosen interactively (by testing the documents and looking for the best results, here: t = 189). In both cases, the chosen threshold then holds for all documents as a whole. DDW determines the threshold value individually for each character. Figure 7 shows the enhancement of quality processing 50 different inventory cards as displayed in figure 1 with fixed thresholds vs. with an adaptively determined threshold (upper curve, resp. right column). The obtained average value of ca. 80 is not yet satisfying for the majority of archivists, but advanced methods will increase the results in further projects.

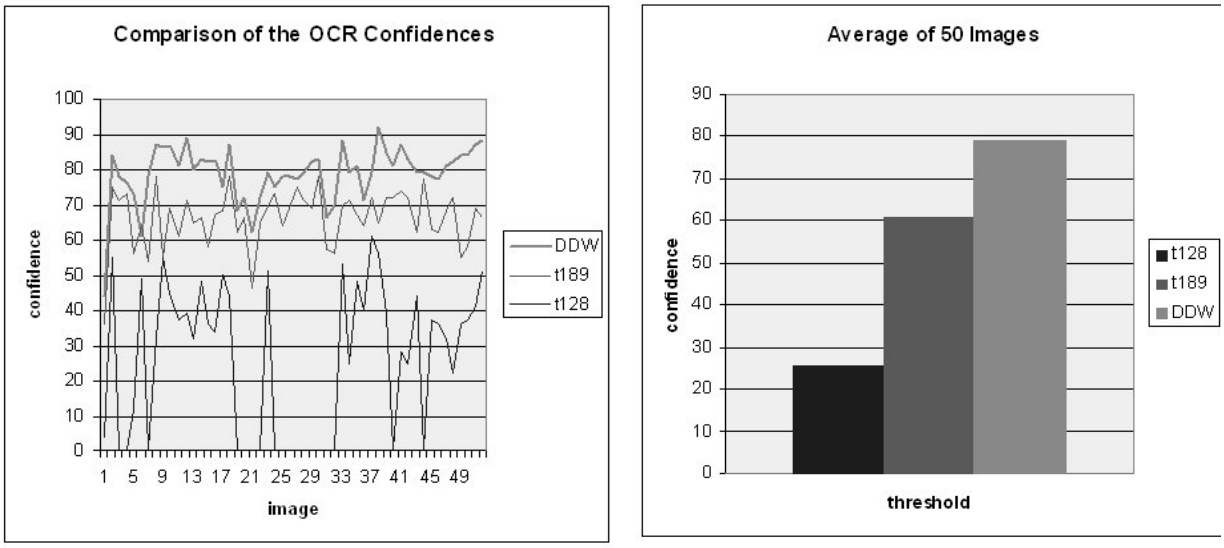


Figure 7 – Influence of binarisation parameters on the OCR results

The DDW toolkit now is suitable for its exploitation as a beta-release product. During two workshops archivists could act as model users and test DDW with a selected set of documents.

Acknowledgements

The MEMORIAL project was funded by the European Commission within the Information Society Technologies (IST) Program from 2002-02-01 to 2004-10-31. The consortium implementing the goal statements was constituted by GfaI (Society for the Promotion of Applied Computer Science)– Germany, Moreshet Holocaust Study and Research Center– Israel, National Museum Stutthof in Sztutowo - Poland, Technical University of Gdańsk – Poland, University of Liverpool – Great Britain.

Bibliography

[1] Antonacopoulos, A., Karatzas, D., Krawczyk, H., Wiszniewski, B.; The Lifecycle of a Digital Historical Document: Structure and Content.; ACM symposium on Document Engineering, Milwaukee, USA, October 28-30, 2004

[2] A. Antonacopoulos, D. Karatzas; A Complete Approach to the conversion of typewritten Historical Documents for Digital Archives; Sixth IAPR International Workshop on Document Analysis Systems, Florence, Italy, 8-10 September 2004

[3] Szwoch M., Szwoch W.; Preprocessing and Segmentation of Bad Quality Machine Typed Paper Documents; Sixth IAPR International Workshop on Document Analysis Systems, Florence, Italy, 8-10 September 2004

[4] Apostolos Antonacopoulos; Document Image Analysis for World War II Personal Records; International Workshop on Document Image Analysis for Libraries, Palo Alto, Jan 2004

[5] Wolfgang Schade, Karola Witschurke, Cornelia Rataj; Improved character recognition of typed documents from middling and lower quality based on application depending tools Processes, results, comparison; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003

[6] Alexander Geschke, Eva Fischer; Memorial Project - A complex approach to digitisation of personal records; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003

[7] Henryk Krawczyk, Bogdan Wiszniewski; Definition gleichartiger Dokumententypen zur Verbesserung der Erkennbarkeit und ihre XML-Beschreibung; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003

[8] Henryk Krawczyk, Bogdan Wiszniewski; Digital Document Life Cycle Development; International Symposium on Information and Communication Technologies, ISICT 2003, Dublin, Ireland

[9] Henryk Krawczyk, Bogdan Wiszniewski; Visual GQM approach to quality-driven development of electronic documents; Second International Workshop on Web Document Analysis, WDA2003, Edinburgh, UK

[10] Bogdan Wiszniewski; Projekt IST-2001-33441-MEMORIAL: Zestaw narzędziowy do tworzenia dokumentów cyfrowych z zapisów osobowych; I Krajowa Konferencja Technologii Informacyjnych 2003 TUG

[11] Alexander Geschke; MEMORIAL Project Overview; Proc. EVA Harvard, Symposium about Collaboration of Europe, Israel and USA, Harvard Library 1-2.10.2003

[12] Jacek Lebiedź, Arkadiusz Podgórski, Mariusz Szwoch; Quality Evaluation Of Computer Aided Information Retrieval From Machine Typed Paper Documents; Third conference on Computer Recognition Systems KOSYR'2003

[13] Witold Malina, Bogdan Wiszniewski; Multimedialne biblioteki cyfrowe; Sesja 50-lecia WETI-PG

[14] Dr. Alexander Geschke, Dr. Wolfgang Schade; The EU Project Memorial - Digitisation, Access, Preservation; Electronic Imaging Events in the Visual Arts - EVA 2002, Berlin 2002

[15] S. Rogerson, B. Wiszniewski; Legislation and regulation: emphasis on European approach to Data Protection, Human Rights, Freedom of Information, Intellectual Property, and Computer Abuse; PROFESSIONALISM IN SOFTWARE ENGINEERING PSE'03

Author's Information

Karola Witschurke - GFaI, Rudower Chaussee 30, 12489 Berlin, Germany; e-mail: witschurke@gfai.de

EXPERIMENTS IN DETECTION AND CORRECTION OF RUSSIAN MALAPROPISMS BY MEANS OF THE WEB

Elena I. Bolshakova, Igor A. Bolshakov, Aleksey P. Kotlyarov

***Abstract**: Malapropism is a semantic error that is hardly detectable because it usually retains syntactical links between words in the sentence but replaces one content word by a similar word with quite different meaning. A method of automatic detection of malapropisms is described, based on Web statistics and a specially defined Semantic Compatibility Index (SCI). For correction of the detected errors, special dictionaries and heuristic rules are proposed, which retains only a few highly SCI-ranked correction candidates for the user's selection. Experiments on Web-assisted detection and correction of Russian malapropisms are reported, demonstrating efficacy of the described method.*

***Keywords:** semantic error, malapropism, error correction, Web-assisted error detection, paronymy dictionaries, correction candidates, Semantic Compatibility Index.*

Introduction

Modern computer text editors and spellers readily detect spelling errors and some syntactic errors, primarily, mistakes in word agreement. Step by step, editing facilities of computers are being extended, in particular, by taking into account specificity of the particular text style and genre [4]. The topical problem is now semantic mistakes, which are hardly detectable because they violate neither orthography nor grammar of the text.

Malapropism is a particular type of semantic mistakes, which replace one content word by another similar word. The latter has the same morpho-syntactic form but different meaning, which is inappropriate in the given context, e.g., *animal word* or *massy migration* given instead of *animal world* and *massive migration*. To correct such mistakes, computer procedures are required that reveal erroneous words and supply the user (human editor) with selected candidates for their correction. However, only few papers (cf. [3, 5]) are devoted to the problem of malapropism detection and correction.

A method for malapropism detection proposed in [5] relies on recognition in the text of words (mainly nouns) distant from all contextual ones in terms of WordNet semantic relations (of synonymy, hyponymy, hyperonymy, etc.). Syntactic relations between words are ignored, and words from different sentences or even paragraphs are analyzed.

In the paper [3] malapropism detection is based on syntactico-semantic relations between content words, thereby much smaller context – only one sentence – is needed for error detection. Specifically, sentences are considered consisting of syntactically related and semantically compatible combinations of content word, the so-called *collocations*. It is presumed that malapropisms destroy collocations they are in: they violate semantic compatibility of word combinations while retaining their syntactic correctness.

In order to detect errors in a sentence, all pairs of syntactically linked content words in it are verified as collocations: their semantic compatibility is tested. Words of four principal parts of speech (POS) – nouns, verbs, adjective, and adverb – are considered as collocation components. To test whether a word pair is a collocation, three types of linguistic resources are proposed: a precompiled collocation database like CrossLexica [1], a text corpus, or a Web search engine like Google or Yandex.

This paper develops the latter method on the basis of experiments with Yandex as a resource for collocation testing. The Web is widely considered now as a huge, but noisy linguistic resource [6]. For the Web, it proved necessary to revise heuristic rules for malapropism detection and correction.

Following [3] we consider only malapropisms that destroy collocations. A malapropism is detected if a pair of syntactically linked content words in a sentence exhibits the value of a specially defined *Semantic Compatibility Index* (SCI) lower than a predetermined threshold. Below we call malapropism the whole pair detected as erroneous.

For malapropism correction, in contrast with the blind search of editing variants used in [3, 5], we propose to use beforehand compiled dictionaries of paronyms, i.e. words differing in some letters or in some morphs. The dictionaries provide all possible candidates for correction of a malapropos word, and the candidates are then tested in order to select several highly SCI-ranked correction candidates for ultimate decision by the user.

The proposed method was examined on two sets of Russian malapropisms. The first set of a hundred of samples was used to adjust heuristic threshold values, whereas the second justified these values. Since collocation components (hereafter *collocatives*) may be adjacent or separated by other words in a sentence, in the experiment we took into account the most probable distances between collocatives, which were previously determined through Yandex statistics.

Dictionaries of Paronyms

For correction of malapropos words, quick search of similar words are required. Words similar in letters, sounds or morphs, are usually called paronyms. In any language, only a limited portion of words has paronyms, and paronymy groups are on an average rather small. Hence, it is reasonable to gather paronyms before their use.

For our purposes, we consider only ***literal*** (e.g., Eng. *pace* Vs. *pact*, or Rus. *краска* Vs. *каска*) and ***morphemic*** paronyms (e.g., Eng. *sensible* Vs. *sensitive*, or Rus. *человечный* Vs. *человеческий*). Russian paronyms of these two types were compiled in corresponding dictionaries, which were preliminary described in [2].

The dictionary of Russian literal paronyms consists of word groups. Each group includes an entry word and its one-letter paronyms. Such paronyms are obtained through applying to the entry word of an elementary editing operation: insertion of a letter in any position, omission of a letter, replacing of a letter by another one, and permutation of two adjacent letters. For example, Russian word *белка* has one-letter paronymy group {*булка*, *елка*, *телка*, *челка*, *щелка*}.

The paronymy groups include words of the same part of speech (POS). Moreover, nouns of singular number and nouns of plural number, as well as nouns for different genders of singular have separated groups. Similar division is done for personal and other forms of verbs. Such measure is necessary for retaining syntactic links between words in the sentence while correcting an erroneous word. For this purpose, we extract malapropos word from the text (e.g., Rus. *белкой*), reconstruct its dictionary morphological form (*белка*), take from the dictionary corresponding paronym (e.g., *булка*), change its morphological form (taking it the same as for sourse malapropos word), and, finally, replace the erroneous word by the resulted word (*булкой*).

An entry of dictionary of Russian morphemic paronyms presents a group of words of the same POS that have the same root morph but differ in auxiliary morphs (prefixes or suffixes), e.g. Rus. {*бегающий*, *беглый*, *беговой*, *бегущий*}.

By now, the developed dictionary of literal paronyms comprises 17,4 thousands of paronymy groups with the mean size 2,65, while the dictionary of morphemic paronyms contains 1310 groups with the mean size 7,1.

Method of Malapropism Detection and Correction

To facilitate understanding of key ideas of the method we should first clarify the notion of collocation adopted in the paper. Collocation is a combination of two syntactically linked and semantically compatible content words, such as the pair's *main goal* and *moved with grace*. Syntactic links are realized directly or through an auxiliary word (usually a preposition). If any of conditions indicated above does not hold, the corresponding word combination is not collocation, for example, *the forest*, *river slowly*, *boiling goal*.

There are several syntactic types of collocations in each language. The most frequent types in European languages are: "the modified word $\rightarrow$ its modifier"; "noun $\rightarrow$ its noun complement"; "verb $\rightarrow$ its noun complement"; "verb predicate $\rightarrow$ its subject"; and "adjective $\rightarrow$ its noun complement". Directed links reflect syntactic dependency "head $\rightarrow$ its dependent".

The most frequent types and subtypes of Russian collocations are given in Table 1. They are determined by POS of collocatives and their order in texts; **N** symbolizes noun, **Adj** is adjective or participle, **V** and **Adv** are verb and adverb correspondingly, and **Pr** is preposition. Subindex ***comp*** means noun complement, while subindex ***sub*** means the noun subject in nominative case. Subindex ***pred*** symbolizes specifically Russian predicative short form of adjectival.

Within a sentence, collocatives may be adjacent either distant from each other. The distribution of possible distances depends on the collocation type and specific collocatives. For example, collocatives of subtypes 2.1 and 2.2 are usually adjacent, whereas the 3.1-collocation such as *give* → *message* can contain intermediate contexts of lengths 0 to 4 and even longer, e.g. *give her a short personalized message*.

Table 1. Frequent types and structures of Russian collocations

Type title	Type code	Type structure	English example	Russian example
modified → its modifier	1.1 1.2	Adj ← N Adv ← Adj	*strong tea* *very good*	*крепкий чай* *очень хороший*
noun → its noun complement	2.1 2.2	N → N_{comp} N→ Prep → N_{comp}	n/a *signs of life*	*огни города* *вызов в суд*
verb → its noun complement	3.1 3.2 3.3	V → N_{comp} V → Prep → N_{comp} N_{comp} ← V	*give message* *go to cinema* n/a	*искать решение* *идти в кино* *здание затушили*
verb predicate → its subject	4.1 4.2 4.3 4.4	N_{sub} ← V V → N_{sub} Adj_{pred} → N_{sub} N_{sub} ← Adj_{pred}	*light failed* (*there*) *exist people* n/a n/a	*мальчик пел* *пропали письма* *отправлен груз* *порт открыт*
adjective → its noun complement	5.1 5.3	Adj → Prep → N_{comp} Adj → N_{comp}	*easy for girls* n/a	*красный от стыда* *занятый трудом*

Our definition of collocations ignores their frequencies and idiomatically. As for frequencies, the advance of the Web shows that any semantically compatible word combination eventually realizes several times, thus we can consider as collocations all those exceeding a rather low threshold.

The main idea of our method of malapropism detection is to look through all pairs of content words within the sentence under revision, testing its syntactic links and its semantic admissibility. If the pair (*V, W*) is syntactically connected but semantically incompatible, a malapropism is signaled.

When a malapropism is detected, it is not known which collocative is erroneous, so we should try to correct both of them. The situation is clarified in Fig. 1. The upper two collocative nodes form the malapropism. The nodes going left-and-down and right-and-down are corresponding paronyms for malapropism's nodes. Each paronym should be matched against the opposite malapropism's node, and any pair may be admissible, but only one combination corresponds to the intended collocation; we call it *true correction*.

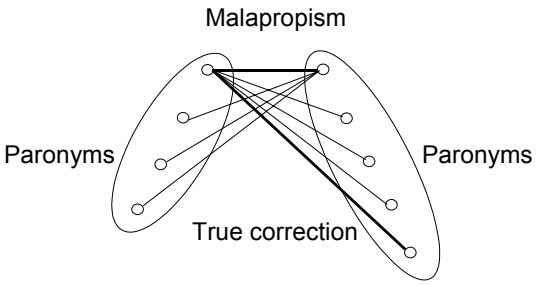


Fig. 1. Correction candidates and true correction

In such a way, all possible pairs of a collocative and its counterpart's paronym are formed, and we call them *primary candidates* for correction. The candidates are tested on semantic compatibility. If a pair fails, it is discarded; otherwise it is included into a list of *secondary candidates*. Then this list is ranked and only the best candidates are kept.

Obviously, for testing pairs (*V, W*) on semantic compatibility, using the Web as a text corpus, a statistical criterion is needed. According to one criterion, the pair is compatible if the relative frequency *N*(*V,W*)/*S* of the

co-occurrence of its words in a short distance in the whole corpus is greater than the product of relative frequencies $N(V)/S$ and $N(W)/S$ of occurrences of V and W taken separately (N means frequency; S is the size of the corpus). Using logarithms, we have the following threshold rule of pair compatibility:

$$\mathrm{MII}(V, W) \equiv \ln(N(V, W)) + \ln(S) - \ln(N(V)) - \ln(N(W) > 0,$$

where $\mathrm{MII}(V, W)$ is the mutual information index [7].

Since any search engine automatically delivers statistics about the queried word or the word combination measured in numbers of pages, to heuristically estimate the pair compatibility we propose a Semantic Compatibility Index (SCI) similar to MII:

$$\mathrm{SCI}(V,W) \equiv \begin{cases} \ln(P) + \ln(N(V,W)) - (\ln(N(V)) + \ln(N(W))) / 2, & \text{if } N(V,W) > 0, \\ NEG, & \text{if } N(V,W) = 0, \end{cases}$$

where N is the number of relevant pages; P is a positive constant to be chosen experimentally; and NEG is a negative constant. A merit of SCI as compared to MII is that the total number of pages is not to be estimated. Similarly to MII, SCI does not depend on monotonic or oscillating variations of all statistical data in the search engine because of the divisor 2.

If $\mathrm{SCI}(V_m,W_m) < 0$, the pair (V_m,W_m) is malapropism, whereas the primary candidate (V,W) is selected as a secondary one according to the following threshold rule:

$$(\mathrm{SCI}(V_m,W_m) = NEG) \textbf{ and } (\mathrm{SCI}(V,W) > Q) \quad \textbf{or} \quad (\mathrm{SCI}(V_m,W_m) > NEG) \textbf{ and } (\mathrm{SCI}(V,W) > \mathrm{SCI}(V_m,W_m)),$$

where Q ($NEG < Q < 0$) is a constant to be chosen experimentally.

The resulted set of secondary candidates is ranked by SCI values. The *best candidates* are all with positive SCI (let be n of them), whereas only one candidate with a negative SCI value is admitted, if n=1, and two candidates, if n=0.

Experimental Sets of Malapropisms

For experiments, we use two experimental sets – both of them consist of hundred of Russian sample malapropisms, which were mainly formed with the aid of the Web newswire. Specifically, we extracted collocations from the news messages, and one collocative of each collocation was then falsified using one of paronymy dictionaries, as a rule, the dictionary of literal paronyms (since literal errors are much more frequent in any language than morphemic ones). While falsifying, the morphological features of the word being changed (number, gender, person, case, etc.) were retained.

Then we again used paronymy dictionaries to make all possible correction candidates for each formed malapropism: through replacing of one word of the malapropism by its paronym we obtained the corresponding primary candidate. Among primary correction candidates, the before mentioned true correction (identical with intended collocation) was necessarily appeared.

The resulted sets consist of enumerated sample groups, each group corresponding to a malapropism and its primary candidates. Several sample groups are given in the first column of Fig. 2. Headlines of groups begin with the number of the changed collocative (1 or 2) and the symbol of the used paronymy dictionary (**L**iteral or **M**orphemic). The next is code $n_1.n_2$ of syntactic type of the collocation (cf. Table 1), and then goes the malapropism string, may be with a short context given in parentheses. Lines with correction candidates begin with the number of the changed word (1 or 2) and the symbol of the used paronymy dictionary; true corrections are marked with '**!!**'. The translation of malapropisms and their correction variants in the second column of Fig. 2. exhibits the nonsense of wrong corrections.

In total, the first malapropism set includes 648 primary correction candidates, and the second, 737 candidates. So the mean number of primary correction candidates is ≈ 7 candidates per error.

Among the samples, the sets include also errors named *quasi-malapropisms* (their total number equals 16). A quasi-malapropism transforms one collocation to another semantically legal collocation, which can be rarer and contradict to the outer context, e.g., *normal* ***manner*** changed to *normal* ***banner*** or *give* ***message*** changed to *give* ***massage***. An example of Russian *quasi-malapropism* is presented in Fig. 2, it is marked with '**!**' (cf. the second

sample group). The detection of quasi-malapropisms (if possible) sometimes permits one to restore the intended words, just as for malapropisms proper.

1)1L 1.1 (проявил) кассовое сознание	'cash consciousness'
1L массовое сознание	'mass consciousness'
1L!! классовое сознание	'class consciousness'
1L кастовое сознание	'caste consciousness'
2L кассовое создание	'cash creature'
2M кассовое знание	'cash knowledge'
2M кассовое признание	'cash confession'
2M кассовое осознание	'cash perception'
2L кассовое познание	'cash cognition'
2)2L! 1.3 (песня)явно сдалась	'evidently capitulated'
2L!! явно удалась	'evidently succeed'
2M явно задалась	'evidently preset'
2M явно далась	'evidently given'
2M явно продалась	'evidently sold'
1L ясно сдалась	'clearly capitulated'
2M явно подалась	'evidently gone'
3)1L 2.1 (занят)смирением террористов	'by submission of terrorists'
1L!! усмирением террористов	'by pacification of terrorists'
1M примирением террористов	'by reconciliation of terrorists'
4)1L 2.2 кастеты с кадрами	'knuckledusters with frames'
1L!! кассеты с кадрами	'cassettes with frames'
2L кастеты с карами	'knuckledusters with retributions'
2L кастеты с кедрами	'knuckledusters with cedars'
1L катеты с кадрами	'legs with frames'
5)2L 3.3 протокол подманили	'protocol is dangled'
2L!! протокол подменили	'protocol is replaced'
2L протокол поманили	'protocol is drown on'
2L протокол подранили	'protocol is injured'

Fig. 2. Several malapropisms and their correction candidates with translations

Experiments with Yandex and Their Results

A specific collocation or malapropism met in a text has its certain distance between collocatives. However, to reliably detect malapropisms by means of the Web and the selected statistic criterion, we should put each word pair being tested in its most probable distance.

For this reason, we initially explored frequencies of various Russian collocative co-occurrences against the distance between them on the base of Yandex statistics, cf. Table 2 and Table 3. The statistics of co-occurrence frequencies (measured in the number of relevant pages) were accumulated for twelve collocations of various frequent types. The used queries contained collocatives in quotation marks separated with */n* indicating the distance *n* between the given words, for example, *+"столбы"/2+"дыма"*. Such queries give frequencies of the words encountered within the same sentence with distance between them equal to *n* (or the number of intermediate words equals to *n-1*).

The statistics show that, for all collocations, frequency maximums correspond to the numbers 0 or 1 of intermediate words, and such cases cover more than 60% of encountered word pairs (cf. the last column of Table 2). Since we cannot determine automatically whether counted Web co-occurrences are real collocations or mere encounters of words without direct syntactic links between, we look through the first fifty page headers, mentally analyzing their syntax. Thereby we ascertained that the most of the co-occurrences with adjacent collocatives or those separated by one word are real collocations.

Thus, we can deduce that for frequent types of Russian collocations the most probable distance between collocatives (measured in the number of intermediate words) equals 0 or 1. As to collocatives linked through prepositions, the most probable distances at both interval are equal to 0, cf. Table 3. The table shows the distribution of frequencies for possible combinations of two distances: between the first collocative and the preposition and the preposition and the second collocative (e.g., combination 0-1 means that the first collocative and the preposition is adjacent, whereas the preposition and the second collocative are separated by one word).

Table 2. Yandex statistics of collocative co-occurrences

Collocation	Type	Number of intermediate words:					Percents in 0 and 1
		0	1	2	3	4	
Уделить внимание	3.1	52248	**72433**	9111	3537	1335	90%
Отправлен груз	4.3	779	**3408**	100	17	8	97%
Сбор информации	2.1	**141395**	32342	54354	31326	13566	64%
Спасатели обнаружили	4.1	**18534**	2440	929	524	740	91%
Здание потушили	3.3	**48**	7	14	10	0	70%
Сроки рассмотрения	2.1	**31517**	2918	2302	891	1075	89%
Затонувшее судно	1.1	**10250**	496	189	642	128	92%
Оценка деятельности	2.1	**29276**	22847	20373	5370	4183	64%
Занятый трудом	5.3	40	**413**	215	16	11	65%
Столбы дыма	2.1	**4382**	1420	507	79	93	90%
Сделать оговорки	3.1	355	**660**	269	44	14	76%
Приведем пример	3.1	**30665**	13106	6343	1376	580	84%

Table 3. Yandex statistics of co-occurrences for collocations with prepositions

Collocation	Number of intermediate words									
	0 –0	1–0	0–1	2–0	1–1	0–2	3–0	2–1	1–2	0–3
ворвались в здание	**9869**	123	156	69	0	74	24	1	9	2
тайники с оружием	**4775**	1	43	10	1	29	3	0	0	38
вызов в суд	**3633**	281	91	90	4	15	120	6	0	15
справиться с управлением	**5744**	16	177	2	0	11	2	0	0	12

```
1)1L 1.1 кассовое сознание       2, кассовое:354955,сознание:4770500
  1L массовое сознание           32973, массовое:916455
  1L!! классовое сознание        2927, классовое:38924
  1L кастовое сознание           56, кастовое:11799
  2L кассовое создание           10, создание:32199807
  2M кассовое знание             1, знание:7120311
  2M кассовое признание          0, признание:2437390
  2M кассовое осознание          0, осознание:823650
  2L кассовое познание           0, познание:605134
2)2L! 1.3 явно сдалась           13, явно:9871866, сдалась:198061
  2L!! явно удалась              6703, удалась:610646
  2M явно задалась               386, задалась:46599
  2M явно далась                 38, далась:88177
  2M явно продалась              2, продалась:24594
  1L ясно сдалась                2, ясно:10816398
  2M явно подалась               0, подалась:298216
3)1L 2.1 смирением террористов   0, смирением:79063,террористов:2762914
  1L!! усмирением террористов    3, усмирением:1787
  1M примирением террористов     0, примирением:17515
4)1L 2.2 кастеты с кадрами       0, кастеты:42266,кадрами:481878
  1L!! кассеты с кадрами         21, кассеты:2923258
  2L кастеты с карами            0, карами:19351
  2L кастеты с кедрами           0, кедрами:5666
  1L катеты с кадрами            0, катеты: 3151
5)2L 3.3 протокол подманили      0, протокол:7635243, подманили:3521
  2L!! протокол подменили        36, подменили:86957
  2L протокол поманили           0, поманили:7545
  2L протокол подранили          0, подранили:946
```

Fig. 3. Several malapropisms and their correction candidates with Yandex statistics

Then we have applied our method to the both experimental sets by means of the computer program that gathers statistics of word pairs co-occurrences with the distance between them (measured in the number of intermediate words) equal to 0 or 1 (collocatives linked through a preposition were tested as adjacent triples). This in no way means that collocations cannot have more distant collocatives, but the Web is not suited for collocation testing at greater distances. The frequencies of word occurrences and co-occurrences gathered for several malapropisms are given in the second column of Fig. 3 (the repeating data for the collocatives are omitted).

We used the first experimental set to adjust the necessary constants of our method. To obtain all negative SCI values for all proper malapropisms from the first set, we take $P = 1200$. The constant $NEG = -100$ is taken lower than SCI values of all occurrences counted as non-zero events. The constant $Q = -7.5$ is adjusted so that all candidates with non-zero occurrences have SCI values greater then this threshold.

Though all eight quasi-malapropisms were excluded while selecting the constant P, our method detects seven of them as malapropisms proper: their SCI values proved to be too low to be acknowledged as collocations. Our program selects 169 secondary candidates from 648 primary ones and then reduces them to 141 best correction candidates. Among the best candidates for the 99 malapropisms signaled, as many as 98 have true correction options, and only two of them are not first-ranked.

While testing the method and the determined constants on the second experimental malapropisms set, all its malapropisms and even all eight quasi-malapropisms were detected. 165 secondary candidates were selected from 737 primary ones; and the secondary ones were reduced to 138 best candidates. But for five detected malapropisms their true corrections do not enter corresponding lists of the best candidates, and three true corrections among them were not selected as secondary candidates (we admit that these collocations are rather infrequent in texts).

1)	1L 1.1 кассовое сознание	-6,29	Detected
	1L массовое сознание	2,95	Best
	1L!! классовое сознание	2,11	Best
2)	2L! 1.3 явно сдалась	-4,49	Detected
	2L!! явно удалась	1,20	Best
	2M явно задалась	-0,37	Best
3)	1L 2.1 смирением террористов	-100,00	Detected
	1L!! усмирением террористов	-2,96	Best
4)	1L 2.2 кастеты с кадрами	-100,00	Detected
	1L!! кассеты с кадрами	-6,28	Best
5)	2L 3.3 протокол подманили	-100,00	Detected
	2L!! протокол подменили	-2,93	Best

Fig. 4. Several malapropisms and the best candidates with their SCI values

We should note that the occasional omission of a true correction does not seem too dangerous, since the user can restore it in the case of error detection. Nevertheless, the most commonly used collocations among primary correction candidates always enter into the list of the best candidates, as true corrections or not.

For both experiments, the lists of the best candidates contain 1 to 4 entries, usually 1 or 2 entries; cf. the detected malapropisms with corresponding SCI values and decision qualifications in Fig. 4. The total decrease of correction candidates, from the primary to the best, exceeds 5.

Hence the results of our experiments are rather promising: for our experimental sets, the method of testing semantic compatibility through the Web has the recall 0.995. The proposed SCI is a quite good measure for detecting malapropisms, and the proposed heuristic rule for selection of secondary correction candidates is appropriate, whereas the heuristic rule for selection of best candidates may be slightly improved.

Conclusions and Further Work

A method is proposed for automatic detection and computer-aided correction of malapropisms. Experimental justification of the method was done on two representative sets of Russian malapropisms with the aid of Yandex search engine. While testing word pairs on their semantic compatibility through the Web, the most probable distances between the Russian words were taken into account.

Since the experiments gave good results, the problem of the Web statistics validity for collocation testing deserves to be investigated deeper. It would be worthwhile to extend the results of our study to broader experimental data and to other Web search engines. Of course, it is quite topical to develop a local grammar parser appropriate for malapropism detection, since for our experiments we extracted collocation components manually.

Since the Web proved to be adequate for testing semantic compatibility of collocations, we hope to use the method to develop procedures of automatic acquisition of collocation databases.

Bibliography

1. Bolshakov, I.A. Getting One's First Million…Collocations. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 5th Int. Conf. on Computational Linguistics CICLing-2004, Seoul, Korea, February 2004. LNCS 2945, Springer, 2004, p. 229-242.
2. Bolshakov, I.A., A. Gelbukh. Paronyms for Accelerated Correction of Semantic Errors. International Journal on Information Theories & Applications. V. 10, N 2, 2003, p. 198-204.
3. Bolshakov, I.A., A. Gelbukh. On Detection of Malapropisms by Multistage Collocation Testing. In: A. Düsterhöft, B. Talheim (Eds.) Proc. 8th Int. Conference on Applications of Natural Language to Information Systems NLDB´2003, June 2003, Burg, Germany, GI-Edition, LNI, V. P-29, Bonn, 2003, p. 28-41.
4. Bolshakova, E.I. Towards Computer-aided Editing of Scientific and Technical Texts. International Journal on Information Theories & Applications. V. 10, N 2, 2003, p. 204-210.
5. Hirst, G., D. St-Onge. Lexical Chains as Representation of Context for Detection and Corrections of Malapropisms. In: C. Fellbaum (ed.) WordNet: An Electronic Lexical Database. MIT Press, 1998, p. 305-332.
6. Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on the Web as Corpus. Computational linguistics, V. 29, No. 3, 2003, p. 333-347.
7. Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

Authors' Information

Elena I. Bolshakova – Moscow State Lomonosov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department, Docent; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: bolsh@cs.msu.su

Igor A. Bolshakov – Center for Computing Research (CIC), National Polytechnic Institute (IPN), Researching Professor; Av. Juan Dios Bátiz esq. Av. Miguel Othon Mendizabal s/n, U.P. Adolfo Lopez Mateos, Col. Zacatenco, C.P. 07738, Mexico D.F., Mexico; e-mail: igor@cic.ipn.mx

Aleksey P. Kotlyarov – Moscow State Lomonosov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department, Postgraduate student; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: koterpillar@yandex.ru

VERBAL DIALOGUE VERSUS WRITTEN DIALOGUE

David Burns, Richard Fallon, Phil Lewis, Vladimir Lovitskii, Stuart Owen

***Abstract**: Modern technology has moved on and completely changed the way that people can use the telephone or mobile to dialogue with information held on computers. Well developed "written speech analysis" does not work with "verbal speech". The main purpose of our article is, firstly, to highlights the problems and, secondly, to shows the possible ways to solve these problems.*

***Keywords**: data mining, speech recognition, natural language processing*

Introduction

There are several problems, which distinguish Verbal Dialogue (VD) from Written Dialogue (WD).

Problem 1. *Sensible VD restricts us to include in our reply to a user enquiry a very restricted number of alternative (3, 4, or maximum 5)* choices, because we should take into account our restricted ability to store information temporarily in our memory [1]. For example, assume the user is searching for *"Security Services Company in UK"*. Any information retrieval engine (e.g. Google) will assumes a maximal matching between enquiry words and documents. In the search result user almost immediately has got 14,600,000 documents and now it is the users' problem to identify the most appropriate one. In the case of VD such an approach is unacceptable even for the first 10 documents.

Problem 2. *Wrong recognition of Verbal Enquiry (VE)* (only *speaker independent* speech recognition is considered). As distinct from WD VD does not have any misspelling problem (we assume that grammar includes only correctly spelled words) because any VE will be converted into a sequence of "correct words". For example, a user said his postcode is: *"PL3 4PX "*. There are a minimum of 4 different possible results of recognition of this VE: (1) *"PL3 4PX "*; (2) *"BL3 4PX"*; (3) *"PL3 4BX"*; (4) *"BL3 4BX"*. The last two results: (3) and (4) represent non existing postcodes and Natural Response (NR) system [2] will simply ask user to repeat postcode. In the case of the second result *"BL3 4PX"* represents *existing postcode but wrong result*.

Problem 3. Barging (VD interruption) is absolutely natural style of VD. This is a fragment of VD:

NR: *Where would you like to go?*
User: *The best beaches, of course.*
NR: OK, *We can offer you Bahamas beach or Mia…*
User: *Bahamas, please.*

The solution of this problem is more technical than scientific regarding the natural language processing and therefore will not be considered in this paper.

Problem 4. TD verification and mining (TDVM). This problem is not just VD problem but is also a problem of WD. Real customer's DB's very often contain product names like: *"Wht thk slcd brd"* which is intended to mean "*White thick sliced bread"*, or *"Dietary spec gltn/fr & wht/frchoc/orge waf x5"*. A good TDVM analogy is to a body scanner. TDVM software can scan text and identify things that need to be looked at; it will show where there are anomalies in the TD. There is some specific feature of TDVM for VD: TDVM should provide selection of phonetically similar words to try to avoid the use of similar sounding words with different meaning.

Problem 5. Grammars creation. Speaker independent speech recognition has to be 'primed' around a particular vocabulary of terms and phrases, known as Grammars. To achieve reasonable recognition accuracy and response time multiple grammars should be generated automatically as a result of text data (TD) analysis. This overcomes a conflict that exists within speech recognisers that limits the effective vocabulary size without impairing the quality of recognition (normally, quality reduces as vocabulary size increases). Here we mean text data from an Application Domain (AD), which is represented by a database (DB). Let us distinguish two classes of grammars: independent and AD dependent grammars. Examples of independent grammars are *YesNo*, *Date*, *Time*, *Currency*, *CreditCard,* etc. Here we will discuss problem related only to AD dependent grammars creation. A grammar's "accuracy" depends on the solution of the previous problem.

VD: Alternative Choice Minimizing

The Alternative Choice Minimizing Mechanism (ACMM) is involved only when VD problem 1 arises i.e. when in reply to VE the number of "alternative choices" exceeds some pre-specified amount of permissible alternatives (APA). APA depends on the contents of AD values e.g. if values are represented by single words then APA might equal 5: *"What colour do you prefer blue, brown, white, magnolia or crème?"* If values are represented by long sequences of words APA would be better assigned as 3: *"We've got Mothers pride plain medium white bread 800g, Kingsmill square cut medium sliced white loaf 400g, or Hovis classic cut medium white bread 800g. Which do you prefer?"*. The best way to describe the solution to Problem 1 is to use some AD and show step by step the method of the solution. An AD database "Company-Addresses" with 2,629 companies is used (Figure 1). Knowing the TD, the user can assign an appropriate APA size before they start to analyse TD and create clusters.

To show the role of ACMM let us compare two VD.

User: *I'm looking for a security company address.*
NR: *There are 96 different security companies. Could you please be more precise?*

and **User:** Please help me to find a security company address.
NR: *We found for you Security services, Security systems, Blue security or Security group. Any good for you?*
User: *I'm afraid, not.*
NR: *Ok, what about Security Management, Security solutions, or Reliable security. Is it better?*
User: *Oh, yes, thank you. Security solutions, please.*
NR: *There are Perimeter Security Solutions Ltd. and Security Solutions (Northern) Ltd. Please make your choice.*
User: *Northern.*
NR: *Ok, fine. The address of Security Solutions (Northern) Ltd. is Unit 7/8, Bolton, Lancashire, BL1 8TU.*

If user could not select anything suitable (e.g. instead of select *"Security Solutions"* user said *"No"*) NR will continue dialogue:

NR: *Never mind. Let's try another way. In which city you are looking for Security Company?*
User: *Bolton.*
NR: *We have Blue Print Security Systems Ltd, Cell Security Ltd, Duval Security Plc or Jaymac Security Products Ltd. Any good for you?*
User: *Yes, thank you. I'm looking for Duval Security.*
NR: *Address of Duval Security Plc is 602 Chorley New Road, Bolton, Lancashire, BL6 6EX.*

The difference between 1st and 2nd VD is provided by ACMM. The main features of ACMM are:

- As a result of TD analysis a Pyramidal Structure of Clusters (PRC) is created [3] (fragment of PRC is shown in Figure1). The criterion for each "parent node" of PRC is very simple: "striving for having no more than APA child nodes" (APA = 4). Let us distinguish three possible amounts of cluster/sub-clusters nodes (NA): (1) NA <= APA e.g. for sub-cluster "*Blue Security*" there are just three members {"*Blue Lamp Security Ltd*", "*Blue Line Security Ltd*", "*Blue Print Security Systems Ltd*"}; (2) NA <= APA*APA. Assume there are APA sub-clusters and each of them has no more than APA nodes. In that case replay includes "sub-cluster name" instead of "company name" e.g. "*Security services", "Security systems", "Blue security or "Security group*"; (3) NA > APA*APA.
- In case (1), if the user is looking for "*Blue Security*" NR immediately offers him/her three companies (see above). Case (2) is represented in given VD.
- As for case (3) **supplementary field** (or fields) (SF) are involved. In the considered VD "*City*" is selected as SF. SF is defined at the stage of Scenario of Dialogue creation. In real AD "Companies-Addresses" except "*City*" SF "*Postcode*" is used.

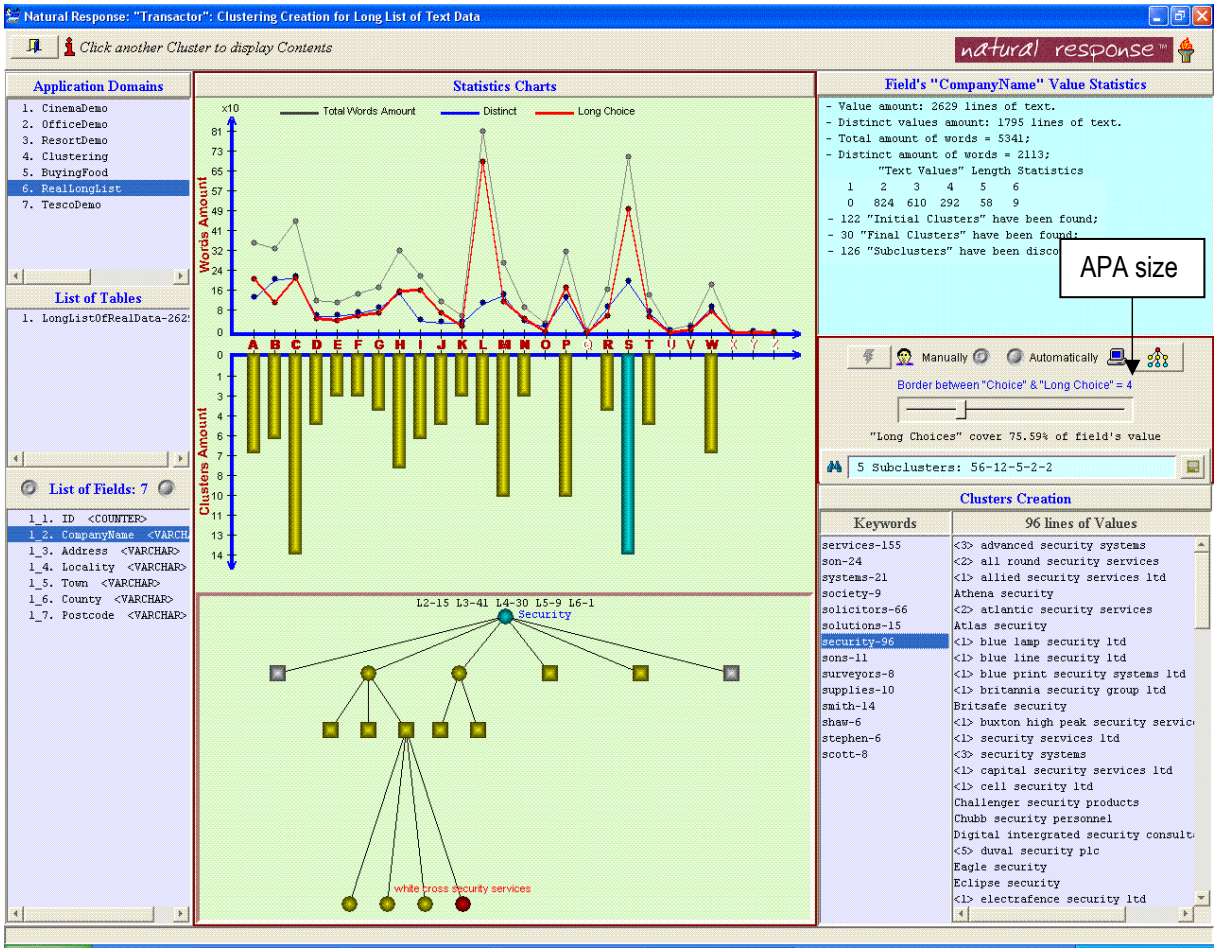


Figure 1. "Companies-Addresses" Application Domain

Wrong Recognition of Verbal Enquiry

Let us consider two different variants of wrong VE recognition. Assume that in the result of TD analysis vocabulary $\mathcal{V}$ of meaningful words and set $\mathcal{S}$ of words sequences (each of which represents a TD line) have been created, and then grammar $\mathcal{G} = \mathcal{V} \cup \mathcal{S}$ has been generated. First of all let us consider the situation when $\mathcal{G} = \mathcal{V}$. In this case when the user said word w_u speech recogniser (SR) using $\mathcal{G}$ produced w_r. Theoretically there are four different results of such event:

(1) $w_u \in \mathcal{V}$ and $w_u = w_r$ (naturally, $w_r \in \mathcal{V}$. But equation $w_u = w_r$ requires some explanation. In the set of recognition results w_i and s_j are distributed in between three intervals in accordance with their scores: "Yes", "Clarifying Question" and "Don't understand" intervals. w_r denotes the top scoring word from "Yes" interval);

(2) $w_u \in \mathcal{V}$ and $w_u \neq w_r$;

(3) $w_u \notin \mathcal{V}$ and $w_u \neq w_r$;

(4) $w_u \notin \mathcal{V}$ and $w_r = nil$ ($w_r = nil$ means that w_r belongs to the "Don't understand" interval).

When $\mathcal{G} = \mathcal{V} \cup \mathcal{S}$ and VE is represented by a sequence of words e.g. $s_u = w_{u1} \rightarrow w_{u2}$ the situation of VE recognition is more complicated. We are not going to discuss here results of recognition for all possible combinations of input VE. We just consider a couple of the most significant cases:

(5) $w_{u1} \in \mathcal{V}$, $w_{u2} \in \mathcal{V}$, $s_u \in \mathcal{S}$, $w_{u1} \neq w_{r1}$, $w_{u2} \neq w_{r2}$ and $s_u \neq s_r$;

(6) $w_{u1} \in \mathcal{V}$, $w_{u2} \in \mathcal{V}$, $s_u \in \mathcal{S}$, $w_{u1} \neq w_{r1}$, $w_{u2} \neq w_{r2}$ and $s_r = nil$;

(7) $w_{u1} \notin \mathcal{V}$, $w_{u2} \notin \mathcal{V}$, $s_u \notin \mathcal{S}$, $w_{u1} \neq w_{r1}$, $w_{u2} \neq w_{r2}$ and $s_u \neq s_r$ (but $s_r \in \mathcal{S}$).

Cases (1) and (4) are obvious. For the solutions of cases (2) and (3) it is better to show on example. Suppose we have AD "Operator", which is the alternative to a receptionist, providing customers call reception. Assume the next VD took place:

User: *Can I speak to Vlad?* (where $w_u = Vlad$. **SR** recognised *Vlad* as *Wlad* from "Yes" interval, i.e. $w_r = Wlad$ and $w_r \in \mathcal{V}$ - case (2)).

NR: *Certainly, connecting you now* (i.e. connecting to the wrong person).

Solution. To avoid annoying "clarification questions" e.g. *"Did you say Vlad?"*, recognised word from "Yes" interval is accepted as a "correct word" but such an approach might be cause a wrong result of VD. To sort out this problem a set of heuristics production rules has been developed to check words from the "Yes" interval e.g. **"If 1st letter of w_r belongs to $\{b, p, v, w, \ldots\}$ ask *"clarifying question"*.**

Case (3) is more complicated. Suppose that the position of someone from the customer's company (AD "Operator") is not *"Director"* but instead *"Chief Executive Officer"* been used. AD "Operator", as well as connecting users, can navigate them i.e. describe the direction to the company's office. Another words, "*director*" $\notin \mathcal{V}$ and "*direction*" $\in \mathcal{V}$. The next VD might take place:

User: *Can I speak to a director?* (where w_u = *director*. **SR** recognised *director* as *direction* from "Yes" interval, i.e. w_r = *direction* and $w_r \in \mathcal{V}$).

NR: *With pleasure. Where are you now?* (completely wrong VD).

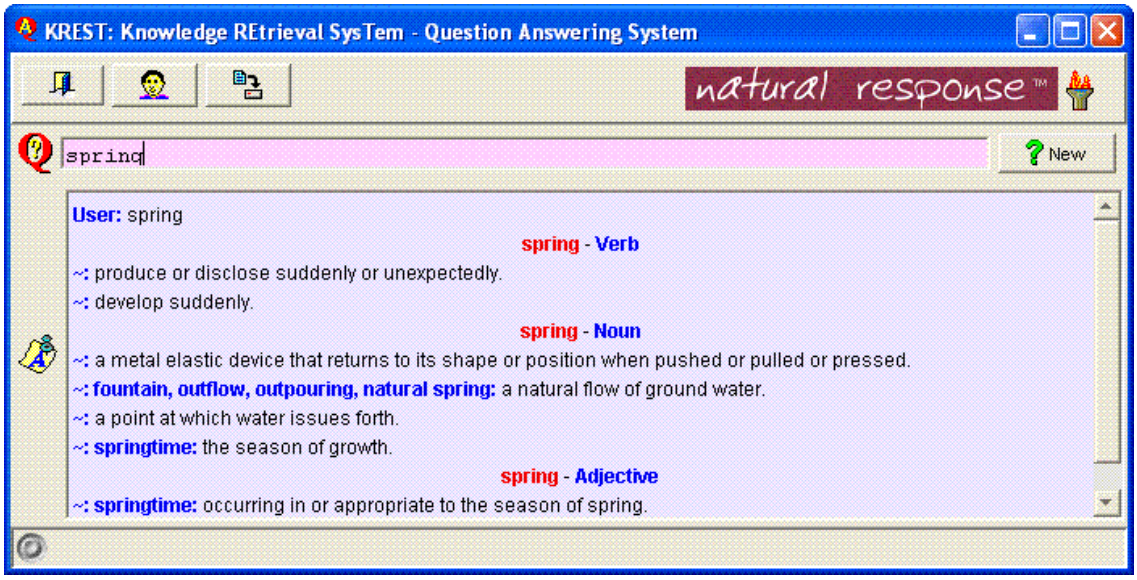


Figure 2. Synonyms and meaning of word "*spring*"

Solution. When AD created it is not enough to include in the knowledge base (KB) of AD only meaningful TD from customer's DB. It is important to extend text values by adding corresponding synonyms. First of all let us distinguish "**closed**" and "**open**" TD. For example, in AD "Operator" field "*Name*" represents ***closed*** data because caller should pronounce a name exactly from the existing list of Names. But field "*Position*" represents ***open*** data because caller might say: *"I would like to talk with a director (chairman, boss, managing director etc.)"*. In the case

of considered field "Position" the solution is very simple because the set of valid positions in UK is well defined and very restricted. It is enough just to provide a link between the existing position i.e. *"Chief Executive Officer"* and set of synonyms i.e. {*director, chairman, boss, managing director etc.*}. Now the previous VD will be changed:

User: *Can I speak to director?*
NR: *Our Company does not have a director position. We have chief executive officer instead. Would you like to speak to him?*

To find an appropriate set of synonyms for an open field is not easy when field represents TD like "*Product*" containing thousands and thousands of different products. On the one hand, it is not difficult to find synonyms for meaningful words of product using, for example, WordNet [6]. WordNet is a lexical database for the English language. Using WordNet's dictionary makes it easy to extract synonyms and meanings for any word e.g. *spring* (see Figure 2). But on the other hand, it is very difficult to automatically select appropriate synonyms for the current AD. To solve this problem the list of VE regarding AD should be analysed.

In the result of VE processing a language model (LM) of VE will be created. LM contains statistical information about which words or word sequences are used mostly by callers of AD. In addition to the statistical information, words adjacent within VE are analysed. The idea of adjacent words is based on the assumption that with the successive presentation of a number of words the strongest relation in it is the relation between the nearest neighbour words. *"Their succeeding one after another presents evidently an important condition of structuring"* [4, p.231]. It is important to underline that this information is always AD dependent. LM is stored in Sequential/Simultaneous Structure (SSS) [5]. LM of VE contains not only a set of synonyms but also patterns of VE. An example of a pattern for AD "Operator" is:

$$(speak \oplus talk \oplus connect) \rightarrow \langle Name \rangle \oplus \langle Position \rangle;$$

where $\oplus$ **-** denotes "*exclusive OR*"; $\rightarrow$ means "*followed by*"; **<Name>** and **<Position>** denote any value from DB fields "*Name*" and "*Position*" respectively. Words *speak*, *talk* and *connect* (such words we call "**action words**") are included in $\mathcal{V}$ along with words from fields "*Name*" and "*Position*". If a pattern of VE is available the initial VD will be have in a more "*natural way*" (even if word *director* is not added to $\mathcal{V}$)**:**

User: *Can I speak to a director?* (**SR** recognised *director* as *direction* and correctly recognised word *speak* i.e. both *speak* and *direction* have been placed in the same "Yes" interval. Action words have priority and that is why the corresponding VE pattern has been involved to continue VD)
NR: *Do you know the name of who you want to speak to?*
User: *I'm afraid, not.*
NR: *Ok, in that case would you like to talk with chief executive officer?*
User: *Sorry, I change my mind. Can I speak to marketing manager?*
NR: *Certainly, connecting you now.*

To discuss cases (5), (6) and (7) AD "Postcodes", which contains postcodes and cities, will be used. There are **1,741,488** postcodes in the UK. Each Postcode consists of two parts. The first part is the Outward Postcode, or **Outcode**. There are **2,918** outcodes. Outcode is separated by a single space from the second part, which is the Inward Postcode, or **Incode**. There are **4,000** incodes and **27,089** cities and towns in UK. Postcodes recognition is very important because as soon as a postcode has been recognised correctly a very restricted set of full addresses (~30 – 60) is allocated. In order to provide postcode recognition, grammar $\mathcal{G} = \mathcal{V} \cup \mathcal{S}$ must be generated but the size of $\mathcal{G}$ is huge (i.e. $|\mathcal{G}|$ = 1,741,481+2,918+4,000 = 1,748,399 items) and modern SR cannot provide reliable and acceptable recognition for such size of grammar. If we exclude postcodes from $\mathcal{G}$ and add cities it is possible significantly reduce grammar's size i.e. $|\mathcal{G}|$ = 2,918+4,000+27,089 = 34,007 items but increase "intelligence of VD". Below several VD are represented to show the solution of cases (5), (6) and (7).

User: *PL3 4PX.*
NR: *You said PL3 4BX. Is it correct?*
User: *No.*
NR: *Never mind. Let's try another way. Is outcode PL3 correct?*

User: *Yes.*
NR: *Ok, fine. In that case you said PL3 4PX* (because only three different incodes with X link to PL3: 4PX, 5PX and 6PX). *Is it right?*
User: *Absolutely.*

The next VD is more complicated. Town and city are involved to clarify the wrong result of postcode recognition.

User: *PL11 3BX.*
NR: *It is not right postcode. Did you really say BL11 3BX?*
User: *No.*
NR: *No matter. Let's try another thing. Tell me your town, please.*
User: *Seaton.*
NR: *There are 8 towns Seaton in different areas like Exeter, Sunderland, etc. What is your nearest city?*
User: *Plymouth.*
NR: *Great. The outcode of Seaton in Plymouth area is PL11. Is it correct?*
User: *Of course.*
NR: *Consequently you expressed your postcode as PL11 3BX. Isn't it?*
User: *Oh, at last. I really said so.*

Text Data Verification and Mining

By default two different kind of DB exist in reality – with correct TD and "wrong" TD. For example, DB, which contains TD about *people*, *addresses*, *companies* etc. as a rule have correct TD and require minimum effort for TD verification. Result of TDVM for such TD is shown on Figure 1. The typical example DB with "wrong" TD is any DB with "Product name" in it. Let us consider the real AD "Product" (Figure 3). We will call TD "wrong" if they contain *wrong words* (WW) and/or *unknown words* (UW). Two different kind of WW will be distinguished:

(1) abbreviation*, which is a sequence of consonants letter.* For example, ***"Wht thk slcd brd"*** means "*White thick sliced bread"*. (2) Sequence of symbols when sequence includes not only letters but also some symbols e.g. *"Dietary spec gltn/fr & wht/frchoc/orge waf x5"* (We don't know the meaning of that).

The definition of UW is very simple: Let us call word UW if it absent in Oxford dictionary. There are three possible reasons for it: (1) word shortening e.g. "***org"*** for ***organic*** (***org*** was used in 261 names of product e.g. "*Doves Farm Org Lemon Cookies Non Gluten"*); (2) new words (= brand) e.g. *Nestle*, *Cadburys* etc.; (3) misspelling. Result of DTVM of AD "Product" is shown on Figure 3.

At the stage of TDVM several tasks need to be solved:

- **Quantity analysis** – count different kinds of frequencies (see Figures 1 and 3);
- **Clustering** – to provide the classification of information into clusters.
- **Data Verification** of TD i.e. extracting WW and UW (see Figure 3). For AD "Product" **76.55%** represent wrong data.
- **Production Rules (PR) creation**. It is very important to underline the fact that **a customer's DB cannot be corrected or reconstituted**. ***It means that conversion of correct user's enquiry to wrong customer's data needs to be provided.*** For such conversions PR need to be created. **Antecedent** of PR might be a single word or sequence of words. Several different WW or UW might represent **consequent** of PR. (+) stands for exclusive OR (see PR on Figure 3).
- **Data Reconstitution** needs to be provided at the step of KB creation when the link between TD and DB Field is made. At this stage each product description is converted using PR from wrong TD to correct TD.

User's Enquiry Conversion. For an explanation it is better to consider some examples. Assume user's enquiry was "W*hite thick sliced bread*". Four appropriate PR have been allocated ("*white* $\Rightarrow$ *wht* ", "*thick* $\Rightarrow$ *thk* ", "*sliced*

$\Rightarrow$ *slcd* $\oplus$ *sld*", and "*bread* $\Rightarrow$ *brd*"). Customer's data might contain any combination of wrong and correct words. Theoretically for the considered example there are 16 possible combinations of data, namely**:** (1) "*White thick sliced bread*", (2) "*White thick sliced brd*", … , (16) "wht thk (slcd $\oplus$ sld) brd". Result of such conversion is shown on Figure 3.

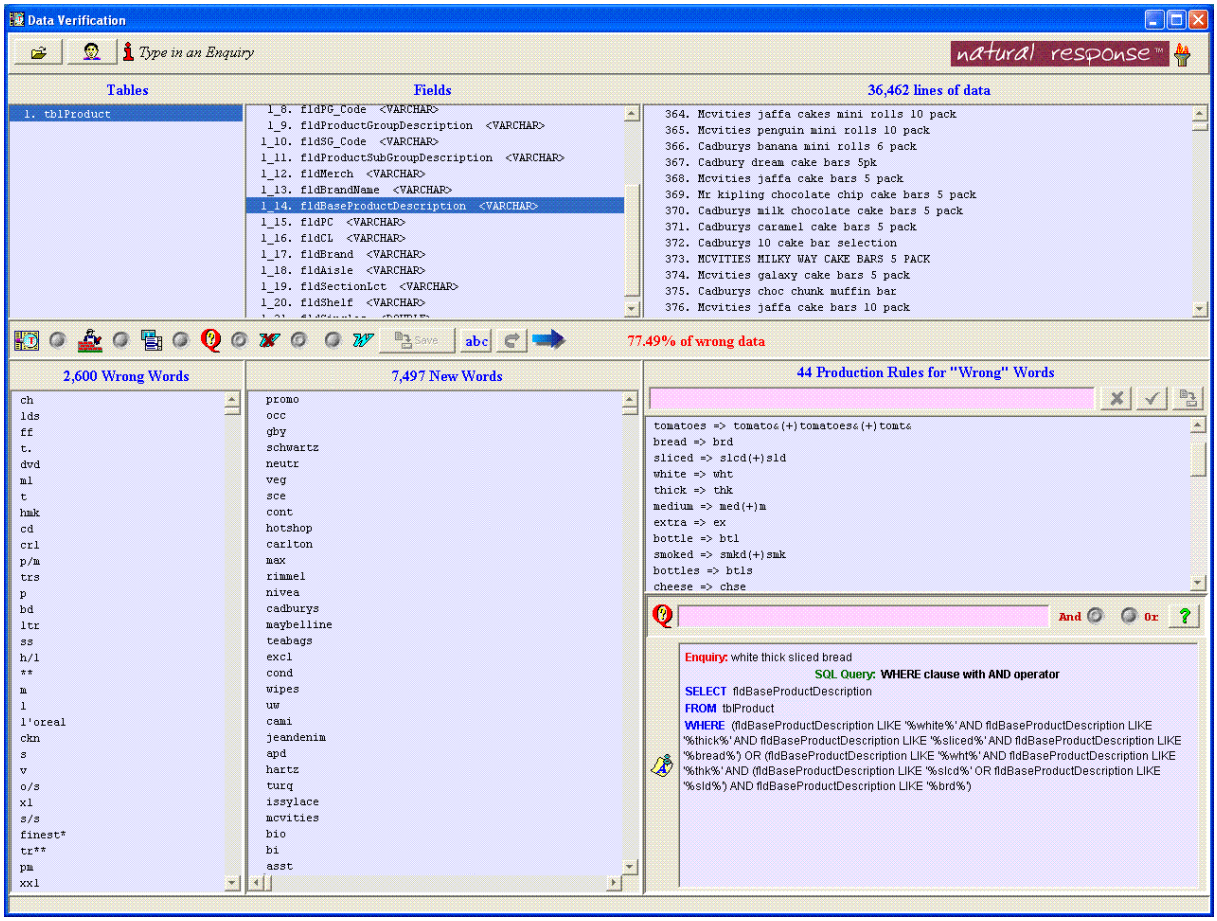


Figure 3. Text data verification and mining

Grammar Creation

Let us consider three aspects of grammar creation. The first one is regarding the ability of SR to better recognise (i.e. more reliable, with higher score) a sequence of words than single words e.g. $\mathcal{G} = \{white\ bread\}$ provides better recognition of VE "*white bread*" than $\mathcal{G} = \{white, bread\}$. But for reliable recognition it is not enough to include in $\mathcal{G}$ just product names because the user is absolutely free to ask whatever he/she wants about, for instance, "W*hite thick sliced bread*", namely, "*white bread*", "*sliced bread*", "*thick sliced bread*" etc. On the one hand SSS (Sequential/Simultaneous Structure [5]) allows us to generate sensible sequences, on the other hand, the size of $\mathcal{G}$ is critical. A complex grammar usually causes a less accurate recognition because of the larger number of possible word sequences. The solution is to extract frequently used utterances from the users' VE. The important point is – **any NR system, which provides VD, should be a self-learning system**. Indeed, during the working day thousands and thousands of telephone calls to AD "Product" are taking place. "Sleeping time" is more than enough to update: (1) set of synonyms; (2) enquiry patterns and (3) grammar.

The second aspect is regarding the NR reaction. It goes without saying that VD must be in real time. Any delay of VD makes NR unacceptable. The obvious way of VE processing is: (1) recognise spoken utterance; (2) using PR add WW and UW; (3) convert VE+WW+UW to corresponding SQL query; (4) run SQL query; (5) synthesize reply; (6) provide text-to-speech conversion. We can avoid step (2) and include PR in grammar. Any line of grammar

might contain two parts: ***(spoken word or utterance) {result,*** *or* ***tag}***. The tag result is interpreted as a string and is delimited by braces "**{}**". All characters within the braces are considered as parts of the tag, including white space. The line of $\mathcal{G}$ for utterance *white sliced bread* will look as

(white_sliced_bread) {white, sliced, bread, wht, slcd ⊕ sld, brd}.

The third aspect is devoted to dynamically created grammars. Assume user said his outcode ***PL10***. Now instead of using the full size grammar (= 4,000 items) for Incodes it is better to allocate a set of Incodes regarding ***PL10*** (see Figure 4) and create a grammar just with 168 items.

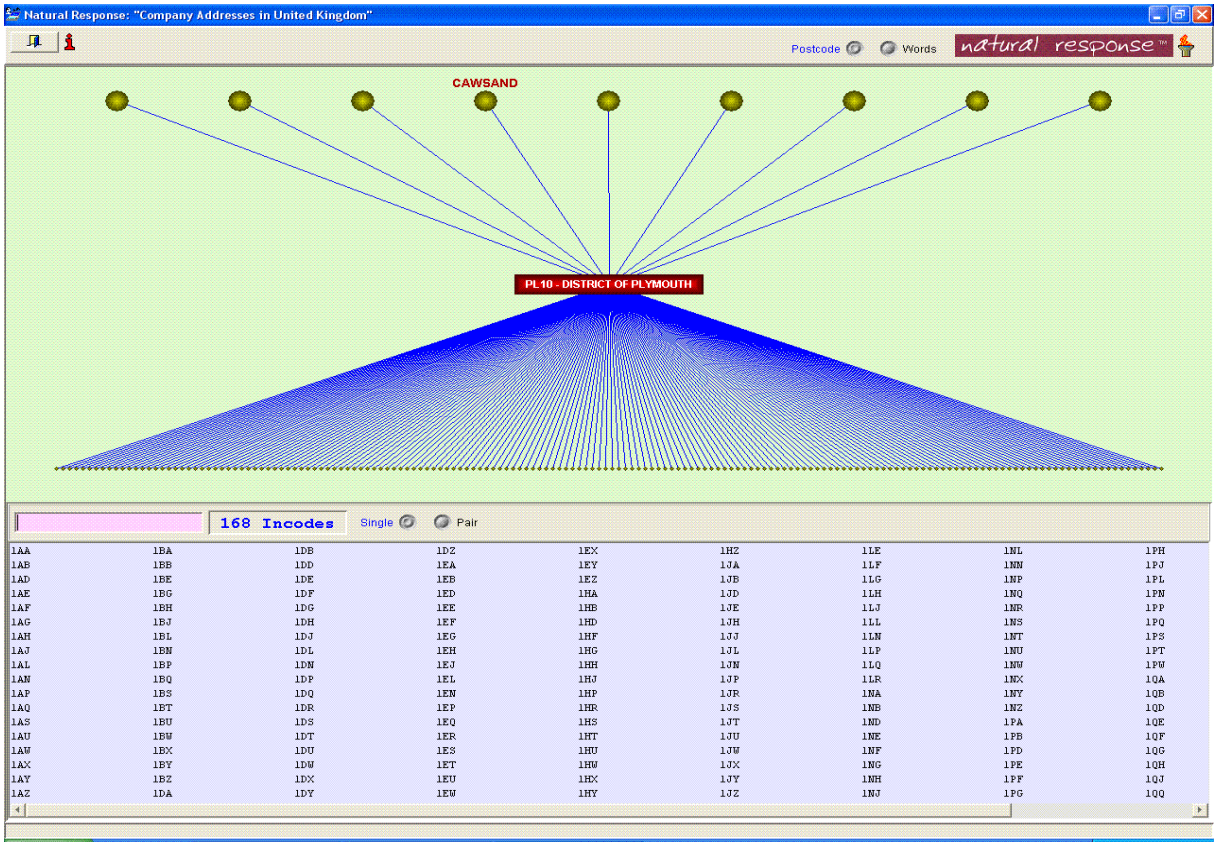


Figure 4. Towns and Incodes allocation

Conclusion

VD and WD are *"the two sides of a coin"*, the features of which should complement each other. The main purpose of our report was to call attention of scientists to distinctive features of VD. As we know from our scientific experience it is always easier to find the best solution when you have an access to some practical results. That is why we included in our report real results but not ideal, or "dream" ones.

Bibliography

[1] G.A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity to Process; Information. *Psy.Rev.*,63,81-97,1956.

[2] Natural Response™. www.naturalresponse.net, 2002.

[3] T.M. Khalil, V. Lovitsky, "A Structure of Memory in Concept Formation", Proc. of the IEEE Conference on *Systems, Man and Cybernetics*, Boston, Mass,509-512,1973.

[4] J.Hoffmann, Das Aktive Gedachtnis. Psycologische Experimente und Theorien zur Menschlichen Gedachtnistatigkeit, *VEB Deutscher Vergal der Wissenschaften*, Berlin, 1982.

[5] В.А.Ловицкий, Классификация структур памяти, *Проблемы бионики*, Харьков,8,138-153,1972.

[6] WordNet®. http://wordnet.princeton.edu.

Authors' Information

David Burns – Natural Response (Natural Response is a trading style and trademark of RSVP Dialogue Limited), The i-zone, Bolton Institute, Dean Road, Bolton, Great Manchester, BL3 5AB, United Kingdom, e-mail: david@naturalresponse.net

Richard Fallon – Natural Response, The i-zone, Bolton Institute, Dean Road, Bolton, Great Manchester, BL3 5AB, United Kingdom, e-mail: richard@naturalresponse.net

Phil Lewis – Natural Response, The i-zone, Bolton Institute, Dean Road, Bolton, Great Manchester, BL3 5AB, United Kingdom, e-mail: phil@naturalresponse.net

Vladimir Lovitskii – Natural Response, The i-zone, Bolton Institute, Dean Road, Bolton, Great Manchester, BL3 5AB, United Kingdom, e-mail: vladimir@naturalresponse.net

Stuart Owen – Natural Response, The i-zone, Bolton Institute, Dean Road, Bolton, Great Manchester, BL3 5AB, United Kingdom, e-mail: stuart@naturalresponse.net

КОНСПЕКТИРОВАНИЕ ЕСТЕСТВЕННОЯЗЫКОВЫХ ТЕКСТОВ

Виктор П. Гладун, Виталий Ю. Величко

Abstract: *Some applications of Natural language texts need such form of text representation that is a result of reasonable compromise between the wish to make the text shorter, saving its fundamental thematic purposes, and the wish to retell the source text more fully. Some degree of this compromise should be achieved at text abstracting when creation of different storage of textual information, for example archives, personal libraries and so on. What are the ways of this compromise achievement? The report deals with this problem. The method, realized in program system KONSPEKT is considered.*

Введение

Конспект является формой представления текстов, в которой объединяются два противоречивых требования к обобщенному представлению текстовых знаний:

1) сокращение объема текста при сохранении основных тематических целей (реферирование, аннотирование);
2) пересказ основных положений содержания при сохранении связности текста.

По отношению к исходному тексту конспект должен быть результатом разумного компромисса между желанием сократить его объем, жертвуя некоторыми элементами содержания (например, в целях экономии памяти или совершенствования поисковых операций), и желанием как можно полнее передать детали содержания. При удачном разрешении этого компромисса конспект становится наиболее рациональной формой представления текстов. Каковы пути выполнения указанных требований? В докладе рассматривается метод конспектирования естественно-языковых текстов, реализованный в программной системе KONSPEKT (Ин-т кибернетики им. В.М. Глушкова НАН Украины)

Совершенно очевидной является необходимость использования средств синтаксического и семантического анализа.

Семантический анализ текста

Все фразы исходного текста подвергаются синтактико-семантическому анализу. Основной операцией синтактико-семантического анализа является распознавание синтаксических и семантических отношений, связывающих слова текста.

Распознавание синтаксических и семантических связей между знаменательными словами осуществляется путем анализа флексий и предлогов без использования категорий и правил традиционной грамматики. По результатам синтактико-семантического анализа в исходном тексте отбираются предложения, содержащие "ядерные конструкции".

Термин “ядерные конструкции” используется в трансформационной грамматике для обозначения простого базового суждения, путем трансформации которого формируется предложение в целом. В данном случае ядерной конструкцией служит предложение, состоящее из подлежащего, сказуемого и соединяющей их связки. Алгоритм синтактико-семантического анализа на основе лексических моделей описан в [1-4].

Таким образом, на этом этапе из исходного текста для дальнейшего анализа отбираются полносоставные предложения. Для каждого из отобранных предложений формируется n - шаговое расширение ядра – часть предложения, содержащая его ядро, а также слова, связанные в дереве зависимостей с элементами ядра путями, длина которых не превышает n (заданный параметр).

В результате синтактико-семантического анализа формируется массив n -шаговых расширений полносоставных предложений исходного текста. Далее осуществляется тематический анализ текста. Тематический анализ текста выполняется на основе онтологии ассоциаций.

Онтология ассоциаций

В дальнейшем будем пользоваться следующими определениями. Ассоциация понятий представляет собой множество понятий, имеющих общую семантическую характеристику (ассоциативный признак). Ассоциативным признаком может быть вид деятельности или отдельное действие, с которым связано понятие, причастность понятия к какому-либо типу событий, явлений, ситуаций, (например, свадьба, рождество, выборы и т.п.), временной период или интервал (зима, лето, июль и т.п.).

В данной разработке используются три группы ассоциаций: A (action)-по виду деятельности; S (situation)-по типу ситуаций; T (time)-по времени.

Онтологией ассоциаций является словарь терминов, в котором термины индексируются наборами ассоциативных признаков, обозначающих ассоциации понятий. Термины, обозначающие деятельность или действие, входят в ассоциации группы A и индексируются сокращенным обозначением деятельности или действия, по отношению к которому индексируемый термин является подвидом. Например, термины "искусственный интеллект", "принятие решений", "извлечение знаний" индексируются индексом "ин". Это означает, что все они связаны отношением "подвид" с термином "информатика" и, таким образом, входит в ассоциацию понятий этого термина. При индексах ассоциаций записывается целое число, указывающее "вес" ассоциации.

Термины, обозначающие объекты, входят в ассоциации группы A, если они связаны с каким-либо видом деятельности или действием отношением типа "семантический падеж", "агент", "объект", "инструмент", "среда", и т.п. Соответствующие индексы терминов обычно имеют вес 2. Например, термины "прораб", "топор", "данные" получают индексы соответственно ст2, ст2, ин2. Это означает, что они входят в ассоциации "строительство" или "информатика" с весом 2.

Связь термина с ассоциациями групп S и t указывается путем ввода в индекс термина сокращенных обозначений ассоциаций, входящих в эти группы. Обычно таким ассоциациям присваивают веса 2 или 3. Например, в индекс термина "елка" могут входить обозначения ро3 (рождество), зи2 (зима).

Индексирование терминов в онтологии ассоциаций выполняется экспертами и, естественно, отображает их субъективное понимание индексируемых терминов. Для удобства выполнения индексирования в распоряжении экспертов должна быть некоторая классификация видов деятельности и отдельных действий, а также список характерных классов событий, явлений, ситуаций, а также привычных терминов, именующих временные периоды.

Тематический анализ текста

Для поддержки связности конспекта (abstract) авторы предлагают подход, при котором текст рассматривается как совокупность текстовых фрагментов, раскрывающих отдельные связанные или независимые темы. Процесс тематического анализа организован циклически. На каждом прогоне цикла из исходного текста в формируемый конспект отбираются предложения, n -шаговые расширения которых содержат ключевой термин, который при запуске цикла задается пользователем системы, а на последующих его прогонах формируется автоматически. В качестве очередного ключа в оставшихся (еще не отобранных в конспект) предложениях выбирается термин, имеющий наибольший суммарный вес ассоциаций, которые являются общими для этого слова и ключа, используемого на предыдущем прогоне цикла. Таким образом осуществляется упорядоченный перебор тематики исходного текста, причем за счет выбора в качестве очередного ключа знаменательного слова, наиболее близкого по ассоциациям предыдущему ключу, каждая очередная тема оказывается связанной с предыдущей.

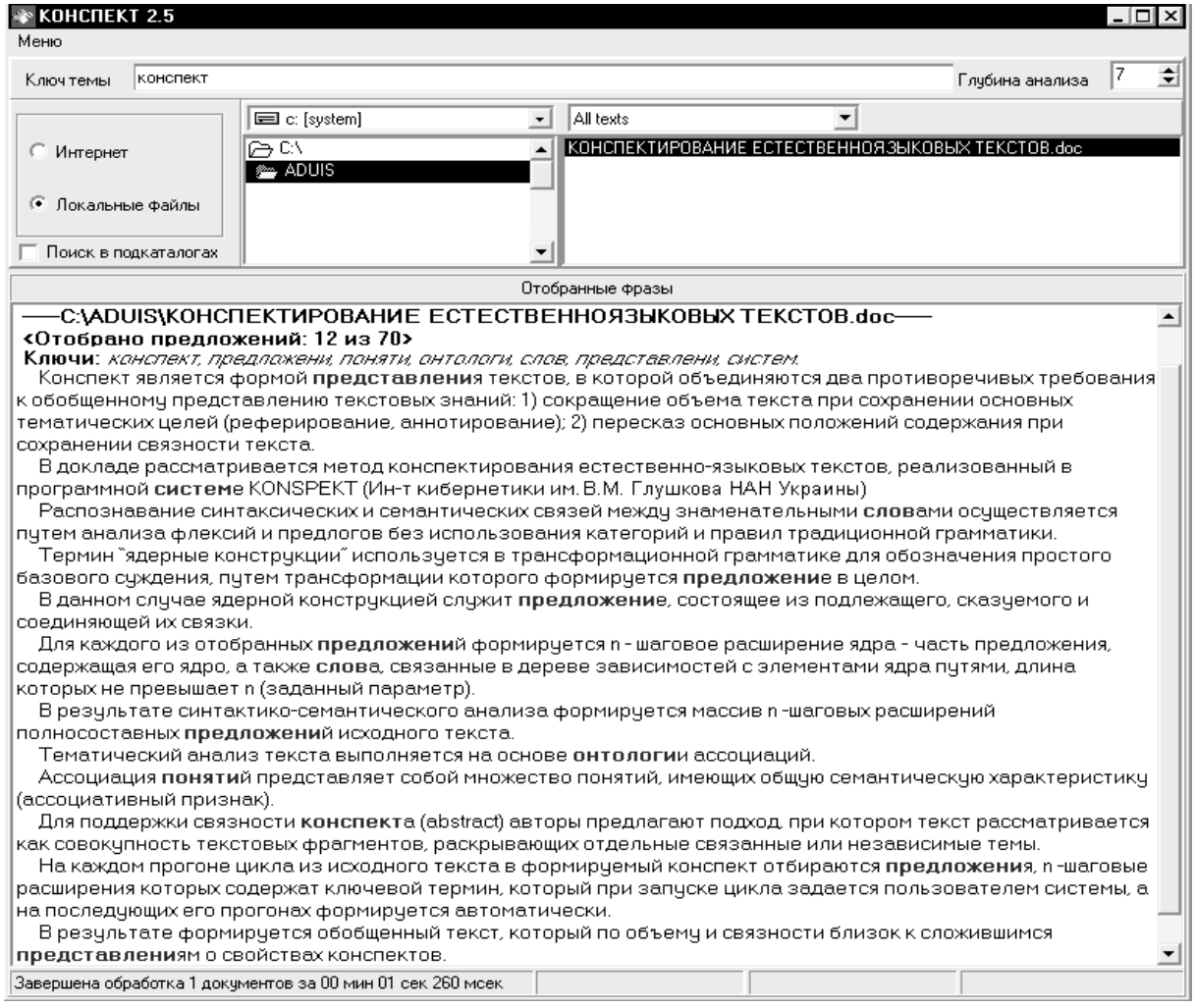


Рис.1 Пример работы системы KONSPEKT.

В результате формируется обобщенный текст, который по объему и связности близок к сложившимся представлениям о свойствах конспектов. На рисунке 1 приведен пример конспекта, полученного с помощью системы KONSPEKT по тексту настоящей статьи.

Система KONSPEKT является эффективным инструментом архивирования при создании различных хранилищ текстовой информации. Онтология ассоциаций позволяет эффективно реализовать различные процессы тематического анализа и синтеза естественно языковых текстов.

Литература

1. Гладун В.П. Процессы формирования новых знаний. - София: СД "Педагогъ". 1994г. - 192с.
2. Гладун В.П. Планирование решений. - Киев: Наукова думка, 1987. -168 с.
3. Гладун В.П. Формирование тематических знаний на основе анализа ЕЯ текстов сети Интернет //Диалог'2003. Прикладные проблемы.
4. V.Gladun, A.Tkachev, V.Velichko, N.Vashchenko. Selection of Thematic NL-Knowledge from the INTERNET // International Journal "Information Theories & Applications" Vol.10 /2003 N.2 FOI-COMMERCE - Sofia.- pp.123-125.

Информация об авторах

Виктор Поликарпович Гладун – Институт кибернетики имени В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua

Виталий Юрьевич Величко – Институт кибернетики имени В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua

О ЗАДАЧЕ СЕМАНТИЧЕСКОГО ИНДЕКСИРОВАНИЯ ТЕМАТИЧЕСКИХ ТЕКСТОВ

Надежда Мищенко, Наталья Щеголева

***Abstract**: The creation of domain ontology is of high importance in information systems. This paper discusses software tools for concept extraction from domain texts in natural languages. The approach presented by authors allows eliminate the irrelevant combination of words from the very beginning of text processing.*

***Keywords**: ontology, combination of words, terms, vocabulary of auxiliary words and links, morphological analysis.*

Введение

Неэффективность поиска текстовой информации в сети Интернет по ключевым словам общепризнана. Стало очевидным, что для того чтобы результатом поиска были документы, релевантные запросу, необходимо анализировать семантику, представленную в текстовых документах. Средством описания семантики текстов являются онтологии – формальные спецификации семантики, ориентированные на задачи компьютерной обработки текстов. Онтология специальной (тематической) области включает концепции области, их определения и семантические связи (отношения) между ними, средства представления отношений, позволяющие устанавливать связь между текстами запросов на естественном языке и семантическими индексами документов.

Составление онтологий для того или иного естественного языка – процесс достаточно трудоемкий. Например, тезаурус русского языка RuTez [Лукашевич, 2002], формируется с 1997 года и по данным публикации ([Dobrov, 2003]) содержит 43 тис. концепций, 166 тис. слов и словосочетаний, 166 тис. связей и более миллиона связей с учетом транзитивных.

Онтологии тематических областей меньшие по объему благодаря ограниченному количеству лексики и потенциальной однозначности терминов. В случае появления нового значения уже существующего термина однозначность обычно достигается за счет присоединения к этому термину определений, служащих уточнением значения термина. В этом случае в качестве нового термина выступает словосочетание. Важным стимулом для разработки тематических онтологий служит тот факт, что поиск тематических текстов составляет значительную часть всех поисковых работ по Интернету.

Как известно, полный цикл обеспечения поиска документов с помощью семантического индексирования тематических текстов должен состоять из следующих основных этапов.

1. Создание онтологии, осуществляемое на основе извлечения знаний из коллекции текстов заданной тематики на естественном языке и их формализации, то есть, их записи на специальном языке, который следует разработать.

2. Семантическое индексирование текстов по мере их введения в Интернет с использованием онтологии.
3. Автоматическое преобразование запроса пользователя на естественном языке в поисковый образ документа на языке онтологии.

4. Поиск документов путем сравнения полученного образа с семантическими индексами документов заданной тематики.

Авторам неизвестны практические поисковые системы, в которых были бы автоматизированы все перечисленные этапы семантического поиска текстов. Однако, извлечение семантики из текстов присутствует во многих системах, поскольку оно необходимо как для построения онтологии, так и для поиска на основе семантического анализа текстов при отсутствии онтологии.

Извлечение знаний из текстов

Прежде чем характеризовать наш подход к задаче извлечения знаний из тематических текстов, рассмотрим две системы, в которых реализованы аналогичные подходы к решению данной конкретной задачи.

Интеллектуальная метапоисковая система “Сириус” [Куршев, 2002] не располагает онтологией и предлагает осуществлять поиск документов в два этапа: на первом этапе использовать линейный поиск документов по ключевым словам, а на втором – уточнить результаты первого этапа, подвергнув тексты найденных документов структурному анализу, цель которого состоит в том, чтобы обнаружить семантические связи между синтаксемами – понятиями данной предметной области. Для этого выполняются морфологический анализ, на результатах которого синтаксический анализ находит синтаксемы – согласованные группы слов, которые затем подвергаются семантическому анализу с целью установления семантической связи между синтаксемами. Под семантической связью понимается отношение понятий в понятийной системе предметной области. Структурному анализу подвергается и запрос, после чего осуществляется линейный поиск соответствующих семантических связей и сравнение связанных синтаксем.

В публикации нет четкого ответа на два важных вопроса.

Во-первых, почти каждое предложение дает минимум две синтаксемы, поэтому не ясно, на каком этапе отбрасываются те, которые не несут семантическую нагрузку, относящуюся к тематике текстов, то есть не являются понятиями.

Во-вторых, запоминается ли полученная семантическая информация в виде троек: (связь, первая синтаксема, вторая синтаксема) на случай повторного запроса по этой же тематике. Не релевантные и поэтому отброшенные документы, для которых был по сути сформирован семантический индекс в результате структурного анализа, могут стать релевантными для других запросов.

Кроме того, в публикации отсутствуют данные о практическом применении системы.

Ответ на первый поставленный вопрос имеется в работе [Dobrov, 2003], которая посвящена построению тезауруса русского языка для предметных областей. Для отбора слов и беспредложных словосочетаний используются частотные списки слов. Для удаления из списка общеупотребительных словосочетаний созданы специальные словари общеупотребительных слов, играющих роль фильтра для исключения из рассмотрения этих слов при морфологическом анализе текстов. Для выявления длинных словосочетаний с предлогами применяются специальные методы, основанные на просмотре соседних слов и анализе частоты их встречаемости в тексте. Цель работы: создание технологии для построения онтологий предметных областей (в статье рассматривается Авиа-Онтология). Для этого используется тезаурус русского языка RuTez [Лукашевич, 2002], подходящие понятия которого включаются в специальную онтологию, а остальные формируются по мере семантического анализа текстов предметной области.

Разработанная нами лингвистическая система FEST для поиска терминов [Мищенко, 1999] преследует исключительно практическую цель, поставленную заказчиком, а именно, формирование ответов на вопросы: какова тема документа или принадлежит ли документ теме, заданной пользователем в виде набора отдельных ключевых слов или связанных отношениями, которые выражаются одной или

несколькими лексемами. Система используется с 1999 года для анализа и классификации текстов по юриспруденции на русском и украинском языке. Работающий вариант является промежуточным. Система находится в развитии за счет создания средств автоматизации различных аспектов процесса семантического анализа текстов.

Отличия системы FEST от рассмотренных выше в аспекте извлечения знаний из текстов состоят в следующем.

1. Система не привязана ни к одному естественному языку, поэтому использует минимальное лексическое обеспечение каждого языка в виде словаря служебных слов (союзы, предлоги, частицы) и таблицы окончаний. Составление словаря и морфологических таблиц осуществляется один раз и представляет собой процесс настройки системы на конкретный язык. Для этого слова описываются на специальном формальном языке. Полученное описание (спецификация) автоматически преобразуется системой в словарь. Кроме описания служебных слов в рамках настройки на конкретный язык составляется формальное описание морфологии для ее использования программой морфологического анализа. На основании этого описания строятся морфологические таблицы. Подробное изложение процесса спецификации лексики и окончаний имеется в [Мищенко, 2003]. На практике анализируются тексты на русском и украинском языке. В статье [Мищенко, 2000] описан экперимент по анализу текстов на английском языке.

2. Морфологический анализ совмещен с поиском словосочетаний, который выполняется в пределах одного предложения. Чтобы найти словосочетания полнозначных слов текста необходимо построить словарь таких слов. При этом достаточно иметь словарь только тех полнозначных слов, которые потенциально являются терминами или входят в термины-словосочетания. В словаре системы FEST каждое полнозначное слово должно быть подано основой и указателем на кортеж окончаний. Таким образом, каждая статья словаря представляет все словоформы с данной основой. Кроме того, слова в словаре сопровождаются грамматическими характеристиками, необходимыми для успешного поиска словосочетаний (синтаксем) и связей между ними.

Итак, поиск словосочетаний выполняется последовательностью таких действий.

2.1. Отбор полнозначных слов для их включения в словарь основ проводится на основе частотного списка основ, поскольку общепризнанным является факт высокочастотного вхождения терминов в тематический тексты. Частотный список формируется по результатам морфологического анализа всего документа без словаря, но с использованием таблицы окончаний. Результат анализа – отметка каждой неизвестной (не имеющейся в словаре) словоформы с разделением основы и окончания.

2.2. Используя разметку словоформ, система строит частотный список основ неизвестных словоформ по убыванию частоты их употребления. Каждая основа сопровождается списком окончаний и адресами употребления словоформ с данной основой. По сути, каждая статья списка определяет лексему. Результат анализа без словаря основ может содержать ошибки – неправильное отделение окончания вследствие ложной омонимии окончаний, но статистически это не влияет на качество частотного списка.

2.3. Верхняя часть частотного списка представляет слова – кандидаты на термины. Среди них могут быть и общеупотребительные слова, которые предположительно могут войти в словосочетания, и глаголы, связывающие словосочетания. Какая часть частотного списка должна быть взята для последующих действий, определяется пользователем системы. Отобранные лексемы описываются формальным языком системы с последующим автоматическим преобразованием описания в компьютерный словарь. Алгоритм автоматизированного формирования спецификаций изложен в [Щоголева, 2004]. По полученной спецификации автоматически строится словарь часто встречающихся основ, который дополняет имеющийся словарь вспомогательных слов.

2.4. Морфологический анализ текста со словарем служебных слов и основ часто используемых полнозначных слов сопровождается распознаванием словосочетаний, в основном именных групп, и связей между ними в пределах одного предложения. Словоформы, принадлежащие словосочетаниям, отмечаются для составления частотного списка словосочетаний.

2.5. Формируется частотный список словосочетаний с предварительным преобразованием основного слова в нормальную форму. Словосочетания располагаются в порядке уменьшения числа словоформ в них, а в пределах словосочетаний одной длины – в порядке уменьшения частоты встречаемости.

Верхняя часть частотного списка словосочетаний дает представление о тематике текста и запоминается, равно как и словарь входящих в словосочетания отдельных лексем, для дальнейшего использования. Эта же часть частотного списка после определения связи между словосочетаниями может служить основой для семантической индексации текста. В качестве средства связи между словосочетаниями используются глаголы.

Заключение

Представлена лингвистическая система FEST, особенностями которой являются мобильность и ориентация на практическое применение. Мобильность обеспечивается универсальностью системы и наличием в ней языковых и программных средств для автоматической настройки системы на конкретный естественный язык. Средства для построения словаря наиболее часто встречающихся в тексте слов, которые потенциально являются терминами или входят в термины-словосочетания, позволяет избавиться от нерелевантных словосочетаний с самого начала процесса извлечения знаний из текста, что обеспечивает практическое применение системы в любой тематической области независимо от наличия в ней словаря терминов.

Библиография

[Лукашевич, 2002] Н. В. Лукашевич, Б.В. Добров. Тезаурус русского языка для автоматической обработки больших текстовых коллекций // /Труды Межд. Сем. “ДИАЛОГ”2002. Компьютерная лингвистика и ее приложения” (6 – 11 июня 2002, РФ, Протвино) в 2-х т. Т.2. Прикладные проблемы, Протвино, 2002, с. 338-346.

[Dobrov, 2003] B. Dobrov, N. Loukashevich, O. Nevzorova. The technology of new domeins’ ontologies development // Abstracts of the X-th International Conference KDS 2003 (June 16-26, 2003, Varna, Bulgaria), FOI-COMMERCE, Sofia, 2003, p. 283-290.

[Куршев, 2002] Е.П. Куршев, Г.С. Осипов, О.В. Рябков, Е.И. Самбу, Н.В. Соловьева, И.В. Трофимов. Интеллектуальная метапоисковая система // Труды Межд. Сем. “ДИАЛОГ”2002. Компьютерная лингвистика и ее приложения” (6 – 11 июня 2002, РФ, Протвино) в 2-х т. Т.2. Прикладные проблемы, Протвино, 2002, с. 320-330.

[Мищенко, 1999] Н.М. Мищенко, В.В. Федюрко, Н.М. Щоголева. Мобильная система статистической обработки профессиональных текстов в целях определения их тематики // Проблемы программирования. Киев, 1999, №2, с. 11-18.

[Мищенко, 2000] Н. Мищенко, А. Мищенко. Об автоматизации поиска терминов в научно-технических текстах // ж. Вестник Гос. ун-та "Львовская политехника", Материалы. 6-ой Межд. науч. конф. "Проблемы украинской терминологии", (19-21 сентября 2000 г., г. Львов), Львов, изд-во Гос. ун-та "Львовская политехника", 2000, № 402, с. 92-97 (на украинском языке).

[Мищенко, 2003] Н.М. Мищенко, Н.Н. Щеголева. О лексико-статистическом анализе научно-технических текстов //. Proc. X-th Intern. Conf. KDS 2003 (June 16-26, 2003, Varna (Bulgaria). FOI-Commerce, Sofia, 2003, с. 315-321.

[Щоголева, 2004] Мищенко Н.М., Щоголева Н.Н. Об определении грамматических характеристик лексем текстов для поиска словосочетаний-терминов // Вестник Нац. ун-та "Львовская политехника" Проблемы украинской терминологии, №503, Львов, 2004, с. 40-42 (на украинском языке).

Authos' Information

Надежда Михайловна Мищенко – Киев, Украина, e-mail: nady@dolphin.icyb.kiev.ua

Наталья Николаевна Щеголева – Институт кибернетики имени В.М. Глушкова Национальной Академии Наук Украины, Киев, 03187, просп. Академика Глушкова, 40, Украина, e-mail: nat@d105.icyb.kiev.ua

RESOLUTION OF FUNCTIONAL HOMONYMY ON THE BASIS OF CONTEXTUAL RULES FOR RUSSIAN LANGUAGE

Olga Nevzorova, Julia Zin'kina, Nicolaj Pjatkin

***Abstract**: Applied problems of functional homonymy resolution for Russian language are investigated in the work. The results obtained while using the method of functional homonymy resolution based on contextual rules are presented. Structural characteristics of minimal contextual rules for different types of functional homonymy are researched. Particular attention is paid to studying the control structure of the rules, which allows for the homonymy resolution accuracy not less than 95%. The contextual rules constructed have been realized in the system of technical text analysis.*

***Keyword**: natural language processing, functional homonymy, resolution of homonymy*

Introduction

There is no common opinion on the phenomenon of homonymy in linguistic literature. The content of the phenomenon, principles of classification, classificatory schemes are being discussed. The most common classification divides homonyms into lexical ones, i.e. referring to the same part of speech, and grammatical ones, i.e. referring to different parts of speech. A term "functional homonyms" is also spread in linguistic literature. This term, offered by O.S. Ahmanova, has been accepted in the works by V.V. Babajceva [Babajceva, 1967] and her followers. V.V. Babajceva defines functional homonyms as "words with homophony, historically coming from one word family, referring to different parts of speech" and spreads the occurrence of functional homonymy not only on autosemantic parts of speech, but also on synsemantic ones. In the actual work methods of automatic resolution of functional homonymy of different types are researched.

Success of applied research in computer linguistics depends remarkably on the availability of appropriate linguistic resources, lexicographical ones being the most important. In the recent years, dictionaries of homonyms of Russian language by different authors have been published. In these dictionaries, the phenomenon of homonymy has been represented with various degree of fullness. Thus, for example, in N.P. Kolesnikov's dictionary [Kolesnikov, 1978] the phenomenon of homonymy is understood in an extended sense, homoforms, homophones and homographs being included into the circle of the occurrences studied besides lexical homonyms. Attempts on describing functional homonyms have been made in separate homonym dictionaries by O.S. Ahmanova [Ahmanova, 1984], O.M. Kim [Kim et al., 2004]. The appearance of an Internet-resource by N.G. Anoshkina [Anoshkina, 2001], attempting to gather all grammatical homonyms, stimulated further development of theoretical and applied research, including that on the classification of functional homonymy types. In the work [Kobzareva et al., 2002] a classification of 58 homonymy types is given. This classification served as a base for working out rules of contextual resolution of functional homonymy. The rules are offered in the article. In the course of research, some changes have been made to the basic classification of functional homonymy types. The changes were connected with the appearance of new subtypes, addition and exclusion of some types.

Method of Contextual Resolution of Functional Homonymy (Problems of Building up Contextual Rules)

Theoretical research on the problem of functional homonymy resolution in texts has a long history. At the end of the 50-s in works by K.E. Harper [Harper, 1956], A. Caplan [Caplan, 1955], studying and describing contextual conditions in which some or another meaning of the word would be realized was accepted as the main way of homonymy resolution. Either the surroundings of the word in the text or the words with which the word given was used would be implied under "context". The question of minimal resolving context was also actual for the researches. Noteworthy are the results obtained by A. Caplan [Caplan, 1955] in his research on minimal resolving context. About 140 frequently used polysemous English words (lexical homonyms mainly) in different contextual surroundings have been analyzed in the work.

The following types of contexts have been selected:

1). Combination with the preceding word – P1.
2). Combination with the following word – F1.
3). Combination with both preceding word and following word – B1.
4). Combination with two preceding words – P2.
5). Combination with two following words – F2.
6). Combination with two preceding words and two following words – B2.
7). The whole sentence – S.

The main conclusion implied that B1 chain was more productive as regards the effect of reducing polysemy (the proportion of the quantity of word meanings in concrete context to their number in zero context) than P2 and F2 contexts and would be almost equal to the effect given by S. Another conclusion emphasized the importance of material context type. That is, whether autosemantic parts of speech or so-called "particles" (including prepositions, conjunctions, auxiliary verbs, articles, pronouns, adverbs like *there* etc) exist in the direct surroundings of the word. The context with autosemantic parts of speech gives much better results than that with "particles". General conclusions by A. Caplan imply that the most useful context is the one consisting of one word to the left and one word to the right from the word given. If one word of the context is a particle, the context should be extended to two words on both sides.

Despite numerous references to the results quoted above in Western literature of the 60-s, their practical usage for Russian language in real contexts is hardly possible. The real situation with the resolution of functional and lexical homonymy in Russian language is far more complicated and cannot be resolved on the basis of simplified rules. In the work [Kobzareva et al., 2002], a special dictionary of diagnostic situations (DDS) has been developed. These situations assign linear structure description of the minimal context necessary for the identification of the homonym's part of speech.

DDS situations consist of two parts:

- component chain of the sentence (may be discontinuous) marks the situation;

- conditions put on the markers and their surroundings, defining the meaning of the homonym.

While analyzing the examples given in the article [Kobzareva et al., 2002], it is remarkable that the borders of the minimal resolving context become moveable. Symbols belonging to a certain multitude (varying for different homonymy types) act as borders. Besides, the authors allow for discontinuous contexts. This complicates the situation even further. Besides, yet another thing should be pointed out. With a set of rules used for the resolution of some homonymy type one of the main problems is building up a control structure for the order of the usage.

Saying it another way, the **method of contextual resolution of functional homonymy** offered by the authors of the present article includes:

- Establishing a full classification of functional homonymy types.

- Selecting the minimal set of resolving contexts for each type.

The minimality of the set means that for every functional homonymy type it should be estimated, how difficult it would be to recognize each part of speech belonging to the type. Then the set of resolving contexts (SRC) with minimal difficulty of resolution should be built up. The algorithmic form of this demand will look as follows:

If a rule from SRC has been applied to a functional homonym of T1 or T2 type, then the type of the homonym is defined by the rule applied, else the alternative type is given to X.

- Building up a control structure for the SRC, allowing for the maximum resolution accuracy.

Classification of the Functional Homonymy Types

The ground for basic classification of functional homonymy types was the classification offered in the article [Kobzareva et al., 2002]. It had been built up on the basis of N.G.Anoshkina's dictionary of grammatical homonyms [Anoshkina, 2001]. The classification proposes 58 types of homonyms, the first ten types being the most numerous (the whole number of homonyms is 2965). Less than 5 homonyms are included into each of 26 other types. In spite of the undoubted importance of the classification, it requires further development. The latter concerns the addition of new types as well as additional subdivision of types and creation of subtypes.

For example, a new type "short form of adjective # category of state" like *gotovo (*in Russian*)/ ready*. Second kind of expansions is connected, for instance, with three subtypes in **Vf/N*** type, where **Vf** stands for a verb form and **N*** ={**N** - a noun, **Npr** **-** pronominal noun}.

The example in Russian is:

1) bereg *2) bereg (inf. berech')*

N - a brink *V, Past Tense - saved*

The subtypes have been selected on the basis of a syntactic criterion, i.e. the recognition of their representatives requires working out remarkably different rules and control structures (the examples are in Russian):

<**Vf/N***>$_1$: *pojmu* = 1) N - *bottom-land*; 2) V, Future Tense - *I will understand*;

<**Vf/N***>$_2$: *l'nu* = 1) N – *flax*; 2) V, Present Tense - *I am clinging to sth*;

<**Vf/N***>$_{3.1}$: *stali* = 1) V, Past Tense - *we became*; 2) N - *steel* ;

<**Vf/N***>$_{3.2}$: *zarosli* = 1) V, Past Tense - *became overgrown*; 2) N – *bush.*

Rules of Resolving Functional Homonymy of Some Types (Realization)

In this part, an example of resolving functional homonymy of type **N*/A*** , here **N*** ={**N** =a noun, **Npr**=pronominal noun}, **A***={**A**=full adjective form, **Av**=participle, **Apr**=pronominal adjective}. Let the following system of denotating object classes be introduced:

X – functional homonym; **P**– preposition; $\underline{Con}^{*YX}_{XY}$ – general form of government model, where either **X** controls **Y** or **Y** controls **X**. For each **X** special government models are given. Actually, they define a set of rules regarding stable syntactic connections of **X**. The term $X_{\bigcap_{pgn} N^*}$ means that X is complied with **N*** by the grammatical characteristics given (p – case, g – gender, n – singular/plural).

A **Z** may be present in the rule. It means that an inserted construction of some type may be present here. In Russian language, the discontinuance of a linear sequence because of such constructions is possible almost everywhere in the sentence. We have used a typology of such constructions that allows identifying them. For example, constructions expressing feelings and emotions – *k schast'ju (in Russian)/ fortunately* –, one expressing the degree of trustworthiness – *nesomnenno (in Russian)/ assuredly* – etc. The recognition of them is a major point, as punctuation marks belong to the multitude of context restrainers. Consequently, recognizing the function of the punctuation mark is important. That is, we should know the one separating an inserted construction from all the others, as it is not considered a context constraint.

While working out the rules of resolution of functional homonymy of different types, we have offered a method of multiplex resolution of functional homonymy. It allows resolving groups of homogeneous homonyms. These rules are particularly typical for some certain types of homonymy, like N*/A* type. An example of contextual rules for this type follows, with rules 1 and 4 using multiplex resolution. Rules testing was carried on the text group from Moshkov's library (http://www.aot.ru/search1.html).

Let us exemplify the group of contextual rules for **N*/A*** type. The minimal set of resolving contexts was selected for resolving a homonym as **A***. The control structure of the general rule defines the order of usage and the results (in rectangles on the scheme).

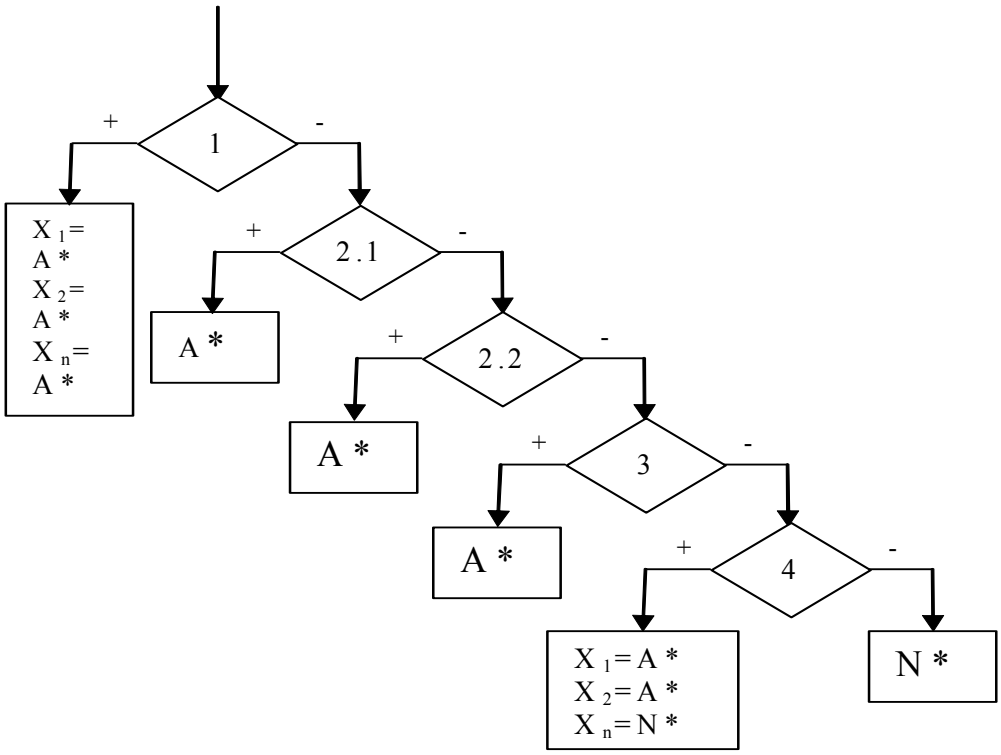
(1) $if[X_{1\bigcap_{pgn} N^*} X_{2\bigcap_{pgn} N^*} \cdots X_{m\bigcap_{pgn} N^*} N^*]__then__X_1 = X_2 = \cdots = X_m = A^*$

(2) $if[X \frac{(Z)\overline{P}N^*_{\bigcap_{pgn} X}}{\le 4/,}]/[\frac{N^*_{\bigcap_{pgn} X}(Z)\overline{P}}{\le 2/,} X]__then__X = A^*$

(3) $if[\underline{Con}^{*YX}_{XY}]__then_X = A^*$

(4) $if[X_{1\bigcap_{pgn} X_n} X_{2\bigcap_{pgn} X_n} \cdots X_n]__then__X_n = N^*, X_1 = X_2 = \cdots = A^*__else_X = N^*$

Picture 1. Rules of resolution of functional homonymy of N*/A* type



Picture 2. The control structure of the general N*/A* rule

Conclusion

Program realization of presyntactic processing module of technical texts analysis system is currently being completed. The module is to comprise the process of resolution of functional homonymy. Real technical texts, however, contain not all of the types. There are no types with interjections, for example, or with stylistically substandard words. Other limitations are connected with the vocabulary of technical texts itself (there are very few imperative verb forms, for instance). Nevertheless, contextual rules allow such verbs to be recognized. Module setting operations allow amount of homonymy types liable to resolution to be changed.

Testing of the program module for the resolution of functional homonymy has given good results on the types realized. For some types, the accuracy of resolution is 100%, in the worst cases it is not less than 95%. The reasons for erroneous situations' appearance are accidental concord in the context analyzed, context insufficiency or resource insufficiency (absence of a case frame dictionary for different parts of speech). Some mistakes in the resolution can be sorted out in the course of further analysis.

Acknowledgements

The work has been completed with partial support of Russian Foundation Basic Research (**grant № 04-06-87501**).

Bibliography

[Ahmanova, 1984] A.S. Ahmanova Slovar omonimov russkogo jazyka. M., 1984. (In Russian).

[Kolesnikov, 1978] N.P. Kolesnikov Slovar omonimov russkogo jazyka. Tbilisi, 1978. (In Russian)

[Anoshkina, 2001] J.G. Anoshkina Slovar omonimichnyh slovoform russkogo jazyka. M: Mashinnyj fond russkogo jazyka Instituta russkogo jazyka RAN, 2001. (In Russian) (http://irlras-cfrl.rema.ru:8100/homoforms/index.htm).

[Caplan, 1955] A. Caplan An Experimental Study of Ambiguity and Context // Mech. Translation, vol. 2, No 2, Nov. 1955.

[Harper, 1956] K.E. Harper Contextual Analysis // Mech. Translation, vol. 4, No 3, Dec. 1956.

[Kobzareva et al., 2002] T.U. Kobzareva, R.N. Afanasiev Universalnyj modul predsintaksicheskogo analiza omonimii chastej rechi b RJa na osnove slovarja diagnosticheskih situacij // Trudy mezhdunar. konferencii Dialog'2002. M., 2002. S. 258-268. (In Russian).

[Kobzareva et al., 2002] O.M. Kim, I.E. Ostrovkina Slovar graamaticheskih omonimov russkogo jazyka. M., 2004. (In Russian).

[Babajceva, 1967] V.V. Babajceva Perehodnye konstrukcii v sintaksise. Voronezh, 1967. (In Russian).

Author's Information

Olga Nevzorova – Research Institute of Mathematics and Mechanics, Kazan State Pedagogical University, Kazan, Russia; e-mail: olga.Nevzorova@ksu.ru

Julia Zin'kina – Kazan State University

Nicolaj Pjatkin – Research Institute of Mathematics and Mechanics, Kazan, Russia; e-mail: nikolaip@mail.ru

INFORMATION PROCESSING IN A COGNITIVE MODEL OF NLP

Velina Slavova, Alona Soschen, Luke Immes

Abstract*: A model of the cognitive process of natural language processing has been developed using the formalism of generalized nets. Following this stage-simulating model, the treatment of information inevitably includes phases, which require joint operations in two knowledge spaces – language and semantics. In order to examine and formalize the relations between the language and the semantic levels of treatment, the language is presented as an information system, conceived on the bases of human cognitive resources, semantic primitives, semantic operators and language rules and data. This approach is applied for modelling a specific grammatical rule – the secondary predication in Russian. Grammatical rules of the language space are expressed as operators in the semantic space. Examples from the linguistics domain are treated and several conclusions for the semantics of the modeled rule are made. The results of applying the information system approach to the language turn up to be consistent with the stages of treatment modeled with the generalized net.*

Keywords*: Cognitive model, Natural Language Processing, Generalized Net, Language Information System*

Introduction

Natural language processing (NLP) is a complex cognitive function, representing a complicated subject for modelling and formal description. The trial in this wok is to elaborate a reliable formal model of NLP in order to propose a tool for analyzing the process and to examine the possibilities for further implementations.

The formal model of NLP, presented here, is elaborated using a cognitive science approach. The intention is to take into consideration as much as possible the essential cognitive principles that most cognitive scientists agree with:

1). The mental system has a limited capacity - the amount of information that can be processed by the system is constrained.

2) A control mechanism is required to oversee the encoding, transformation, processing, storage, retrieval and utilization of information.

3) The constructing of meaning is a dynamic process resulting of a two-way flow of information – the flow, gathered through the senses (Bottom-up processing) and the flow of the information, which is stored and classified in the memory (Top-down processing)[1].

4) The human organism has been genetically prepared to process and organize information in specific ways. In the further description we'll consider that all these functions are performed within a system, called *"cognitive system"*.

1 Concerning the cognitive aspects of language processing, the constraints on linguistic performance come mostly from the top-down information processing.

A formal model of NLP has been elaborated, using the mathematical formalism of Generalized Nets (GN). The obtained Net, called AGN (Figure 1) gives a formal description of the cognitive process of treatment of a language message, arriving on the input of the auditory system (Bottom), processed stage by stage and conduced to the mind (Top) through several parallel pathways. AGN[1] corresponds to the cognitive system, providing a control mechanism for overseeing the transformation, processing, storage and retrieval of information.

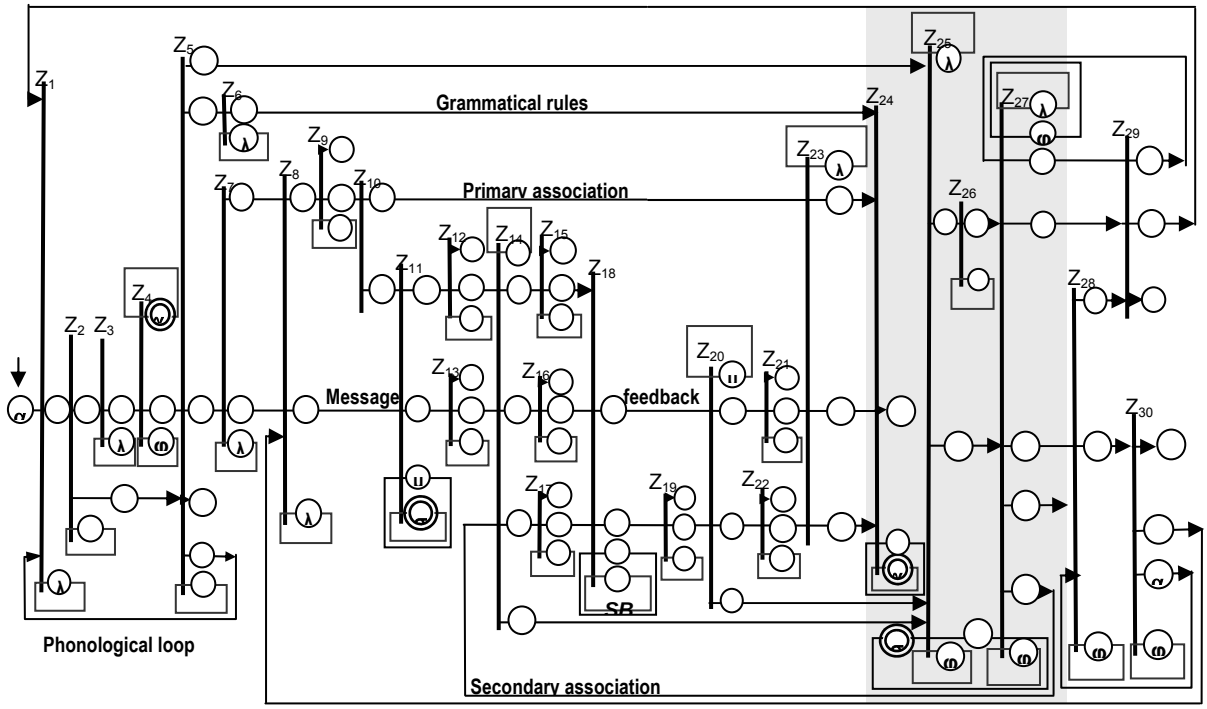


Figure 1. AGN - Generalized Net model of process of message acquisition

AGN treats language message, consisting in sentence-fragments, formally expressed as a sequence of α-tokens. Each α-token, travelling from the input to the final position of the net, is submitted to consequent treatments, performed on the transitions Z_i. AGN transitions $Z_1 – Z_{29}$ imitate the phases of the process of speech perception. The information, obtained by the system, is formalized as α-token's characteristics, which are acquired when crossing the transitions. The net al.ows modelling the interaction between the Bottom-up and Top-down information flows. The Top-down information flow assigns new characteristics to the "travelling up" signal. To imitate the parallel "emergence" of the gathered information of different type, the α-token splits (see for example Z_5), follows different pathways and terminates by fusing all obtained characteristics into an internal lexical and semantic representation of the message content.

The Top-knowledge, stored in Long Term Memory (LTM) is organized in two related spaces[2] – the language space (as a system of lexical units and rules) and the semantic space (the semantic representation of the world as a system of semantic primitives and rules). They have respectively two underlying structures - the word-forms graph WG, expressed by the γ-token, and the semantic net NSet, expressed by the σ-token. The result of the

[1] AGN has been presented in a series of papers in the domain of information technologies and cognitive modelling in linguistics (see for example [Kujumdjieff, Slavova, 2000]. Here we follow the numbering and the names, accepted in the complete formal description of the net, given in [Slavova, 2004].

[2] Most cognitive researchers agree on the different nature of the language knowledge and the conceptual knowledge, including their separate localization on the cortex.

Top-down flow is stored in WG as "*expectation*" of word-forms[1]. Four sources of expectation are modeled. Two are related to the language - the memorized language practice, called "*primary association*" and the knowledge of *grammatical rules*. Two other sources of expectation are due to semantic activation: the listening-comprehension *message feedback* (caused directly by the word-forms in the message) and the "*secondary association*" (semantic activation, accumulated because of the sequence of the message word-forms). Tokens γ and σ are thought as structures, which elements accumulate expectation/activation. AGN has access to each of them on two places - one 'retrieval' place (Z_4 and Z_{11}) and one place for storing expectation/activation (Z_{24} and Z_{25})

The knowledge of language is represented by a number of λ-tokens, each skilled with a treatment procedure as characteristic. They are on transitions: Z_1 - with: "Segmentation procedure *Seg*"; Z_3 - with: "Phonemes recognition procedure *Rec*"; Z_6 - with: "Grammatical features and dependencies procedure *Gr*"; Z_7 - with: "Lexeme representative retrieval procedure *InL*"; Z_8 - with: "Primary association procedure *Ass1*"; Z_{23} - with: "Lexeme members retrieval procedure *Lexemize*"; Z_{25} - with: "word-forms concordance procedure *TreeBranches*"; and Z_{27} - with: "Syntax structure discovery procedure *Parse*". The cognitive processes are expressed by φ-tokens, and the "mental dictionary" (relations on WG x NSet) - by μ-tokens. All this tokens, with procedures as characteristics, are 'turning' over the corresponding transitions during the time-period of AGN functioning.

Initially, token α enters the net with characteristics "Phonological features". At transition Z_1, a λ–token "Language knowledge – prosody" skilled with the Segmentation procedure transforms the input to a "Sentence segmented into word-form segments", given to the α-token as characteristics. Transition Z_2 simulates auditory sensory memory. Transition Z_3 corresponds to the stage of phoneme recognition. Transition Z_4 corresponds to the comparison of the recognized phonetic content with the lexical knowledge retrieved from WG. Transition Z_5 accepts or rejects the retrieved word-form for further treatment (AGN is supposed to identify the input word-forms using the *expectation,* gathered in the units of WG). Transitions Z_6 to Z_{24} represent a Working Memory (WM) Sub-Net[2] where the two information flows meet[3] and generate expectation. Multiple transitions in this part are connected to LTM-tokens, producing lists of LTM knowledge (lexeme representatives, synonyms, homonyms, concepts, attributes etc.) The sub-net imitates limited WM resource and 'concentration' on the message content by retaining only the heads of the lists, sorted following the accumulated activation. On Z_{24} the expectation from all pathways is overlapped and stored in the elements of WG (see also figure 2, token γ on place l_{82}). Transition Z_{25} simulates the activation of NSet by the message word-forms. Transition Z_{26} is a working memory for lexical units. Transition Z_{27} expresses the process of analyzing the entire sentence. Transition Z_{28} simulates the extraction of basic semantic roles (with a feedback from the message memory content). Transition Z_{29} simulates rechecking (if the semantic roles can not be properly discovered). The α-token is finally stored with all obtained characteristics in the message memory, represented on transition Z_{30}.

The Generalized Net approach has allowed formalizing, on a high level of abstraction, the cognitive process of message acquisition. This representation allows incorporation of sub-nets and separate modules, such as databases, neuron nets etc. Such an approach starts to be used in hybrid nets in AI [see Atanassov, K., 1998].

The Problem – Parallel Language and Semantic Treatment

A big part of the treatment procedures, introduced on AGN transitions, such as *"segmentation procedure"* or *"expectation dependent retrieval from WG"*, are easy to be imagined. The problem is to conceive the procedures, which run on the transitions after the WM sub-net. The tracking of the cognitive process has lead to base them on semantic and language knowledge at once. The presented work gives further development of AGN by analyzing the procedures, which perform simultaneously in the language and the semantic space.

[1] It is known that the capacities of speech perception do not allow capturing all the pronounced phonemes. In fact, the cognitive system constructs the missed, but 'expected' content of the message. The same top-down phenomenon is available when reading texts.

[2] This sub-net is presented in details in [Slavova, Atanassov, 2004).

[3] According to the most part of the existing in cognitive science theories and models of memory, the Top-down and the Bottom-up flows meet using Working Memory resources.

Transition Z_{25} (figure 2) simulates two processes, which run in parallel. The first is the activation of the semantic space by the message word-forms W and corresponds to building a mental image of W in terms of concepts and features. The second one is the detection of related words in the sentence and occurs at the moment when the grammatical features **and** the semantic image of W are discovered. It is supposed that the cognitive system first assembles a fractional representation of the sentence-meaning structure (coupled words for example) by consulting the semantic net for incompatibilities, as the grammatically determined word-chains have to be coherent with the meaning of the corresponding concepts and/or features. The formal description of Z_{25} is:

Z_{25} = < {I_{12}, I_{49}, I_{69}, I_{70}, I_{83}, I_{89}}, {I_{83}, I_{85}, I_{87}, I_{88}, I_{89}},

	I_{83}	I_{85}	I_{87}	I_{88}	I_{89}
I_{12}	*false*	*true*	*true*	*false*	*false*
I_{49}	*false*	*false*	*true*	*true*	*false*
I_{69}	*false*	*false*	*false*	*true*	*false*
I_{70}	*false*	*false*	*true*	*false*	*false*
I_{83}	*true*	*false*	*true*	*false*	*false*
I_{89}	*false*	*false*	*false*	*true*	*true*

, $\vee(\wedge(I_{70}, I_{83}, I_{12}), \wedge(I_{89}, \vee(I_{49}, I_{69})))$>.

The α-token enters transition Z_{25} with the following characteristics, coming from:

position I_{12} – W, word-form assumed to be perceived (on transition Z_5) with its grammatical feathers GrFtrs;

position I_{49} - Nt ct (from the message feed back pathway) – the head of the list of semantic net elements NSet, which correspond to W (the correspondence is found on Z_{11} using the mental dictionary - μ-token);

position I_{69}, - NtSBlist - the first n of list of NSet elements, which correspond to the received up to the moment W - s, arranged following the number of their manifestations in a semantic buffer SB (secondary association).

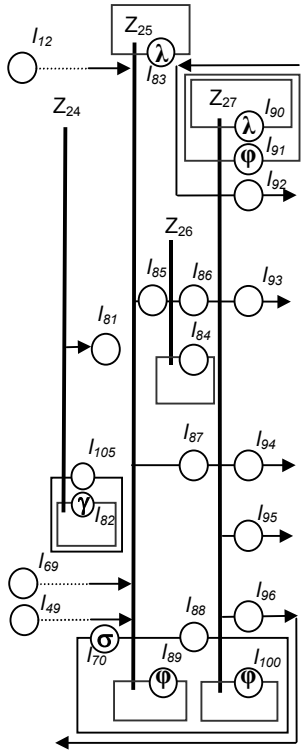


Figure 2. Two of the transitions, based on language and semantics

On transition Z_{25}:

λ-token "Language knowledge - syntax and grammar" turns on place I_{83} with:

"word-forms concordance - Procedure *TreeBranches* ";

σ-token, the Semantic net NSet stays on place I_{70}:

"Semantic net elements – NSet";

φ-token "Cognitive process - semantic activation" turns on place I_{89} with

"Storage of activation in NSet nodes - Procedure *SemA*";

After Z_{25} , on place I_{88} σ-token takes the characteristic:

"ANet = *SemA* (Nt ct, NtSBlist) – activation of NSet elements"

Token α obtains in place I_{87} the characteristic:

"ParSynStr = *TreeBranches* (NSet, GrFtrs) – Partial syntax structure."

Tokens do not change their characteristics on places I_{83}, I_{85} and I_{89}.

Transition Z_{26} is WM buffer for W, queued on I_{84} and transmitted to I_{86}.

On transition Z_{27}, the following procedures are running:

λ-token "Language syntax knowledge" stays on place I_{90} with:

"Syntax structure discovery - Procedure *Parse*";

φ-token "comparing semantics and syntax" stays on I_{91} with:

"Comparing – Procedure *Comp*";

φ-token "focus determination" stays on place I_{100} with:

"Semantic center localization - Procedure *DetSC*".

Transition Z_{27} expresses the mental process of analyzing the entire sentence after its last word-form has been perceived. It is assumed that two parallel processes take place at this time-moment: the sentence syntax structure is clarified and the semantic focus Nt^1 of the sentence is detected (NSet element, staying higher in NSet's hierarchy). The brought by α-token information, acquired before entering Z_{27}, consists of: partial syntax representation, word-forms W in the lexical buffer content and activation of the corresponding nodes of NSet. It is supposed that the syntax structure of the sentence is recognized with semantic justification.

$Z_{27} = <$ {l_{86}, l_{87}, l_{88}, l_{90}, l_{91}, l_{99}, l_{100}}, {l_{70}, l_{90}, l_{91}, l_{92}, l_{93}, l_{94}, l_{95}, l_{96}, l_{100}},

	l_{70}	l_{90}	l_{91}	l_{92}	l_{93}	l_{94}	l_{95}	l_{96}	l_{100}
l_{86}	*false*	*false*	*false*	*true*	*true*	*true*	*false*	*false*	*false*
l_{87}	*false*	*false*	*false*	*false*	*true*	*true*	*false*	*false*	*false*
l_{88}	*true*	*false*	*false*	*false*	*true*	*true*	*true*	*true*	*false*
l_{90}	*false*	*true*	*false*	*false*	*true*	*true*	*false*	*false*	*false*
l_{91}	*false*	*false*	*true*	*false*	*true*	*true*	*false*	*false*	*false*
l_{99}	*false*	*false*	*false*	*true*	*true*	*true*	*false*	*false*	*false*
l_{100}	*false*	*false*	*false*	*false*	*false*	*false*	*true*	*true*	*true*

$, \vee(\wedge(\vee(\ l_{86}\ ,\ l_{99}\ ),\ \wedge(l_{88}, l_{90}, l_{91})),\ \wedge(\ l_{88}, l_{100})) >$

In places l_{93} and l_{94} the α-token obtains the characteristic:

"TRes = *Comp* (*Parse*(Buff, ParSynStr), NSet) - Obtaining complete syntax structure"

and in places l_{95} and l_{96} the α-token obtains the characteristic:

"Nt^1 = *DetSC* (NSet, ANet) – momentary semantic center"

Tokens do not change their characteristics on places l_{90}, l_{91}, l_{92}, l_{70} and l_{100}. The procedure *Comp* may have two results: 1. TCF (Tree Construction Failed); 2. TREE+ WsC (Syntax tree with W corrected - WsC).

The procedure *Comp* (*Parse*(Buff, ParSynStr), NSet) on Z_{27} is applied on the retained in a STM buffer for W and on the result of *TreeBranches* (NSet, GrFtrs), running on Z_{25}. Both procedures are based on NSet and on the grammatical features of W (discovered on Z_4).

The Language as an Information System

The generalized net model suggests that we have to formalize Z_{25} and Z_{27} through procedures, running by using in parallel semantic and language knowledge and related to the elements and to the rules in the two knowledge spaces. A sequence of questions appears concerning the simultaneous operations in the two spaces. In AGN, the correspondences between them are given by the 'mental dictionary', containing the relations between the lexical units in WG and the elements of the semantic net NSet, but the structures of the two spaces are independent. Are there rules that allow joint operations in the two spaces? Are they established on principles for mapping the structures of the two spaces? That needs to examine the human language in a general way.

Let us present human language as an Information System (IS) – a Language Information System (LIS). One of the primary goals of a human language is to assure the information exchange between individuals. Information, residing as *internal cognitive representation* of the individual H_1 is first presented as language-coded information, communicated to another individual H_2, and interpreted to *internal cognitive representation* of individual H_2. It is intriguing to provide the example of one home-made sign language, created and utilized by two deaf sisters. The used pointing gestures were found to be part of lexical terms referring to present and non-present objects, persons and places. Some gestures occupied fixed positions in sentences, apparently used as grammatical terms. Oral movements were frequently used together with manual signs, and their functions may be classified as lexical, adverbial, and grammatical (Torigoe et al., 2002). This strongly suggests the existence of the *innate* mechanism for mental representations (Hauser 2002, Chomsky 2004).

Imagine we have to construct a LIS. Let us apply the used in the technology domain procedure. For conceiving an IS, the representation "*input - treatment block – output"* is used. On its input, an IS receives *data* and *resources* and on its output - obtains *informative products*. The treatment block runs on the bases of a particular *model and method for data-processing,* which includes a number of *rules* and *operators* on data. So:

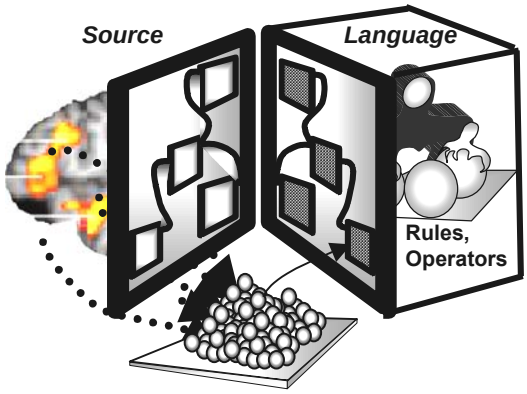


Figure 3. The language, seen as information system

1. The *resources* of LIS are the human cognitive resources – *static* (long term memory and working memory), and *operational* (operators that human mind performs on the available operable substances).

2. Data. The mind-operable content of the data-source (figure 3) has to be transmitted as data, operable in language. The functioning of the data-source has to be presented as a model, a system of elements with determined roles, reproducing how the cognitive system operates when performing its tasks. Let us call the components on this model "semantic primitives". A structure of data-containers has to be assembled in the language and matched to the structure of the semantic primitives. Data-values have to be accorded to all distinct entities, available and operable in the source, and stored in the corresponding data-containers. This approach is well-known in the IS domain (see for example Codd, 79).

3. The *treatment block* requires conceiving a set of *rules* and *operators* on data. This set has to allow generating larger units, reproducing the processing in the source. The cognitive system will operate data following these rules, so they must be expressible by means of the operators, which run on the semantic primitives.

The *final product* of the language has to create an accurate internal representation of individual H_2, who receives the LIS output. The 'decoding' is done with the active participation of H_2 (the 'Top-down" information processing), which presupposes that H_2 knows the language **and** possess a semantic description of the world.

The information system reasoning shows that the accurate functioning of a LIS strongly relies on the internal semantic representation. The language is constructed on the bases of the semantic primitives and the mind operators on them. The claim is that in all languages must exist interactions between the purely language features and their semantic fundamentals. This could give a basis for the joint semantic-language operations, leading to a solution of the problem concerning the procedures on Z_{25} and Z_{27}. The next step is to examine the relationships between the grammatical level and the semantic one, starting from concrete working examples.

Cognitively Based LIS on the Example of Russian

We assume that there is a common general underlying semantic scheme for all languages. Then it will follow that any grammatical rule can be represented as consisting of some semantic primitives as internal representations, which are mind-operable. We did a trial to show how LIS operates on the example of a concrete syntactic representation in one specific language. Secondary predication is a grammatical particularity in Russian, allowing variability of case marking (Instrumental/Nominative) on secondary predicates. Example:

(1) a. Maria prišla ustalaja-nom.

b. Maria prišla ustaloj-instr.

Mary arrived tired.

In Russian, we can make the notion of arriving tired, simultaneously being contrasted with some non-tired state, using the instrumental case. In the nominative case, we only know that 'she arrived tired', with no past reference to any other possible state. The semantics of Russian secondary predication has been examined a lot by the specialists in linguistics and the obtained results and explanations are not uniform.

We have followed the LIS reasoning and we constructed a database (DB) in which exists simultaneously the language level and the semantic level, with their structures, interconnected.

Examples of statements, taken from the linguistics studies of secondary predication, (53 sentences) wore stored in the table Examples. The cases in Russian are used as markers for the grammatical annotation of the examples. Data, expressing the language and semantics spaces, have been organized in tables as follows (figure 4): Table Objects stores all ***concepts*** from the examples, with their names in different languages. The ***attributes***

- characteristics or possible states, are stored in a separate table. The verbs with their grammatical features are stored in the table 'Verbs' (they are introduced in the field "Matrix Verb" of the table Examples as foreign key). The ***events***, providing the underlying semantic features of the verbs, are expressed as attributes of the verbs (foreign key). Events are stored separately with their semantic features following some of the existing classifications. With this construction of the DB, the grammatical rules of case marking can be examined independently of the events' structure. For the purposes of this analysis, we employ a revised version of the 'event structure' (Davidson 1967, Vendler1957, Verkuyl 2001), examining the language semantics primitives.

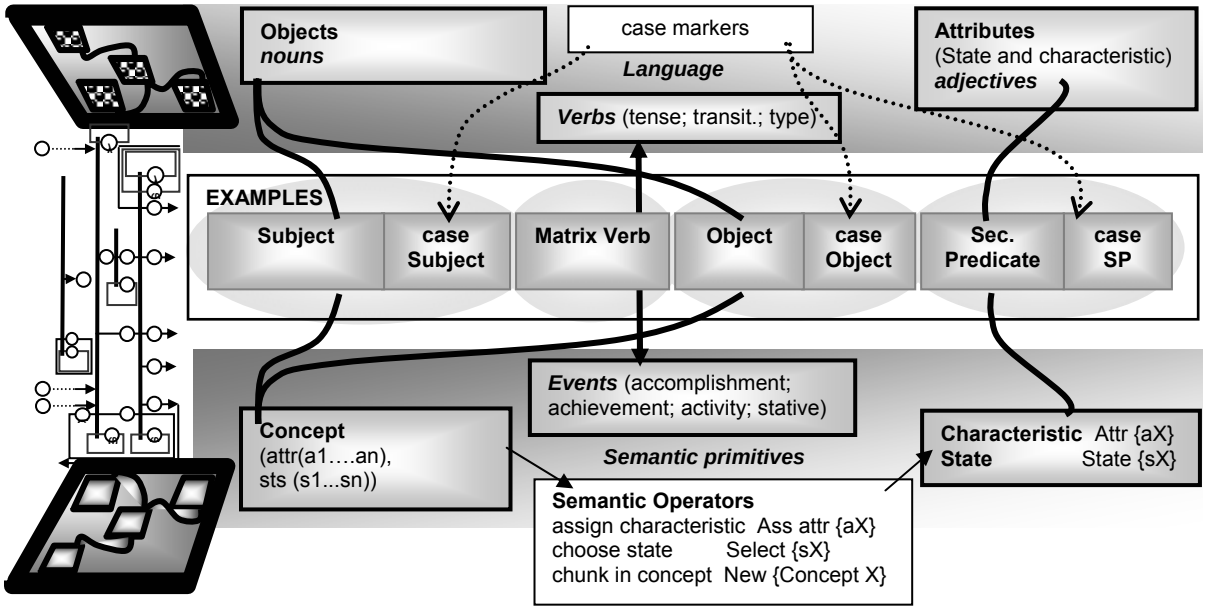


Figure 4. Language – semantics database design

The statements are assigned two levels of representation - phases of semantic-levels-translation: The first step accords to a lexical item its basic semantic category. The categories that we used are *concept*, *characteristic*, *state* and *event.* We assume that the grammatical level, expressed by means of case markers, implies running of semantic operators. For our examples we took as basic operators the following set: "*assign characteristic*: Ass attr {aX}", "*choose state:* Select {sX}" and "*chunk in concept*: New {Concept X}".

The second representation of the statements gives the result of applying the semantic operator. Using queries over the modeled in the DB parameters, we checked several guesses about the semantic interpretation, coming from the linguistics domain. For example, as the running of queries over the events characteristics does not give any indication of changing states, so the conclusion is that the meaning of the matrix verb events is not influenced by the case marking. It is interesting to see the result of queries, which put together (taking data from the corresponding tables) all language labels (Russian and English in Table 1), the semantic markers and the verb-event information. Here are a few of the examples for transitive verbs (*Re* - the results of the first phase of semantic-levels-translation, *Se* – the semantics of the sentence, obtained after the second phase):

Table 1

	13a.			
Ex	**Ja/-nom**	**Pokupaju**	**banany/-acc**	**spely/-instr**
	I/-nom	*Buy*	*bananas/-acc*	*ripe/-instr*
Re	**Ja (concept)**	**Pokupaju**	**banany (concept)**	**spely (state)**
	I (concept)	*Buy*	*bananas (concept)*	*ripe (state)*
	attr(a1….an), sts (s1….sn)		attr(a1….an), sts (s1….sn)	
Se	**Ja (concept)**	**Pokupaju**	**banany-spely (selected state)**	
	I (concept)	*Buy*	*bananas-ripe (selected state)*	
	attr(a1….an), sts (s1….sn)		in state {sX}	
	Activity	(<…,<sn,sn+k,k=1>,…>)	'I buy bananas ripe.'	

	16a1			
Ex	**Don/-nom**	**Pišet**	**pis'mo/-acc**	**ustal/-instr**
	Don/-nom	*Writes*	*letter/-acc*	*tired/-instr*
Re	**Don (concept)**	**Pišet**	**pis'mo (concept)**	**ustal (state)**
	Don (concept)	*Writes*	*letter (concept)*	*tired (state)*
	attr(a1….an), sts (s1….sn)		attr(a1….an), sts (s1….sn)	state {sX}
Se	**Don-ustal (selected state)**	**Pišet**	**pis'mo (concept)**	
	Don-tired (selected state)	*Writes*	*letter (concept)*	
	in state {sX}		attr(a1….an), sts (s1….sn)	
	Activity	(<…,<sn,sn+k,k=1>,…>)	'Don writes letter tired.'	

It come out that the use of a canonical underlying semantic scheme of events, objects, states and attributes explains the semantics of the grammatical rule of secondary predication in a clear way:

The case marking of the secondary predicate implies meaning of a "*choose state:* Select {sX}" operator in the case of instrumental and an "*assign characteristic*: Ass attr {aX}" operator in the case of nominative. The grammatical rule of secondary predication is applied to concepts and plays a role of a "choice of state" operator without influencing the structure of the matrix verb's event. The semantic-level interpretation of all statements shows that the event structure plays its role for the meaning of the sentences in an independent way.

This clear representation of secondary predication may be implemented in several ways. We constructed a Neuron Net (figure 5), which performs the treatment of secondary predication grammatical rule, as the aim is to include further the treatment of other rules.

Our task is to model the more explicit Russian syntax. So, we choose our essential cognitive features, with just a sufficient enough neural network. The nodes of our neural network are conventional symbols: nouns, adjectives, adverbs or verbs. Activation of the node occurs when a sufficient threshold value is reached. The sum is used for AND operation, and 1 is use for OR operation. Initially, all inputs are OFF. There is no learning component.

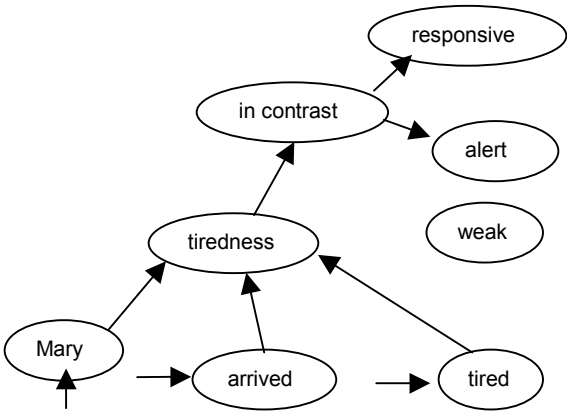


Figure 5. Neuron Net

Word order is not part of the model, but left up to the actual language. Because Russian has a more flexible word order, and the net meaning is the same for both Russian and English (as concerns ordering of words), then assume words have been entered in the right order.

The neuron net imitates: Instr. case marking on *tired* in *Mary arrived tired* triggers a state of *tiredness.* Being in a state of tiredness gives rise to an explicit state *in contrast*, then non-tired states: of: *alert*, and *responsive*.

The advantages of this representation are 1) semantics and syntax are combined; 2) knowledge engineering is much easier, including maintenance, because symbols are used, instead of dynamic numeric values. Clearly, the neuron net has to have a much richer set of semantics for a practical system.

Conclusion

The supposal that the cognitive system treats in parallel the semantic and the language knowledge space was made on the bases of two formal representations: the AGN model of the cognitive process and the representation of the language as an information system. The results of these two formal approaches are in agreement. The 'semantic' database representation of primary/secondary predication on the example of Russian was used in the

analysis of the links between basic semantic units and grammar. Grammatical rules of the language space are expressible as operators in the semantic space. Some important linguistics conclusions wore made on this base.

It is interesting to analyze the content of Table 1. It became clear that for intransitive verbs the *choice of state* operator is applied always to the state of subject, but for transitive verbs it can also be applied to the state of object. Obviously in the statement 13a the state 'ripe' can not be accorded to the subject 'I' and in 16a1 the state 'tired' is not for the object 'letter'. But these two statements are absolutely correct, they are not ambiguous and in use in both languages. From the point of view of AGN treatment, on Z_{25} the procedure *TreeBranches* has to 'attach' the word-form "ripe" to "bananas" and the word-form "tired" to "I", as it consults the semantic space NSet, where the concepts and their attributes are known. The further development of this work necessitates formalizing in a detailed way the AGN-s part Z_{25} Z_{27}, associated to LIS on order to implement this part of the model.

Bibliography

[Atanassov, 1998] K. Atanassov. Generalized Nets in Artificial Intelligence. Vol. 1: Generalized nets and Expert Systems. "Prof. M. Drinov" Academic Publishing House, Sofia

[Becker, Immes, 1987] L.A. Becker, L. Immes. An approach to automating knowledge acquisition for Expert Systems: annotated traces -> Diagnostic Hierarchies. In: ACM Conference on Computer Science, p133-137.

[Chomsky, 2004] N. Chomsky. The Generative Enterprise Revisited. Mouton de Gruyter.

[Codd, 1979] E. F. Codd. Extending the Database Relational Model to Capture more Meaning. In: ACM TODS, Vol 4, n 4.

[Davidson, 1967] D. Davidson. The Logical Form of Action Sentences. In: N. Rescher, ed., The Logic of Decision and Action, Pittsburgh University Press, Pittsburgh.

[Hauser et al., 2002] M. Hauser, N. Chomsky and W.T. Fitch. The Faculty of Language: What is it, who has it, and how did it evolve? In: Science Vol. 298.

[Kujumdjieff, Slavova, 2000] T. Kujumdjieff, V. Slavova. An association-based model of natural language comprehension. In: International Journal "Information theories & applications", Vol. 7 Number 4, 169-174.

[Torigoe et al., 2002] Torigoe, Takashi, Takei, Wataru. Descriptive Analysis of Pointing and Oral Movements in a Home Sign System. In: Sign Language Studies 2.3, Spring 2002;

[Slavova, 2004] V. Slavova. A generalized net for natural language comprehension. In: Advanced Studies in Contemporary Mathematics, vol 8, Ku-Duk Press, 131-153.

[Slavova, Atanassov,2004] V. Slavova, K. Atanassov. A generalized net for working memory and language. In: Text Processing and Cognitive Technologies, VII International conference «Cognitive modelling in linguistics», Varna, 90-101;

[Soschen, 2003] A. Soschen. On Subject and Predicates in Russian. Ph.d. Dissertation, Univ. Ottawa.

[Vendler, Zeno.,1957] Vendler, Zeno. Verbs Verbs and Times. In: Philosophical Review 66, 143-160.

[Verkuyl, H.J., 2001] H.J. Verkuyl. Aspectual Composition: Surveying the Ingredients. In: Proceedings of the Utrecht Perspectives on Aspect Conference December 2001, OTS.

Author's Information

Velina Slavova – New Bulgarian University, Department of Computer Science, 21 Montevideo str., 1618 Sofia, Bulgaria, e-mail: vslavova@nbu.bg

Alona Soschen – Massachusetts Institute of Technology, MIT Department of Linguistics and Philosophy 77 Massachusetts Ave. 32-D808Cambridge, MA 02139-4307, USA, e-mail: soschen@mit.edu

Luke Immes – Multisegment Co., 20 Williamsburg Court, suite 14 Shrewsbury, MA 01545, USA, e-mail: li@multisegment.com

4.2. Application of AI Methods for Prediction and Diagnostics

APPLICATION OF ARTIFICIAL INTELLIGENCE METHODS TO COMPUTER DESIGN OF INORGANIC COMPOUNDS

Nadezhda N. Kiselyova

***Abstract**: In this paper the main problems for computer design of materials, which would have predefined properties, with the use of artificial intelligence methods are presented. The DB on inorganic compound properties and the system of DBs on materials for electronics with completely assessed information: phase diagram DB of material systems with semiconducting phases and DB on acousto-optical, electro-optical, and nonlinear optical properties are considered. These DBs are a source of information for data analysis. Using the DBs and artificial intelligence methods we have predicted thousands of new compounds in ternary, quaternary and more complicated chemical systems and estimated some of their properties (crystal structure type, melting point, homogeneity region etc.). The comparison of our predictions with experimental data, obtained later, showed that the average reliability of predicted inorganic compounds exceeds 80 %. The perspectives of computational material design with the use of artificial intelligence methods are considered.*

***Keywords:** artificial intelligence, computer design of materials, databases on properties of inorganic materials, information-analytical system.*

Introduction

Now the search for new inorganic materials is carried out, for the most part, on the basis of the experience and intuition of researchers. The problem of *a priori* prediction of compounds that have not yet been synthesized and evaluations of their properties is one of the most difficult problems of modern inorganic chemistry and materials science. Here the term "*a priori* prediction" means predicting yet unknown substances with predefined properties from only the properties of constituent components - chemical elements or more simple compounds.

The problem of predicting new compounds can be reduced to the analysis of the multidimensional array of the property values and the column vector of the desired property. Each row corresponds to some known chemical system, whose class is indicated by the row position of the column vector. For example, the multidimensional array can include the properties of chemical elements for set of known chemical systems with formation or without formation of compound with predefined composition (desired property). The process of analyzing all this information allows finding the classifying regularity. By substituting the values of the properties of the elements for unknown chemical system into the regularity thus found, it is possible to determine the class (compound formation or non-formation). So, the problem of *a priori* prediction of new compounds can be reduced to the classical task of computer learning. The chemical foundations of material computer design are based on Mendeleev's law which asserts that the periodic nature of changes in the properties of chemical systems depends on the nature and properties of the elements which makes these systems (compounds, solutions, etc). Another premise justifying the proposed approach is the existence of good classification schemes for inorganic substances.

Databases on Properties of Inorganic Materials and Substances as a Foundations of Material Computer Design

The application of computer learning methods for finding regularities is rather put into use in case of complete and qualitative initial data. Our experience of computer learning applications to chemistry shows that the number of erroneous predictions varies proportionally with ratio of number of errors in experimental data to number of learning set to be processed and the reliability of prediction grows with an increase of initial data volume (reliability mounts to a limit with an increase of size and representatives of learning set). Consequently, the

application of the computer learning methods to chemistry implies the use of databases (DBs), containing extensive bulks of qualitative information, as a basis. With this aim in mind, we develop the DBs containing data with the qualified expert assessment. The most interesting of them are DBs on materials for electronics with completely assessed information [Kiselyova *et al.*, 2004] and an inorganic compound properties DB containing partially assessed information [Kiselyova, 2002; Kiseleva *et al.*, 1996]. These DBs are integrated. The all our DBs have Internet-access (http://www.imet-db.ru).

1). A phase diagram DB of material systems with semiconducting phases "Diagram" [Kiselyova *et al.*, 2004] contains information on physical and chemical properties of the intermediate phases and the most important Pressure-Temperature-Concentration phase diagrams of semiconducting systems evaluated by qualified experts. Now the DB contains detailed information on several tens of semiconducting systems.

2). DB on acousto-, electro-, and nonlinear optical properties "Crystal" [Kiselyova *et al.*, 2004] contains detailed information regarding substances of these types evaluated by experts. In addition, DB includes extensive graphical information about properties of the materials.

3). A DB on inorganic compound properties "Phases" [Kiselyova, 2002; Kiseleva *et al.*, 1996] contains information about thermo-chemical and crystal chemical properties on more than 41,000 ternary compounds taken from more than 13,000 publications. Some of the data have been assessed by materials experts. This DB is a main data source for material computer design.

On the one hand, the use of DBs increases the reliability of predicting inorganic substances, and, on the other hand, the application of artificial intelligence methods extends the capabilities of databases owing to the search for regularities in the known information and the use of these regularities for prediction of new substances not yet synthesized.

Artificial Intelligence Methods for Material Computer Design

The search for, and development of effective material computer design systems were aimed at the creation of more powerful programs capable of analyzing, on the one hand, very large arrays of experimental information, and, on the other hand, of allowing construction of multidimensional classified regularities under the conditions of small sets. Improvements in electronics allowed the development of systems for chemical applications with a user-friendly interface, working in real time, for example, [Gladun, 1995; Gladun *et al.,* 1995; Chen *et al.,* 1999; Chen *et al.,* 2002; Pao *et al.,* 1999]. The trend has been a transition from the simplest algorithms of pattern recognition [Gulyev *et al.,* 1973; Kutolin *et al.,* 1978; Savitskii *et al.,* 1968; Talanov *et al.,* 1981; Vozdvizhenskii *et al.*, 1973] toward more powerful methods based on the use of neural and semantic networks [Kiselyova, 1987, 1993a, 1993b, 2002; Kiselyova *et al*., 1998, 2000; Manzanov *et al*., 1987; Pao *et al*., 1999; Savitskii et al., 1979; Villars *et al*., 2001; Yan *et al*., 1994].

It must be pointed out that since there are very many aspects in the domain of the artificial intelligence no criteria for selecting the most suitable algorithm for a particular application were available. After testing many algorithms intended for computer learning applications we formulated the principal criteria of a choice of the programs of computer learning ensuring the most effective decision of chemical tasks:

- possibility of the analysis of the large data volumes;
- possibility of obtaining qualitative classifying regularities at the analysis of the small learning sets;
- automatic exception of properties, which are no important for classification;
- possibility of decision of problems in conditions of weak fulfillment of the principal hypothesis of pattern recognition - hypothesis of compactness;
- fast learning and predicting;
- possibility of analysis of properties with the gaps of some values;
- possibility of analysis of properties having a qualitative nature (e.g., color, type of incomplete electronic shell: s, p, d, or f);
- high accuracy at the decision of chemical tasks;
- convenient interface of the user.

Taking into account these criteria we fixed on the class of algorithms in which all classifying regularities to be found could be presented in the form of a Boolean expression [Gladun, 1995]. This system of concept formation represents information about known chemical systems like - growing pyramidal networks (GPNs). A pyramidal

network is an acyclic oriented graph having no vertices with one entering arc. If the processes of concept formation are determined in the network then the pyramidal network is designated as a growing one [Gladun, 1995]. GPN is built during the process of objects' input. Each object (chemical system) is put in as a set of values of the component properties with an indication of the class to which the system belongs. The nearby values of components' properties are united into one interval using a special program or the experience of a researcher. Concept formation process consists of the analysis of vertices in built network and the choice of those ones that are the most typical for each class [Gladun, 1995]. These vertices became the checking vertices. The resultant concepts (classifying regularities) can be stored in computer memory and printed or read out in the form of a learned GPN or an equivalent Boolean expression, which the values of the component properties make the variables. During the prediction process, the computer receives only the atomic numbers of the elements or designations of simple compounds, while the values of the properties of the appropriate elements or simple compounds are automatically extracted from the DB. They are substituted into the GPN and the researchers can easily obtain the necessary prediction.

Application of Artificial Intelligence to the New Inorganic Materials Computer Design

Using this approach we solved problems of the following types [Kiselyova, 1987, 1993a, 1993b, 2002; Kiselyova *et al*., 1998, 2000; Savitskii *et al.*, 1979]:

- prediction of compound formation or non-formation for binary, ternary and more complicated systems;
- prediction of the possibility of forming ternary and more complicated compounds of desired composition;
- prediction of phases with defined crystal structures;
- estimation of phase properties (critical temperature of transition to superconducting state, homogeneity region, etc.).

Shown in Table 1 is a part of the table illustrating predictions of the Heusler phases with composition $ABCu_2$ [Kiselyova, 1987]. These results were obtained in the process of searching for new magnetic materials. All 6 checked predictions agreed with the new experimental data.

Table 1. Part of a table illustrating the prediction of a crystal structure type resembling the Heusler alloys for compounds with the composition $ABCu_2$

A / B	Li	Be	Al	K	Sc	V	Cr	Fe	Co	Ni	Ga	Ge	Y	Nb	Mo	Ru	Rh	Pd
Zn			+									-	-	-	-	-	-	-
Ga	+	+		+	⊕	+	+	+	+	+		O						
In	+	+		+	©	+	+	+	+	+	+	+	©	+	+	+	+	+
Sn	O	-	O	-	-	-		⊕	⊕	⊕	-	-	-	↔	-	-	-	-
Lu	-	-		-	-	-	-	-	-	-		-	-	-	-	-	-	-
Ta		-			-	-	-	-	-	↔		-	-	↔	-	-	-	-
Au	-	-		-	-	-	-	-	-	-		-	-	-	-	-	-	↔
Tl												-						
Pb	-	-	O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Designations:

+ formation of a compound with the composition AB_2Cu and a crystal structure type resembling the Heusler alloys is predicted;
– formation of a compound with a crystal structure type resembling the Heusler alloys is not predicted;
⊕ a compound with a crystal structure type resembling the Heusler alloys was synthesized and appropriate information was used in the computer learning process;
↔ formation of a compound with a crystal structure type resembling the Heusler alloys is known from experiment and appropriate information was used in the computer learning process**;**
© predicted formation of a compound with a crystal structure type resembling the Heusler alloys which is confirmed by experiment;
O predicted absence of a compound with a crystal structure type resembling the Heusler alloys which is confirmed by experiment; empty square - indeterminate result.

In Table 2 the comparison between the results after predicting the compounds with composition ABX_2 (A and B – various elements; X – S or Se) [Savitskii *et al.*, 1979] and the new experimental data. These compounds were predicted in the process of the search for new semiconductors. Only two predictions were detected to be in error ($CsPrS_2$ and $TlEuSe_2$).

Table 3 shows the predictions of more complicated compounds - new langbeinites with composition $A_2B_2(XO_4)_3$ [Kiselyova *et al.*, 2000]. These results are importance for searching for new electro-optical materials. Of 17 checked predictions, 12 agreed with the new experimental data.

Table 2. Part of a table illustrating the prediction of compounds with the composition $A^IB^{III}X_2$

X	S							Se							
A^I B^{III}	Li	Na	K	Cu	Rb	Ag	Cs	Li	Na	K	Cu	Rb	Ag	Cs	Tl
B			©	©	©		©			+	©	+	+	+	⊕
Al			©	⊕	+	⊕	+	©	⊕	⊕	⊕	+	⊕	+	⊕
Sc	©	©	+	⊕	+	+	+	+	+	+	⊕	+	⊕	+	+
Ti	⊕	©	©	©	⊕	+	⊕	©	©	+	+	+	+	+	+
V	⊕	©	+	+	+	+	+	©	©	+	+	+	+	+	+
Cr	⊕	⊕	©	⊕	⊕	⊕	+	+	⊕	+	⊕	⊕	⊕	+	⊕
Mn	+	+	+	+	+	+	+	©	©	+	+	©	+	©	+
Fe	+	⊕	⊕	⊕	⊕	⊕	⊕	+	+	©	⊕	©	⊕	©	⊕
Co	+	+	+	+	+		+	+	+	+	+	+	+	+	©
Ni	+	+	+	⊕	+	+	+	+	©	+	+	+	©	+	+
Ga	⊕	©	©	⊕	©	⊕	©	+	+	⊕	⊕	+	⊕	©	⊕
As	+	⊕	©	⊕	+	⊕	+	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Y	©	⊕	©	⊕	+	©	+	©	©	+	⊕	+	⊕	+	©
Rh	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
In	©	⊕	⊕	⊕	⊕	⊕	⊕	©	©	⊕	⊕	©	⊕	+	⊕
Sb	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
La	©	⊕	⊕	⊕	⊕	↔	⊕		⊕		⊕		↔		↔
Ce	©	⊕	⊕	⊕	⊕		⊕	©	⊕	+	⊕		↔	+	+
Pr	⊕	⊕	⊕	⊕	⊕		⊗	+	⊕	+	⊕	©	↔	+	©
Nd	⊕	⊕	⊕	⊕	⊕		+	©	⊕	+	⊕	©	↔	+	©
Pm	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
Sm	⊕	⊕	⊕	⊕	⊕	⊕	+	©	⊕	+	⊕	©	↔	+	⊕
Eu	⊕	⊕	⊕	⊕	⊕	⊕	+	+	⊕	+	⊕	+	↔	+	⊗
Gd	⊕	⊕	⊕	⊕	⊕	⊕	+	©	⊕	+	⊕	©	©	+	⊕
Tb	⊕	⊕	⊕	⊕	©	⊕	+	©	⊕	+	⊕	©	⊕	+	©
Dy	⊕	⊕	⊕	⊕	©	⊕	+	©	⊕	+	⊕	+	⊕	+	©
Ho	⊕	⊕	⊕	⊕	©	⊕	+	©	⊕	+	⊕	©	⊕	+	©
Er	⊕	⊕	⊕	⊕	©	⊕	+	©	⊕	+	⊕	©	⊕	+	©
Tm	⊕	⊕	⊕	⊕	©	⊕	+	+	+	+	⊕	+	⊕	+	©
Yb	⊕	⊕	⊕	⊕	©	⊕	+	+	+	+	⊕	+	⊕	+	⊕
Lu	⊕	⊕	⊕	⊕	©	©	+	+	+	+	⊕	©	⊕	+	©
Tl	+	+	©	⊕	⊕	©	⊕	©		©	⊕	⊕	⊕	+	

Designations:

+ predicted formation of a compound with composition ABX_2; **-** - prediction of no formation of a compound with composition ABX_2;

⊕ compound ABX_2 is known to be formed and this fact is used in the computer learning process;

↔ compound ABX_2 is not known to be formed and this fact is used in the computer learning process;

© predicted formation of a compound with composition ABX_2 which is confirmed by experiment;

⊗ predicted formation of a compound with composition ABX_2 which is not confirmed by experiment; empty square - indeterminate result.

Predicted compounds were then searched for new magnets, semiconductors, superconductors, electro-optical, acousto-optical, nonlinear optical and other materials required for new technologies. The comparison of these predictions with the experimental data, obtained later, showed that average reliability of predicted compounds exceeds 80 %.

Table 3. Part of a table illustrating prediction of a crystal structure type for compounds with the composition $A_2B_2(XO_4)_3$

X	S					Cr					Mo					W				
A / B	Na	K	Rb	Cs	Tl	Na	K	Rb	Cs	Tl	Na	K	Rb	Cs	Tl	Na	K	Rb	Cs	Tl
Mg	L¢	(L)	(L)	(*)	L	L	L©		L©	L©	K¢	(K)	(L)	(L)	(L)		↔	(L)	L©	L
Ca	(*)	(L)	L©	(L)	(*)			L	L	L	(*)	?	?	?	?	(*)	*	?	?	?
Mn	(*)	(L)	(L)	L	(L)	L		(L)	L©	L	K	↔	(L)	(L)	(L)					
Fe	*	L©	L©		(L)	L	K	L	L	L	K	K	?	?	?		K			
Co	(*)	(L)	L©		(L)	L	K	L	L	L		(K)	(L)	(L)	↔		K			
Ni	(*)	(L)	L©	L	L	L		L	L	L	K	(K)	(L)	(L)	(L)					
Cu	(*)		L	*	L	L	K	L	L	L	K	(K)	?	?	?		K			
Zn	*	(L)	L	*	L	L	K	L	L	L	↔	(K)	(K)	-	(K)		K¢			
Sr	(*)	?	?		(*)	*	*	?	?	?	(*)	?	?	?	?	(*)	*	*	*	*
Cd	(*)	(L)	(L)		(L)							K	↔	(L)	K¢	(*)		L	L	L
Ba	(*)		(*)	(*)	(*)	*	*				(*)		*	*	*	*	*			
Pb					*	(*)	*	*©	*	*	(*)	*©	(*)	*¢	*	(*)	*	(*)	(*)	*

Designations:

L formation of a compound with the langbeinite crystal structure type is predicted;

K formation of a compound with the crystal structure type $K_2Zn_2(MoO_4)_3$ is predicted;

– the crystal structure differing from those listed above is predicted;

(L),(K) a compound with corresponding type of crystal structure was synthesized and appropriate information was used in the computer learning process;

↔ a compound with the crystal structure differing from those listed above does not exist at normal conditions and this information was used in the computer learning process;

(*) a compound $A_2B_2(XO_4)_3$ is not formed and this fact was used in the computer learning process;

¢ predicted formation of a compound with this structure type which is not confirmed by experiment;

© predicted formation of a compound with this structure type which is not confirmed by experiment; empty square - indeterminate result.

Information-Analytical System for Materials Computer Design

The promising line of materials computer design is associated with development of information-analytical system [Kiselyova, 1993a, 2002]. This system is intended for data retrieval on known compounds, the prediction of new inorganic compounds, not yet synthesized, and the forecasting of their properties. This system includes learning and predicting subsystems based on artificial intelligence methods - method of concept formation using growing network CONFOR [Gladun, 1995; Gladun *et al.,* 1995] and other methods of computer learning [http://www.solutions-center.ru]. For increasing reliability of predicting the voting of predictions, which were obtained using various algorithms and component property descriptions will be carried out. The information-analytical system employs also the databases on properties of inorganic compounds, described above, and a DB on chemical elements, a knowledge base, a conversational processor and monitor (Figure 1).

The *knowledge base* of information-analytical system stores the regularities already obtained for various classes of inorganic compounds. They can be use in the prediction of phases and estimation of the phase properties, unless the databases have no such information about the particular chemical systems. Rules in the knowledge base are represented, for example, in the form of growing pyramidal networks, neural networks, logical expressions, etc.

The *conversational processor* manages the conversation of the user with the information-analytical system. It provides an expert in the given application domain a dialog with the information-analytical system also.

In the future, the employment of a linguistic processor in the software or software-hardware support can be expected. It will allow the system to understand the problem-oriented language of the user.

The *monitor* controls the computation process and provides the interface between the functional subsystems as well as the Internet-access to the system. In addition, the monitor signals whenever new experimental data contradict existing classification regularities. Such contradictions will be eliminated by including the new data in the computer learning and modifying the regularity in the knowledge base.

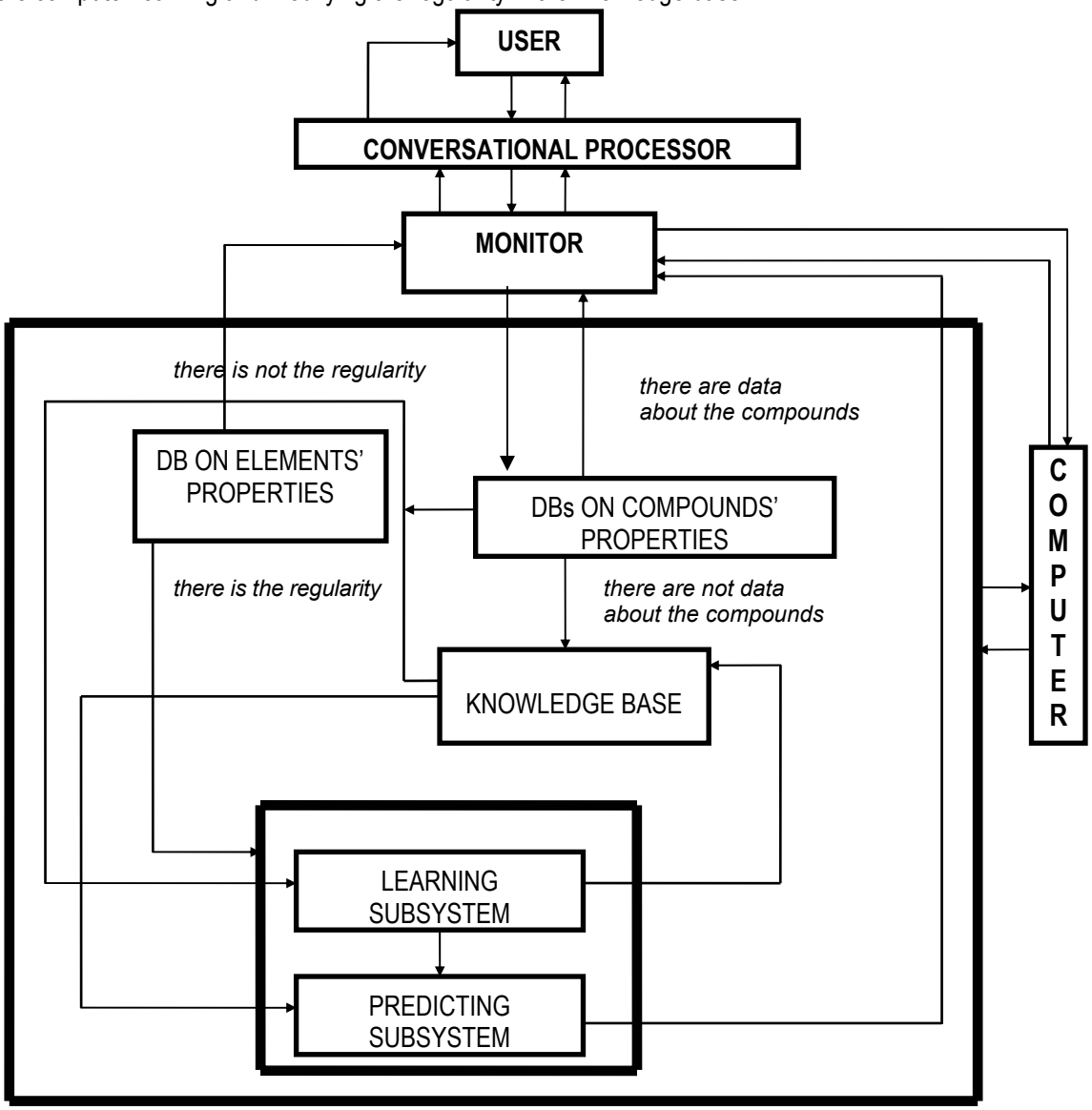


Figure 1. Schematic Diagram of an Information-Analytical System

The information-analytical system operates as follows (Figure 1). The user requests information about a compound of a certain composition. If data about this phase are stored in the databases, they can be extracted for user. If no information about the compound is stored in the databases, or if the information available is incomplete, the computer determines whether the regularity corresponding to the desired property for a compound of a certain type is present in the knowledge base. If the regularity is present, the databases supply the appropriate set of component properties for prediction of the desired characteristic. If the knowledge base does not have the desired regularity, then examples for the computer learning process are extracted from the databases. The correctness of these examples is estimated by the user once more; and, if the samples are found adequate for computer learning, the learning and prediction subsystems process them. The user receives the

resultant prediction, while the new classifying regularity is stored in the knowledge base. The above example is the simplest of the problems that can be solved by an information-analytical system. A more complicated problem would be, for example, predicting all possible phases in ternary and multi-component systems, combined with the estimation of their properties. The previous problem can be solved by real-time processing. The latter problem requires much more time.

Conclusion

We have applied artificial intelligence methods for discovering regularities to the design of new inorganic materials. The effectiveness of the proposed approach is illustrated, for example, in Table 1, 2 and 3. The approach discussed in this paper for predicting new materials is based on a process, which is analogous to the search for new materials by an inorganic chemist. It has allowed predicting hundreds of new compounds in ternary, quaternary and more complicated chemical systems, not yet investigated. These compounds are analogs to known substances with important for industry properties. The predictions have been based on the interaction of growing pyramidal networks, computer learning and databases on properties of inorganic compounds. The proposed approach not only simulates the process of design of new inorganic materials but also extends the capabilities of the chemist in discovering new materials with powerful multi-dimensional data analysis tools.

Acknowledgements

Partial financial support from RFBR (Grant N.04-07-90086) is gratefully acknowledged. I should like to thank my colleagues Prof. Victor P.Gladun, Dr.Neonila D.Vashchenko, and Dr.Vitalii Yu.Velichko of the Institute of Cybernetics of the National Academy of Sciences of Ukraine for their help and support.

Bibliography

[Chen et al., 1999] N.Y.Chen, W.C.Lu, R.Chen, P.Qin. Software package “Materials Designer” and its application in materials research. Proc. of Second Int. Conf. Intelligent Processing & Manufacturing of Materials. Honolulu, Hawaii, v.2, July 10-15, 1999.

[Chen et al., 2002] N.Y.Chen, W.C.Lu, C.Ye, G.Li. Application of Support Vector Machine and Kernel Function in Chemometrics. Computers and Applied Chemistry, 2002, v.19:

[Gladun, 1995] V.P.Gladun. Processes of formation of new knowledge. SD "Pedagog 6”, Sofia, 1995 (Russ.).

[Gladun et al., 1995] V.P.Gladun, N.D.Vashchenko. Local Statistic Methods of Knowledge Formation. Cybernetics and Systems Analysis, 1995, v.31.

[Gulyev et al., 1973] B.B.Gulyev, L.F.Pavlenko. Simulation of the search for components of alloy. Avtomatika i Telemekhanika, 1973 v.1 (Russ.).

[http://www.solutions-center.ru] http://www.solutions-center.ru.

[Kiselyova, 1987] N.N.Kiselyova. Prediction of Heusler-phases with composition ABD_2 (D =Co, Ni, Cu, Pd). Izvestiya Akad. Nauk SSSR, Metalli, 1987, v.2 (Russ.).

[Kiselyova, 1993a] N.N.Kiselyova. Information-predicting system for the design of new materials. J.Alloys and Compounds, 1993, v.197.

[Kiselyova, 1993b] N.N.Kiselyova. Prediction of inorganic compounds: experiences and perspectives. MRS Bull., 1993, v.28.

[Kiseleva et al., 1996] N.N.Kiseleva, N.V.Kravchenko, and V.V.Petukhov. Database system on the properties of ternary inorganic compounds (IBM PC version). Inorganic Materials, 1996, v.32.

[Kiselyova et al., 1998] N.N.Kiselyova, V.P.Gladun, N.D.Vashchenko. Computational materials design using artificial intelligence methods. J.Alloys and Compounds, 1998, v.279.

[Kiselyova et al., 2000] N.N.Kiselyova, S.R.LeClair, V.P.Gladun, N.D. Vashchenko. Application of pyramidal networks to the search for new electro-optical inorganic materials. In: IFAC Symposium on Artificial Intelligence in Real Time Control AIRTC-2000. Preprints, Budapest, Hungary, October 2-4, 2000.

[Kiselyova, 2002] N.N.Kiselyova. Computer design of materials with artificial intelligence methods. In: Intermetallic Compounds. Vol.3. Principles and Practice / Ed. J.H.Westbrook & R.L.Fleischer. John Wiley&Sons, Ltd., 2002.

[Kiselyova et al., 2004] N.N.Kiselyova, I.V.Prokoshev, V.A.Dudarev, et al. Internet-accessible electronic materials database system. Inorganic materials, 2004, v.42, №3.

[Kutolin et al., 1978] S.A.Kutolin, V.I.Kotyukov. Chemical affinity function and computer prediction of binary compositions and properties of rare earth compounds. Zh.Phys. Chem., 1978, v.52 (Russ.).

[Manzanov et al., 1987] Ye.E.Manzanov, V.I.Lutsyk, M.V.Mokhosoev. Influence of features system selection on predictions of compound formation in systems A_2MoO_4-$B_2(MoO_4)_3$ and A_2MoO_4-$CMoO_4$. Doklady Akad. Nauk SSSR, 1987, v.297 (Russ.).

[Pao et al., 1999] Y.H.Pao, B.F.Duan, Y.L.Zhao, S.R.LeClair. Analysis and visualization of category membership distribution in multivariant data. Proc.Second Int.Conf.Intelligent Processing&Manufacturing of Materials, Honolulu, Hawaii, v.2, July 10-15, 1999.

[Savitskii et al., 1968] E.M.Savitskii, Yu.V.Devingtal, V.B.Gribulya. About recognition of binary phase diagrams of metal systems using computer. Doklady Akad. Nauk SSSR, 1968, v.178 (Russ.).

[Savitskii et al., 1979] E.M.Savitskii, and N.N.Kiselyova. Cybernetic prediction of formation of phases with composition ABX_2. Izvestiya Akad.Nauk SSSR, Neorganicheskie Materialy, 1979, v.15 (Russ.).

[Talanov et al., 1981] V.M.Talanov, L.A.Frolova. Investigation of chalco-spinels formation using method of potential functions. Izvestiya VUZov. Khimiya i Khimicheskaya Technologiya, 1981, v.24 (Russ.).

[Villars et al., 2001] P.Villars, K.Brandenburg, M.Berndt, et al. Binary, ternary and quaternary compound former/nonformer prediction via Mendeleev number. J.Alloys and Compounds. 2001. V.317-318.

[Vozdvizhenskii et al., 1973] V.M.Vozdvizhenskii, V.Ya.Falevich. Application of computer pattern recognition method to identification of phase diagram type of binary metal systems. In: Basic Regularities in Constitution of Binary Metal Systems Phase Diagrams. Nauka, Moscow, 1973 (Russ.).

[Yan, 1994] L.M.Yan, Q.B.Zhan, P.Qin, N.Y.Chen. Study of properties of intermetallic compounds of rare earth metals by artificial neural networks. J.Rare Earths, 1994, v.12.

Author's Information

Nadezhda N.Kiselyova – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, senior researcher, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: kis@ultra.imet.ac.ru

К ВОПРОСУ О РАЗВИТИИ ИНТЕРФЕЙСА «РАЗРАБОТЧИК-ЗАКАЗЧИК»

Леонид Святогор

В своем выступлении я хотел бы коснуться вопроса: каким образом упростить дорогу от наших готовых разработок к потенциальному потребителю?

Речь не идет обо всех сложностях, которые неизбежно нужно преодолевать разработчику интеллектуального программного обеспечения для внедрения результатов. Их очень много. Наиболее трудной частью этого пути, по нашему опыту, является поиск партнера – потребителя продукта, или посреднической организации, инвестора или других. Однако, как нам кажется, можно облегчить контакт. Для этого необходимо, по крайней мере, уметь изложить результаты на таком языке, который был бы более привычным для потребителей с разнообразным спектром запросов. К сожалению, часто приходится наблюдать, что разработчик излагает возможности и преимущества программного обеспечения на строго-научном уровне своей предметной области, в то время как партнёр пытается понять, что из этого ему прагматически полезно.

Как правило, потребитель видит свою конкретную задачу под узким «целевым» углом зрения. Свои требования он выражает в формате технического задания. Ответной реакцией, как правило, служит техническое предложение, которое является нормативным документом. Однако на начальной стадии общения (до разработки нормативных документов) мы предлагаем предельно упростить описание возможностей того программного комплекса, который хотим предложить заказчику. Для этого достаточно изложить возможности программы на предметно-ориентированном языке, широко используемом в разных технических приложениях. Достоинства такого подхода очевидны:

а) разработчик первым делает шаг навстречу потребителю, упреждая и привлекая его интерес и сообщая априори сведения о новой разработке и

б) при этом излагает свои результаты на языке и в терминологии заказчика.

За основу возьмем общепринятую схему описания модели явления или процесса: «вход» – «выход» – «функция» – «параметры». Не отступая от неё, мы проиллюстрируем на примерах, как привязать широкие возможности конкретного программного комплекса к подклассам задач, существующих в конкретной

предметной области, – например, ***задач технической диагностики, химии, материаловедения, медицины, геологии, астрономии, бизнеса*** и др**.** В качестве примера программной системы («функции») выбран программный комплекс КОНФОР.

Комплекс КОНФОР базируется на методах искусственного интеллекта. Он разработан для выявления существующих закономерностей, которые проявляются в стабильных связях (признаках) между входными и выходными данными. Входными данными обычно служат некоторые числовые измерения или логические значения заданных исходных признаков. Выходными данными являются заранее указанные человеком ситуации или классы множеств, которые имеют содержательный для человека смысл.

Алгоритм КОНФОРА использует различные процедуры искусственного интеллекта – метод обучения, генерацию информационной семантической структуры типа растущих пирамидальных сетей, формирование признаков, построение логических выражений и другие. Теоретические обоснования приведены в ряде публикаций. Однако заказчика на первоначальном этапе общения интересует возможность применения алгоритмического комплекса КОНФОР к его конкретной задаче.

Способы представления потребительской информации в форме, удобной для пользователя, ясны из приведенных ниже примеров.

В области ХИМИИ и МАТЕРИАЛОВЕДЕНИЯ

Задача 1. Прогнозирование существования химических соединений с наперёд заданными свойствами

Решаемая задача: на основании ряда известных химических соединений, обладающих необходимым свойством, предсказать новые соединения с данным свойством.

Входные данные: различные физические, химические и атомарные характеристики элементов таблицы Менделеева.

Искомые (потребительские) характеристики: свойство соединения (например, сверхпроводимость, ферромагнетизм, полупроводниковые свойства, кристаллическая структура).

Оценка результата: синтезируемое вещество обладает (не обладает) заданным свойством. Синтез вещества возможен (невозможен).

Метод настройки на решаемую задачу: обучение по образцам. В обучающую выборку включаются химические соединения, которые имеют заданное свойство, а также соединения, в которых оно отсутствует, несмотря на сходство или аналогию. Программа устанавливает объективные закономерности – связи между химическими элементами соединения и его потребительскими свойствами.

Результаты прогнозирования: прогнозируемое химическое соединение, – возможно (не возможно).

Применение: в Институте металлургии им. А.А.Байкова осуществлен прогноз новых химических соединений (с 1976 г. по настоящее время).

Задача 2. Выбор режима синтеза материалов с заданными свойствами (синтез искусственных алмазов).

Решаемая задача: найти новые, ранее не известные технологические режимы, приводящие к созданию новых высококачественных материалов (технических алмазов).

Входные данные: измеряемые и контролируемые характеристики режимов технологического процесса (давление, температура, циклы и др).

Искомые (потребительские) характеристики: твёрдость, кристаллическая структура, зернистость.

Оценка результата: сортность алмазов удовлетворяет (не удовлетворяет) техническим требованиям.

Метод настройки на решаемую задачу: обучение по образцам. В обучающую выборку включаются данные по синтезу качественных, средних и низкосортных алмазов. На основе выборки формируются скрытые связи между режимом синтеза и качеством продукта.

Результаты: технические характеристики режимов синтеза алмазов, отвечающих предъявленным требованиям. Оптимальные режимы технологического процесса.

В области ТЕХНИЧЕСКОЙ ДИАГНОСТИКИ и РАСПОЗНАВАНИЯ ОБРАЗОВ

Задача 1. Определение степени износа двигателей и элементов механических конструкций

Решаемая задача: определить техническое состояние двигателя или механического узла с целью предотвращения аварийных ситуаций и определения ремонтопригодности.

Входные данные: измерения вибраций при работе двигателя в разных режимах (отсчеты датчиков за период времени).

Параметры: различные режимы работы (под нагрузкой, на высоте, на холостом ходу).

Потребительские характеристики: степень износа двигателя или узла.

Оценка результата: степень износа агрегата соответствует определенному классу (исправен, неисправен, подлежит ремонту).
Метод настройки на решаемую задачу: по обучающей выборке, которая включает наборы измерений совместно с известными показателями износа.
Выполняется поиск скрытых закономерностей (логических функций), связывающих входные сигналы и состояние агрегата.
Результаты диагностики: определение состояния и ремонтопригодности двигателя или узла.
*Использование***:** на предприятии «КАМОВ» (Росия).

Задача 2. Нахождение характерных неисправностей в радиотехническом оборудовании
Решаемая задача: на основании электрических измерений в различных точках схемы и показаний приборов сигнализации определить состояние оборудования, указать тип неисправности и её место.
Входные данные: значения контролируемых электрических параметров и показания приборов сигнализации.
Параметры: различные нагрузки.
Потребительские характеристики: тип неисправности и её место.
Метод настройки на решаемую задачу: обучение по прецедентам. Обучающая выборка содержит показания измерений совместно с указаниями на тип и место неисправности. Программа обучения выполняет поиск причинно-следственных связей между измеряемыми (контролируемыми) электрическими величинами и типом (местом) неисправности.
Результаты диагностики: характер (тип) неисправности и её место.

В области МЕДИЦИНЫ

Задача 1. Диагностика заболевания кровеносных сосудов
Решаемая задача: на основании медицинских показателей определить стадию заболевания кровеносных сосудов пациента.
Входные данные: численные и качественные данные анализов, осмотра и субъективные показатели больного.
Искомые характеристики: стадия предполагаемой болезни.
Оценка результата: пациент здоров, болен, заболевание находится в определённой стадии.
Метод настройки на решаемую задачу: обучение по медицинской статистике данного заболевания. Выборка состоит из пациентов трёх групп: больных, с подозрением на болезнь и здоровых (с похожей симптоматикой). Устанавливаются связи между симптомами заболевания и заключением врача.
Результаты диагностики: попадание пациента в одну из трёх групп и определение стадии болезни. Выбор методики лечения.
Апробация: эффективность методики подтверждена в лечебном учреждении 4-го управления Минздрава СССР.

Задача 2. Прогнозирование различных форм бронхита и астмы
Решаемая задача: на основе медицинских данных пациента определить разновидность заболевания и указать степень развития болезни.
Входные данные: таблица численных измерений на выходе газоанализатора выдоха пациента.
Искомые характеристики: тип болезни, стадия заболевания.
Оценка результата: пациент страдает данным типом (формой) заболевания; болезнь находится в определённой стадии.
Метод настройки на решаемую задачу: по обучающей выборке, в которую включаются представители разных форм заболевания и на разных стадиях болезни, определяются закономерности проявления заболевания. Находятся ненаблюдаемые признаки и их связь с заключением врача.
Результаты диагностики: пациент попадает в одну из медицинских групп. Рекомендуемая стратегия лечения.
Апробация: диагностика бронхиальных заболеваний по совокупности признаков выполнялась в НИИ «Медстатистика» (1983 г.)

В области ГЕОЛОГИИ, АСТРОНОМИИ

Задача 1. Прогнозирование месторождений нефти
Решаемая задача: Определить перспективность нового месторождения нефти по данным георазведки.
Входные данные: измерения химического состава и концентрации газового конденсата на месте бурения.
Искомые характеристики: эксплуатационные параметры месторождения: качество нефти, объёмы, концентрация, глубина залегания, сопутствующие вещества.

Оценка результата: искомые характеристики попадают (не попадают) в требуемую область или находятся вблизи границ.
Метод настройки на решаемую задачу: обучение по известным месторождениям. Обучающая выборка составляется из перспективных, малоперспективных и неперспективных месторождений. Программа синтезирует признаки, связывающие качество нефтяного пласта с характеристиками газоконденсата.
Результаты прогнозирования: найденное месторождение нефти является перспективным, ограниченно-перспективным или неперспективным.
Апробация: система применялась для прогнозирования нефтяных оторочек в тресте «Полтаванефтегазразведка» (1979 г.).

Задача 2. Прогнозирование солнечной активности
Решаемая задача: составить прогноз активности солнца на основании многолетних наблюдений.
Входные данные: графики плотности солнечного излучения (в различных спектрах), периодические и апериодические составляющие процессов, электрические, магнитные, температурные и иные параметры.
Искомые характеристики: интегральный показатель солнечной активности.
Оценка результата: солнечная активность соответствует определённому уровню по шкале интенсивности излучения (энергетической плотности).
Метод настройки на решаемую задачу: составление обучающей выборки по астрономическим и другим наблюдениям. Поиск скрытых закономерных связей между наблюдениями и показателями солнечной активности.
Результаты прогнозирования: показатель солнечной активности прогнозируется на уровне, найденном в процессе работы программы.
Апробация: проводились исследования в Астрономической обсерватории Киевского Госуниверситета.

В области БИЗНЕСА

Задача 1. Прогнозирование биржевого курса валют
Решаемая задача: На основании изучения биржевых показателей курсов основных валют предсказать вероятный курс заданной валюты.
Входная информация: (факторы влияния): тренды валютных курсов разных стран за значимый период; показатели спроса и предложения; текущие курсы и индексы валют на межбанковской валютной бирже; тенденции отклонения курса данной валюты от среднего; показатели товарооборота, инвестиционной активности и экспорта государства и его банковские активы; курсы акций крупнейших мировых компаний; другие факторы.
Параметры: интервалы анализа и интервалы предсказания.
Оценка результата: валютный курс укладывается в одном из заданных интервалов.
Настройка на решаемую задачу: входная информация на интервале анализа в обучающей выборке сопоставляется с реальными курсами валют на интервале предсказания. Находятся скрытые закономерности, формирующие валютный курс.
Результаты прогнозирования: прогноз валютного курса на заданный интервал времени (текущий, на неделю, квартал, год).

Задача 2. Прогнозирование цены на новый продукт или услугу. Сфера электронного бизнеса
Решаемая задача: оценить рыночную стоимость нового товара.
Входные данные: потребительские характеристики нового продукта и аналогов. Цены аналогов (информация из справочников, каталогов и Интернет). Отношения «цена – качество». Маркетинговые исследования (спрос – предложение).
Оценка результата: предполагаемая цена нового товара находится в определённом интервале.
Метод настройки на решаемую задачу: обучающая выборка включает пары соотношений «рыночные характеристики товаров – их цены». Выявляются скрытые тенденции и закономерности ценообразования.
Результаты: определение рыночной (продажной) цены нового продукта или услуги с обоснованием их ценообразования.

4.3. Planning and Sheduling

TWO-MACHINE MINIMUM-LENGTH SHOP-SCHEDULING PROBLEMS WITH UNCERTAIN PROCESSING TIMES

Natalja Leshchenko, Yuri Sotskov

***Abstract**: The problem of non-preemptive job-shop scheduling with bounded random processing times is studied. In contrast to deterministic two-machine job-shop problem, it is assumed that processing time t_{jm} of job $j \in J = \{1, 2, ..., n\}$ by machine $m \in M = \{1, 2\}$ is not fixed before scheduling. Moreover, probability distribution of random processing time t_{jm} between given lower and upper bounds is unknown. In such an uncertain version of a shop-scheduling problem, there may not exist a unique schedule that remains optimal for all possible realizations of processing times t_{jm}. Therefore, we have to consider a set S^* of schedules (permutations) that dominates all other feasible schedules for makespan criterion. We find necessary and sufficient conditions when transposition of two jobs may be used to minimize makespan for the uncertain version of a flow-shop problem with two machines. We study the easiest case of a flow-shop problem if there exists a single permutation that is optimal for all possible realizations of processing times t_{jm}. We also study hardiest case if cardinality $|S^*|$ of set S^* is maximal.*

***Keywords**: Scheduling, flow-shop, job-shop, makespan, uncertainty.*

Introduction

We consider the following non-preemptive job-shop scheduling problem with bounded random processing times. Two machines $M = \{1, 2\}$ have to process *n* jobs $J = \{1, 2, ..., n\}$ with different two-stage machine routes. All *n* jobs are available to be processed from time 0. Let $J_{12} \subset J$ be set of jobs with route (1, 2) (first, job $j \in J_{12}$ has to be processed by machine 1, and then by machine 2), $J_{21} \subset J$ be set of jobs with route (2, 1), and $J_m \subset J$ ($m \in \{1, 2\}$) be set of jobs that are processed only by one machine $m \in M$. So, we have $J = J_1 \cup J_2 \cup J_{12} \cup J_{21}$ with $n_l = |J_l|$, $l \in \{1,2,12,21\}$. In contrast to deterministic job-shop problem, it is assumed that processing time t_{jm} of job $j \in J$ by machine $m \in M$ is not fixed before scheduling. In a realization of the process, t_{jm} may be equal to any real value between lower bound t_{jm}^L and upper bound t_{jm}^U being given before scheduling. Moreover, the probability distribution of random processing time is unknown before scheduling (Fig.1). In such a case, there may not exist a unique schedule that remains optimal for all possible realizations of job processing times. Therefore, we are forced to consider a set of schedules that dominate all feasible schedules. In [4], such a set of schedules is called a solution to scheduling problem with uncertain job processing times.

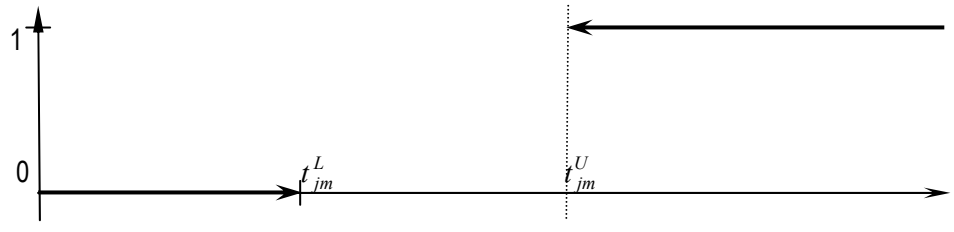


Fig.1. Probability distribution of random processing time t_{jm}

Let $C_j(s)$ denote the completion time of job $j \in J$ in schedule $s \in S$, where *S* is a set of feasible schedules including at least one optimal schedule for makespan criterion, and let criterion $C_{\max}$ denote minimization of schedule length: $\max_s \{C_j(s) \mid j \in J\}$. Using three-field notation, this problem may be denoted as $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$.

We have to emphasize that random processing times in the problem $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ are due to *external forces* in contrast to scheduling problems with *controllable* processing times [6-8]. In the latter problem, the objective is *to choose* optimal processing times (which are under the control of the decision-maker) and optimal schedule with *chosen* processing times.

The rest of the paper is organized as follows. First, we give motivation with two main definitions for a solution of a job-shop problem with uncertain processing times. Then we show how to use Johnson's permutations for solving a flow-shop problem with uncertain and bounded processing times. Extreme special cases to such a flow-shop problem are considered and general case of a two-machine flow-shop problem is treated. Finally, the approach is generalized for a job-shop problem with uncertain and bounded processing times, and illustrative examples are given.

Motivations and Definition

We address the stochastic flow-shop and job-shop problems for the cases when it is hard to obtain exact probability distributions for random and bounded processing times, and when assuming a specific probability distribution is not realistic before scheduling. Commonly, schedules obtained after assuming a certain probability distribution may be not close to the optimal schedule. It has been observed that, although the exact probability distribution of job processing times may not be known in advance, upper and lower bounds on job processing times are easy to obtain in many practical cases. In what follows, we show that information on the bounds of job processing times is important and should be utilized in finding a solution to shop-scheduling problem with uncertain numerical input data.

If equality $t_{jm}^L = t_{jm}^U$ holds for each job $j \in J$ and each machine $m \in M$, then problem $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ turns into a deterministic job-shop problem (denoted as $J2 \mid\mid C_{\max}$) that is polynomially solvable [2]. Let *T* denote a set of feasible vectors $t = (t_{11}, t_{12}, ..., t_{n1}, t_{n2})$ of job processing times:

$$T = \left\{ t \mid t_{jm}^L \le t_{jm} \le t_{jm}^U, j \in J, m \in M \right\}.$$

Jackson [2] proved that the optimal schedule for the problem $J2 \mid\mid C_{\max}$ may be defined by the two permutations (π', π''), where π' is a permutation of jobs $J_1 \cup J_{12} \cup J_{21}$ on machine 1, and π'' is a permutation of jobs $J_2 \cup J_{12} \cup J_{21}$ on machine 2, the jobs within set J_{12} (set J_{21}) being ordered by SPT rule (by LPT rule) on machine 1 (on machine 2, respectively). Rule SPT (shortest processing time) means that the job set have to be processed in non-decreasing order of job processing times. Rule LPT (longest processing time) means that the job set have to be processed in non-increasing order of job processing times.

Thus, we can find an optimal schedule in the following form: ($\pi' = (\pi_{12}, \pi_1, \pi_{21})$, $\pi'' = (\pi_{21}, \pi_2, \pi_{12})$), where job *j* belongs to permutation π_l if and only if $j \in J_l$, $l \in \{1,2,12,21\}$. Since the order of the jobs in set J_1 and in set J_2 may be arbitrary, we can fix both permutations π_1 and π_2, e.g., we can order jobs in these permutations with respect to their job numbers. We denote $S_{12} = \{\pi_{12}^1, \pi_{12}^2, ..., \pi_{12}^{n_{12}!}\}$ a set of all permutations (sequences) of n_{12} jobs from set J_{12}, and $S_{21} = \{\pi_{21}^1, \pi_{21}^2, ..., \pi_{21}^{n_{21}!}\}$ a set of all permutations (sequences) of n_{21} jobs from set J_{21}. Let $S = \langle S_{12}, S_{21} \rangle$ be such a subset of Cartesian product $(S_{12}, \pi_1, S_{21}) \times (S_{21}, \pi_2, S_{12})$ that elements of set *S* are ordered pairs of two permutations (π', π''), where $\pi' = (\pi_{12}^i, \pi_1, \pi_{21}^j)$, and $\pi'' = (\pi_{21}^j, \pi_2, \pi_{12}^i)$, $1 \le i \le n_{12}!$, $1 \le j \le n_{21}!$. Since both permutations π_1 and π_2 are fixed, and index *i* (index *j*) is the same in each permutation from pair $(\pi', \pi'') \in S$, then we have $|S| = n_{12}! \cdot n_{21}!$. We use the following definition of a solution to the problem $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ (see [1]).

Definition 1 *A set of pairs of permutations* $S^* \subseteq S$ *is called a solution to the problem* $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$, *if for each feasible vector* $t \in T$ *set* S^* *contains at least one optimal Jackson's pair of permutations for the problem* $J2 \mid\mid C_{\max}$ *with vector t of job processing times and any proper subset of set* S^* *being not a solution to the problem* $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$.

Flow-Shop Problem

First, we consider a flow-shop problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with set of jobs $J = J_{12}$, i.e., when $J_{21} = J_1 = J_2 = \emptyset$ and hence all $n = n_{12}$ jobs have the same machine route. So, if we deal with problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$, then set of all feasible permutations is set S_{12}.

Johnson [3] proposed the $O(n \log n)$ algorithm for constructing an optimal schedule for deterministic flow-shop problem $F2 \mid\mid C_{\max}$. Permutation $\pi = (i_1, i_2, ..., i_{n_{12}})$, satisfying conditions $\min\{t_{i_l 1}, t_{i_{l+1} 2}\} \le \min\{t_{i_{l+1} 1}, t_{i_l 2}\}$, $l = \overline{1, n_{12} - 1}$, is called a Johnson's permutation. At least one optimal permutation for problem $F2 \mid\mid C_{\max}$ is Johnson's permutation π. However, for problem $F2 \mid\mid C_{\max}$ optimal schedule may be defined by permutation that does not satisfy the latter conditions. Nevertheless, hereafter set of optimal schedules treated for problem $F2 \mid\mid C_{\max}$ is restricted by set of Johnson's permutations. Set of all feasible permutations has cardinality $\mid S_{12} \mid = n_{12}!$ (since each of the two machines has to process jobs in the same order).

Since our aim is to construct a solution to the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with minimal cardinality, we can remove all the redundant permutations. In particular, if there is no a feasible vector of the processing times such that the permutation is a Johnson's one for this vector, then we can remove this permutation from a problem solution. Thus, for problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ we use the following special definition instead of Definition 1.

Definition 2 S*et of permutations* $S_{12}^* \in S_{12}$ *is a solution to the problem* $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$, *if for any vector* $t \in T$ *set* S_{12}^* *contains at least one permutation, which is Johnson's one (and so it is optimal permutation) for the problem* $F2 \mid\mid C_{\max}$ *with vector t of job processing times, any proper subset of set* S_{12}^* *being not a solution to the problem* $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$.

Extreme Cases of a Flow-Shop Problem

Let us consider the easiest case of problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with respect to its solution cardinality, i.e., when there exists a single permutation π, defining the solution to the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$. In other words, this permutation π is optimal for each problem $F2 \mid\mid C_{\max}$ with any feasible vectors $t \in T$ of job processing times.

We consider set $N_a \subseteq J_{12}$ such that job *i* belongs to set N_a, if inequality $t_{i1}^U \le t_{i2}^L$ holds, and set $N_b \subseteq J_{12}$ such that $i \in N_b$, if inequality $t_{i2}^U \le t_{i1}^L$ holds.

Theorem 1. *For the problem* $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ *there exists a solution* S_{12}^* *that is a singleton if and only if the following three conditions hold:*

a) *for each job j from set* J_{12} *(perhaps except one job* $j_0 \in J_{12}$*), one of the following two inclusions* $j \in N_a$ *or* $j \in N_b$ *holds;*
b) *for any pair of different jobs from set* N_a *their intervals on machine 1 have no more than one common point, and for any pair of different jobs from set* N_b *their intervals on machine 2 have no more than one common point;*
c) *if there exists job* $j_0 \in J_{12}$, *then inequalities* $t_{j_0 1}^L \ge \max\{t_{i1}^U \mid i \in N_a\}$ *and* $t_{j_0 2}^L \ge \max\{t_{j2}^U \mid j \in N_b\}$ *hold.*

Next, we consider the worst (hardiest) case of a problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with respect to its solution cardinality. Namely, we consider the case when solution of the problem has maximal cardinality, i.e., $\mid S_{12} \mid = n_{12}!$.

Theorem 2 *If condition* $t_{\max}^L = \max\{t_{jm}^L \mid j \in J, m \in M\} < \min\{t_{jm}^U \mid j \in J, m \in M\} = t_{\min}^U$ *holds for all intervals of job processing times for the problem* $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$, *then* $S_{12}^* = S_{12}$.

Theorems 1 and 2 characterize the extreme cases (easiest and hardiest) of a solution to the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$, i.e., when $\mid S_{12}^* \mid = 1$ and when $S_{12}^* = S_{12}$, respectively.

General Case of a Flow-Shop Problem

We show how to delete all redundant permutations from set S_{12} while solving uncertain version of a flow-shop problem. We use transposition of two jobs, when a concrete order of these jobs may be fixed in the desired solution since there exists at least one optimal permutation with this fixed order of two jobs. In [1], the following sufficient conditions have been proven for transposition of two jobs.

Theorem 3 [1] *If inequalities* $t_{w2}^U \le t_{w1}^L$ *and* $t_{j1}^U \le t_{j2}^L$ *hold, then for each vector* $t \in T$ *of job processing times, there exists a sequence* $\pi = (s_1, j, s_2, w, s_3) \in S_{12}$ *that is a Johnson's permutation for the problem* $F2 \mid\mid C_{\max}$ *with vector t of job processing times.*

Theorem 4 [1] *If inequalities* $t_{j1}^U \le t_{w1}^L$ *and* $t_{j1}^U \le t_{j2}^L$ *hold, then for each vector* $t \in T$ *of job processing times, there exists a sequence* $\pi = (s_1, j, s_2, w, s_3) \in S_{12}$ *that is a Johnson's permutation for the problem* $F2 \mid\mid C_{\max}$ *with vector t of job processing times.*

Theorem 5 [1] *If inequalities* $t_{w2}^U \le t_{w1}^L$ *and* $t_{w2}^U \le t_{j2}^L$ *hold, then for each vector* $t \in T$ *of job processing times, there exists a sequence* $\pi = (s_1, j, s_2, w, s_3) \in S_{12}$ *that is a Johnson's permutation for the problem* $F2 \mid\mid C_{\max}$ *with vector t of job processing times.*

We can show that these sufficient conditions put together necessary condition for fixing the optimal order of two jobs in the sense of Definition 2.

If for a pair of jobs $j \in J_{12}$ and $w \in J_{12}$ condition of at least one of Theorems 3, 4 or 5 holds, then we can fix the order of these two jobs ($j \to w$) in a solution to the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ (i.e., job *j* has to be processed before job *w*). After testing all the pairs of inequalities from Theorems 3, 4 or 5 we can construct (partial) strong order relation on set S_{12} that defines problem solution in condense form.

We can prove that if no condition of Theorems 3, 4 or 5 holds for two jobs $j \in J_{12}$ and $w \in J_{12}$, then the order of these two jobs cannot be optimally fixed (in the sense of Definition 2). In any solution S_{12} in such a case, due to Definitions 1 and 2, at least one permutation with order $j \to w$ and at least one permutation with order $w \to j$ have to be included in any solution to the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$.

Theorem 6 *For any fixed vector* $t \in T$ *of job processing times there exists a permutation of the form* $\pi = (s_1, j, s_2, w, s_3) \in S_{12}$ *that is a Johnson's permutation for the problem* $F2 \mid\mid C_{\max}$ *if and only if at least one of the following three conditions holds:*

$$t_{w2}^U \le t_{w1}^L \text{ and } t_{j1}^U \le t_{j2}^L;$$

$$t_{j1}^U \le t_{w1}^L \text{ and } t_{j1}^U \le t_{j2}^L;$$

$$t_{w2}^U \le t_{w1}^L \text{ and } t_{w2}^U \le t_{j2}^L.$$

Job-Shop Problem

Finally, we can show that the above results for a flow-shop problem may be used for solving a job-shop problem with uncertain numerical data. In particular, we proved the following theorem.

Theorem 7 *If* S_{12}^* *is a solution to the flow-shop problem* $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ *with job set* J_{12}*, and* S_{21}^* *is a solution to the flow-shop problem* $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ *with job set* J_{21}*, then* $\langle S_{12}^*, S_{21}^* \rangle$ *is a solution to the job-shop problem* $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ *with job set* $J = J_1 \cup J_2 \cup J_{12} \cup J_{21}$.

In Theorem 7, inclusions $S_{12}^* \in S_{12}$, $S_{21}^* \in S_{21}$ hold, and set $\langle S_{12}^*, S_{21}^* \rangle$ denotes subset of set $\langle S_{12}, S_{21} \rangle$. Using the above results one can construct solution (with respect to Definitions 1 and 2) for a job-shop problem with uncertain numerical input data in polynomial time. The full proofs of Theorems 1, 2, 6 and 7 are given in preprint [5].

Examples

To illustrate the above results let us consider the following examples.

Example 1 Let bounds of job processing times for problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ be given in Table 1. These bounds define polytope *T* of feasible vectors of job processing times.

Table 1: Input data for Example 1

j	t_{j1}^L	t_{j1}^U	t_{j2}^L	t_{j2}^U
1	2	4	5	6
2	3	4	5	6
3	7	8	7	8

It is easy to see that Theorem 6 implies the following dominance relations between jobs *J* = {1, 2, 3}. Job 1 precedes job 3 and job 2 precedes job 3. However, we cannot order jobs 1 and 2. Therefore, instead of considering 3! = 6 permutations of three jobs, it is sufficient to consider two permutations of three jobs (namely, permutations (1, 2, 3) and (2, 1, 3)) that present the solution S_{12}^* to the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with numerical input data given in Table 1.

Example 2. Let us consider the job-shop problem $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with input data given in Table 2.

Jobs 1, 2 and 3 have the same route (1, 2), jobs 6, 7 and 8 have route (2, 1), jobs 4 and 5 have to be processed only by machine 1 and machine 2, respectively.

It is easy to see that all bounds of processing times of jobs 1, 2 and 3 are the same as in Example 1. Thus, we have to consider jobs 6, 7 and 8. These jobs have the same route, and due to Theorem 7, we can find the following dominance relations between these jobs. Job 6 precedes job 8, and job 7 precedes job 8. However, we cannot order jobs 6 and 7. Therefore, instead of considering 3! = 6 permutations of these three jobs, it is sufficient to consider two permutations: (6, 7, 8) and (7, 6, 8).

So, for the problem $J2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with input data given in Table 2 we obtain a solution that consists of four pairs of permutations:

π' = (1, 2, 3, 4, 6, 7, 8), π'' = (6, 7, 8, 5, 1, 2, 3);
π' = (1, 2, 3, 4, 7, 6, 8), π'' = (7, 6, 8, 5, 1, 2, 3);
π' = (2, 1, 3, 4, 6, 7, 8), π'' = (6, 7, 8, 5, 2, 1, 3);
π' = (2, 1, 3, 4, 7, 6, 8), π'' = (7, 6, 8, 5, 2, 1, 3);

instead of set S with $|S| = n_{12}! \cdot n_{21}! = 3! \cdot 3! = 36$ permutations of 8 jobs.

Table 2: Input data for Example 2

j	t_{j1}^L	t_{j1}^U	t_{j2}^L	t_{j2}^U
1	2	4	5	6
2	3	4	5	6
3	7	8	7	8
4	2	3	-	-
5	-	-	3	5
6	4	5	1	3
7	5	6	2	3
8	7	8	7	8

Example 3 Let us consider the problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ with numerical input data given as in Example 1 but with one exception: $t_{11}^U = 3$. Then condition of Theorem 1 holds, and this problem has the solution which consists of only one permutation π = (1, 2, 3).

Example 4 Let bounds of job processing times for problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ be given in Table 3.

Table 3: Input data for Example 3

j	t_{j1}^L	t_{j1}^U	t_{j2}^L	t_{j2}^U
1	2	4	2	6
2	3	4	2	6
3	2	8	2	8

It is easy to see that condition of Theorem 2 holds for this numerical input data with $t_{\max}^L = 3$ and $t_{\min}^U = 4$. So, the solution to this problem has the maximal cardinality 3! = 6 and we have to consider six permutations of three jobs in solution S_{12}.

Conclusion

In this paper, we characterize the easiest case of a two-machine flow-shop problem with uncertain job processing times, i.e., we prove necessary and sufficient conditions for existence of a single schedule, which dominates all the other schedules for makespan criterion (Theorem 1). On the other hand, we describe the set of the hardest problems for which any semiactive schedule may be unique optimal one for some feasible realizations of job processing times (Theorem 2).

We also find necessary and sufficient conditions when it is possible to fix optimal order of two jobs to minimize makespan for the uncertain version of a flow-shop problem with two machines. If there is no possibility to construct unique dominant schedule, we propose to use sufficiency proof of Theorem 6 in order to construct more general schedule form on the basis of partial job order. The partial order defined by Theorem 6 may be extended later in on-line mode when more information about job processing times will be available.

If partial order defined by Theorem 6 will be extended to complete order of *n* jobs (finally, to unique permutation), then the process will be realized in the optimal way. Otherwise, using solution S^* one can look for the schedule minimizing the lost of the objective function value.

Along with off-line scheduling problem, one can consider realization stage of flow-shop scheduling, when a part of the schedule is already realized. We are interesting in how to use additional information about realized operations in order to obtain better solution than that constructed for off-line version of scheduling problem. It is interesting to find sufficient conditions for choosing the unique permutation that is optimal for any feasible processing times of the remaining operations.

There are other directions for future research, e.g., to consider another objective functions for uncertain scheduling problem (total completion time, criteria with job due dates, etc.), along with uncertain processing times it is useful to consider uncertain setup times, job release times, due dates, etc.

Note that all conditions proven in this paper may be tested in polynomial time. However it is clear, that for the most generalizations of the above scheduling problem existence of polynomial algorithms is unlikely (e.g., if deterministic scheduling counterpart is *NP*-hard). So, for future more general and so more practical investigation

it will be reasonable to look for methods of branch and bound type for solving off-line scheduling problem and it will be also useful to consider heuristic solution instead of optimal one both for off-line problem and for realization stage of a flow-shop scheduling problem.

The research was partially supported by INTAS (project 03-51-5501) (both authors) and by World Federation of Scientists (first author).

Bibliography

[1] Allahverdi A., Yu.N. Sotskov, Two-machine flowshop minimum-length scheduling problem with random and bounded processing times. Intl. Trans. Oper. Res., 2003, 10, 65—76.
[2] Jackson J.R., An extension of Johnson's results on job lot scheduling. Naval Res. Logist. Quart., 1956, 3, 201—203.
[3] Johnson S.M., Optimal two- and three-stage production schedules with setup times included. Naval Res. Logist. Quart., 1954, 1, 61—68.
[4] Lai T.-C., Yu.N. Sotskov, N. Sotskova, F. Werner, Optimal makespan scheduling with given bounds of processing times. Math. Comput. Modelling, 1997, 26, 67—86.
[5] Leshchenko N.M., Yu.N. Sotskov, Optimal makespan schedule for processing *n* jobs with uncertain durations of operations on two machines. - Minsk, Belarus. (Preprint: United Institute of Informatics Problems of National Academy of Sciences of Belarus, Nr. 3) , 2003 (in Russian).
[6] Ng C.T., T.C.E. Cheng, M.Y. Kovalyov, Batch scheduling with controllable setup and processing times to minimize total completion time. J. Oper. Res. Society, 2003, 54, 499—506.
[7] Ng C.T., T.C.E. Cheng, M.Y. Kovalyov, Single machine batch scheduling with jointly compressible setup and processing times. Eur. J. Oper. Res., 2004, 153, 211—219.
[8] Strusevich V.A., Two machine flow shop scheduling problem with no wait in process: Controllable machine speeds. Discrete Appl. Math., 1995, 59, 75—86.

Authors' Information

Natalja M. Leshchenko – United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganova str., 6, 220012, Minsk, Belarus; e-mail: yanush@newman.bas-net.by
Yuri N. Sotskov – United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganova str., 6, 220012, Minsk, Belarus; e-mail: sotskov@newman.bas-net.by

LEARNING TECHNOLOGY IN THE SCHEDULING ALGORITHM BASED ON THE MIXED GRAPH MODEL

Yuri Sotskov, Nadezhda Sotskova, Leonid V. Rudoi

***Abstract**: We propose the adaptive algorithm for solving a set of similar scheduling problems using learning technology. It is devised to combine the merits of the exact enumerative algorithm based on the mixed graph model, and heuristic algorithm oriented on the real-world scheduling problems. The former one may ensure high quality of the solution by means of implicit exhausting enumeration of all the feasible schedules. The latter one may be developed for certain type of problems using their peculiarities. The main idea of the learning technology is to produce effective (in performance measure) and efficient (in computational time) heuristics by adapting local decisions for scheduling problems under consideration. Algorithm adaptation is realized at the stage of learning while solving a set of sample problems (using branch and bound algorithm), accumulating knowledge, and structuring it (using artificial intelligence apparatus).*

***Keywords**: Scheduling, mixed graph, adaptation, learning.*

Introduction

Scheduling is defined as an optimal assignment of the scarce resources (machines) to competitive activities (jobs) over time. Having arisen from the needs of practice, scheduling theory attracts much attention of management workers and operations research (OR) practitioners. The most general problems may be described in a framework of the multistage system with machine set M including both different machines and identical machines that should process set of jobs $J = \{J_1, J_2, \ldots, J_n\}$. Each job consists of processing ordered set of operations. Before scheduling for each operation i, we know its processing time p_i, and the machine type in set M that has to be used to process operation i. Let $M = \bigcup_{k=1}^{w} M_k$ where M_k denotes a set of all identical machines of type k. If $\mu_i \in M_k$ and $\mu_j \in M_l$ with $k \neq l$, then machines μ_i and μ_j are of different types. Otherwise (if $k = l$), machines μ_i and μ_j are identical. We assume that operation preemption is not allowed. As usual, each machine cannot process two (or more) operations simultaneously. The objective is to assign all the operations to machine set M, and to define for each machine $\mu_j \in M$ the sequence of operations assigned to μ_j in such a way that the value of the given objective function is minimal.

The complexity research has shown that even special cases of the scheduling problem (e.g., problems with either two different or identical machines, or problem with three jobs) are NP-hard [19], and the use of exact scheduling algorithms in practice is limited. The general methodology for developing exact algorithms may be based on a *Branch and Bound* (B&B) method [12,19]. The exhausting enumeration while *branching* set of feasible schedules ensures the optimality of the schedule obtained. The lower and upper *bounds* of the given objective function allow to eliminate schedules, which are dominated, and to reduce enumeration process. However, we are forced to recognize that current improvements of B&B algorithms cannot change radically the limit on the size of the scheduling problems solvable within practically reasonable running time. Moreover, it is unlikely to construct fast exact algorithm (or even approximate one with good performance measure) for scheduling problems of real-world size.

A common way of dynamically scheduling jobs is by means of dispatching rules [2]. “The problem of this method is that the performance of these rules depends on the state the system is in at each moment, and no single rule exists that is better than the rest in all the possible states that the system may be in. It would therefore be interesting to use the most appropriate dispatching rule at each moment. To achieve this goal, a scheduling approach that uses machine learning can be used. Analyzing the previous performance of the system (training examples) by means of this technique, knowledge is obtained that can be used to decide which is the most appropriate dispatching rule at each moment in time.” A review of the main machine learning-based scheduling approaches described in the literature is presented in [2].

Meta-heuristics [3] are general combinatorial optimization techniques, which are not dedicated to the solution of a particular problem, but are rather designed with the aim of being flexible enough to handle as many combinatorial problems as possible. These general techniques have rapidly demonstrated their usefulness and efficiency in solving hard problems.

Local decision rules are the backbone of heuristic algorithms, and the outcome of scheduling algorithm depends on the underlying strategy. As it was mentioned in [1], artificial intelligence (AI) provides often a better basis for modelling and solving the scheduling problem. *"Many real-world problems seem to be solved by 'semilogical' methods, such as recognizing one of a thousand familiar patterns applying to current situation and recalling the appropriate thing to do when that pattern occurs."* In [7], it was shown *"that by combining a learning system with simulation, a manufacturing control system can be developed that learns from its historical performance and makes its own scheduling and control decisions by simulating alternating combinations of dispatching rules."* The objective of the research described in [8] was to study the feasibility of automatically scheduling multi-machine complexes and adjusting the schedule on a real-time basis by a unified computer system. In [9], a probabilistic learning strategy has been developed based on the principles of evolutions. In [4-6], it was shown how the algorithm could be extended to consider local decision rules of any number and complexity in order that the resulting program will do better than deterministic learning approach.

In this paper, both OR and AI apparatuses are used. Our approach may be considered as a generalization of what proposed in [11] for special case of the problem when all machines are different

(job shop). We use mixed graph model [16], and so-called conflict resolution strategy [19]. The main issue of this paper is the classification of local conflicts due to computational results of exact and approximate versions of the B&B method.

Next, we formulate the scheduling problem in terms of the mixed graph model and give an overview of the conflict resolution strategy being a basis for exact, approximate and heuristic algorithms. Modules 1, 2 and 3 of the knowledge-based system are discussed in the following sections. Some remarks are given in Conclusion.

Mixed Graph Model

A general multistage system with both different and identical machines may be described via mixed graph $G = (V, A, E)$ [16], which is an appropriate model for constructing various types of scheduling algorithms [15]. In such a model, operation set V is represented as a set of vertices $V = \{1, 2, \ldots, n\}$, a weight prescribed to vertex $i \in V$ being equal to the processing time p_i of operation i. As usual, precedence constraints between operations are represented via arc set A. Let equality $V = \bigcup_{k=i}^{w} V_k$ hold, where V_k denotes the set of all operations processed by machines of type $k \in \{1, 2, \ldots, w\}$, and let $w_k = |M_k|$ be the number of identical machines of type k with $|M| = \sum_{k=1}^{w} w_k$. Competitions among operations that have to be processed by machines of the same type are represented via edge set E. Namely, edge $[i,j]$ belongs to set E if and only if operations i and j belong to the same set V^k.

On the one hand, if $r \in V_k$ and $s \in V_l$ with $k \neq l$, then operations r and s have to be processed by different machines (namely, by machines of different types: one machine from set M_k, and another one from set M_l). On the other hand, for each pair of operation i and operation j with edge $[i, j] \in E$ there exist three possibilities represented in Fig. 1. So, assigning operations V to machines M and sequencing operations assigned to the same machine correspond to removing or orienting edges in set E in one of the three possibilities. If all such edge transformations do not cause conflicts in the resulting digraph, then they define assigning operations V_k to machines M_k, $k = 1, 2, \ldots, w$, and define the sequence of operations assigned to each machine $\mu_m \in M$. Acyclic digraph constructed in such a way corresponds to a feasible schedule if the number w_k of available machines of each type k is sufficient for such a transformation of edges in the set E.

More precisely, selecting edges $[i, j]$ one by one from set E of the mixed graph G, and applying to each edge $[i,j]$ one of the three possible transformations (removing edge $[i, j]$, or substituting it either by arc (i, j) or by arc (j, i)), we obtain transformation sequence $\alpha^h = (a_1^h, a_2^h, \ldots, a_{|E|}^h)$ of all the edges in the mixed graph G. Let $G[\alpha^h] = (V, A[\alpha^h], \varnothing)$ denote a digraph obtained from the mixed graph G as a result of transformation sequence α^h. In [16], the following claim has been proved.

Theorem. *Digraph $G[\alpha^h]$ defines a feasible schedule if and only if:*

1) *digraph $G[\alpha^h]$ contains no circuits;*
2) *for each $k \in \{1, 2, \ldots, w\}$, the number of components in subgraph $(V_k, A_k[\alpha^h], \varnothing)$ of digraph $G[\alpha^h]$ is not greater than w_k;*
3) *each component in digraph $(V_k, A_k[\alpha^h], \varnothing)$ is a tournament, $k \in \{1, 2, \ldots, w\}$.*

Let $P(G)$ denote the set of all digraphs $G[\alpha^h]$ satisfying all the above conditions 1), 2) and 3). To find optimal schedule we may enumerate (implicitly via B&B method) digraphs of set $P(G)$, and choose the *optimal digraph,* i.e. the one with minimal value of the objective function.

Edge $[i,j] \in E$ is called *conflict* if both its orientations (i, j) and (j,i) in the mixed graph G cause the increase of the starting time of at least one operation from set V. Using *conflict resolution strategy,* we deal with one conflict edge at each iteration of the B&B algorithm, and test three possible transformations of this edge. The rule for choosing the edge from the set of conflict ones may be different, but the common idea is to choose the one with maximal value of conflict measure [12,19]. To find the solution of the scheduling problem, exact B&B algorithm may be fully used to realize implicit enumeration of all the feasible schedules. An approximate version of a B&B algorithm may be obtained from the exact one using some error bound allowed for the desired objective function value. A heuristic algorithm may be constructed by considering only one branch

of the solution tree via choosing at each iteration only one transformation of the conflict edge (e.g., due to some priority rule used for choosing operation from the set of two operations involved in the conflict). Thus, mixed graph model with conflict resolution strategy may be a foundation for constructing the variety of algorithms differing in the running time and in the accuracy of the solution obtained.

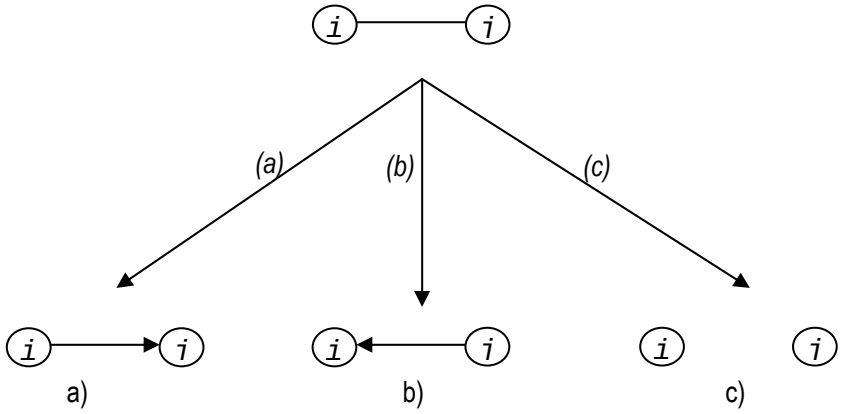


Fig. 1. Case a: $(i,j) \in A[\alpha^h]$ (case b: $(i,j) \in A[\alpha^h]$, respectively) if operations *i* and *j* are assigned to the same machine, and operation *i* (operation *j*) is processed before operation *j* (operation *i*); case c: edge [*i, j*] is removed from the mixed graph *G* if operations *i* and *j* are assigned to two different machines of the same type.

In our learning technology, we use the exact (or approximate) B&B algorithm in Module 1, while in Module 3 we use adaptive algorithm with logical rules produced within Module 2 by inductive inference. In the following sections, we show how to realize such an adaptation process using mixed graph model *G*.

Stage of Learning (Module 1 and Module 2)

At the stage of acquiring knowledge, we accumulate the knowledge of problems under consideration. Module 1 solves the sample problems via exact B&B algorithm (or via approximate B&B algorithm with allowed upper bound of the objective function value), and stores information on the successful local decisions that have led to optimal schedule (or to the best schedule obtained). The main idea for acquiring knowledge is to accumulate (in the form of so-called *learning table*) the information that was obtained during the use of the time-consuming exact (or approximate) B&B algorithm. In Section 4, we show how this knowledge could be used when solving other problems that are close to previous ones solved by Module 1 (i.e., problems with a similar structure of the processing system, close numerical data, the same objective function, etc.).

The aim of constructing the learning table is to accumulate the knowledge on successful local decisions used in the historical performance of the B&B algorithm. The stage of learning consists in accumulating information and tuning the parameters of the B&B algorithm to some properties of the class of the problems of real-world size to be solved. The learning table is analogous to those used in *pattern recognition,* and it describes which transformation of a conflict edge seems to be preferable. While solving the sample problem, information on successful transformations of the conflict edges has to be written in the learning table (see Table 1). Each raw of this table corresponds to one conflict edge. After B&B algorithm developed in [16] is stopped, the path in the solution tree is known that has led to the optimal schedule (or to the best schedule obtained). So, we can consider the set of conflict edges tested on this path within B&B algorithm. If B&B algorithm chooses the most conflict edge $[i_1,j_1]$ as the first one, Module 1 considers $[i_1,j_1]$ as the first one as well. Table columns corresponding to the characteristics of the object (conflict edge [*i, j*]) are filled by calculating characteristic values x^i of the object $[i,j]$. The last column in Table 1 contains the classes $\Omega_1, \Omega_2, ..., \Omega_7$ attributed to object $[i,j]$. Let $T_{u,v}$ denote the learning table with *v* objects and *u* their characteristics. In our software, number *u* is the parameter of the program fixed at the initial step, while number *v* of objects may grow with the growth of the number of sample problems solved by Module 1.

Notion "characteristic" x^i is very important for object description in any knowledge-based system. Next, we demonstrate how characteristics of the edges may be introduced by generalizing some priority rules. While creating the learning table for the first time, the characteristics are chosen from the given list and ordered with respect to their

"theoretical" informativeness measure. Each characteristic may correspond to a priority rule for a conflict resolution. After words having adopted on the class of problems, Module 1 and Module 3 may use also more complex (and often more informative) logical rule as a new characteristic created by Module 2. In this case, some useless (non-informative) characteristic has to be rejected from table $T_{u,v}$ (since number u of characteristics used in table $T_{u,v}$ is fixed). Replacement of characteristics is realized within Module 2.

In our software, we used first the characteristics corresponding to the priority rules given in [14]. A priority rule allows comparing priority values of two operations competing on machines of the same type. It is useful to introduce the corresponding characteristic equal to difference between these priority values for two operations i and j for the conflict edge $[i,j]$. To make such characteristics more universal and so applicable to other scheduling problems, it is reasonable to use relative difference of these priority values obtained via dividing them by some common value (e.g., by the maximal value of this characteristic in the mixed graph G). Next, we demonstrate typical examples of characteristics used in our computational experiments.

SPT rule recommends to process first the operation with the "shortest processing time", and then the operation with the greater processing time. So, we can introduce corresponding characteristic as follows: $x_1 = (p_i - p_j) / \max\{p_k : k \in V\}$. Value x_1 characterizes the edge $[i,j]$ more adequate than the corresponding priority rule: the sign of this value shows which operation i or j has the shortest processing time, and absolute value of x_1 makes it possible to compare edges of the initial mixed graph G with those of other mixed graphs.

FIFO rule recommends to process first the operation that has the shortest starting time ("first in, first out"). The starting time s_i of operation i can be easily calculated as the maximal weight of the path ending in vertex i of digraph $(V, A, \varnothing)$. Value x_2 of characteristic corresponding to priority rule *FIFO* may be calculated for edge $[i,j]$ as follows: $x_2 = (s_i - s_j) / \max\{s_k : k \in V\}$.

In our computational studies (realized for the special case of the problem with $|M_k| = 1$ for each $k = 1,2,\ldots,w$) especial attention has been paid to priority rule *LIOF* ("least increase in the objective function value"). Considering two operations i and j processed by the same machine, two values F_{ij} and F_{ji} of the objective function F are calculated for two possible orders of these operations. Value F_{ij} is calculated for partial schedule with operation i being processed first and operation j being processed second (in this case, edge $[i,j]$ has to be substituted by arc (i,j)). Value F_{ji} is calculated for partial schedule with operation j being processed first and operation i being processed second (edge $[i,j]$ has to be substituted by arc (j,i)). The *LIOF* rule recommends to choose the order defined by arc (i,j) for operations i and j if $F_{ij} < F_{ji}$, and to choose the opposite order defined by arc (j, i) otherwise.

The computational studies with three famous test problems from [13] showed that *LIOF* rule is a very informative one. More precisely, it was shown (by experiments) that while constructing an optimal schedule for the makespan criterion $C_{max} = \max\{C_i : i \in V\}$ (where C_i means the completion time of job $J_i \in J$) via B&B algorithm, the most part of the conflict edges was successfully oriented just due to the *LIOF* rule. Indeed, for some problems number of edges substituted by arcs with respect to *LIOF* rule formed more than 60 percent, and the number of exceptions formed only 1 percent of the whole number of conflict edges tested (for the rest 39 percent of conflict edges $[i,j]$, operations i and j were often equivalent in the sense of equality $F_{ij} = F_{ji}$).

Table 1: The form of the learning table $T_{u,v}$

Objects	Characteristic values				Classes
(conflict edges)	x^1	x^2	….	x^u	$\Omega_1, \Omega_2, \ldots, \Omega_7$
$[i_1,j_1]$	b_1^1	B_1^2	…	b_1^u	Ω_{k_1}
$[i_2,j_2]$	b_2^1	B_2^2	…	b_2^u	Ω_{k_2}
…	…	…	…	…	…
$[i_v,j_v]$	b_v^1	B_v^2	…	b_v^u	Ω_{k_v}
$[i,j]$	b^1	b^2	…	b^u	?

Thus, at the learning stage we consider the conflict edge of the mixed graph G as *object* using pattern recognition term. All columns in table $T_{u,v}$ (except the first and the last ones) contain characteristic values

of the objects. Namely, they form the description-vector for each conflict edge tested while constructing optimal schedule via Module 1. The first column enumerates conflict edges in the order as they have been tested in B&B algorithm on the path to optimal schedule (or to the best schedule obtained). The last column is used to classify objects in order to compare them with new objects within Module 3. At the solution stage (Module 3), the description-vector of the object is used to decide what edge substitution or removal has more probability to be successful.

Classes of Objects

To identify what transformation of conflict edge has been successful in the historical performance of the B&B algorithm, we introduce seven classes of objects: Ω_1, Ω_2, ..., Ω_7 (see last column of Table 1). In what follows, it is assumed that inequality $i < j$ holds for each edge $[i,j] \in E$.

Object $[i,j]$ belongs to class Ω_1 if it is preferable to substitute this edge by arc (i,j) in order to obtain effective schedule with respect to the given objective function. (In such a case, we say that arc (i,j) *dominates* both arc (j,i) and edge $[i,j]$).

Object $[i,j]$ belongs to class Ω_2 if it is preferable to substitute it by arc (j,i), and we say that arc (j,i) dominates both arc (i,j) and edge $[i,j]$.

Object $[i,j]$ belongs to class Ω_3 if it is preferable to remove edge $[i,j]$ from the mixed graph G, and we say that edge $[i,j]$ dominates both arcs (i,j) and (j,i).

If there is no enough information to decide what sole edge transformation is preferable, we have to include object $[i,j]$ to one of the classes Ω_4, Ω_5, Ω_6 or Ω_7 depending on available information. Namely, if arcs (i,j) and (j,i) simultaneously dominate edge $[i,j]$ but not each other, then $[i,j] \in \Omega_4$.

If arc (i,j) and edge $[i,j]$ simultaneously dominate arc (j,i) but not each other, then $[i,j] \in \Omega_5$.

If arc (j,i) and edge $[i,j]$ simultaneously dominate arc (i,j) but not each other, then $[i,j] \in \Omega_6$.

Finally, if there is no information on arc (edge) domination, object $[i,j]$ belongs to class Ω_7. Next, we render concrete the edge classification rules realized within Module 2.

The learning process is based on the solution tree constructed for the sample problem. Let scheduling problem S be solved by exact B&B algorithm within Module 1. Then we obtain optimal digraph $G[a^*] = (V,A[a^*],\varnothing)$ that define optimal schedule. Moreover, the path from vertex G to vertex $G[a^*]$ in the solution tree is known. Each vertex of this path is defined by some conflict edge $[i,j]$. If $(i,j) \in A[a^*]$, then $[i,j] \in \Omega_1$. If $(j,i) \in A[a^*]$, then $[i,j] \in \Omega_2$. If $(i,j) \notin A[a^*]$ and $(j,i) \notin A[a^*]$, then $[i,j] \in \Omega_3$. Thus, each object obtained due to the exact solution of the sample problem via exact B&B algorithm belongs to one of the three classes Ω_1, Ω_2, or Ω_3.

If the sample problem is solved only approximately, object classification is based on the relation between lower and upper bounds of the values of the given objective function (for short, we omit words "of the values of the given objective function").

Let scheduling problem S be solved by approximate *B&B* algorithm, then the best digraph $G[\alpha^\circ] = (V,A[a^\circ],\varnothing)$ constructed gives upper bound *UB* that is greater than the minimal lower bound calculated for all the pendant vertices of the solution tree. We obtain again path $\pi(G - G[\alpha^0])$ from vertex G to vertex $G[\alpha^0]$. Each intermediate vertex G' in this path defines some conflict edge $[i,j]$ for further bunching. Object $[i,j]$ has to be included in one of the seven classes. We describe the rule for this classification in detail only for the case when path $\pi(G - G[\alpha^0])$ includes arc (a): (a) $\in \pi(G - G[\alpha^0])$ (see Fig. 1). In this case, edge $[i,j]$ in the mixed graph G is substituted by arc (i,j). Classification rules for other two cases ((b) $\in \pi(G - G[\alpha^0])$ or [c] $\in \pi(G - G[\alpha^0])$ are analogous.

Let (a) $\in \pi(G - G[\alpha^0])$. In the solution tree, three possible transformations have been realized for edge $[i,j]$ (see Fig. 1). Let $LB(j,i)$ ($LB[i,j]$ respectively) denote the minimal lower bound calculated for all the pendant vertices G'' to which there exists a path $\pi(G' - G'')$ from vertex G' starting with arc (b): (b) $\in \pi(G' - G'')$ (starting with arc (c): (c) $\in \pi(G' - G'')$). If $LB(j,i) \geq UB$ and $LB[i,j] \geq UB$, then $[i,j] \in \Omega_1$. If $LB(j,i) < UB$ and $LB[i,j] \geq UB$, then $[i,j] \in \Omega_4$. If $LB(j,i) \geq UB$ and $LB[i,j] < UB$, then $[i,j] \in \Omega_5$.

If (b) $\in \pi(G - G[\alpha^0])$, then object $[i,j]$ belongs to one of the classes Ω_2, Ω_4 or Ω_6. If (c) $\in \pi(G - G[\alpha^0])$, then object $[i,j]$ belongs to one of the classes Ω_3, Ω_5 or Ω_6.

At the end of the stage of learning, the information accumulated in table $T_{u,v}$ is used in Module 2 to construct the *recognition table, which* is used at the solution stage within Module 3. A recognition table has the similar form as table $T_{u,v}$ has, except the last row added to recognition table. But we use the same notation $T_{u,v}$ for recognition table as well. To construct the recognition table a subset of the most informative characteristics is selected and ordered with respect to their informativeness. Some characteristics that turned out to be not essential for sample problems have to be rejected from table $T_{u,v}$. In [11], it is shown how to select the most informative characteristics. Module 2 systematizes the knowledge obtained using the pattern recognition apparatus, and constructs the generalized logical rule for conflict resolution. Such a rule describes successful local decisions and may be used in Module 3 for more efficient solving large scale scheduling problems via adopting successful local decisions by analogy.

Solution Stage (Module 3)

Scheduling algorithm used as Module 3 occupies intermediate position between heuristic and approximate B&B algorithms depending on the number of sample problems solved exactly or approximately at the stage of learning. Conflict resolution strategy is used in Module 3 as well. If Module 3 chooses conflict edge $[i,j]$, it has to recognize to which class from set $\{\Omega_1, \Omega_2, ..., \Omega_7\}$ edge $[i,j]$ has to belong. Thus, a recognition problem arises, and it is reasonable to adopt to a new situation the strategy that has led to successes at the learning stage.

To solve recognition problem Module 3 uses the procedure based on the bound calculation. A resemblance bound is calculated that characterizes the "distance" between the description vector of recognizable object $[i,j]$, and that of sample objects with respect to a combination H of characteristics x^1, x^2,..., x^u used in the recognition table $T_{u,v}$. We have to determine edges $[i_k,j_k]$ given in table $T_{u,v}$ such that the corresponding vectors $(b^1_k, b^2_k, .., b^{\hat{u}}_k)$ are the closest ones to vector $(b^1, b^2,...,b^u)$ with respect to the characteristics $x^1, x^2.., x^u$. Two objects $[i,j]$ and $[i_k,j_k]$ are considered to be similar with respect to the system H of characteristics if at least δ inequalities $|b^j - b^j_k| \le \varepsilon_k$ with $j = 1,2,\ldots,u$ and $x^j \in H$ are satisfied where ε_1, $\varepsilon_2,\ldots, \varepsilon_u$, δ being parameters of Module 3. Bounds are calculated for the number of objects of the learning table that are similar with respect to the system H. The analysis of such bounds allows to decide to which class $\Omega_1, \Omega_2, \ldots, \Omega_7$ the edge $[i,j]$ can be assigned.

Due to this we can decide what transformations are preferable for the conflict edge $[i,j]$, and how many transformations of this edge have to be treated (one, two or three). If $[i,j] \in \Omega_1$, then solution tree contains only arc (a) (see Fig. 1). If $[i,j] \in \Omega_2$, then solution tree contains only arc (b). If $[i,j] \in \Omega_3$, then solution tree contains only arc (c). If $[i,j] \in \Omega_4$, then solution tree contains two arcs (a) and (b). If $[i,j] \in \Omega_5$, then solution tree contains two arcs (a) and (c). If $[i,j] \in \Omega_6$, then solution tree contains two arcs (b) and (c). If $[i,j] \in \Omega_7$, then solution tree contains three arcs (a), (b) and (c).

Conclusion

The use of an exact scheduling algorithm is computationally expensive and so impracticable. Therefore, at present rather common methodology carried out by practitioners is the use of simple heuristics and weakly grounded dispatching which does not ensure solution accuracy. To overcome these difficulties, we propose a knowledge-based system that may be used for the adaptation of general B&B algorithm for specific practical scheduling problems. The process of problem solving is partitioned into two stages: the *learning* stage realized by Module 1 and Module 2, and the *solution* stage realized by Module 3. Module 1 creates the database of learning information for the certain class of scheduling problems and enlarges it while solutions of new problems become available. At the stage of learning, we can use practical schedules obtained via commercial packages, e.g., [10]. The volume of information, its correctness and absence of noises depends on the quality of scheduling algorithms applied in the frameworks of this module and on the time given on learning process. More exact algorithms requiring large computing time may ensure more informative knowledge. So, it is important to use a good model and exact algorithms within Module 1. The algorithm used in Module 3 should consider at each step the same question as the algorithm used in Module 1. The bridge linking

Module 1 and Module 3 is Module 2, which automates the learning of peculiarities of the problems under consideration and gives foundation for creating generalized logical rule for conflict resolution. Moreover, Module 2 produces a generalized logical rule, which may be considered as a new comprehensive heuristic developed just for problem class under consideration.

While the reasons for applying certain heuristics in dispatching algorithms are weakly grounded and are based mainly on the experiment, the choice of heuristics in the knowledge-based system is more precise (adequate) and is computer-aided. Results show that a solution can always be obtained for changing the dispatching rule so that the system is better than one using fixed scheduling rules [17]. Obtained in [18] results demonstrate that proposed approach produces an improvement in the performance of the system when compared to the traditional method of using dispatching rules.

Bibliography

[1] Wiers V.C.S. A review of the applicability of OR and AI scheduling techniques in practice. Omega, International Journal of Management Science, 1997, 25(2), 145-153.
[2] Priore P.,De La Fuente D.,Gomez A., Puente J. A review of machine learning in dynamic scheduling of flexible manufacturing systems, AI EDAM, 2001,15(30), 251-263.
[3] Hertz A., Widmer M. Guidelines for the use of meta-heuristics in combinatorial optimization. European Journal of Operational Research, 2003, 151, 247–252.
[4] Wang J.M., Usher J.M. Application of reinforcement learning for agent-based production scheduling. Engineering Applications of Artificial Intelligence, 2005, 18(1), 73-82 .
[5] Paternina-Arboleda C.D., Das T.K. A multi-agent reinforcement learning approach to obtaining dynamic control policies for stochastic lot scheduling problem. Simulation Modelling Practice and Theory, 2005 (to appear, available online).
[6] Whiteson S., Stone P. Adaptive job routing and scheduling. Engineering Applications of Artificial Intelligence, 2004, 17(7), 855-869.
[7] Kim M.H., Kim Y-D. Simulation-based real-time scheduling in a flexible manufacturing system. Journal of Manufacturing Systems, 1994, 13(2), 85-93.
[8] Nof S.Y., Grant F.H. Adaptive/predictive scheduling: review and a general framework. Production Planning and Control, 1991, 2(4), 298-312.
[9] Dorndorf U., Pesch E. Evolution based learning in a job shop scheduling environment. Computers Operations Research., 1991, 22(1), 25-40.
[10] Janicke W. Excel planning interface for the SAP R/3 System. Chemical Engineering & Technology, 2001, 24(9), 903-908.
[11] Shakhlevich N.V., Sotskov Yu.N., Werner F. Adaptive scheduling algorithm based on mixed graph model. IEE Proc.-Control Theory Appl., 1996, 143(1), 9-16.
[12] Jain A.S., Meeran S. Deterministic job-shop scheduling: past, present and future. European Journal of Operational Research, 1999, 113, 390-434.
[13] Muth J.F., Thompson G.L. Industrial scheduling. Englewood Cliffs, N.Y., Prentice Hall. 1963.
[14] Panwalkar S.S., Iskander W. A survey of scheduling rules. Operations Research, 1977, 25, 45-61.
[15] Sotskov Yu.N. Software for production scheduling based on the mixed (multi)graph approach. Computing & Control Engineering Journal, 1996, 7(5), 240-246.
[16] Sotskov Yu.N. Mixed Multigraph approach to scheduling jobs on machines of different types. Optimization, 1997, 42, 245-280.
[17] Chan F.T.S., Chan H.K. Dynamic scheduling for a flexible manufacturing system - the pre-emptive approach. International Journal of Advanced Manufacturing Technology, 2001, 17(10), 760-768.
[18] Priore P., D.L.Fuente, Gomez A., Puente J. Dynamic scheduling of manufacturing systems with machine learning. International Journal of Foundation of Computer Science 2001, 12(6) 751-762.
[19] Tanaev V.S., Sotskov Yu.N., Strusevich V.A. Scheduling theory. Multi-stage systems. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

Authors' Information

Yuri N. Sotskov – United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganova str., 6, Minsk, Belarus; e-mail: sotskov@newman.bas-net.by

Nadezhda Sotskova – Hochschule Magdeburg-Stendal, Fachbereich Wasserwirtschaft, PSF 3680, 39011 Magdeburg; Germany; e-mail: drsotskova@hotmail.com

Leonid V. Rudoi – United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganova str., 6, Minsk, Belarus; e-mail: margol@newman.bas-net.by

4.4. Intelligent Technologies in Control

АВТОМАТНЫЙ МЕТОД РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ ОГРАНИЧЕНИЙ В ОБЛАСТИ {0,1}

Сергий Крывый, Людмила Матвеева, Виолета Гжывач

***Аннотация**: Рассматривается автоматный метод построения базиса множества всех решений системы линейных диофантовых уравнений с коэффициентами из множества {-1,0,1} в области {0,1}.*

***Ключевые слова**: системы линейных диофантовых уравнений, базис множества решений, конечный автомат без выходов.*

Введение

Алгоритмы построения базиса множества всех решений системы линейных однородных диофантовых уравнений (СЛОДУ) в множестве натуральных чисел встречаются во многих приложениях. Среди таких приложений находятся сети Петри (поиск инвариантов, ловушек и дедлоков) [Murata,1989], проверка противоречивости множества дизъюнктов [Baader, 1994], арифметика Пресбургера [Common, 1999], логический вывод в исчислении предикатов [Lloyd,1987] и т.д. Частным случаем СЛОДУ являются СЛОДУ, решения которых ищутся в множестве {0,1}, а коэффициенты берутся из множества {-1,0,1}. В данной работе описывается алгоритм построения базиса множества всех решений СЛОДУ этого частного вида. Этот алгоритм применим к СЛОДУ общего вида, а также к системам неоднородных уравнений и неравенств.

1. Необходимые сведения

СЛОДУ, которые будут рассматриваться ниже, имеют вид

$$S = \begin{cases} a_{11}x_1 + \ldots + a_{1q}x_q = 0 \\ a_{21}x_1 + \ldots + a_{2q}x_q = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{p1}x_1 + \ldots + a_{pq}x_q = 0 \end{cases}, \qquad (1)$$

где $a_{ij} \in \{-1,0,1\}$, $x_j \in \{0,1\}$, $i \in [1,p]$, $j \in [1,q]$. Вектора вида $e_1 = (1,0,\ldots,0,0)$, $e_2 = (0,1,\ldots,0,0)$, $\ldots$, $e_q = (0,0,\ldots,0,1)$ называются векторами канонического базиса. Вектор $c = (c_1,\ldots,c_q)$, $c_i \in \{0,1\}$, называется решением СЛОДУ S, если имеют место тождества

$$\forall \quad i \in [1,p] \quad a_{i1}c_1 + a_{i2}c_2 + \ldots + a_{iq}c_q \equiv 0,$$

где знак $\equiv$ означает отношение тождества. Сравнение двух векторов $x = (x_1,\ldots,x_q)$ и $y = (y_1,\ldots,y_q)$ из множества N^q определяется следующим образом

$$x \le y \Leftrightarrow \forall \quad i \in [1,q] \quad x_i \le y_i.$$

Это отношение является отношением частичного порядка и минимальные вектора-решения СЛОДУ S относительно данного порядка составляют базис множества всех решений СЛОДУ S. Опишем алгоритм построения этого базиса на методами теории автоматов.

2. Описание алгоритма и его обоснование.

Предлагаемый ниже алгоритм базируется всецело на теории конечных автоматов без выходов. Для того чтобы описать идею построения алгоритма, рассмотрим поочередно случаи одного однородного диофантового уравнения, представление базиса множества его решений в виде автомата, а затем рассмотрим случай системы таких уравнений.

2.1. Случай одного уравнения

Пусть имеем уравнение вида

$$a_1x_1 + a_2x_2 + \ldots + a_qx_q = 0, \qquad (2)$$

$a_j \in \{-1,0,1\}$, $x_j \in \{0,1\}$, $j \in [1,q]$. Применив TSS-алгоритм [Krivoi, 2003] для построения множества решений уравнения (2), получаем базис множества всех решений уравнения (2). Это вытекает из следующего утверждения,

Теорема 1. Если уравнение $a_1x_1 + a_2x_2 + \ldots + a_qx_q = 0$ такое, что $a_j \in \{-1,0,1\}$, $j \in [1,q]$, то TSS этого уравнения совпадает с базисом множества всех его решений в области {0,1}.

Доказательство. Пусть $c = (c_1,\ldots,c_q)$, $c_i \in \{0,1\}$ – произвольное решение уравнения (2) и $c_{j1},\ldots,c_{jr}$ все его ненулевые координаты. Удалим из этого множества ненулевых координат все те, которые соответствуют нулевым коэффициентам уравнения (2). Это значит, что строится вектор $c' = c - e_{j1} - \ldots - e_{jr}$, где e_{ji} $(i \in [1,r])$ – вектора канонического базиса, которые по построению являются элементами TSS. Очевидно, что построенный таким образом вектор c' тоже есть решением уравнения (2). Если $c'=0$, то все доказано и $c = e_{j1} + \ldots + e_{jr}$. Если $c' \neq 0$, то число его ненулевых координат должно быть четным и, более того, одна половина из его координат соответствует коэффициентам –1 в уравнении (2), а вторая половина коэффициентам +1. Поскольку TSS содержит все вектора, соответствующие комбинациям коэффициентов –1 и +1, то ясно, что c' представим в виде неотрицательной линейной комбинации вида

$$c' = e_{l1} + \ldots + e_{lt},$$

где $e_{ji} \in TSS$ уравнения (2). Если учесть, что $c' = c - e_{j1} - \ldots - e_{jr}$, то $c = e_{j1} + \ldots + e_{jr} + e_{l1} + \ldots + e_{lt}$.

Теорема доказана.

Следствия.

(а) Число базисных решений уравнения (2) равно k*m + r, где k, m – число отрицательных и положительных коэффициентов соответственно, а r – число нулевых коэффициентов в этом уравнении;

(б) Построение базиса множества всех решений уравнения (2) выполняется за время $O(q^2)$.

Пример 1. Рассмотрим уравнение

$$1x_1 + 0x_2 - 1x_3 - 1x_4 + 1x_5 + 1x_6 - 1x_7 = 0.$$

TSS этого уравнения, составляющие базис множества всех его решений, состоит из таких векторов:

$s_1 = (0,1,0,0,0,0,0)$, $s_2 = (1,0,1,0,0,0,0)$, $s_3 = (1,0,0,1,0,0,0)$, $s_4 = (1,0,0,0,0,0,1)$, $s_5 = (0,0,1,0,1,0,0)$,
$s_6 = (0,0,1,0,0,1,0)$, $s_7 = (0,0,0,1,1,0,0)$, $s_3 = (0,0,0,1,0,1,0)$, $s_9 = (0,0,0,0,1,0,1)$, $s_{10} = (0,0,0,0,0,1,1)$.

Число нулевых коэффициентов в этом уравнении, т.е. число r=1, числа k=3 и m=3.соответственно. Следовательно, k*m + r = 3*3 + 1 = 10.

2.2. Автоматное представление базиса решений

Пусть A = (A, X, f, $\{a_0\}$, F) – конечный автомат без выходов, где A – конечное множество состояний, X - конечное множество входных символов, f: $A \times X \to A$ – (частичная) функция переходов, $a_0 \in A$ – начальное состояние, $F \subseteq A$ - множество заключительных состояний.

Представление базиса множества решений уравнения (2) с помощью автомата имеет следующий вид. Множество состояний автомата $A = \{0,1,-1\}$, $X = \{e_1,e_2,\ldots,e_q\}$, где e_i $(i \in [1,q])$ – вектора канонического базиса множества N^q, начальным и заключительным состоянием является состояние 0, а функция переходов f определяется так:

$f(0,e_i) = 0$, если e_i соответствует коэффициенту 0;
$f(0,e_i) = 1$, если e_i соответствует коэффициенту 1;
$f(0,e_i) = -1$, если e_i соответствует коэффициенту -1;
$f(1,e_i) = 0$, если e_i соответствует коэффициенту –1;
$f(-1,e_i) = 0$, если e_i соответствует коэффициенту 1;

иначе функция f неопределена.

Пример 2. Для уравнения из примера 1 имеем:

$A = \{0,1,-1\}$, $X = \{e_1,e_2,\ldots,e_7\}$, $a_0 = 0$, $F = \{a_0\}$, а функция переходов f определяется такой таблицей переходов:

f	0	1	-1
e_1	1	-	0
e_2	0	-	-
e_3	-1	0	-
e_4	-1	0	-
e_5	1	-	0
e_6	1	-	0
e_7	-1	0	-

Вектор-решение строится по слову p, которое переводит автомат из состояния 0 в состояние 0, следующим образом. Если $p = e_{i1}e_{i2}\ldots e_{ik}$, то $x_p = e_{i1} + e_{i2} + \ldots + e_{ik}$, где $e_{ij} \in X$, $j \in [1,k]$. Ясно, что нас интересуют слова, соответствующие простым циклам в автомате и переводящие автомат из состояния 0 в состояние 0. Например, в вышеприведенном автомате слова $p_1 = e_1e_3$, $p_2 = e_1e_4$, $p_3 = e_1e_7$, $p_4 = e_5e_3$, $p_5 = e_5e_4$, $p_6 = e_5e_7$, $p_7 = e_6e_3$, $p_8 = e_6e_4$, $p_9 = e_6e_7$ и $p_{10} = e_2$ порождают базис множества всех решений, состоящий из векторов s_1-s_{10} *(см. пример 1).*

Теорема 2. Автомат, определенный описанным выше способом, представляет все множество решений уравнения (2).

Доказательство следует непосредственно из способа построения автомата. Действительно, все базисные вектора-решения, соответствующие нулевым коэффициентам, представлены в данном автомате, поскольку соответствующие им слова переводят автомат из состояния 0 в состояние 0. Далее все базисные вектора-решения, соответствующие комбинациям –1 и 1 тоже представлены словами, которые переводят данный автомат из состояния 0 в состояние 0. Поскольку этими векторами исчерпываются все базисные вектора-решения уравнения (2), то отсюда и следует справедливость данной теоремы.

2.3. Случай системы уравнений

Не ограничивая общности, будем рассматривать СЛОДУ, которая состоит из двух уравнений. Представим каждый из базисов множества решений каждого из уравнений в виде автомата, описанным выше способом. Построим прямое произведение полученных автоматов и найдем все слова, соответствующие простым путям из состояния (0,0) (как единственного начального и заключительного состояния) в это же состояние. Утверждается, что данное множество включает все базисные вектора множества решений СЛОДУ. Прежде чем доказать это утверждение, рассмотрим пример.

Пример 3. Рассмотрим СЛОДУ

$$S = \begin{cases} 1x_1 + 0x_2 - 1x_3 - 1x_4 + 1x_5 + 1x_6 - 1x_7 = 0 \\ 1x_1 - 1x_2 + 0x_3 + 0x_4 - 1x_5 - 1x_6 + 1x_7 = 0 \end{cases}.$$

Таблица переходов автомата, соответствующего первому уравнению приведенной СЛОДУ приведена выше (см. пример 1), а таблица переходов автомата, соответствующего второму уравнению имеет вид:

f_1	0	1	-1
e_1	1	-	0
e_2	-1	0	-
e_3	0	-	-
e_4	0	-	-
e_5	-1	0	-
e_6	-1	0	-
e_7	1	-	0

Таблица переходов прямого произведения полученных автоматов после удаления недостижимых и тупиковых состояний выглядит следующим образом:

f_2	(0,0)	(0,-1)	(-1,0)	(1,-1)	(-1,1)	(1,0)	(0,1)
e_1	-	(1,0)	(0,1)	-	-	-	-
e_2	(0,-1)	-	-	-	-	(0,0)	(0,0)
e_3	(-1,0)	-	-	-	-	-	-
e_4	(-1,0)	-	-	-	-	-	-
e_5	(1,-1)	-	(0,-1)	-	(0,0)	(1,0)	(1,0)
e_6	(1,-1)	-	(0,-1)	-	(0,0)	(1,0)	(1,0)
e_7	(-1,1)	(-1,0)	-	(0,0)	-	-	-

Начальным и заключительным состоянием данного автомата является состояние (0,0). Поиск простых циклов в полученном автомате приводит к таким словам:

$p_1 = e_2e_1e_3$, $p_2 = e_2e_1e_4$, $p_3 = e_3e_1e_5e_4$, $p_4 = e_3e_1e_6e_4$, $p_5 = e_5e_7$,

$p_6 = e_6e_7$, $p_7 = e_7e_6$, $p_8 = e_2e_7e_1e_5e_3$, $p_9 = e_2e_7e_1e_6e_3$

и так далее. Эти слова порождают следующие базисные решения:

$s_1 = (1,1,1,0,0,0,0)$, $s_2 = (1,1,0,1,0,0,0)$, $s_3 = (1,0,1,1,1,0,0)$,

$s_4 = (1,0,1,1,0,1,0)$, $s_5 = (0,0,0,0,1,0,1)$, $s_6 = (0,0,0,0,0,1,1)$.

Анализируя рассмотренные выше примеры и все вышесказанное, приходим к таким выводам:

- размеры автоматов, соответствующие уравнениям СЛОДУ, не превосходят величины *3*q*;
- память, необходимая для представления исходных автоматов, пропорциональна величине *3*p*q;*
- длина самого длинного простого цикла в прямом произведении исходных автоматов не превосходит величины *q.*

Перейдем теперь к обоснованию данного алгоритма.

Теорема 3. Множество решений СЛОДУ, соответствующих словам, представляющим простые циклы в автомате, который есть прямым произведением исходных автоматов, содержит все базисные решения данной СЛОДУ.

Доказательство. Не ограничивая общности, предположим, что СЛОДУ имеет два уравнения. Пусть x – произвольное решение СЛОДУ, тогда вектору x в автомате, представляющем первое уравнение, соответствует слово p и этому вектору x соответствует такое же слово p и в автомате, представляющем второе уравнение данной СЛОДУ. Это вытекает из того, что исходные автоматы представляют множество всех решений СЛОДУ. Но тогда слово p, будучи елементом пересечения двух регулярных языков, представленным в прямом произведении исходных автоматов, должно переводить этот автомат из начального состояния в начальное состояние (которое является и единственным заключительным состоянием в этом автомате). Это значит, что слову p соответствует некоторый цикл в этом автомате. Представим это слово как конкатенацию его подслов $p_1, p_2,\ldots,p_k$, соответствующих простым циклам. Поскольку слова $p_1, p_2,\ldots,p_k$ генерируют базисные решения, то пусть это будут решения $e_1, e_2,\ldots,e_k$. Тогда становится очевидным, что $x = e_1 + e_2 + \ldots + e_k$. Теорема доказана.

3. Временные характеристики алгоритма

Как следует из общей теории автоматов, количество состояний в автомате, который представляет множество всех решений СЛОДУ, имеющей p уравнений, не превосходит величины 3^p. Эта оценка говорит о том, что время построения такого автомата и поиск всех простых циклов с выделением базисных решений пропорционально величине 3^p, что не является оптимистической оценкой. Полученная оценка не противоречит результату о экспоненциальной сложности алгоритма построения базиса множества решений СЛОДУ. Однако, в большинстве случаев число состояний результирующего автомата имеет значительно меньше состояний, чем 3^p. Очевидной избыточностью вычислений предложенного алгоритма является то, что некоторые решения вычисляются многократно. Например, решение (0,0,0,0,1,0,1) вычисляется дважды, как и другие решения, а автомат, таблица переходов которого приведена ниже, представляет то же множество решений, что и автомат из примера 3, но имеет меньшее число состояний.

f_3	(0,0)	(0,-1)	(-1,0)	(-1,1)	(1,0)
e_1	-	(1,0)	-	-	-
e_2	(0,-1)	-	-	-	-
e_3	(-1,0)	-	-	-	-
e_4	(-1,0)	-	-	-	-
e_5	-	-	(0,-1)	(0,0)	(1,0)
e_6	-)	-	(0,-1)	(0,0)	(1,0)
e_7	(-1,1)	-	-	-	-

Необходимо также заметить, что поскольку решения СЛОДУ ищутся в множестве {0,1}, то все вектора, фигурирующие в данном алгоритме являются булевскими, что сильно упрощает вычисления.

Наиболее существенным преимуществом данного алгоритма является то, что построение базиса множества всех решений можно вести не прибегая к явному построению автомата, являющегося прямым произведением исходных автоматов, а строить базис непосредственно по исходным автоматам путем параллельного движения по соответствующим состояниям исходных автоматов. Это позволяет сильно экономить память и, следовательно, более эффективно выполнять вычисления. Например, явно строить результирующий автомат в примере 3 нет необходимости. Построение базиса можно вести по исходным автоматам, представляющим базисы множеств решений первого и второго уравнений соответственно.

Наконец, как отмечалось выше, в предлагаемом алгоритме имеется возможность оптимизации его работы. Эта возможность состоит в том, что при построении базиса множества решений по автомату представляющему все множество решений, многие решения находятся по несколько раз. Эта избыточность связана с тем, что операция конкатенации в данном случае является коммутативной. В связи с этим возникает ***проблема минимизации конечного автомата с учетом коммутативности операции конкатенации слов во входном алфавите автомата.***

4. Другой подход к построению автомата, представляющего базис множества решений

Опишем еще один способ построения автомата, представляющего базис множества всех решений СЛОДУ, который не требует построения прямого произведения автоматов.

Пусть X = {s_1, s_2,…, s_n} – базис множества решений первого уравнения СЛОДУ S и $L_2(x) = 0$ – второе уравнение системы S. Используя множество X и функцию $L_2(x)$, построим конечный автомат A = (A, X, {0}, f, {0}), в котором состояние 0 есть начальным и заключительным, ва функция переходов определяется следующим образом:

$$f(a,s_i) = \begin{cases} L_2(s_i), если a = 0, \\ a + L_2(s_i), если a \neq 0 \& |a + L_2(s_i)| < |a|, \\ неопределено, иначе \end{cases}$$

По построенному таким образом автомату строим базис множества всех решений подсистемы $L_1(x) = 0$ & $L_2(x) = 0$ системы S. Элементы этого базиса играют роль нового алфавита. Этот алфавит и функция $L_3(x)$ (если она имеется), которая стоит в левой части уравнения $L_3(x) = 0$, служат для построения следующего автомата и так далее до тех пор, пока не будет построен базис множества всех решений СЛОДУ S.

Рассмотрим СЛОДУ S из примера 3. Значениями функции $L_2(x)$ на векторах из множества X = {s_1,s_2,…,s_{10}} являются числа: -1, 1, 1, 2, -1, -1, -1, -1, 0, 0. Автомат A = (A, X, {0}, f, {0}), соответствующий алфавиту X и функции $L_2(x)$, имеет вид:

f	0	-1	1	2
s_1	-1	-	0	1
s_2	1	0	-	-
s_3	1	0	-	-
s_4	2	-	-	-
s_5	-1	-	0	1
s_6	-1	-	0	1
s_7	-1	-	0	1
s_8	-1	-	0	1
s_9	0	-	-	-

s_{10}	0	-	-	-

Строим вектора-решения СЛОДУ. Этим векторам соответствуют простые циклы из состояния 0 в это же состояние. В данном случае это будут такие вектора:

$s_1 + s_2 = (1,1,1,0,0,0,0)$, $s_1 + s_3 = (1,1,0,1,0,0,0)$, $s_5 + s_2 = (1,0,2,0,1,0,0)$, $s_6 + s_2 = (1,0,2,0,0,1,0)$,

$s_6 + s_3 = (1,0,1,1,0,1,0)$, $s_7 + s_2 = (1,0,1,1,1,0,0)$. $S_7 + s_3 = (1,0,0,2,1,0,0)$, $s_8 + s_2 = (1,0,1,1,0,1,0)$,

$s_8 + s_3 = (1,0,0,2,0,1,0)$, $s_4 + s_1 + s_5 = (1,1,1,0,1,0,1)$, $s_4 + s_1 + s_6 = (1,1,1,0,0,1,1)$, $s_4 + s_1 + s_7 = (1,1,0,1,1,0,1)$.

$s_4 + s_1 + s_8 = (1,1,0,1,0,1,1)$, $s_4 + s_5 + s_6 = (1,0,2,0,1,1,1)$, $s_4 + s_5 + s_7 = (1,0,1,1,2,0,1)$,

$s_4 + s_5 + s_8 = (1,0,1,1,1,1,1)$, $s_4 + s_6 + s_7 = (1,0,1,1,1,1,1)$, $s_4 + s_6 + s_8 = (1,0,1,1,0,2,1)$,

$s_4 + s_7 + s_8 = (1,0,0,2,1,1,1)$, $s_9 = (0,0,0,0,1,0,1)$, $s_{10} = (0,0,0,0,0,1,1)$.

Базисными векторами будут вектора:

$s_1 = (1,1,1,0,0,0,0)$, $s_2 = (1,1,0,1,0,0,0)$, $s_3 = (1,0,1,1,1,0,0)$,

$s_4 = (1,0,1,1,0,1,0)$, $s_5 = (0,0,0,0,1,0,1)$, $s_6 = (0,0,0,0,0,1,1)$,

что подтверждается решением этой СЛОДУ предыдущим способом.

4. Заключение

Предложенные алгоритмы могут применяться как к СЛОДУ общего вида, так и к системам неоднородных уравнений и системам линейных неравенств. Для этого вводится одна (в случае системы неоднородных уравнений) или несколько (в случае системы неравенств) дополнительных переменных и таким образом задача сводится к поиску решений СЛОДУ в множестве {0,1}.

Данные алгоритмы особенно эффективны для определения несовместности СЛОДУ и для нахождения первого решения СЛОДУ. Действительно, если СЛОДУ несовместна, то прямое произведение автоматов или автомат, строящийся во втором алгоритме, которые представляют базисы множеств решений данной СЛОДУ, будут пустыми автоматами. Как следует из способа построения таких автоматов, они будут построены достаточно быстро. А если в процессе построения таких автоматов появился первый цикл, то это свидетельствует о совместности данной СЛОДУ и, при необходимости, можно найти решение СЛОДУ, соответствующее этому циклу в автомате и прекратить дальнейшие вычисления.

Отметим, наконец, что и во втором подходе возникает та же проблема минимизации конечного автомата с учетом коммутативности операции конкатенации слов в алфавите X.

Литература

[Murata, 1989] T. Murata. Petri Nets: Properties, Analysis and Applications. In Proceedings of the IEEE, 1989, Vol. 77, № 4, p. 541-580.

[Baader, 1994] Baader F., Ziekmann J. Unification Theory. In Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford University Press. 1994. p. 1-85.

[Common, 1999] Comon H. Constraint solving on terms: Automata techniques (Preliminary lecture notes). Intern. Summer School on Constraints in Computational Logics: Gif-sur-Yvette, France, September 5–8. 1999. 22 p.

[Lloyd,1987] Lloyd J. Foundations of Logic Programming (2-d adition). Springer Verlag. 1987. 276 p.

[Krivoi, 2003]. Krivoi S. Об алгоритмах решения систем линейных диофантовых констрейнтов в области {0,1}, ж. Кибернетика и системный анализ. 2003, № 5. с. 58-69.

Информация об авторах

Sergiy Kryvyy – Institute of cybernetics, NAS of Ukraine, Pr. Glushkova, 40, Kiev-03187, Ukraine; e-mail: krivoi@i.com.ua

Lyudmila Matvyeyeva – Institute of cybernetics, NAS of Ukraine, Pr. Glushkova, 40, Kiev-03187, Ukraine; e-mail: luda@iss.org.ua

Wioletta Grzywac – Technical University of Czestochowa, Czestochowa, Poland, Dabrowskiego str. 73; e-mail: wiola@icis.prz.czest.pl

LOGICAL MODELS OF COMPOSITE DYNAMIC OBJECTS CONTROL

Vitaly J. Velichko, Victor P. Gladun, Gleb S. Gladun, Anastasiya V. Godunova, Yurii L. Ivaskiv, Elina V. Postol, Grigorii V. Jakemenko

Abstract: *The questions of designing multicriteria control systems on the basis of logic models of composite dynamic objects are considered.*

Keywords: *control, logical model, composite dynamic object, balancing network.*

1. Introduction

The special attention in the class of control systems is deserved by multicriteria control systems of composite dynamic objects (CDO). As composite dynamic objects we shall consider such technical objects with a broad spectrum of states and plenty of permissible transitions between states, the control by which one consists in implementation of often and miscellaneous control actions in connection with often changes modes of exploitation, purposes, conditions of operation, preferences of the user, estimation's criterions.

It is necessary to take into account the following conditions at designing control systems of such objects.

1. A basic function of control systems is the selection of control actions ensuring optimization of a state of control object. For implementation of this function in multicriteria systems, the methods of vectorial optimization are used. In the basis, which one lay selection operations of alternatives by confrontation of their vectorial estimations by given criterions. The run time of these base operations depends first of all on organization of structure presenting the information on alternatives and criterions of estimation in memory of a management system. The structures used for representation of information should provide at all stages of control procedure selection of indispensable alternatives or criterions of estimation without looking through appropriate sets. The factor of processing speed is specially important for control in extreme situations. The designing of information structures accelerating processes of selection, is a problem of logical designing [1-3].

2. The problems of multicriteria vectorial optimization have no precise solution more often. Usually process of the initial problem solution ends getting sequence of effective not predominant alternatives (Pareto set). Further "improvement" of the solution can be obtained only at the expense of calling the additional information from person who takes decision [Decision taking Person](PTD). Attempts to "improve" the solution are implemented by means of sort out alternatives or criterions reflecting preferences PTD. An effective means of selection the "best" solutions become testing of versions proposed PTD. PTD preferences should be envisaged in a structure of a control system at a phase of its logical designing. For preliminary modelling and selection of the control solutions, which are taking into account preference PTD, it is useful in design organizations to use a special modelling solutions stand.

3. Formation of the control solutions is a complex process including such operations, as monitoring and estimation of a situation, modelling, estimation and selection of control operations.

Logical model of control of objects has to:

- Allow effective execution of the basic procedures of control;
- Assist the acceleration of the search-sorting operations;
- Contain the information about actions, variants of actions, sequence of actions, connections of actions with objectives, i.e. as a matter of fact, to show all the totally procedural knowledge about objects of control;
- Represent the states of the object of control that arise on the separate steps of solution of the task of control;
- Allow the description to terminology of the ending user;
- Allow the enter of the changes under the influence of information that comes in from without.

In the time of developing means of automation of complex control procedures, it is desirable to find a unified effective technical principle, which one would provide enough fast response time all components of complex process. In concordance with methodology of an artificial intelligence, such determining principle can be idea of organization representation and looking up of knowledge.

In modern intellectual systems such successful idea are traditionally connected to usage of network ways of a data representation, that allows to avoid having overcome large amounts of information during execution of search operations.

Report deals with logical designing of multicriteria control systems of composite dynamic objects on the basis of balancing networks [4-6].

2. Problem of Control

It is comfortably to consider the processes of choice of handling actions as a task of satisfaction of limitations (CSP - Constraint Satisfaction Problem). In the tasks of this class, states are described by sets of values of variables, and purposes are a set of limitations, with which these values must satisfy.

Let's consider a problem of such type.

There is an object of control

$$<X, P, D, Y, F, L>, \quad (1)$$

where:

X - set of entrance variables - set of numerical, Boolean or linguistic variables the values of which are not determined with the help of some functions or rules through other variables. In set *X* the subset *C of controlling* variables by which change it is possible to influence states of object of control is allocated. The controlling action changing its value is connected to each variable of set *C*;

P - set *of parameters* - numerical, Boolean or linguistic variables which values are considered set and constant during the decision of one task;

D - set *of derivative* variables - numerical, Boolean or the linguistic variables determined through other variables functionally or with the help of rules such as « if - that »;

Y - set *of target* variables, $Y \subseteq D$;

F - set of functions and rules « if - that », of derivative variables determining dependence of derivative variables on other variables;

L - set of restrictions of type $a_y \le y \le b_y$, where a_y, b_y - numerical constants, $y \in Y$.

The task will consist in a finding in space of controlling variables of a point in which all restrictions of set *L* are carried out.

The target variable refers to *normalized* if its value corresponds to the set restrictions. The target variables, which are not corresponding to the restrictions, are considered *not normalized*. Controlling variables, which change causes change of value of a target variable $y \in Y$, refer to *relevant* as this target variable.

If mapping $X \to Y$ it has been set by some system of the equations, for a choice of the decision it would be possible to use methods of mathematical programming. In (1) mapping $X \to Y$ directly it is not set. The decision should be received only on the basis of the semantic information describing the environment and the purposes. The tool of the decision of such problems is balancing networks.

3. Balancing Models

A balancing network [1-3] is the oriented graph $< V, U >$ (illustrated by Fig.1), in which:

V - set of the vertices corresponding to variables of all types in model (1).

U - set of arches. The vertex corresponding to a variable $d \in D$, incorporates coming arches to the vertices representing variables which are used for calculation *d*.

To the left parts rules correspond Boolean variables which accept value 1 at performance of conditions of applicability of rules; the expressions which are included in the right parts of rules, are calculated only at value "1"

of this variable. At use of a balancing network for representation of model (1) vertices, presenting controlling and target variables refer to accordingly controlling and target vertices of the network. Vertices proper to the derivative variables are provided with pointers, indicative functions or rules, which are used for determination of values of these variables.

The proper limitations from the set *L* contact with target vertices.

Alongside with such structures as «trees of decisions», «and-or graphs», «networks of utilities», balancing networks are the effective instrument of decision-making process. In the systems using balancing model, great value has the block providing automatic construction of the balancing network under the information (1), and also operative reorganization of the network at changes of the environment.

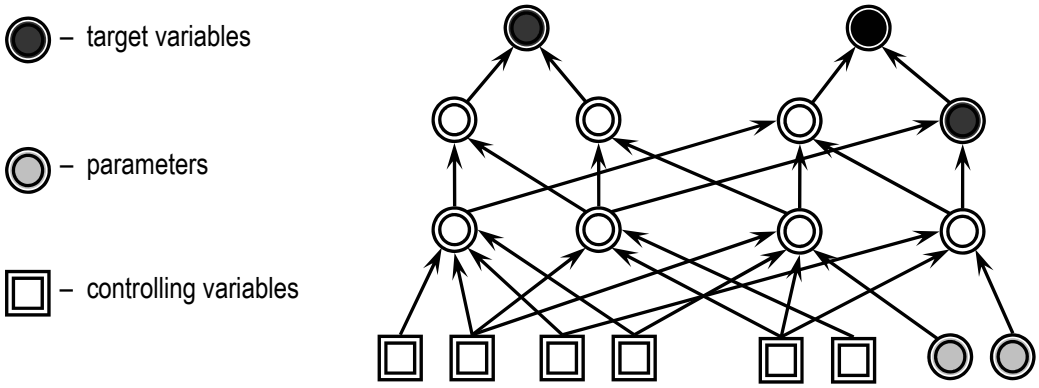


Fig. 1. A balancing network.

The balancing network is development of settlement - logic networks [4].

4. Choice of Controlling Actions

Constant watching of values of entrance variables, and also calculation of values of derivative variables are carried out in the system of control. If the set of values of target variables corresponds to a known class of situations, the system specifies the recommended sequence of actions fixed in its memory. Each carried out action causes change of value of a corresponding controlling variable and, as consequence, change of values of all variables dependent on it. If the arisen extreme situation does not belong to any known class of situations, procedure of a choice of controlling actions as a matter of fact solves a problem of multicriterion decision-making.

For the description of procedure, we shall enter the following designations:

Y' - set of not normalized target variables;

R - set of controlling variables, relevant to not normalized target variables;

h_r^y - a variable designating necessary "direction" of change by controlling variable $r \in R$ for normalization of a target variable y; the variable h_r^y - appropriates value 1 if value y is necessary to increase and value - 1 if value y needs to be reduced;

H_r - the variable designating "direction" of change by controlling variable r in view of all not normalized target variables.

In procedure of a choice of controlling actions, the following operations are used.

Operation 1. Search of relevant controlling variables.

For each not normalized target, variable relevant controlling variables and the necessary direction of their change are defined by viewing against arrows of the ways coming into corresponding target top. As a result of performance of operation 1 set R is formed and values h_r^y are defined for all $y \in Y'$.

Operation 2. Calculation of values of variables H_r.

For everyone $r \in R$ value of a variable H_r is calculated $H_r = \sum_{y \in Y'} h_r^y$.

If $H_r > 0$, it is expedient to increase value of a controlling variable *r*, if $H_r < 0$ - to reduce.

Operation 3. A choice of the "best" controlling variable.

As the "best" the controlling variable having **max**$|H_r|$ can be chosen. This integrated criterion is used for a choice of the controlling action rendering normalizing influence on the greatest number of not normalized target variables. Applying various integrated criteria, it is possible to change strategy of control.

Operation 4. Modelling of application of the controlling action corresponding to the chosen controlling variable.

Controlling action changes value of the chosen controlling variable *r*. Depending on a sign H_r value *r* increases or decreases for some fixed size which is a priori underlined by experts for each controlling variable.

Values of all variables corresponding to vertices, which are connected by coming ways to the changed controlling variable, are recalculated.

Result of application of controlling action is normalization available and, probably, appearance of the new not normalized target variables.

For a new state of the network if in it still there are not normalized variables, all set above operations repeat.

In case of success procedure forms sequence of controlling actions, i.e. some plan of achievement of the normalized situation.

If changes of each of relevant controlling variables causes both positive, and negative changes of not normalized variables, it means, that full normalization of a situation with the help of the procedure is impossible, as, according to the theory of decisions, the space of not dominating alternatives (set of Pareto) is achieved.

Preferences of the person accepting the decision can be taken into account with the help of weights, which are entered for controlling or target variables and are used at a choice of controlling variables.

The described method is realized as program system.

Let's consider his application for the decision of the problems arising at operation of the helicopter.

The helicopter at absence of visibility as shows sad experience, is on occasion deduced on the modes leaving on corners and speeds of flight for established restrictions that result sometimes in destruction of the device and death of people. It is necessary on parameters of spatial position of the helicopter, modes of flight, works of the screw and engines to warn crew about danger and to give recommendations on controlling actions.

The specified task enters a class of problems where the system - on the basis of a balancing network can be used.

For decision of the task model (1) is formed.

1. Set of entrance variables: *X*. $X_\theta, X_\beta, X_\gamma, \ldots$ - controlling variables which, influencing on the screw, change parameters of the helicopter on corners θ, β, γ , speeds *V*.
2. Parameters, which for separate modes of flight because of their slow change can be accepted as constants, among them are weight of the helicopter, height, external temperature, sometimes speed, etc.
3. Derivative variables - speed, corners, angular speeds - *V*, $\theta, \beta, \gamma, \omega_x, \omega_{at}, \omega_z$, … which can be determined through controlling variables and parameters. These variables are connected by the differential equations of the second order (the equation of movement of a firm body under action of external forces).
4. Restrictions on corners are set θ, β, γ , to speeds, etc.

The decision of the task can be divided into two stages.

- Finding of a point in space of derivative variables in which all restrictions on corners and angular speeds are carried out.
- On the basis of limiting restrictions of dynamics of changes on corners and speeds, taking into account functional dependence of corners and speeds on moving controls the area of change of controlling variables is defined.

5. Summary

In the resulted procedure of formation multicriterion controlling decisions:

- search operations as much as possible become simpler due to the advanced associative opportunities of a balancing network;
- selection of controlling actions is carried out by the analysis of a target situation without taking into account any aprioristic information about utilities, priorities or probabilities of the purposes, conditions and actions though such information can be taken into account by giving weights to vertices and connections of a balancing network;
- the formed plan of action is directed on achievement of balance between conflicting criteria of a choice of the decision;
- the uniform organization of memory for a data storage and knowledge as a balancing network plays a role of the link connecting components of complex process in system and, thus, promoting achievement of compactness hardware and the software;
- processes of formation of a balancing network suppose an opportunity of operative reorganization of a network at a choice of decisions in dynamic environments.

On the basis of the described approach it is expedient to create the stands for the modelling and selecting controlling decisions in the projecting organizations.

Bibliography

1. Zakrevskiy A.D., The Parallel Algorithm of the Logical Management, The Institute of the Logical Management, The Institute of the technical cybernetics NAN of Belarus, Minsk, 1999.
2. Alexeyev A.V., Borisov A.N., Sliadz N.N., Phomin S.A., Intellectual Systems of Decision making, “Zikatne”, Riga, 1997.
3. Valkman U.R., Intellectual Technologies of The Searching Projecting, Port-Royal, Kiev, 1998.
4. Gladun V.P., The Partnership with a Computer. Human-machine Purposeful Systems. Kiev: «Port-Royal», 2000.
5. Gladun V., Vaschenko N., Control on the Basis of Network Models // Artificial Intelligence in Real-Time Control: 7th IFAC Symposium. - Arizona, USA. 1998.
6. Gladun V.P., Vashchenko N.D., Goal Formation and Action Planning on the Basis of the Network Model // Procedings of the 1st Intern. IEEE Symp. “Intelligent Systems”. Varna, Bulgaria, Sept.10-12. 2002, v.1, p.150-153.
7. Tyugue E. H., Conceptual programming. - M.: Nauka, 1984.

Authors' Information

Victor Gladun – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Vitaly Velichko – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Gleb Gladun – Kamov Company, the 8th March 8a, 140007, Lubertsy, Moscow Region, Russia; e-mail: glad@aduis.kiev.ua

Anastasiya Godunova – National Aviation University, Prospekt Kosmonavta Komarova 1, 03058, Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Yurii Ivaskiv – National Aviation University, Prospekt Kosmonavta Komarova 1, 03058, Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Elina Postol – National Aviation University, Prospekt Kosmonavta Komarova 1, 03058, Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Grigorii Jakemenko – Kamov Company, the 8th March 8a, 140007, Lubertsy, Moscow Region, Russia; e-mail: glad@aduis.kiev.ua

THE INFORMATION-ANALYTICAL SYSTEM FOR DIAGNOSTICS OF AIRCRAFT NAVIGATION UNITS

Ilya Prokoshev, Vyacheslav Suminov

***Abstract**: The operation of technical processes requires increasingly advanced supervision and fault diagnostics to improve reliability and safety. This paper gives an introduction to the field of fault detection and diagnostics and has short methods classification. Growth of complexity and functional importance of inertial navigation systems leads to high losses at the equipment refusals. The paper is devoted to the INS diagnostics system development, allowing identifying the cause of malfunction. The practical realization of this system concerns a software package, performing a set of multidimensional information analysis. The project consists of three parts: subsystem for analyzing, subsystem for data collection and universal interface for open architecture realization. For a diagnostics improving in small analyzing samples new approaches based on pattern recognition algorithms voting and taking into account correlations between target and input parameters will be applied. The system now is at the development stage.*

***Keywords**: technical diagnostics, fault detection, inertial navigation system, navigation, aircraft units, supervision, monitoring, fault diagnostics, diagnostic reasoning*

Introduction

Improvements in the reliability and safety of technical systems require advanced methods of supervision, including fault detection and fault diagnostics. Many modern systems are very complex and it is difficult to manually adjust control functions and settings when departures arise between a system and a system model. It is also difficult to respond manually to the onset of faults before they develop into huge failures. This group seeks to develop and apply effective methods to cope with these problems.

In the recent years, activities in the navigation field have been boosting. There are and will be more and more areas where navigation becomes an important part of a system solution. So far navigation systems have been of major importance for aircraft, missiles, ships, etc. In aircraft systems, the accurate navigation plays an important role. Typical tools today for navigation are inertial navigation systems (INS), which essentially means that acceleration and angular velocity measurements are integrated to a position.

The INS is based on the principle that a Schuler-tuned platform will remain aligned to the local vertical regardless of the movement of the vessel carrying it. Three mutually perpendicular sensitive gyros are gimbaled to create a stable platform on which is mounted a two-axis accelerometer.

Nowadays INS development is aimed to achieving navigation parameters in an unlimited range of mobile object orientation corners with the subsequent information digital output to the consumer.

Growth of complexity and functional importance of INS leads to high losses at refusals of the equipment. Development of INS diagnostic and prediction information systems is necessary for prevention of occurrence of refusals and the malfunctions leading to high breakdown of aircraft units and increase of expenses for its major overhaul.

In the most cases the records analysis and data processing of the flight information is performed by operator. That is a human based approach; it is aimed to compare the received data with the set ranges of control parameters to make a decision about system technical condition. The solution of diagnostic task is made practically in a visual form and not takes into account interrelation between parameters. Thus, in this case it's necessary to use the complex solution that considers the general structure of the output data.

Problem Statement

As one of the main research object, the integrated inertial navigation system INS-2000 developed by Ramensky Instrument Engineering Plant is considered. The development of The INS-2000 system provides definition and delivery of navigating parameters and is intended for new and modern helicopters and planes. The INS-2000 is

made as a mono-block consisting of gyrostabilized platform on base of dynamically-tuned gyros, service electronics and computer interface units.

The technical acts analysis research of INS-2000 refusals has shown enough plenty of faults of a product at various production stages (adjustment, trial, refining and so forth).

The experience of inertial navigation systems development shows that the intrinsic error of these units defining their functional reliability is the random parametric drift called by dynamically-tuned gyros, interface electronic cards, control cards and couplers. The given task solution is impossible without more profound analysis of occurrence reasons and influence of design and technological parameters on values and stability of random drift.

According to stated the research of the factors' influential in involuntary drift of system and creation of the effective diagnostic technique permitting to estimate current technical condition of INS-2000 is the actual task.

The main work purpose is development of algorithms for the INS diagnostics, permitting to reveal reasons of refusals and faults on the data on the basis of structural adapting and identification of parameters of navigation model.

The offered technology of solution of the task includes the following stages:

- the structural adapting of the INS equations in view of the detected disorder and model defect in parametric type;
- retrospective estimation of the extended state INS error vector, originating because of defects;
- correlation processing of the received estimations of errors;
- solution of the algebraic equations on parameters, approximating correlation function and included in diagnostic model;
- INS state handle in view of the current state of meters, namely - retargeting of parameters of models of errors and restoring of working capacity of INS.

Given technology will allow solving the following problems:

- optimization malfunctions search strategy;
- separate system units technical condition estimation.

According to the purpose of work, it is possible to solve the following research problems:

- the statistical analysis of INS units parameters accuracy not meeting the quality specifications requirements
- refusals database development of INS interconnected units not past a trial stages;
- open architecture development for processing information from various data sources;
- development realizing automated information capturing for its subsequent processing.

The decision-making task in diagnostic problems starts with observation of behaviour recognized as a deviation from that which is expected or desirable and establishes some hypothesis about the cause of the malfunction. In recent years, two methodologies have been widely applied to approximate the nonlinear assignment rule from the set of observations to the hypothesis: Rule-based systems, characterized by linguistic, logical and cognitively oriented schemes, versus the paradigm of artificial neural networks, characterized by the numeric, associative and adaptive nature.

Fault detection is a key technology in automatic supervision of engineering systems, such as production facilities, machines, airplanes, and appliances. There are a great number of fault detection methods available, ranging from more traditional approaches, such as limit checking, to more advanced model-based methods.

Most model-based methods for fault detection and diagnostics rely on the idea of analytical redundancy that is the comparison of the actual behavior of a system to the behavior predicted on the basis of the mathematical system model. Typical model-based fault detection process consists of two steps: residual generation and residual assessing/classification. Residuals that are the difference between the measurements and the model predictions are nominally zero, and become non-zero because of faults. Residual assessing is to make a detection decision for the monitored system through evaluating the residuals obtained. The decision making is actually a process of classifying the residuals into one of two classes: normal and fault. Technically, after

obtaining residuals, the model-based fault detection becomes a pattern classification problem. Hence, different classification methods can be applied.

It is interesting to observe that almost all of the modern fault detection functions for both unmanned and piloted aircraft are designed by using model-based fault detection methods as described above. This probably contributes to the fact that the model-based fault detection methods have several advantages over the model-free methods, for example, the model-based methods have relatively higher performance and computational straightforwardness, have noise depression capability, and can provide more fault information that can facilitate the subsequent fault isolation and corrections.

Diagnostics Methods

Within the automatic control of technical systems, supervisory functions serve to indicate undesired or unpermitted process states, and to take appropriate actions in order to maintain the operation and to avoid damage or accidents. The following functions can be distinguished:

- (a) *monitoring:* measurable variables are checked with regard to tolerances, and alarms are generated for the operator.
- (b) *automatic protection:* in the case of a dangerous process state, the monitoring function automatically initiates an appropriate counteraction.
- (c) *supervision with fault diagnostics:* based on measured variables, features are calculated, symptoms are generated via change detection, a fault diagnostics is performed and decisions for counteractions are made.

The classical methods (a) and (b) are suitable for the overall supervision of the processes. To set the tolerances, compromises have to be made between the detection size of abnormal deviations and unnecessary alarms because of normal fluctuations of the variables. Most frequently, simple limit value checking is applied, which works especially well if the process operates approximately in a steady state. However, the situation becomes more involved if the process operating point changes rapidly. In the case of closed loops, changes in the process are covered by control actions and cannot be detected from the output signals, as long as the manipulated process inputs remain in the normal range. Therefore, feedback systems hinder the early detection of process faults.

The big advantage of the classical limit-value-based supervision methods is their simplicity and reliability. However, they are only able to react after a relatively large change of a feature, i.e. after either a large sudden fault or a long-lasting gradually increasing fault.

In addition, an in-depth fault diagnostics is usually not possible.

Therefore (c) *advanced methods of supervision and fault diagnostics* are needed, which satisfy the following requirements:

- Early detection of small faults with abrupt or incipient time behaviour,
- Diagnostics of faults in the actuator, process components or sensors.
- Detection of faults in closed loops.
- Supervision of processes in transient states.

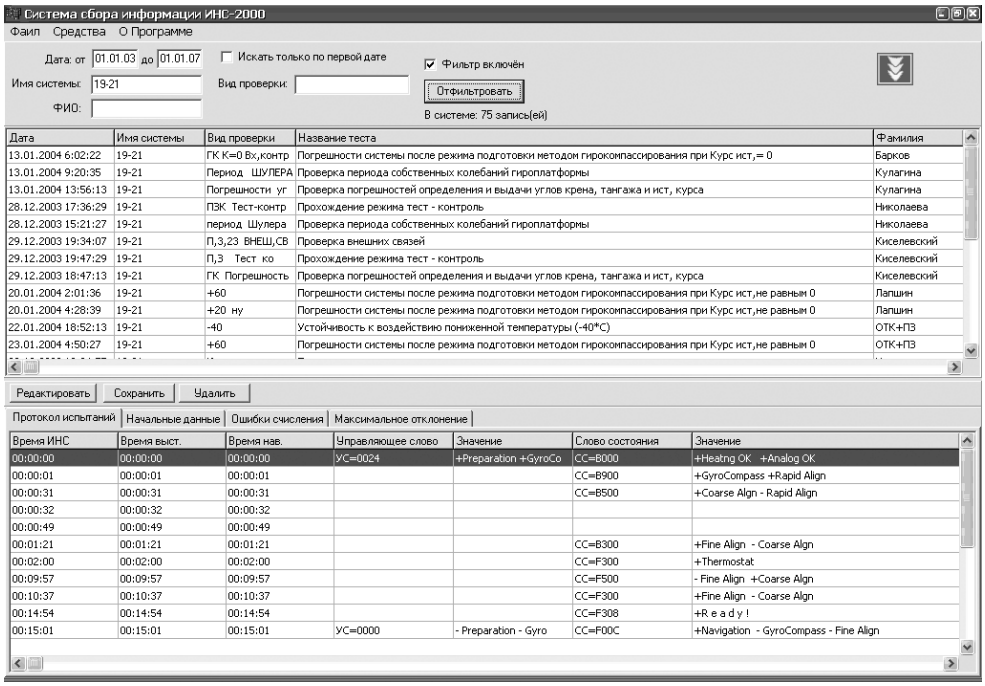
The goal for the early detection and diagnostics is to have enough time for counteractions such as other operations, reconfiguration, maintenance or repair. The earlier detection can be achieved by collection more information, especially by using the relationship between the measurable quantities in the form of mathematical models. For fault diagnostics, the knowledge of cause-effect relations has to be used.

INS Monitoring System Developing

The full analysis of various methods has led to expediency of application of complex monitoring systems which use different by the physical nature research methods that, in turn, will allow excluding lacks of one method and use advantages of other methods to realize thus a principle of "redundancy" increasing reliability of INS systems.

The improvement of quality of diagnostics and prediction in conditions of small analyzed samples new approaches based on voting of algorithms of recognition and the account of correlations between target and output parameters are developing.

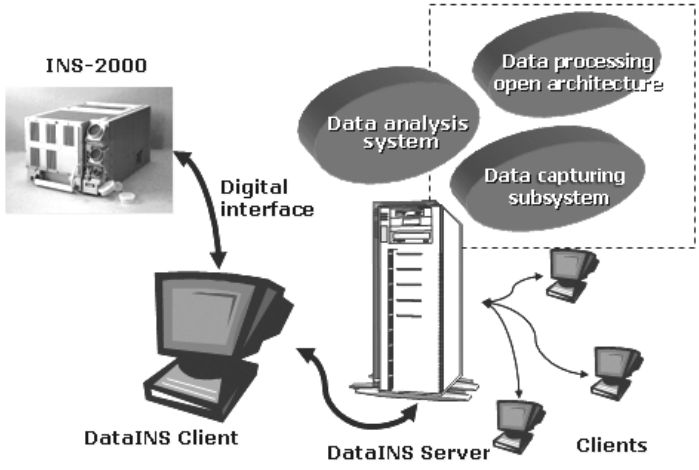
Nowadays, an intellectual system of information capturing and a subsystem of data analysing are developing.



Pic. 1. The screenshot of DataINS system

Open architecture of a system allows using different data source based on Microsoft SQL Server, Microsoft Access, Oracle and others. That approach gives a universal model for data analyzing systems.

The general DataINS structure is shown on the following picture.



Pic. 2. DataINS system structure

Conclusion

An overview of the different approaches to fault diagnostics has been given. So far, none of the methods presented solves the remaining task of completeness. Thus, in practical application, principle of "redundancy" increasing reliability of INS systems is able to solve the defined problem. Complex application of quality monitoring and diagnostics methods for fault detection in units and systems is directed to increase the efficiency, validity check, prolongation of system resources working capacity.

For the first part of work the data capturing system is developed. At stage of development, there is an open architecture data processing system, allowing expanding a set of algorithms without system reconstruction.

Sharing a subsystem of capturing information with a subsystem of data analysis will allow eliminating in time the malfunctions both at a level of test of pre-production models, and in operation and perfection of serial samples.

Bibliography

[Gertler, 1998] Janos Gertler, Fault Detection and Diagnosis in Engineering Systems, Marcel Dekker (May 1, 1998), 484 p.

[Chiang,Russell,Braatz, 1991] Leo H. Chiang, Evan Russell, Richard D. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer-Verlag; 1st edition (February 15, 2001), 296 p.

[Clark,1978] R.N. Clark. Instrument fault detection. *IEEE Trans. Aerospace Electron. Syst.,* Vol. 14, No. 3, pp. 456-465.

[Frank,1990] P.M. Frank. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy. *Automatica,* Vol. 26, pp. 459-474.

[Singh, 1987] M.G. Singh. Fault Detection and Reliability: Knowledge Based and Other Approaches (International Series on Systems and Control, V. 9) by Reliability, and Related knowl European Workshop on Fault Diagnostics, Pergamon Pr; 1st edition (December 1, 1987), 324 p.

[Isermann,1984] R. Isermann. Process fault detection based on modelling and estimation methods - a survey. *Automatica,* Vol. 20, No. 4, pp. 387-404.

Author's Information

Ilya V. Prokoshev – Moscow State Aviation Technological University, 121552, Russia, Moscow, Orshanskaya, 3; post-graduate student, member of teaching staff; +7(926) 5239446; e-mail: mati@e-music.ru

Vyacheslav M. Suminov – Moscow State Aviation Technological University, 121552, Russia, Moscow, Orshanskaya, 3, Doctor of Science, Professor, “Instruments production engineering and aircraft control systems” department header; +7(095)9155441

ДИНАМИЧЕСКИЕ СИСТЕМЫ В ОПИСАНИИ НЕЛИНЕЙНЫХ РЕКУРСИВНЫХ РЕГРЕССИОННЫХ ПРЕОБРАЗОВАТЕЛЕЙ

Микола Ф. Кириченко, **Владимир С. Донченко**, **Денис П. Сербаев**

Abstract*: The task of approximation-forecasting for a function, represented by empirical data was investigated. Certain class of the functions as forecasting tools: so called RFT-transformations, – was proposed. Least Square Method and superposition are the principal composing means for the function generating. Besides, the special classes of beam dynamics with delay were introduced and investigated to get classical results regarding gradients. These results were applied to optimize the RFT-transformation. The effectiveness of the forecast was demonstrated on the empirical data from the Forex market.*

Keywords*: empirical functions, learning samples, beam dynamics with delay, recursive nonlinear regressive transformation, Generalized Inverse, Least Square Method.*

Введение

Аппроксимация функции, представленной своими значениями для того или иного набора аргументов, является классическим для математики направлением исследований, как в детерминированной постановке, так и в статистическом варианте (см., [4,5], а также [1-3]),. В указанных работах достаточно полно представлены классические результаты в этой области. Они свидетельствует о важности суперпозиции - рекуррентности как средства порождения класса аппроксимирующих функций.

Естественным способом использования аппроксимации в прикладных исследованиях является прогноз значений исследуемой функции. Особую важность средства прогнозирования имеют в современных унифицированных системах автоматизации менеджмента фирмы: в так называемых системах Business Intelligence и, в частности, в тех структурных элементах, которые в отечественной литературе обозначаются как СППР.

В настоящей статье предложен метод оптимального конструирования прогнозных средств, построенных на применении классического МНК в комбинации с суперпозицией-рекурсией, линейным преобразованием координат и покомпонентным нелинейным преобразованием в рамках так называемых RFT-преобразователей, вариант построения которых вместе с алгоритмом синтеза основного структурного элемента рассмотрен в работе [6]. Поскольку оптимизация связана с представлением исследуемых объектов системами управления с запаздыванием, в статье введены и исследованы специальные классы систем с запаздыванием, как для одиночных траекторий, так и для их пучка.

Рекурсивный регрессионный нелинейный преобразователь: RFT-преобразователь

Как уже отмечалось, идея рекурсивного регрессионного нелинейного преобразователя (RFT-преобразователя) предложена в работе [6] в варианте, который можно назвать обратной рекурсией. Такие преобразователи: и с обратной рекурсией [6], и прямой, которая будет предложена и рассмотрена ниже, – строятся рекурсивным применением определенного стандартного элемента, который будет обозначаться аббревиатурой ЭРРП (Элементарный Рекурсивный Регрессионный Преобразователь). Процесс построения RFT-преобразователя заключается в присоединении очередного ЭРРП к уже построенному в ходе выполнения предыдущих шагов преобразователю в соответствии с одним из трёх возможных типов соединения. Типы соединения, которые будут обозначаться как parinput, paroutput и seq, – реализуют естественные варианты использования входного сигнала: параллельного или последовательного по входу, – и параллельного по выходу. В варианте обратной рекурсии [6], входом вновь присоединяемого ЭРРП является вход всего RFT, а в варианте прямой рекурсии, предлагаемом ниже, вход всей системы является входом уже построенного RFT.

Основным структурным элементом RFT-преобразователя является ЭРРП – элемент [6], который определяется как отображение из R^{n-1} в R^m вида:

$$y= A_+\Psi_u\left(C\begin{pmatrix}x\\1\end{pmatrix}\right), \quad (1)$$

аппроксимирующее зависимость, представленную учебной выборкой

$$(x_1^{(0)},y_1^0),\ldots,(x_M^{(0)},y_M^{(0)}),\ x_i^{(0)}\in R^{n-1}\ ,\ y_i^{(0)}\in R^m\ ,\ i=\overline{1,M}\ ,$$

где:

- C–(n×n))-матрица, осуществляющая аффинное преобразование вектора $x\in R^{n-1}$, являющегося входом системы; на этапе синтеза ЭРРП считается заданной;
- Ψ_u – нелинейное отображение с R^n в R^n, которое состоит в покомпонентном применении скалярных функций скалярного аргумента $u_i\in\Im, i=\overline{1,n}$ из заданного конечного множества $\Im$ допустимых преобразований, включающего тождественное преобразование: в ходе синтеза ЭРРП выбирается так чтобы минимизировать невязку между входом и выходом на обучающей выборке;
- A_+– минимальное по следовой норме решение A матричного уравнения

$$AX_{\Psi_u C}=Y, \quad (2)$$

в котором матрица $X_{\Psi_u C}$ составлена из векторов -столбцов $\Psi_u(C\begin{pmatrix}x_i^{(0)}\\1\end{pmatrix})=\Psi_u(z_i^{(0)})$, а Y – из столбцов $y_i^{(0)}, i=\overline{1,M}$.

В сущности, ЭРРП представляет собой эмпирическую регрессию для линейной регрессии y на $\Psi_u(C\begin{pmatrix} x \\ 1 \end{pmatrix})$, построенную с помощью метода наименьших квадратов, с предварительным аффинным преобразованием системы координат для векторного регрессора x и последующим нелинейным преобразованием каждой полученной координаты в отдельности.

Замечание 1. В дальнейшем будем предполагать, функции покоординатных преобразований из $\Im$ имеют необходимую степень гладкости там, где это необходимо.

В уже цитированной выше работе [6] поставлена и решена задача синтеза ЭРРП по оптимальному выбору нелинейных преобразований координат на заданной обучающей выборке. Решение задачи синтеза базируется на методах анализа и синтеза псевдообратных матриц, развитых в работах [7-9]. В частности, в этих работах получено обращение формулы Гревиля [10], что позволяет рекуррентно пересчитывать псевдообратные матрицы при замене строки или столбца соответствующей матрицы.

Обобщённые рекурсивные регрессионные преобразователи с прямой рекурсией

Рекурсия в построении RFT-преобразователя в предлагаемом ниже варианте прямой рекурсии будет рассматриваться в обобщённом варианте, в котором в рекуррентном присоединении к уже имеющейся RFT-структуре используется не один ЭРРП, а несколько. Общее количество рекуррентных обращений будем обозначать через N, а количество используемых на шаге m ЭРРП – через k_m, $m = \overline{1,N}$. Общее число ЭРРП, используемых для построения всего преобразователя будет обозначаться через T: $T = \sum_{m=1}^{N} k_m$.

Варианты обобщённой прямой рекурсии в зависимости от типа соединения присоединяемых ЭРРП: parinput, paroutput и seq,– представлены рисунками 1-3. На них $\hat{y}$ обозначает выход уже имеющейся RFT-структуры или выход той же структуры после присоединения очередного ЭРРП из общего числа k_m таких элементов, присоединяемых на шаге рекурсии m, $m = \overline{1,N}$. Каждый рисунок сопровождается системой соотношений, определяющих преобразование сигнала на очередном шаге рекурсии.

- **Тип parinput:**

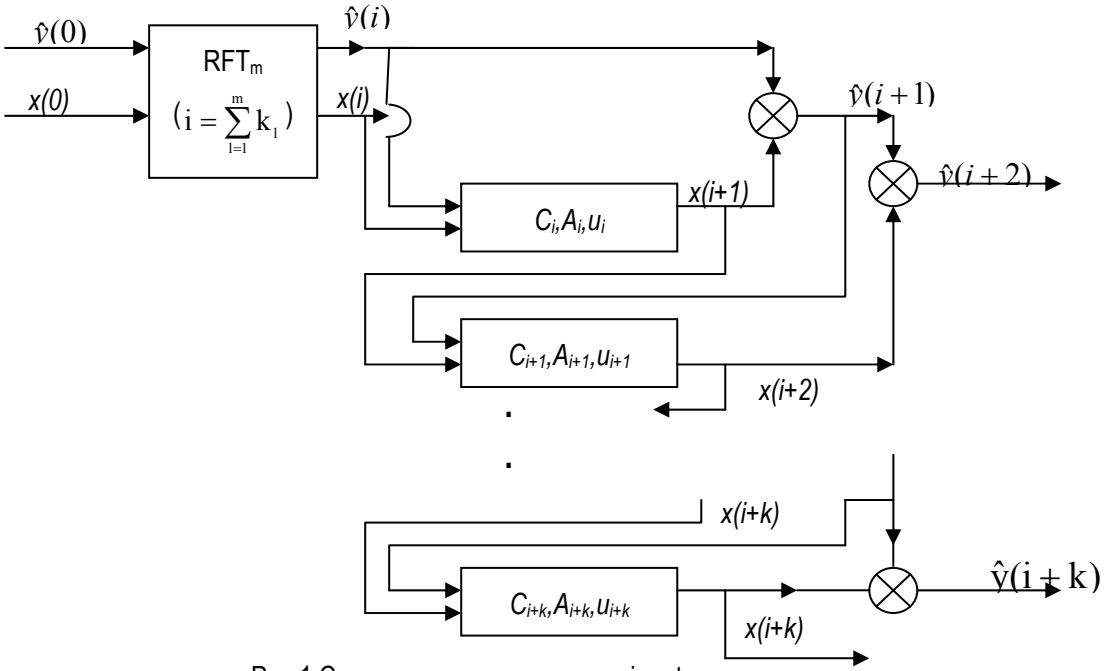


Рис.1.Схема соединения типа parinput –пряма рекурсия.

В соединении типа parinput уже построенная структура аппроксимирует выход обучающей выборки по её входу, а набор ЭРРП – результирующую невязку такой аппроксимации в зависимости от входа обучающей выборки. Преобразование информации при этом типе соединения описывается следующей системой:

$$\begin{aligned} &x(i+j) = A_{i+j-1}\Psi_{u_{i+j-1}}(C_{i+j-1}x(i)), \\ &\hat{y}(i+j) = \hat{y}(i+j-1) + A_{i+j-1}\Psi_{u_{i+j-1}}(C_{i+j-1}\cdot x(i)). \\ &i = \sum_{l=1}^{m} k_l,\ j = \overline{1,k_{m+1}} \end{aligned} \tag{3}$$

- **Тип paroutput:**

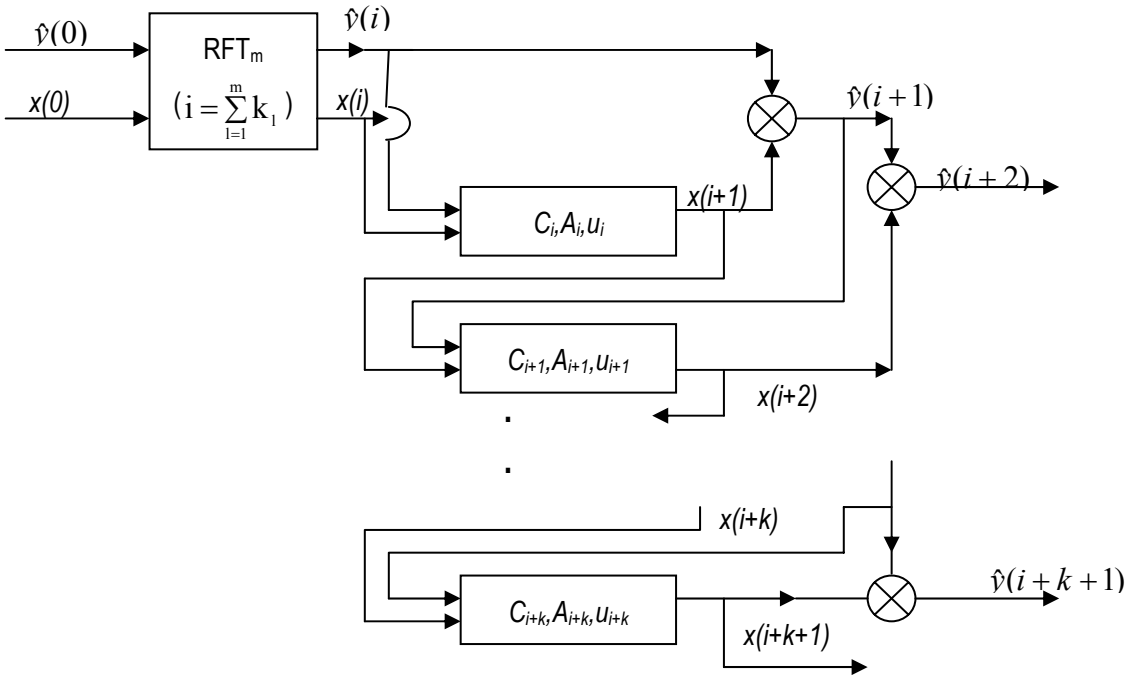


Рис.2. Схема соединения типа paroutput – прямая рекурсия

Система, описывающая преобразование внутри преобразователя и связь между входом и выходом, при этом типе соединения имеет вид:

$$\begin{aligned} &x(i+j) = A_{i+j-1}\Psi_{u_{i+j-1}}(C_{i+j-1}x(i+j-1)), \\ &\hat{y}(i+j) = \hat{y}(i+j-1) + A_{i+j-1}\Psi_{u_{i+j-1}}(C_{i+j-1}\cdot x(i+j-1)) \\ &i = \sum_{l=1}^{m} k_l,\ j = \overline{1,k_{m+1}} \end{aligned} \tag{4}$$

- **Тип seq:**

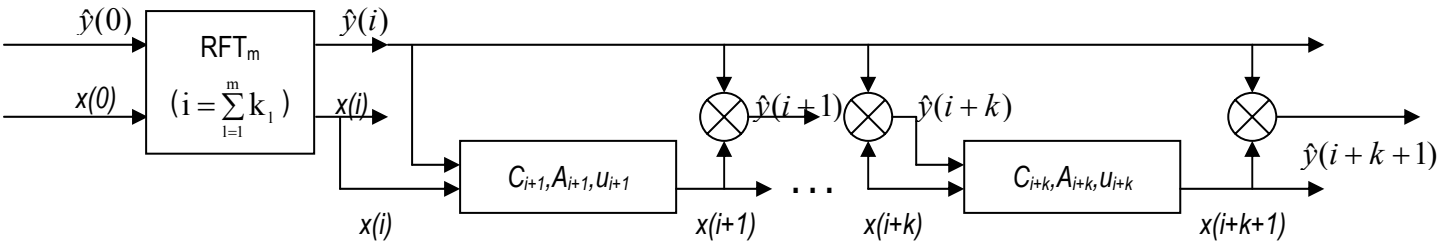


Рис.3. Схема соединения типа seq –прямая рекурсия.

Соотношения, описывающие преобразование информации на очередном шаге рекурсии выглядят следующим образом:

$$\begin{aligned} &x(i+j) = A_{i+j-1}\Psi_{u_{i+j-1}}(C_{i+j-1}x(i+j-1)), \\ &\hat{y}(i+j) = \hat{y}(i) + A_{i+j-1}\Psi_{u_{i+j-1}}(C_{i+j-1}\cdot x(i+j-1)) \\ &i = \sum_{l=1}^{m} k_l,\ j = \overline{1,k_{m+1}} \end{aligned} \tag{5}$$

В этой схеме соединения RFT_{m-1} аппроксимирует выходную часть обучающей выборки по входной, а набор ЭРРП – невязку в зависимости от выхода очередного ЭРРП.

Начальные условия для всех типов соединений описываются соотношениями:

$x(0) = x$ – вход всего RFT-преобразователя, (6)

$\hat{y}(0) = 0,\ \hat{y}(1) = x(1)$ для всех типов соединений. (7)

В соответствии с (6) в режиме настройки (обучения) входами RFT – преобразователя являются $\mathrm{X}_{i1}^{(0)}$: $x_i^{(0)} \in R^{n-1}$, $y_i^{(0)} \in R^m$, $i = \overline{1,M}$, а выходами – $y_i^{(0)} \in R^m, i = \overline{1,M}$.

Соединения рекурсивного построения RFT - преобразователя определены так, что в ходе его построения минимизируется стандартный функционал метода наименьших квадратов, т.е.

$$\sum_{i=1}^{M} || y_i^{(0)} - RFT(x_i^{(0)}) ||^2 . \tag{8}$$

Соотношения (3)-(7) для N шагов рекурсии с общим числом T используемых ЭРРП, $\mathrm{T} = \sum_{\mathrm{m}=1}^{\mathrm{N}} \mathrm{k}_{\mathrm{m}}$, k_m - количество ЭРРП, используемых на шаге с номером $m = \overline{1,N}$, а также функционал качества (8) – представляют собою математическую модель RFT – преобразователя.

Дискретные системы управления с запаздыванием

Соотношения (3)-(8) представляют собой систему рекуррентных соотношений, являющуюся определённым обобщением классической системы управления с дискретным временем (подробнее см.например[12-16]) и прежде всего в том, что касается запаздывания.

Простые и комбинированные системы управления с запаздыванием

Определение. Простой, соответственно – комбинированной, – нелинейной системой управления с запаздыванием на временном интервале $\overline{0,\mathrm{N}}$ называется система управления, траектории которой определяются системой рекуррентных соотношений (9), соответственно – (10), начальными условиями (11) и функционалом качества (12) ниже:

$$x(j+1) = f(x(j - s(j)), u(j), j)\ , \tag{9}$$

$$x(j+1) = f(x(j), x(j - s(j)), u(j), j) \tag{10}$$

$$j = \overline{0,N-1}, \qquad x(0) = x^{(0)}, \tag{11}$$

$$I(U) = \Phi(x(\mathrm{N})), \tag{12}$$

где функция s(j), k=0,...,N–1: s(0)=0, $s(j) \in \{2,...,j\}$, $j = \overline{0,N-1}$ – известна.

Очевидным образом определяются системы с запаздыванием для пучка траекторий. Начальные условия траекторий пучка будем обозначать $(x(0))_i = x_i^{(0)}$, $i = \overline{0,M}$, M– количество траекторий пучка. Траектории пучка для обоих типов систем с запаздыванием будем обозначать соответствующей индексацией: $x_i(j)$, $j = \overline{0,N-1}$, $i = \overline{0,M}$.

Функционал качества для системы управления пучком траекторий, который будем считать зависящим только от финальных состояний траекторий пучка, определим соотношением:

$$I_p(U) = \sum_{i=1}^{M} \Phi^{(i)}(x_i(N)). \tag{13}$$

Замечание 2. Очевидным образом функционал качества, как и для классических систем управления, может определяться на всей траектории или траекториях. Однако в связи с RFT - преобразователями будут рассматриваться системы с запаздыванием, у которых функционал качества зависит только от финальных состояний траектории.

Фазовые траектории простых систем с запаздыванием определяются набором функций f(z, u, j), $j = \overline{0, N-1}$ с одним аргументом z, отвечающим за фазовую переменную, а комбинированных – набором f(z,v,u,j), $j = \overline{0, N-1}$ с двумя переменными: z, v, отвечающими фазовым переменным. В этом последнем случае градиенты по фазовым переменным будут обозначаться соответственно $grad_z f$, $grad_v f$.

Задача оптимизация для обоих типов таких систем с запаздыванием решается, как и в классическом случае, построением сопряжённых систем и функций Гамильтона. Предположения гладкости, обеспечивающие корректное построение сопряжённых систем и функций Гамильтона, а также их использования для нахождения градиентов по управлениям полностью совпадают с классическими и в дальнейшем будут считаться автоматически выполненными.

Оптимизация в простых системах управления с запаздыванием

Определение. Сопряжённой системой простой системы управления с запаздыванием (9), (11), (12) будем называть следующее рекуррентное соотношение относительно p(k), $k = \overline{N,0}$:

$$p(k) = \sum_{j \in \{j: j - s(j) = k, \geq kj\}} grad_{x(k)} \{ p^T(j+1) f(x(j), u(j), j) \} \qquad k = \overline{N-1,0} \tag{14}$$

с начальным условием

$$p(N) = -grad_{x(N)} \Phi(x(N)) \tag{15}$$

Соответственно, в случае пучка траекторий сопряжённые системы определяются для каждой траектории соотношениями:

$$p^{(i)}(N) = -grad_{x_i(N)} \Phi^{(i)}(x_i(N)), \tag{16}$$

$$p^{(i)}(k) = \sum_{j \in \{j: j - s(j) = k, \geq kj\}} grad_{x_i(k)} \{ p^{(i)T}(j+1) f(x_i(j), u(j), j) \} \qquad k = \overline{N-1,0}, i = \overline{1, M}. \tag{17}$$

Функция Гамильтона для простой системы с запаздыванием определяется классическим соотношением:

$$H(p(k+1), x(k - s(k)), u(k), k) = p^T(k+1) f(x(k - s(k)), u(k), k), \; k = \overline{N-1,0}. \tag{18}$$

Для пучка траекторий простой системы управления с запаздыванием определяется набор функций Гамильтона $H^{(i)}$, $i = \overline{1, M}$ для каждой из траекторий $x^{(i)}(k)$, $k = \overline{N-1,0}$.

Теорема 1. Градиент по управлению от функционала качества, зависящего только от финальных состояний траекторий пучка, для простой системы управления с запаздыванием определяется соотношениям:

$$grad_{u(k)} I(U) = -\sum_{i=1}^{M} grad_{u(k)} \left\{ p^{(i)T}(k+1) f(x^{(i)}(k - s(k)), u(k), k) \right\} = -grad_{u(k)} \sum_{i=1}^{M} H_v^{(i)}(x_i(k), u(k), k), \qquad k = \overline{0, N-1} \tag{19}$$

Доказательство. Доказательство проводится точно так же, как и в классическом случае: сначала для одной траектории, а потом использованием аддитивности функционала качества по траекториям системы.

Оптимизация в комбинированных системах управления с запаздыванием

Определение. Сопряжённой системой для комбинированной системы управления с запаздыванием называется система, определяемая рекуррентным соотношений:

$$p(k) = grad_z\{ p^T(j+1) f(x(k), x(k - s(k)), u(j), j) \quad + \\ + \sum_{j \in \{j: j - s(j) = k, \geq kj\}} grad_v\{ p^T(j+1) f(x(k), x(j), u(j), j)\}, \\ k = \overline{N-1, 0}, i = \overline{1, M} \tag{20}$$

с начальным условием

$$p(N) = -grad_{x(N)} \Phi(x(N)). \tag{21}$$

Соответственно, функция Гамильтона H(p(k+1),x(k),x(k–s(k),u(k),k)) комбинированной системы определяется соотношением:

$$H(p(k+1),x(k),x(k–s(k),u(k),k)) = p^T(k+1) f(x(k), x(k - s(k)), u(k), k). \tag{22}$$

Как и раньше, верхний индекс (i), $i = \overline{1, M}$: $p^{(i)}(k)$, $H^{(i)}$, $k = \overline{N-1, 0}$, будет определять объекты для траекторий пучка.

Теорема 2. Градиенты по соответствующим управлениям от функционала качества, зависящего только от финальных значений траекторий комбинированной системы управления с запаздыванием, определяются градиентами от соответствующих функций Гамильтона:

$$grad_{u(k)} I_p(U) = - \text{grad}_{u(k)} \text{H}(\text{p}(\text{k}+1), \text{x}(\text{k}), \text{x}(\text{k} - \text{s}(\text{k})), \text{u}(\text{k})), \text{k})). \tag{23}$$

И, следовательно, для пучка траекторий соответствующий градиент определяется соотношением:

$$grad_{u(k)} I_p(U) = - \sum_{\text{i}=1}^{\text{M}} \text{grad}_{u(k)} \text{H}^{(\text{i})}(\text{p}^{(\text{i})}(\text{k}+1), \text{x}^{(\text{i})}(\text{k}), \text{x}^{(\text{i})}(\text{k} - \text{s}(\text{k}), \text{u}(\text{k}), \text{k})). \tag{24}$$

Доказательство. Доказательство проводится стандартным для использования сопряжённых систем и функций Гамильтона способом

RFT-преобразователи и системы управления с запаздываниями

Как уже отмечалось RFT – преобразователь может быть представлен системой управления с запаздыванием. Точнее, справедлива следующая теорема.

Теорема 3. Регрессионный RFT_N- преобразователь с прямым N-кратным рекурсивным обращением к использованию соответственно $k_m, m = \overline{1, N}$ ЭРРП - элементов на каждом из шагов рекурсии представляется нелинейной комбинированной системой управления пучком траекторий с запаздыванием на интервале $\overline{0, T}$, $T = \sum_{l=1}^{N} k_l$:

- с фазовой переменной

$$z(t) = \begin{pmatrix} z_1(t) \\ z_2(t) \end{pmatrix} \text{с } z_i(t) \in R^m, i = 1,2, t = \overline{0, T}; \tag{25}$$

- системой рекуррентных соотношений, определяющих траектории пучка:

$$z^{(i)}(t+1) = f(z^{(i)}(t), z^{(i)}(t-k(t)), C_t, t) = = \begin{pmatrix} f_1(z^{(i)}(t), z^{(i)}(t-k(t)), C_t, t) \\ f_2(z^{(i)}(t), z^{(i)}(t-k(t)), C_t, t) \end{pmatrix}, \quad (26)$$

с $f_1, f_2 \in R^m$, $t = \overline{1, T-1}$, $i = \overline{1, M}$, зависящими от топологии RFT- преобразователя и определяемыми соотношениями (29)-(31);

- начальными условиями:

$$z^{(i)}(0) = \begin{pmatrix} x_i^{(0)} \\ 0 \end{pmatrix}, i = \overline{1, M}, \quad (27)$$

где $z^{(i)}(0), i = \overline{1, M}$ начальные состояния траекторий пучка, а $x_i^{(0)}, i = \overline{1, M}$ – элементы входа обучающей выборки;

- функционалом качества I(C), $C = (C_1, \ldots, C_T)$, зависящим от матриц $C_1, \ldots, C_T$ как от управлений:

$$I(C) = \sum_{k=1}^{M} || y_k^{(0)} - z_2(T) ||^2 \quad (28)$$

где $y_k^{(0)}, k = \overline{1, M}$ компоненты выхода обучающей выборки.

Доказательство можно найти в [8].

Доказанное утверждение даёт возможность использовать методы оптимизации теории управления для оптимизации уже построенного RFT_N-преобразователя по величине невязки в зависимости от матриц $C_0, C_1, ..., C_{T-1}$. Это утверждение и является предметом следующей теоремы.

Теорема 4. При наличии непрерывных производных до второго порядка включительно у функций семейства $\Im$ RFT - преобразователь может быть оптимизирован градиентными методами с градиентами функционала качества по матрицам С, как по параметрам.

Доказательство. В соответствии с теоремой 3 RFT-преобразователя может быть представлен комбинированной системой управления с запаздыванием, а в соответствии с теоремой 2 градиенты по соответствующим управлениям - параметрам RFT преобразователя определяются соотношением (24).

Важность приведенной теоремы заключается в том, что она даёт исчерпывающую интерпретацию алгоритма Back Propagation , очерчивая в то же время границы указанного метода

RFT - преобразователи в различных вариантах: линейном, нелинейном, а также с градиентной оптимизацией – использовались для прогноза значений по дням на таком быстро меняющемся рынке, характеризующемся модельной неопределённостью, как рынок валюты. Такое использование требует решения ряда вопросов, связанных с формированием обучающей выборки, выбором интервала прогнозирования и некоторых других, которые остаются за рамками обсуждения. Отметим только, что прогнозировался стандартный набор характеристик: курс открытия, максимальный и минимальный за день курсы. Следует отметить, что предсказание хорошо улавливает развороты курса и вместе с другими техническими показателями прогнозы на основе подходящего RFT - преобразователя могут быть эффективно использованы на соответствующем рынке.

Литература

1. Ивахненко А.Г. Самоорганизующиеся системы распознавания и автоматического управления. – К.: Техника. – 1969. – 395 с.
2. Колмогоров А.Н. О представлении непрерывных функций нескольких переменных суперпозициями непрерывных функций меньшего числа переменных.// Докл. АН СССР. – 1956. – Т. 108. – № 2. – С. 179- 182.
3. Арнольд В.И. О функциях трех переменных.// Докл. АН СССР. – 1957. – Т. 114. – № 4. – С. 953- 956.
4. Кириченко Н. Ф., Крак Ю. В., Полищук А.А. Псевдообратные и проекционные матрицы в задачах синтеза функциональных преобразователей.// Кибернетика и системный анализ –2004.–№3.

5. Кириченко Н.Ф., Лепеха Н.П. Применение псевдообратных и проекционных матриц к исследованию задач управления, наблюдения и идентификации.// Кибернетика и системный анализ. – 2002. – № 4. – С. 107-124.
6. Бублик Б.Н., Кириченко Н.Ф. Основы теории управления. – К.: Высшая школа, 1975.–328 с.
7. Бублик Б.Н., Гаращенко Ф.Г., Кириченко Н.Ф. Структурно-параметрическая оптимизация и устойчивость динамических пучков .– К.: Научная мысль, 1985.–304 с.
8. Нелінійні регресійні рекурсивні перетворювачі: динамічні системи та оптимізація / Кириченко Н.Ф., Донченко В.С., Сербаєв Д.П. .// Кибернетика и системный анализ.–2005.–№ .– 1.

Информация об авторах

Микола Ф. Кириченко – профессор, Институт кибернетики НАН Украины, e_mail: kir@dept115.icyb.kiev.ua

Владимир С. Донченко – профессор, Киевский национальный университет имени Тараса Шевченко, кафедра Системного анализа и теории оптимальных решений, e_mail: vsdon@unicyb.kiev.ua

Денис П. Сербаев – аспирант, Киевский национальный университет имени Тараса Шевченко, кафедра Системного анализа и теории оптимальных решений, e_mail: sdp@unicyb.kiev.ua

THE MATRIX METHOD OF DETERMINING THE FAULT TOLERANCE DEGREE OF A COMPUTER NETWORK TOPOLOGY

Sergey Krivoi, Miroslaw Hajder, Pawel Dymora, Miroslaw Mazurek

Abstract: *This work presents a theoretical-graph method of determining the fault tolerance degree of the computer network interconnections and nodes. Experimental results received from simulations of this method over a distributed computing network environment are also presented.*

Keywords: *computer network, fault tolerance, coherent graph, regular graph, network topology, adjacency matrix.*

Introduction

Computer networks plays an extremely important role in today's informatic technologies, because by its means it's possible to accelerate processes like i.e. transmission, processing and storage of information in computer systems. In such a process the most crucial issues are related with protecting a correct work of a computer network and its interconnection and node fault tolerance. The solution of these problems is related with examining the network topological characteristics and its topological structures.

In this work the theoretical-graph method of determining the computer network topology critical points which refers to computer network interconnections and nodes failures is proposed.

1 Preliminary Information and Definitions

Computer network (CN) consisted of $n > 1$ computers connected between themselves is presented as a graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ represents a real number of nodes and $E = \{(u, v): u, v \in V\}$ – a real number of interconnections. Under these assumptions the network nodes are represented by graph G nodes, and network interconnections corresponds with the given graph connection links. If not assumed differently a graph definition is always meant as an undirected graph. If $e = (u, v) \in E$, than node u (node v) is called an end of a link e, and such nodes – adjacent. If node u turns out to be the end of link e, than link e and node u, is called incident. Note, that the adjacency relation for a graph node turns out to be symmetric. If $u \in V$, than $n(u)$ a number of graph

links incidental to node u is called a degree of a node u. A path from a node u to a node v of the graph $G = (V, E)$ with a length k is called a link sequence $(u_1, u_2), ..., (u_k, u_{k+1}) \in E$ such as $u = u_1$, $v = u_{k+1}$.

Definition 1. Graph $G = (V, E)$ is called coherent if from any node u there exists a path to its any other node v (symmetric-relation – inverted path: from a node v to a node u).

An operation of link and node removal is considered on examples of a Cartesian product and an isomorphic joint of two graphs.

Graph $G' = (V', E') = G - v$ is called a graph received from a graph $G = (V, E)$, as a result of applying a node $v \in V$ removal operation if $V' = V \setminus \{v\}$, and E' squares with a number E, from which all links incident to a node v were removed.

Graph $G' = (V', E') = G - e$ is named a graph received from a graph $G = (V, E)$ as a result of applying a link $e \in E$ removal operation, if $V' = V$, and $E' = E \setminus \{e\}$.

Notice, that for both operations the following equations are true:

$$(G - u) - v = (G - v) - u; \quad (G - e) - e' = (G - e') - e,$$

i.e. the result graphs doesn't depend on the sequence of links or nodes removal order .

The nodes number M (or links) of a graph $G = (V, E)$ is called crucial, if as a result of its elements removal from a given graph this graph becomes incoherent.

Let graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be given. Graph $G = (V, E) = G_1 \times G_2$ is called a graph received as a result of applying a Cartesian product operation on graphs G_1 and G_2, if $V = V_1 \times V_2$ and $E = \{[(u, u'), (v, v')]: u = v$ in a graph G_1 and $(u', v') \in E_2$ or $u' = v'$ in a graph G_2 and $(u, v) \in E_1\}$. As an example, if this operation is applied on graphs presented in the following figure 1 the result graph is called a toroid.

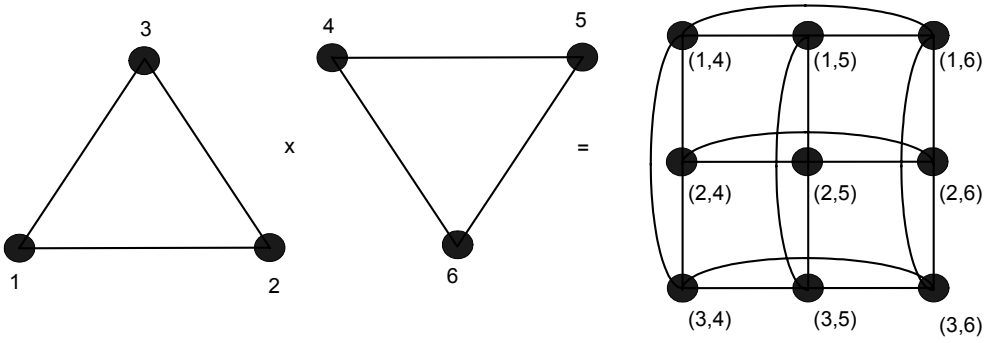


Fig. 1. Example of a Cartesian product $G_1 \times G_2$ operation

Definition 2. Graphs $G_1 = (V_1 E_1)$ and $G_2 = (V_2 E_2)$ are called isomorphic, if between the sets of nodes V_1 and V_2 exists a bijection mapping $f: V_1 \rightarrow V_2$, such that nodes u and v are adjacent to the graph G_1, than nodes $f(v)$ and $f(v)$ are adjacent in the graph G_2. Mapping f is called isomorphic.

Let isomorphic graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ and f – isomorphic mapping are given. Graph $G = (V, E) = G_1 *_f G_2$ is called a graph received as a result of applying an isomorphic joint operation on graphs G_1 and G_2, if $V = V_1 \cup V_2$ and also $E = E_1 \cup E_2 \cup \{(u, u'): u \in V_1, u' \in V_2$ and $f(u) = u'\}$.

As an example, if this operation is applied on graphs presented in the figure 2 with isomorphic $f(i) = i + 8$ the result graph is called a hypercube.

The next theorem follows directly from definitions.

Theorem 1. If graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ – are coherent, than a graph received as a result of applying a Cartesian product operation or an isomorphic joint operation is also coherent.

It's obvious that mentioned links and nodes removal operations in coherent graphs may create incoherent graphs. With each graph $G = (V, E)$ a matrix $A_G = ||a_{ij}||$ is related, where $i, j = 1, 2, ..., n$, and is called an adjacency matrix and is described as follows:

$$a_{ij} = \begin{cases} 1, \; if \;\; (v_i, v_j) \in E, \\ 0, otherwise. \end{cases}$$

Therefore, graph *G* with *n* nodes corresponds to a square matrix *n x n,* filled up with 0 and 1 values. Introduced links and nodes removal operations in graphs may be carried out in an easy manner on these graphs adjacency matrixes.

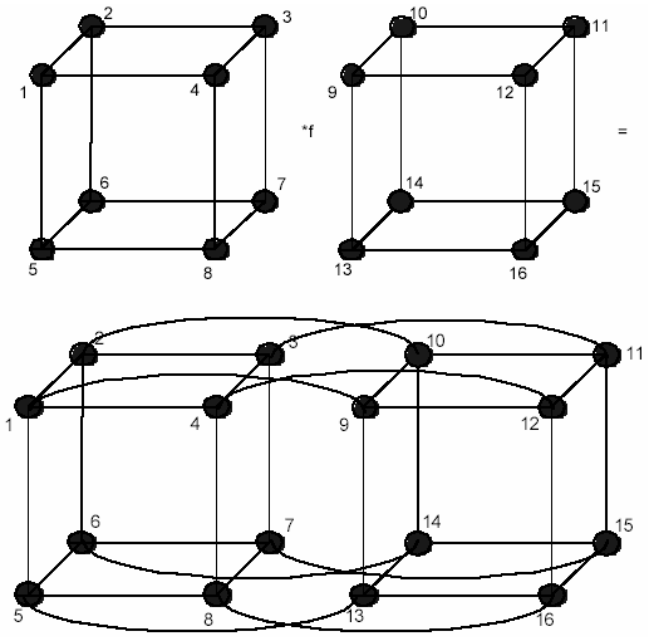


Fig. 2. Example of an isomorphic $G_1 f G_2$ operation

2 The Matrix Method of Determining the Fault Tolerance Degree of a CN

Let a CN consists of *n* computers and is represented as a graph *G = (V, E)*. With a help of a graph *G* analysis for a given CN the most critical points are specified. That's why formal statements are introduced.

Let *G = (V, E)* – is the given coherent graph representing CN, while A_G – is the adjacency matrix of this graph. The nodes number $V' \subseteq V$ (links $E' \subseteq E$) of a graph *G = (V, E)*, is called a computer critical point (failure place) if *V' (E')* the minimal nodes set (links) of a graph *G* becomes crucial, than any element removal from the graph *G* causes that this graph becomes incoherent.

Fulfilling these definitions and an adjacency matrix of a graph *G*, representing a CN, it's possible to create a method of determining the critical points in a CN. In this end the matrix interpretation of introduced before graph operation is considered.

2.1 The Matrix Interpretation of the Graph Operations

As follows from determining critical points this method turns out to be useful in finding the minimal existing subsets. In order to do this a graph adjacency matrix is used because for an inherent graph its adjacency matrix becomes a diagonally-block. With a reorder of an appropriate rows and columns it becomes as follows:

$$A_G = \begin{bmatrix} A \mid 0 \\ 0 \mid B \end{bmatrix},$$

where *A* and *B* are nonzero matrixes, and 0 – represents zero matrixes. On the basis of this simple adjacency matrix property, the matrix method of determining the critical points of a CN is based.

Interpretation of nodes and links removal operation in an adjacency matrix turns out to be very simple. In fact, the removal of links (v_i, v_j) in a graph *G* reduces to changing of an element a_{ij} values from 1 to 0 in a matrix $A_{G,}$ and a removal of a node v_i in a graph *G* corresponds with a removal of an *i*-ary row of this matrix.

For example, for a graph presented in the figure 3a its adjacency matrix has such a structure (rows and columns of the matrix corresponds to enumeration 1,2,3,4,5,6 from the left to the right and from up to down):

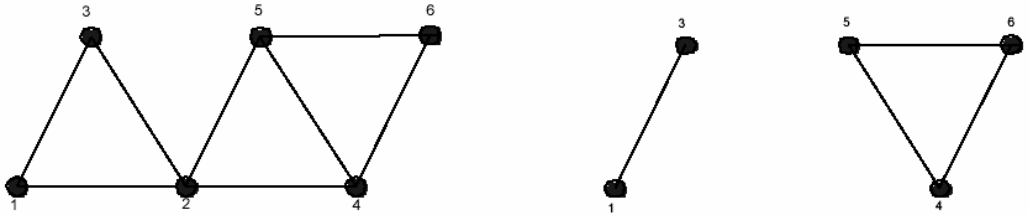


Fig. 3. a) Example of the graph with 6 nodes; b) Graph with removed links and nodes.

$$A_G = \begin{vmatrix} 011000 \\ 101110 \\ 110000 \\ 010011 \\ 010101 \\ 000110 \end{vmatrix}.$$

The node 2 removal from this graph reduces to a matrix $A_G = \begin{vmatrix} 01000 \\ 10000 \\ 00011 \\ 00101 \\ 00110 \end{vmatrix}$, for which graph is shown in the Fig. 3b.

The Cartesian product operation on graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ on the basis of an adjacency matrix is possible to present on an example. From this operation definition it results that an adjacency matrix of a graph $G = G_1 \times G_2$ has a following form:

$$A_G = \begin{vmatrix} A_{G_1} \mid E_a \mid E_a \\ E_a \mid A_{G_1} \mid E_a \\ E_a \mid E_a \mid A_{G_1} \end{vmatrix},$$

where a size of an adjacency matrix A_{G1} is equal to $|V_2|$, while E_a - a diagonal matrix of the size *a* which on its diagonal accepts values 0 or 1 depending on whether the other pair nodes in a graph G_2 are adjacent to it or not. For example, an adjacency matrix of a graph $G = G_1 \times G_{2,}$ where graphs G_1 and G_2 are presented in the figure 1, has a following form (matrix rows and columns correspond to an enumeration (1,4), (2,4), (3,4), (1,5), (2,5), (3,5), (1,6), (2,6), (3,6)):

$$A_G = \begin{vmatrix} 011 \mid 100 \mid 100 \\ 101 \mid 010 \mid 010 \\ 110 \mid 001 \mid 001 \\ 100 \mid 011 \mid 100 \\ 010 \mid 101 \mid 010 \\ 001 \mid 110 \mid 001 \\ 100 \mid 100 \mid 011 \\ 010 \mid 010 \mid 101 \\ 001 \mid 001 \mid 110 \end{vmatrix}.$$

The isomorphic joint $G = G_1 *_f G_2$, of graphs G_1 and G_2 has an adjacency matrix as it follows from the applied operation:

$$A_G = \begin{vmatrix} A_{G_1} \mid C \\ C \mid A_{G_2} \end{vmatrix},$$

where in a diagonal the adjacency matrixes of graphs G_1 and G_2 are placed respectively, and an element a_{ij} of a matrix C is equal 1, if $f(v_i) = v_j$, or equal 0 otherwise. For example, if this operation is applied on a graph from the figure 2, then an adjacency matrix of this graph has a form, shown in figure 4.

$$A_G = \begin{vmatrix} 01011000 \mid 10000000 \\ 10100100 \mid 01000000 \\ 01010010 \mid 00100000 \\ 10100001 \mid 00010000 \\ 10000101 \mid 00001000 \\ 01001010 \mid 00000100 \\ 00100101 \mid 00000010 \\ 00011010 \mid 00000001 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ 10000000 \mid 01011000 \\ 01000000 \mid 10100100 \\ 00100000 \mid 01010010 \\ 00010000 \mid 10100001 \\ 00001000 \mid 10000101 \\ 00000100 \mid 01001010 \\ 00000010 \mid 00100101 \\ 00000001 \mid 00011010 \end{vmatrix}$$

Figure 4.

2.2 The Matrix Method of the CN Topology Analysis

Let $G = (V, E)$ – represents a coherent graph describing a CN and A_G – the adjacency matrix of this graph. As follows, with a simple analysis of a graph adjacency matrix it's possible to determine critical points of a CN, which is represented by this graph.

Generally, in a simple way it's possible to determine nodes, which has the lowest degree because the numbers of 1 in a row (or column) is equal to a node degree, which corresponds with this row (or column). After the nodes with the minimal degree are chosen then the starting number of links which might be suspected of critical links point are received. However, it's not guaranteed that they will appear because the minimal degree doesn't causes that the number of incident to them links is the minimal existing number. Suspected of number of nodes (links) is removed from the adjacency matrixes and after this, a received matrix has a block form. If a matrix has such a form then removed number of nodes (links) turns out to be crucial. Such an analysis requires a consideration of 2^n variants.

The general algorithm, which results from the above-presented analysis, might be presented in a following way (A – an adjacency matrix of a graph G, and n – a sequence presenting a CN):

Critical points (A)

1. **Determine critical points *M* of a graph *G*, consisted of one element of a matrix *A*.**
2. **for *i* = *2* to 2^n fulfill**

   Determine critical points *M'* of a graph *G*, consisted of *i* elements of a matrix *A*

   If $/M'/ < /M/$ than *M:=M'*;

End of cycle.

3. **Return *(M)*.**

The first operator is processed directly from the same matrix *A* (the before presented manner), while the second operator << determines critical points *M'* of a graph *G*, consisted of *i* elements of a matrix *A,* >> presents by itself computations of a transition closure of an achievement relation in a graph. It's common knowledge that a given algorithm is based on an estimation of *n-i-x* (or n) matrix degree received from the matrix *A* as a result of its node (or link) removal.

Obviously, a cycle described in a second paragraph can be parallelized because computations of different iterations are independent from each other. In the following paragraph an experimental results of computations on a cluster with 10 processors are presented.

3 Experimental Results

In order to parallel ours computations the Parallel Virtual Machine (PVM) was used. Next, the PVM was implemented on a cluster with 10 processors. The PVM is a main parallel library, which may process on heterogeneous computer networks. Created in PVM the Message Passing Interface (MPI) which is a library of procedures and functions became a modern standard of building parallel applications. The MPI is independent of operating systems' platform.

Nodes number	Time [s]
5	0,0028
10	0,0047
50	0,0228
100	0,0431
250	0,34
500	0,6402
750	0,7913
1000	1,4252
1500	2,1958
2500	3,47
5000	7,3218
7500	8,9551
10000	22,8534

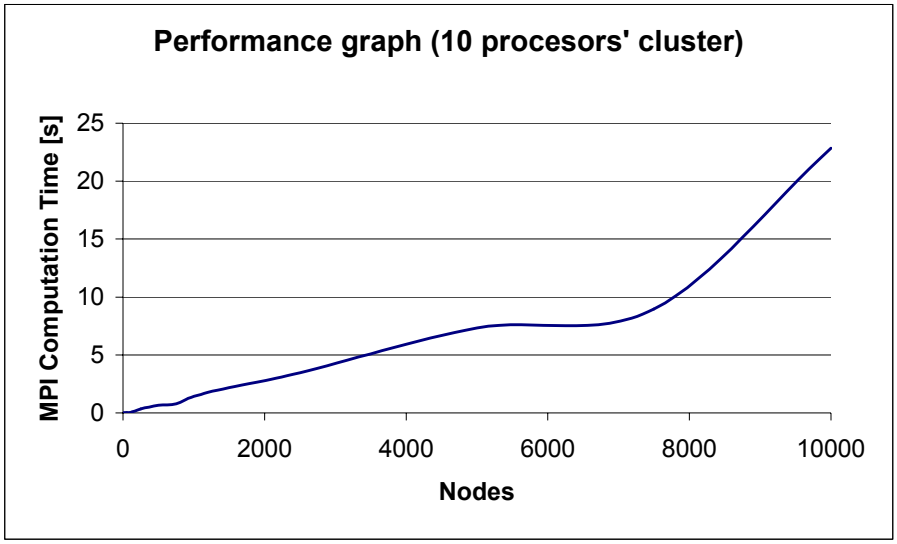


Fig. 5. Experimental results

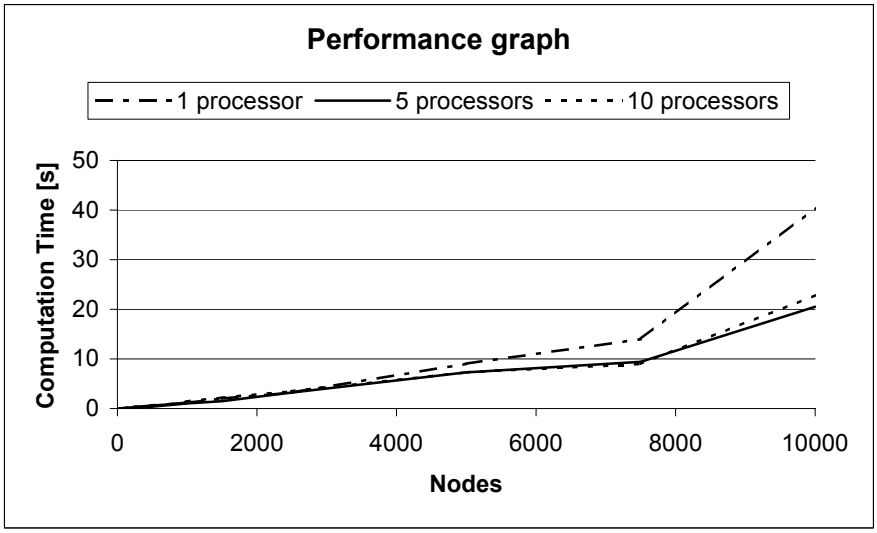


Fig. 6. Experimental results

Conclusions

The presented method of determining the critical points is based on the graph theory and some features of the adjacency matrix, which represents graphs. Searching for critical points in computer networks, as it follows from the above researches is characterized with a large complexity and requires applying of a great computational performance. However, the main advantage of this method is a fact that is uses homogenous structures, and the computations itself are of the same type. Presented experiments were implemented and realized on the multiprocessor cluster and the results of these experiments are presented in the above table. Analyzing these experimental results, the conclusion may be drawn independently.

Bibliography

[Капитонова Ю.В., Кривой С.Л., 1] Капитонова Ю.В., Кривой С.Л., Летичевский А.А., Луцкий Г.М. – *Лекции по дискретной математике*. – Санкт-Петербург: БХВ-Петербург. –2004. 614 стр.

[Свами М., Тхуласираман К., 2] Свами М., Тхуласираман К. Графы, сети и алгоритмы. Москва: Мир. – 1984. 454 стр.

[Hajder M., Loutski G., 3] Hajder M., Loutskij G., Streciwilk W. - *Informatyka*. Rzeszow: WsiZ. - 2002. – 714 s. (на польском языке)

[Hajder M., Dymora P., 4] Hajder M., Dymora P. - *Algorithmical and topological methods of fault tolerace assurance*; Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej; Annales Informatica, Kazimierz Dolny 2004 – 143-153,

[Hajder M., Dymora P., Bolanowski M., 5] M. Hajder, P. Dymora, M. Bolanowski: *Fault tolerance in multichannel scalability systems;* International Conference of Science and Technology Scalnet'04, Scalable Systems and Computer Networks Design and Aplications, Ministry of Education and Science of Ukraine Kremenchuk State Polytechnical University, Kremenchuk 2004, – pp. 52-57,

[Spohn D.,L., 6] Spohn D. L. - *Data Network Design. New York: McGrew-Hill*. – 1997. – 347 p.

[Harsall F., 7] Harsall F. - *Data Communications, Computer Design Networks and Open Systems*. 4-th Eddition. – Addison-Wesley Company. – Harlow. – 1996. – 412 p.

[Mazurek M., 8] Mazurek M. - *Hyper-deBruijn topology as a result of the Network Product*, International Conference of Science and Technology Scalnet'04, Scalable Systems and Computer Networks Design and Aplications, Ministry of Education and Science of Ukraine Kremenchuk State Polytechnical University, Kremenchuk 2004, pp. 62-65.

[Hajder M., Mazurek M., 9] Hajder M., Mazurek M. - *Embending as a method to improve multichanel bus topology parameters*, Annales Universitatis Mariae Curie-Skłodowska, Sectio Ai Informatica, Lublin, 2004.

Authors' Information

Krivoi Sergey – Glushkov Institute of Cybernetics NAS of Ukraine, Ukraine, Kiev, 03187, 40 Glushkova Street, Institute of Cybernetics, e-mail: krivoi@i.com.ua

Hajder Miroslaw – Technical University of Rzeszow, Rzeszow, Poland, e-mail: mhajder@prz.rzeszow.pl

Dymora Pawel – Technical University of Rzeszow, Rzeszow, Poland, e-mail: pdymora@prz.rzeszow.pl

Mazurek Miroslaw – Technical University of Rzeszow, Rzeszow, Poland, e-mail: mmazurek@prz.rzeszow.pl

ROBOT CONTROL USING INDUCTIVE, DEDUCTIVE AND CASE BASED REASONING

Agris Nikitenko

Abstract: *The paper deals with a problem of intelligent system's design for complex environments. There is discussed a possibility to integrate several technologies into one basic structure that could form a kernel of an autonomous intelligent robotic system. One alternative structure is proposed in order to form a basis of an intelligent system that would be able to operate in complex environments.*

The proposed structure is very flexible because of features that allow adapting via learning and adjustment of the used knowledge. Therefore, the proposed structure may be used in environments with stochastic features such as hardly predictable events or elements. The basic elements of the proposed structure have found their implementation in software system and experimental robotic system. The software system as well as the robotic system has been used for experimentation in order to validate the proposed structure - its functionality, flexibility

and reliability. Both of them are presented in the paper. The basic features of each system are presented as well. The most important results of experiments are outlined and discussed at the end of the paper. Some possible directions of further research are also sketched at the end of the paper.

Keywords: *Artificial intelligence, Inductive reasoning, Deductive reasoning, Case based reasoning, Machine learning, Learning algorithms.*

Introduction

During a short period of time (lasting only several decades) there have been developed a lot of different technologies and approaches to solve various types of problems existing in the field of artificial intelligence. A complexity of those tasks that can be performed by intelligent systems is growing from year to year. Therefore, the range of application of artificial intelligence (AI) technologies has been significantly widened. One of the challenging tasks that is always hard to accomplish is building an autonomous intelligent robotic system, because it includes design of software, hardware and mechanics as well. The task is even more challenging if an operation environment is complex and has some stochastic features or entities with stochastic behavior.

Before trying to build a structure of an autonomous intelligent robotic system, it is necessary to define the environment in which the system should be able to operate. A basis of such a definition can be found in the assumption that every object can be described as a system [Lit.1] Obviously, a complex environment can be described as a complex system. There are several very basic features that define a complex system [Lit. 2]:

- uniqueness – usually complex systems are unique or number of similar systems is unweighted.
- hardly predictable – complex systems are very hard to predict. It means that it is hard to calculate the next state of a complex system if the previous states are known. The hard predictability may be related of the mentioned stochastic elements or features of the environment.
- an ability to maintain some progress resisting against some outer influence (including influence of the intelligent system).

Of course, any complex system has every general feature such as a set of elements, a set of relations or links etc… [Lit. 3]

Obviously, if the system operates autonomously in a complex environment, it has to form some model of the environment. It is not always possible to build a complete model of the environment for different reasons. That may be caused by a huge space of possible states of the environment (or even infinite), expanses or other reasons. It means that an intelligent system will use only an incomplete model of the environment during operation and be able to achieve its goals.

The structure presented in the paper exploits an adaptation and uncertain reasoning technique as general methods to deal with an incompleteness of the system's model of its environment.

Basic Feature of the Intelligent System

In this section the basic features of the proposed structure are outlined and explained according to the previous research activities [Lit.4].

Summarizing the basic features of the proposed structure are as following:

- An ability to generate a new knowledge from the already existing in the system's knowledge base. This ability can be achieved by means of deductive reasoning. In order to increase the efficiency a case based reasoning may be used [Lit.6] This feature, obviously, includes also an ability to reason logically. The proposed structure does not state the kind of deductive reasoning that should be used. The only rule is that the selected deductive reasoning method has to address demands of a particular task. As it is described above, complex environments may be very dynamic and even with stochastic features. Therefore, some uncertain reasoning techniques may be the most suitable for complex environments. The described below experimental systems also have uncertain reasoning techniques implemented as a deductive reasoning module.
- An ability to learn. As it is assumed above in complex environments the intelligent system eventually will not have a complete model of the environment. Therefore, the environment will be hardly predictable. Also complex environments are dynamic - in other words the system will face with new situations very often. Obviously, some mechanisms of adaptation should be utilized. From point of view of intelligence, an

adaptation includes the following main capabilities: a capability of acquiring new knowledge and adjustment of the existing knowledge. The inductive reasoning module refers to capability of acquiring new knowledge or learning, in other words. This feature may be implemented by means of inductive reasoning. During an operation, the intelligent system collects a set of facts through sensing the environment that forms input for learning.

- An ability to reason associatively. This feature is necessary due to a huge set of possible different situations that the intelligent system may face with. For example, there may be two different situations that can be described by n parameters (n is big enough number) where only k parameters are different (k is small enough number). Obviously, these situations may be assumed as similar. Therefore an associative reasoning is used – to reason about objects or situations that are observed for the firs time by the intelligent system similarly to reasoning about known situations and using knowledge about the known situations and object. The associative reasoning is realizes through using associative links among similar objects and situations. An issue about which objects and situations should be linked is conditioned by particular tasks or goals of the system's designer.
- An ability to sense an environment. This feature is essential for any intelligent system that is built to be more or less autonomous. This feature also includes an ability to recognize objects / situations that the system has faced with as well as an ability to obtain data about new objects. All sensed data is structured in frames (see below). During the frame formation process the sensed environment's state is combined with system's inner state thereby allowing the system to reason about system itself as well as relation between system's inner state and sensed environment's state. Also, the sensed system's and environment's states are used to realize a feedback in order to adjust system's knowledge. Thus, the system's flexibility is increased.
- An ability to act. This feature is essential for any intelligent system that is designed to do something. If the system (autonomous) is unable to act, it will not be able to achieve its goals. Obviously, the system has to act in order to achieve its goals as well as to obtain the feedback information for readjusting its knowledge or to learn new knowledge. The way of acting and the purpose of acting vary depending of goals of the system's designer or user.

The listed above features form the basis for an intelligent system that operates in sophisticated environment. According to the features of complex systems that are listed above, any of them may be implemented, as it is needed for a particular task. In other words, the implementation methods and approaches are dependent of the purposes of the system itself.

Nevertheless the main question is how to bind all of the listed above features in one whole - one intelligent system.

Obviously, there is a necessity for some kind of integration. There are many good examples of different kinds of integration. For example so called soft computing which combines fuzzy logic with artificial neuron nets [Lit.6] or Case based reasoning combined with deductive reasoning [Lit.7].

In order to adjust an intelligent system for some particular tasks different structures may be used [Lit 14]. This paper presents one of the alternative structures that may be used in order to implement all of the listed above features and may form a kernel of autonomous intelligent system.

The proposed structure is based on intercommunicating architecture. In other words, the integrated modules are independent, self-contained, intelligent processing modules that exchange information and perform separate functions to generate solutions [Lit. 14].

Structure of the Intelligent System

According to the list of very basic features there can be outlined the basic modules that correspond to the related reasoning techniques:

As it is outlined in the figure 1, there are four basic modules that form system's kernel. Each of the modules has the following basic functions

- *Deductive reasoning module.* This module performs deductive reasoning using *if..then* style rules. In order to implement adaptation functionality, this module may exploit some of uncertain reasoning techniques. In the proposed structure the main purpose of this module is to predict future states of the environment as well as the inner state of the system. During the reasoning process *if..then* rules are used in combination with input data obtained from system's sensors. The proposed structure itself does not state what kind of deductive reasoning method should be used. It depends of particular goals of the system's designer. In both practical

implementations a forward chaining certainty factor based reasoning had been used [Lit.17]. If a task requires fuzzy reasoning or other reasoning technique may be used as well.

- *Inductive reasoning module.* This module performs an inductive reasoning or in other words – inductive learning. It learns new rules and adds them to the rule base. Also, the incoming data from system's sensors is used. Again, the proposed structure does not state what kind of inductive learning technique is used. The only limitation is a requirement to produce rules that could be used by the deductive reasoning module. For example, if the fuzzy reasoning is used, then the result has to include fuzzy rules.
- *Case based reasoning module.* Case based reasoning operates with "best practice" information that helps to reduce planning time as well as provides this information for researcher or user in explicit manner. As said above in complex environments there may be a while of unique situations. To extract (or to learn) any rule an intelligent system needs at lest two equal (or similar – the most part of feature (attributes) are equal) situations. It means that in complex environments a while of situations experienced by the intelligent system may remain unused. Obviously, these unique situations (or cases) may be extremely valuable not only for the intelligent system but also for the modeler who uses the system. The case based reasoning module is involved to process and use these unique situations.
- *Associative reasoning module*. This module links objects according to similarities among object features as well as situations, thus allowing to reason associatively [Lit.5]. This module allows to reason about new situations or new objects using knowledge about similar objects or situations. It is an essential ability in complex and dynamic environments in order to increase a flexibility of the intelligent system.

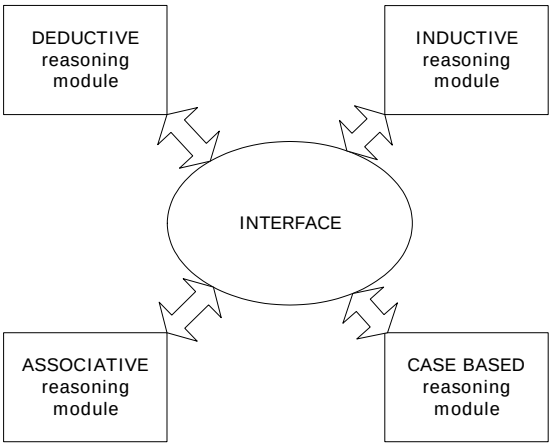


Figure 1. Basic modules

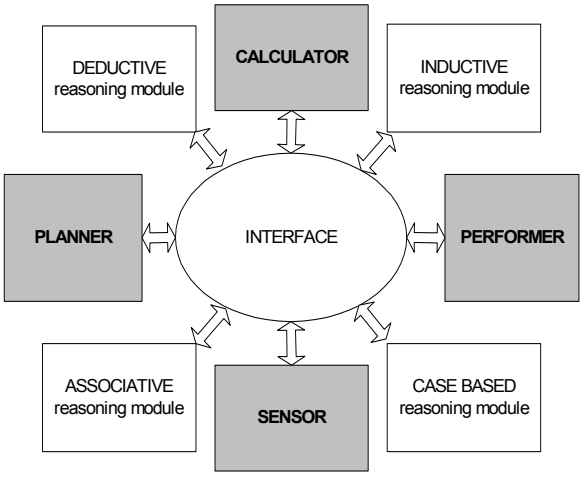


Figure 2 Enhanced structure

Of course, the intelligent system needs additional modules that would supply it with necessary information about the environment and mechanisms to perform some actions.

Therefore, the basic structure shown in figure 1 is complemented with additional modules. The enhanced structure is depicted in the figure 2.

The additional modules are drawn in gray. Each of the additional modules has the following basic functions:

- *Planner module.* This module is one of the central elements of the system. Its main function is to plan future actions that lead to achievement of goals of the system. On author's opinion, an ability to predict future events or situations in the most obvious manner demonstrates an intelligence of any more or less autonomous intelligent system. During the planning process three of the basic reasoning techniques are involved – deductive, case based and associative reasoning. A result of the planner is sequence of actions that are expected to be accomplished by the system thereby achieving its goals.
- *Sensor module.* The module's purpose is to collect information from system's sensors about environment and system's inner state. The sensed information is portioned in separate frames (see below) and forwarded to the interface (discussed later). Once the information is forwarded, it is available for other modules for readjustment of knowledge, for learning new knowledge or other purposes.

- *Performer module.* This module performs a sequence of actions that are listed in the plan. Also, this module uses information about current state of the system an environment in order to determine whether the instant actions can be accomplished. If not appropriate feedback information is sent to the sensor module.
- *Calculator module.* This module collects and produces any reasoning relevant quantitative data. For example, in both implementations (see below) this module is used to calculate certainties of rules including rules newly generated by the inductive reasoning module. Thus, this module is directly involved in knowledge readjustment process. Functionality of the module may be enhanced according to the necessities of the particular tasks or goals of the system's designer.

As it is depicted in figure 1, all of the four modules need some interface to communicate with each other therefore all of the modules use a central element – *Interface* in order to communicate to each other. They are not communicating to each other directly thereby a number of communication links is reduced as well as all of information circulating in the system is available for any module, if there is such a necessity.

A simplified structure of the interface is depicted in the following figure:

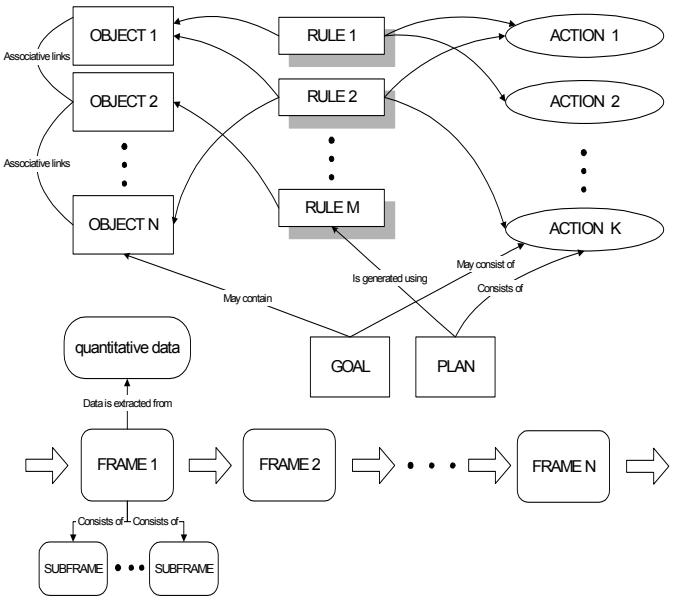


Figure 3 Structure of the interface.

The structure consists of several basic elements. The fundamental element of whole structure is *object*.

Object. Objects are key elements in the interface structure. They correspond to some kind of entities in the environment (or in the intelligent system). Every object is described with a set of features (attributes). Each feature has some value. As it is depicted in the figure 3 objects are linked to each other by associative links. These links form basis for associative reasoning.

When the intelligent system runs into a new situation some subset of objects is activated. These objects map to those entities that the intelligent system currently senses. If there is no rule that can be activated, then the intelligent system may try to activate associated objects. Thus, the system can try to reason about objects by using associated rules. The result may be less feasible, but using association among objects the system can run out of dead end situations.

A mechanism of associative memory is very useful when the system works with a noisy data. This mechanism allows to correct faults of the sensing mechanism [Lit.5,Lit.7]. For example, if the input vector of the sense, which corresponds to some entity, has some uncertain or incorrect elements (attributes of object) then the system would not be able to activate any of the objects. In this case associative memory mechanism will activate the closest object [Lit.5, Lit.7] thus the sensing error will be less significant to the reasoning process.

Rules. Rules are any kind of notation that represents causalities. In the practical experiments were used a well-known *if..then* notation. As it is depicted in the figure 3 rules are linked to objects and actions. When the system activates objects by using associative links linked rules also are activated thus system can scan a set of

"associated" rules as well. This ability improves system's ability to adapt. Rules (for example, those of type *If..Then*) may refer not only to facts but also to actions. Thereby rules through deductive reasoning are used in planning process.

Actions. Actions are some kind of symbolic representation that can be translated by the intelligent system and cause the system to do something. For example "*turn to the right*" causes the system to turn to the right by 90^0. Each action consists of three parts: precondition, body and postcondition. Precondition is every factor that should be true before the action is executed. For example, before opening the door it has to be unlocked. Body is a sequence of basic (or lover level) actions that are executed directly – for example a binary code that forwarded to motor controller causes the motors to turn (in a case of robotic system). Post conditions are factors that will be true after the execution of action. For example after opening the door, the door will be opened. It is important to stress that both implementations of the structure do not have any postcondition information at the beginning. All of the postconditions are learned during the system's runtime.

Frames. Frames are some kind of data structures that contain the sense array from environment and from the system. It means that frames contain snapshots of the environment's and the system's states.

As it is depicted in figure 3 frames are chained one after another thus forming a historical sequence of the environment's and the system's states. Frames can be structured in hierarchies. Hierarchies help to see values of features that can't be seen in a single snapshot. For example: motion trajectories of some object etc. Frames form an input data for learning (induction module) algorithms as well.

Goal. A goal is some kind of task that has to be accomplished by the system. It can be defined in three ways: as a sequence of actions that should be done, as some particular state that should be achieved or as a combination of actions and states. The third option is implemented in robotic systems described below.

Plan. Plan is a sequence of actions that is currently executed by the system (Performer module). It may be formed using both basic and complex actions. After the plan is accomplished, it is evaluated depending on whether the goal is achieved or not thereby forming a feedback information for calculator module.

Quantitative data. This element is used to maintain any kind of quantitative data that is produced by calculator module and is used during the reasoning process. For example, it may contain certainties about facts or rules, possibilities etc. Quantitative data is collected during the reasoning process as well as during the analysis of the input data - feedback data.

All of those components together form an interface for the basic modules: Inductive, Deductive, Case based and Associative reasoning.

Fundamental elements of the structure are implemented in experimental software and robotic systems that are shortly described below.

Experimental Software System

As it is mentioned above, fundamental elements of the proposed structure have found their implementation in experimental software systems. The implemented elements are: Case based reasoning, Inductive reasoning and Deductive reasoning. Deductive reasoning is implemented as a statement logic module based on rules designed in *if…then* manner. The induction module is implemented using very well known algorithm ID3 [Lit.9] It has its more effective successor C4.5 [Lit.10]. The case based reasoning module is implemented using pairs {situation, action}. Each of pair has its value that determines how effective it is in a particular case. During the planning this value determines which actions are selected if more then one action may be selected. The maximum length of the plan is limited in order to avoid infinite planning due to lack of a necessary knowledge for successive planning.

The environment is implemented as world of rabbits and wolf (domain of pray and hunter). There are defined additional objects "obstacles". The number of rabbits and obstacles is not specified thus allowing definition of very complex configurations of the environment. The intelligent system is implemented as wolf. Rabbits may be moving or standing at one place. Wolf can catch rabbits. The wolf is moving according to its plan. Researcher (modeler) can freely change number and place of obstacles and rabbits during the system's runtime thus acting as a stochastic element in the system's environment. The goal also may be defined and changed at any time by the researcher (modeler).

The intelligent system demonstrates flexibility of the proposed structure. The results of experiments and experience accumulated during the implementation shows that new types of objects can be introduced without changing the proposed structure.

It means that even being incomplete this structure demonstrates good ability to adapt and to operate

Experimental Robotic System

The implemented robotic system is the next step of validation of the proposed structure.

Robotic system is a semi autonomous intelligent system that encapsulates all of the mentioned above elements of the proposed structure and interface among basic modules described above.

Mechanics

The system is a two-wheeled semi-autonomous robot. It has two driving or casting wheels and two auxiliary wheels for balancing. Casting wheels are driven by direct current motors equipped with planetary reducers.

The robot is equipped with the following sensors:

- Eight IR (infrared) range measuring sensors;
- Electronic compass;
- Four bump sensors (two front and two rear micro switches)
- Four driving wheel movement measuring resistors (two for each driving wheel in order to achieve reliable enough measurements).

Two Basix-X [Lit 15] microprocessors are used in order to communicate with PC and perform input data preprocessing and formatting. Processors communicate with each other via RS-232 connection.

Prepared and formatted data as frames (see above) are sent to PC also via RS-232 connection. All other modules of the intelligent system are implemented as PC-based software that has user-friendly interface allowing simple following the system's operation, collection of research sensitive data, changing system's goals etc.…

Input

All sensed data is portioned in frames – one frame per second that is sent to PC via RS-232. In order to increase reliability the sensed data is preprocessed. The most important preprocessing is performed with IR ranger data because of the signal errors that should be eliminated. The IR sensors form so-called *situation* that consists of eight integer values. Each of the value describes distance to the closest object at one of eight directions.

Situation data is combined with other sensor data thereby forming an input data array that has the following structure:

[situation, compass, wheel 1 current position, wheel 2 current position,
wheel 1 last position, wheel 2 last position, action, Bumper 1, bumper 2],

where:

Situation – eight integer values of ranges at appropriate directions;
Compass – System's orientation angel according to the Earth magnetic pole;
Wheel 1/2 current position – integer values that describes current position of the casting wheel
Wheel 1/2 last position – integer value that describes the last position of the casting wheel;
Bumper 1/2 – Boolean values that indicates whether an appropriate bumper switch is on (the robot has run onto obstacle);
Action – integer value that indicates the last action that has been performed;

The arrays are sent to PC each second. PC software stores incoming data into frames – one data array per frame.

Used basic algorithms

Inductive reasoning module:

This module is implemented using Quinlan's C4.5 inductive learning algorithm. One of the C4.5 features it that it requires a portion of data with static structure therefore as an input data a portion of frames is used. Each frame has a static structure. A sequence of frames is reformatted in the way that in one input data set record includes two frames – cause frame and consequence frame. For example:

Frame 1; Frame 2
Frame 2; Frame 3:
……
Frame n; Frame n+1;
Frame n+1; Frame n+2;

Thus pairs of causes and consequences are formed therefore rules of type *If Cause Then Consequence* may be produced.

Obviously, the produced rules can describe events that follow right one after another. This is very important limitation of the proposed technical solution. Therefore, this issue is one of the future research directions in order to enhance the system with ability to reason about longer periods.

Deductive reasoning module:

This module is implemented via using forward chaining certainty factor reasoning technique. Rules are described in *if…then* manner. Each rule is evaluated with certainty factor that is obtained each time when new input action is accomplished by the robot and appropriate input frame is received. From the newly added frame the feedback information is obtained and certainty values of appropriate rules (that where used in order to plan the accomplished action during the planning process) are recalculated. Thus, the knowledge of the system is adjusted – via feedback information.

After induction when new rules are added, their certainty values are set to the maximum thus indicating that new knowledge is more valuable then the existing. If the new rule is conflicting with any existing rule, than certainty values of the existing rules are decreased, thus priority of the new rule during the deductive reasoning process is increased. This simple mechanism adds a capability to readjust knowledge according to the changing environment.

Associative reasoning module:

Associative reasoning module uses links among situations (each unique situation is stored in the interface data structure). If in the newly added frame a unique situation appears, than it is stored in the interface data structure. The new situation is linked to the similar situations. Similarities are determined by empirical algorithm.

During the deductive reasoning process, those links are used to select rules about similar situations if none of them is found about the actual situation. Thus, the system can reason associatively about certain situations if there is such a necessity.

Case-based reasoning module:

This module is used during the planning process. Cases are added after the feedback information about accomplished task (achieved goal) is received. Each case consists of the following parts: Goal, Situation and Plan. In other words, each case indicates what plan may be used in order to achieve certain goal in a certain situation.

Planner module:

This module combines deductive, case-based and associative modules in order to plan future actions that lead to achievement of the goal.

As it is possible to define a combined goal consisting of actions, and states sequenced one after another thus forming a sequence of subgoals, for each subgoal a subplan is produced. Planner can produce a plan only for a single goal or subgoal. A special algorithm is built that follows the plan execution during the runtime.

System's features

The PC-based software system implements and demonstrates all of the structure's elements mentioned above. The system is built for research purposes only in other words it is built for experiments in order to examine and validate the proposed structure. Therefore the system's user interface is built to be as flexible as possible allowing its user to manipulate with the robot's state, goals and results at the runtime.

The most important features of the system are:

- Ability to work with multiple goals with mixed structure that may include – actions, states or both;
- Ability to adapt via using inductive learning algorithm C 4.5[Lit.10].
- Case-Based reasoning is used to store information about best-practice cases and to use this information during the planning process.

- Ability to reason via using Certainty theory ideas thus allowing addition of new rules that may be conflicting with existing rules in the rule base.
- Ability to reason using associative links among objects (situations).
- The system's knowledge and system's state relevant data is stored and processed in explicit and easy to follow manner thus demonstrating advantages of the used knowledge based techniques.

It is important to stress that at the very beginning of the system's operation it has no information about consequences of each action – it needs to learn them. But if it is necessary the system's rule base may be filled with rules, cases and other research relevant information thus allowing to model some particular state of the system.

All of the necessary experiments are not finished yet - the system is under research process, but even the first experiments demonstrate very good ability of adaptation and learning new sequences of actions in order to achieve goals. All of the conceived experiments may be split in two major groups – experiments with goals that require matching of one action to one goal and goals that require more than one action in order to achieve the goal. Till now only the first group of experiments has been accomplished.

Possible Advances and Future Research

In order to queue actions one after another thereby building a sequence of actions that lead to achievement of the goal a planning module is used. The planning module is built as a classic single goal planner. If there is more than one goal, then the planner builds plans one by one for each goal. Usually for autonomous systems, there is a necessity to work with more than one goal at the same time, for instance, to follow the charge of batteries and to avoid obstacles. If the system is a team member, then the team's goals should be taken on account as well.

In a common situation, the avoidance of obstacles may have a higher priority than following the charge of batteries. If the battery charge is low then global priorities may change. In other words the system should be able to handle so-called global dynamics [Lit. 16] of the plans and their priorities. The mentioned ability is essential in such domains as robot soccer game or other similar very dynamic environments and complex.

The proposed structure cannot handle globally dynamic planning yet. This is one of directions for future research activities.

Obviously, may be there are tasks that cannot be accomplished using a single intelligent system. For example, simulation of some complex environments such as battlefields, transport systems etc. Therefore, there should be used more than one system, thus forming a multiagent environment.

There are different ways to design multiagent system [Lit.11, Lit 12]. In different domains, different solutions may be applied. Referencing to the said above, there may be outlined another direction of farther research and experiments – adjustments of the proposed structure in order to allow the intelligent system operate in a multiagent environment.

One of the most sophisticated problems in such a multiagent environment is communication because every communication parameter may be variable. It is easy to imagine that two intelligent systems may try to communicate using different knowledge representation schemas, different knowledge, different communication protocols, different type of "conversation" (for example: questioning, answering, argumentation etc.) or even different physical communication channels (radio frequency, verbal communication etc.) [Lit.13].

Another important direction of future research activities is adjusting the basic reasoning modules to allow reasoning about longer periods. That would allow planning actions for longer periods of time and to reason about events that could be caused by the system and would take a place after longer period as well.

Conclusions

Practical experiments show that the proposed structure may be very flexible even in very changing environments with variable goals. In both cases an adaptation and ability to learn is essential and both of them are persistent in the proposed structure. The proposed structure demonstrates capability to operate autonomously that makes it useful for autonomous robot control purposes.

In spite of the first results that are quite promising there are still some open questions that should be answered in the farther research activities.

References

[Lit.1] В.Н. Волкова «Теория систем и методы системного анализа вуправлении и связи» Радио и Связь 1983.
[Lit.2] В.В. Дружинин, Д.С. Конторов «Системотехника» Радио и Связь 1985.
[Lit.3] А.Д.Цвиркун «Структура сложных систем» Советское радио 1975.
[Lit.4] A.Nikitenko, J.Grundspenkis "The kernel of hybrid intelligent system based on inductive, deductive and case based reasoning" KDS2001 conference proceedings, St. Petersburg 2001.
[Lit.5] Bradley J. Rhodes "Margin Notes Building a Contextually Aware Associative Memory", *The Proceedings of the International Conference on Intelligent User Interfaces* (IUI '00), New Orleans, LA, January 9-12, 2000.
[Lit.6] M. Pickering "The Soft Option" 1999.A.R.Golding,
[Lit.7] P.S.Rosenblum "Improving Accuracy by Combining Rule-based and Case based Reasoning" 1996.
[Lit.8] A.M.Wichert "Associative Computation" der Fakultät der Informatik der Univesität Ulm Wrotslav 2000.
[Lit.9] J.R. Quinlan "Comparing Connectionist and Symbolic Learning Methods" Basser Department of Computer Science University of Sydney, 1990.
[Lit.10] J.R. Quinlan "Improved Use of Continuous Attributes in C4.5" Journal of Artificial Intelligence Research 1996.
[Lit.11] P.Scerri, D.Pynadath, M.Tambe "Adjustable Autonomy in Real-world Multi-Agent Environments" Information Institute and Computer Science Department University of Southern California, 2000
[Lit.12] P.Stone, M.Veloso "Multiagent systems: A Survey from a Machine Learning Perspective." IEEE Transactions on Knowledge and Data Engineering. 1996.
[Lit.13] J.Dale, E.Mamdeni "Open Standards for Interoperating Agent-Based Systems" 2000 (last changes)
[Lit.14] S. Goonatilake, S. Khebbal "Intelligent Hybrid systems" 2002 john wiley & sons
[Lit.15] © NetMedia Inc. www.basicx.com
[Lit.16] H. Burkhard, J. Bach, R. Berger, B. Brunswieck, M. Gollin "Mental models for Robot Control" "Advances in Plan – based control of Robotic agents" International seminar proceedings, Dagstuhl Castle, 2001
[Lit.17] J.J.Mahoney, R.J.Mooney "Comparing methods for refining certainty-factor rule-bases" Proceedings of 11th international workshop on Machine Learning, Rutgers, 1994.

Author's Information

Agris Nikitenko – Doctoral student, Riga Technical University, Department of System theory, Meza street 1 k.4, Riga LV-1048, Lavtia, E-mail: agris.nikitenko@rembox.lv

INFORMATION MODELS FOR ROBOTICS SYSTEM WITH INTELLECTUAL SENSOR AND SELF-ORGANIZATION

Valery Pisarenko, Ivan Varava, Julia Pisarenko, Viktoriya Prokopchuk

***Abstract**: It is chosen broad class of complex technical intellectualized systems with adaptation and self-training and some special properties of extremality. As example, the problem of adaptive control system for the scientific experiment with thermonuclear plasma is considered. The second example is connected with creation self-organizing mobile robot for examination emergency technical object on sea bottom.*

***Keywords**: simulation modelling, autonomous mobile robots, self-training, aggressive ambience.*

Introduction

For performing a certain work in extreme condition and aggressive ambience, which is not compatible with health and safety of the person-operator, it is necessary to develop specialized robotics systems, including autonomous mobile robots with element of the artificial intelligence. Herewith for in stage of the designing as well as for the robot usages it is necessary to provide exchange by information between person and robot on language of the natural images about space, to form the certain images about external ambience with creation of knowledge-base, to provide the ability to self-training and adaptive control for a robot.

Class of Complex Technical Intellectualized Systems

In more general event of the object of control (OC) for aim of simulation modelling of the crossing processes between different staff situation in system "ambience-OC" it is necessary to research the ensemble of the decisions initial-marginal problems for dynamical equations of ambiences and OC, from which in mode of the education is formed database and knowledge-base "staff situation". This information is brought in memory information-analytical system IAS, connected with OC (in particular OC can be a mobile robot) and is used then for adaptive control connecting process when change the staff condition "$i \rightarrow j$" of system "ambience-OC".

The multiple problems of the designing object new technology can be united, in the opinion of authors, in a certain broad class so named complex intellectualized systems with adaptation and self-training, which on determination must be an inherent following characteristic:

1. Operation of such object provides independent acceptance to him decisions of the certain types with artificial intelligence element use;
2. Designed object must possess the characteristic to adaptation to changing condition of the external ambience;
3. In the both stages designing and operating the object developers provide simulation modelling of the base processes, accompanying operating the object , including their modelling real-time;
4. Developers provide the possibility of the use the tools of simulation modelling function object for tasks of software of the functioning simulator, intended for learning operator-users by object in space of the realization (both in virtual, and in physical);
5. Project of the future object must be created with provision for minimization of the design work on economic-time criterion;
6. For designing the object and its education (adjustment on elected application domain) necessary to form the corporate information vault given with technology of verification given Data Clearing and technologies of quick searching for of the knowledge's Data Mining.

The Authors refer hereto class, in particular, the following objects:

a) designing autonomous mobile robot for operating in extreme ambience [Golovin, 1996];

b) control of scientific complex experiment

c) creation GIS forecasting of the natural catastrophes

d) scene analysis anomalous weakly contrasting object on air- and cosmic- picture.

For these problems is offered structured scheme of adaptive control object, functioning in extreme condition conflict ambiences, containing, aside from most OC, block of the observation and identifications of the external ambience (OIEA), block to identifications of the condition of the object of management (ICOC) and block executive devises (ED). Information on condition of the ambience from block OIEA is given in block of simulation modelling (SM), with output, which falls into information-analytical system (IAS). Worked out in IAS commands of control enter on block executive device (ED).

We shall consider hereinafter two examples of the application domain, for which actual problem of the designing problem-oriented intellectualized robots with system of the identifications of the current condition and adaptive control system on given criterion. The Information model in each of these examples is formulated absolutely different way, illustrated hereinafter.

The First example is connected with problem of adaptive control system extinguishing the dangerous category in installation T-10 with thermonuclear plasma by use pellet-injection (i.e. injecting admixture macroparticle with big electrical charge). The Second example concerns designing the undersea uninhabited mobile robot for examination of the emergency technical object on sea bottom (for instance, oil gaining platform, sea oil pipeline, drowned ammunitions, aggrieved damage atomic submarine).

The Problem of Scientific Experiment Control with Thermonuclear Plasma

For 50 years of the studies on problem operated thermonuclear syntheses (TS) from the first laboratory experimental installation before complex and road of the installation with thermonuclear plasma, is reached big progress, allowed to get such extreme parameters of the plasma (on the temperature, time of deduction and density), which already provide begin thermonuclear reaction in worker to camera [Pisarenko, 2004].

Alongside with successful study of the thermonuclear plasma in traditional tokamaks, the stellarators and magnetic trap [Golovin, 1996], record parameters of the plasma were reached plasma, retightened in Z-pinch, at inject light ion (N, D, Ve) by means of light laser pulse (1)

For scientific experiment on development perspective thermonuclear reactor presents greater urgency a development fireware facilities of adaptive management work by installation with toroid magnetic floor (tokamak) both in staff, and in before-damage mode [Pisarenko, 2004], [Golovin, 1996].

To emergency mode pertain first of all vagary of the vertical offset of the cord, vagary of the big failure, as well as any other undesirable processes of the quick growing of the current, requiring urgent shutdown (extinguish) of the category in plasma in order to avoid emergency conditions.

One of the technologies of adaptive management pertains perspective technology of the shutdown real-time category in tokamak T-10 method admixture macroparticle injection with big *Z* (so named technology pellet-injection). Designed in work [Timohin 2001] information model is based on decision initial-marginal problem for system following four equations with quotient derived parabolic and hyperbolic type:

$$\frac{\partial n}{\partial t}+\frac{1}{r}\frac{\partial}{\partial r}(r\Gamma)=n_{imp}\frac{\partial Z_{imp}}{\partial t} \tag{1}$$

$$\frac{\partial n_{imp}}{\partial t}+\frac{1}{r}\frac{\partial}{\partial r}(r\Gamma_{imp})=0 \tag{2}$$

$$\frac{3}{2}\frac{\partial nT}{\partial t}+\frac{1}{r}\frac{\partial}{\partial r}\left(r\left[q+\frac{5}{2}\Gamma\right]\right)=n(Q_{oh}-Q_{imp}-Q_{br}) \tag{3}$$

$$\frac{\partial \sigma E}{\partial t}=\frac{c^2}{4\pi r}\frac{\partial}{\partial r}\left(r\frac{\partial E}{\partial r}\right) \tag{4}$$

The system of the equations (1) - (4) describes the speaker an admixture macroparticle (the object of management) and conflict of the ambience (the plasma in toroid camera with dangerous category). The flows electron and ion injected admixture are written in following type:

$$\Gamma=-D\frac{\partial n}{\partial r}+nV_p \tag{5}$$

$$\Gamma_{imp}=-D\frac{\partial n_{imp}}{\partial r}+n_{imp}V_p\,, \tag{6}$$

where $q=-\chi\frac{\partial T}{\partial r}$ - electronic flow of the heat.

The author of work [1] built numerical algorithm and software "PLASMA" for simulation modelling of the process pellet-injection for equations (1) - (6). By means of this program, it is obtained the values of the vector parameter experiment (i.e. coefficient values for these equations, marginal and initial conditions and right parts of the equations), which provide the achievement of the necessary result of the dying down for the emergency mode. The carried out simulation experiments have shown that only under definite sign vector parameter experiment as a result of using pellet-injection occurs extinguishing the dangerous category in toroid magnetic camera of the installation T-10 with stabilization of the process on more low level of density and temperature of the plasma. In our numerical experiment leaving the system on stabilization of the process at values of electronic density, on order less initial (answered source dangerous category), existed, in particular, under the following parameter of the information model: D=100, V_p=10, σ=7.19e+14, Z=0.3, χ=1 and Q_{oh}-Q_{imp}-Q_{br}=2600.

Example of the calculated decisions is shown on fig. 1 for specified parameter for density *N* electronic components of the plasma, density injected mixture N_{imp}, electronic temperature *T*, the tension of the electric field *E*. Herewith for ranking the miscellaneous experiment with different values of the vector parameter experiment was calculated importance of the functional quality of the concrete experiment in the manner of integral:

$$J=\alpha\int_0^{tm}[N(t=0,R)-N(t,r)]dt-\int_a^b rdr+\beta\cdot t_m\,, \tag{7}$$

where t_m -importance of the moment of time, when expression under integral (7) reaches the first minimum, a = 0,2 R; b = 0,8 R; R - a small radius of the toroid camera tokamak T-10.

The coefficients α and β were selected in adaptive mode for ensuring eximatority (i.e. for the optimal expert ranking) of the whole expression (7) on result the most efficient experiment on extinguishing the dangerous category by way pellet- injection. Having Chosen optimum these factors, get the functional a quality (7) with these best values α^* and β^* for ranking (marking) ensemble executed numerical experiment on subject of the choice the most efficient strategy extinguishing the dangerous category, appeared in worker to tokamak's camera T-10 for given initial and marginal conditions.

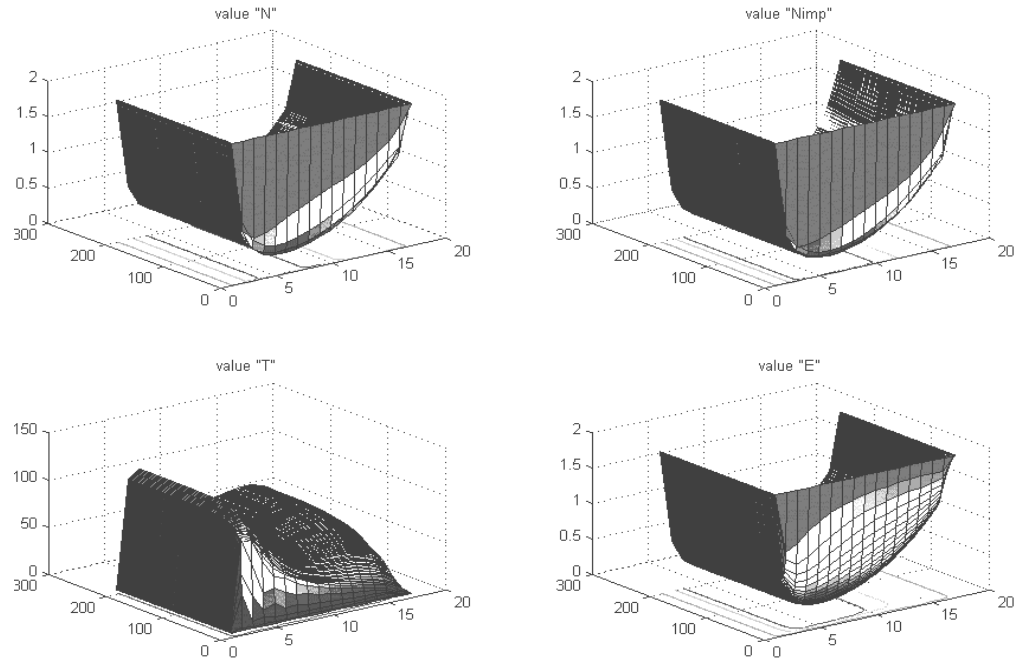


Fig.1. The numerical solutions for electronic density N, density injected mixture N_{imp}, electronic temperature T and values of the electric field E for D=100, V_p=10, σ=7.19e+14, Z=0.3, χ=1 and Q_{oh}-Q_{imp}-Q_{br}=2600

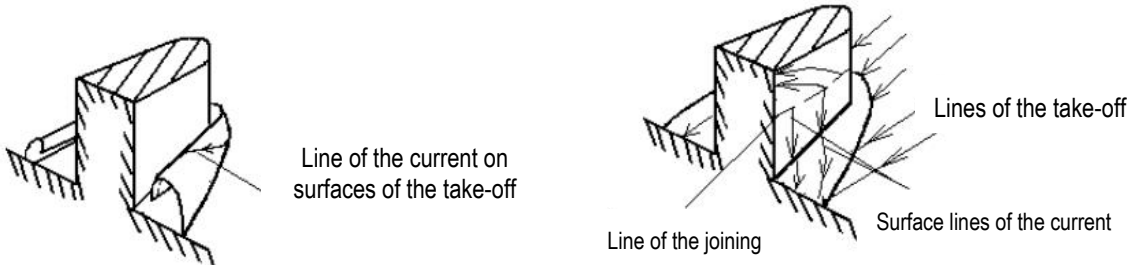
Optimum Path of the Autonomous Mobile Device at Examination of the Object on Sea Bottom

For problem of object management, functioning in fluid ambience or atmosphere (undersea or flying machine) necessary to form information model, requiring solve corresponding to initial-marginal problem for equation Navie-Stokes and equations to continuity [Betchelor, 1973] ,[Gogish, 1990]:

$$-\partial v/\partial t + (v \cdot \partial/\partial x)\, v = grad\,(p/\rho) + \nu \Delta v\,, \tag{8}$$

$$\partial \rho/\partial t = -\partial/\partial x\,(\rho\, v), \tag{9}$$

where ρ- density of the ambience, v - velocity, r - pressure.



a) a surface of the take-off b) lines of the current

Fig.2. Form to surfaces of the take-off and lines of the current under streamlined three-dimensional salient:

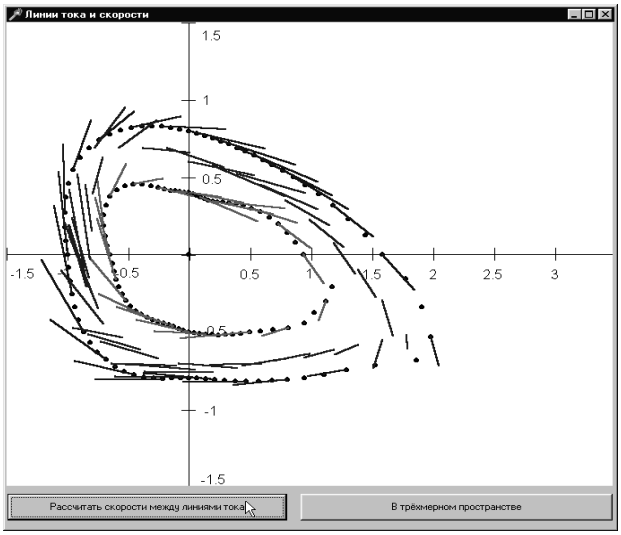


Fig.3. The Field of the vector to velocities in cylindrical curl for cylindrical body, streamlined near-bottom current across axis of the curl.

At results calculation for given numbers staff of situation reasonable structuring in the manner of element database "AMBIENCE", - for instance, in the manner of 3D-flap to velocities of the external ambience, which example is shown on fig. 2. and 3.

Making use of these database it is built algorithm and program complex for adaptive optimum control path of the moving autonomous robot-drill (probe) with the block knowledge-based sensors, which is executing program of the examination of the object (according with the block- diagram on fig.4) and getting in condition near-bottom currents in the curl trace for emergency object on sea bottom (fig. 5).

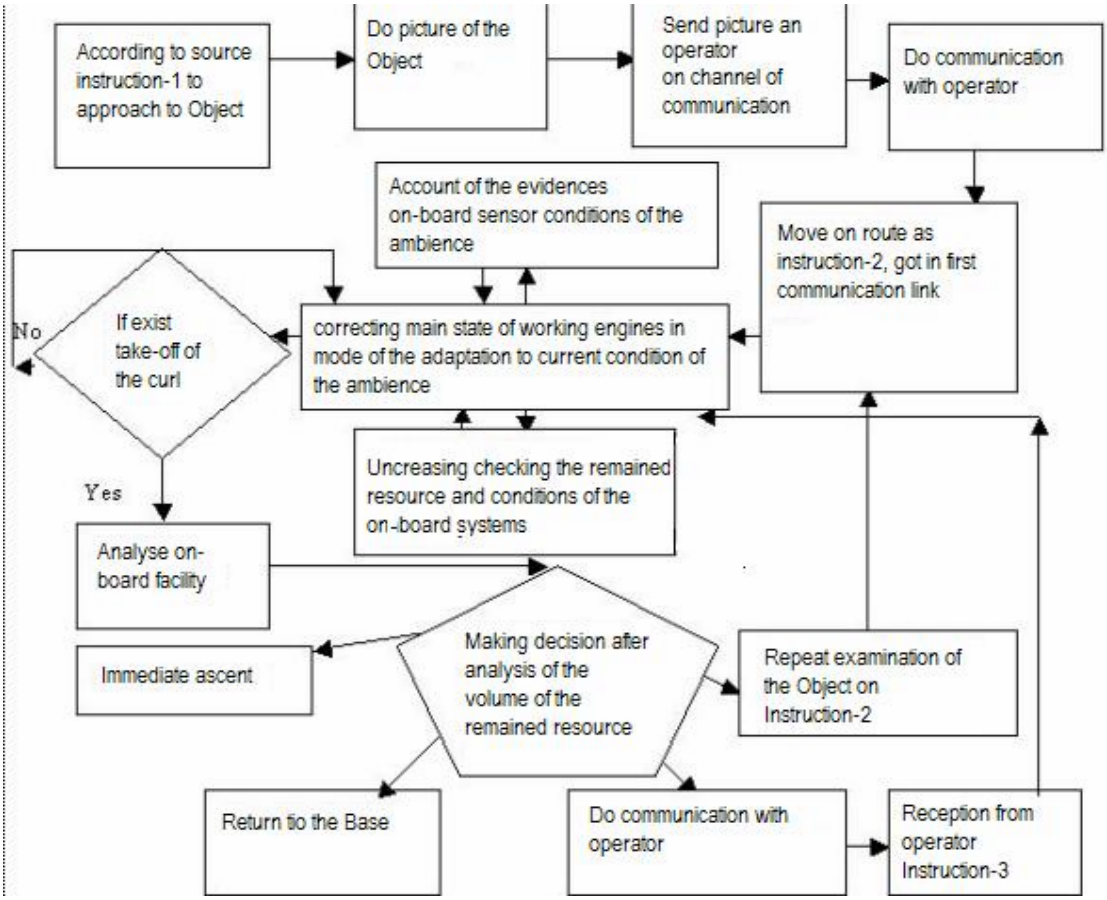
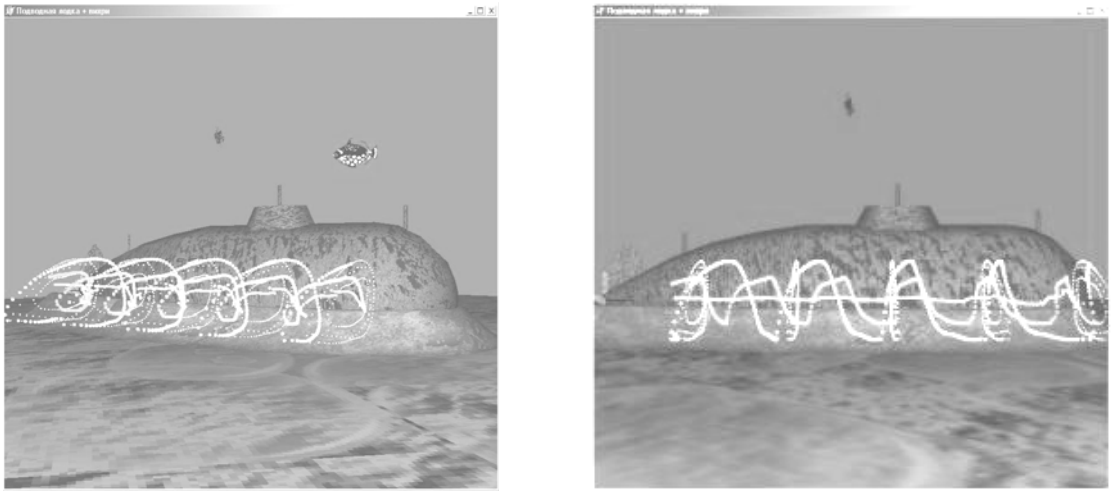


Fig.4. Block-diagram for the operating mobile undersea robot with self-organization



a) b)

Fig.5. The view of the curls in zone of the current near undersea object with calculated family of number of trajectories for undersea robot, which is slowly moving in mode of adaptive control along axis of the object and participating in this curl motion:

a) the family of number of close- to- optimum trajectories;
b) the family of close- to-optimum drill trajectories, which the one of them is the nearest to optimal.

Bibliography

[Pisarenko, 2004] V.G.Pisarenko, I.A.Varava, J.V.Pisarenko, V.I.Semenova Robotics systems with intellectual sensor and multiprocessor simulator of the dynamic condition of the object of control // Artificial intelligence, 3, 2004, s.752 - 758.

[Golovin, 1996] I.N.Golovin, B.B. Kadomcev Condition and prospects operated thermonuclear syntheses // Atomic energy. vol.81, №.5, 1996, p. 364 -

[Timohin 2001] Timohin V.M. and others. The Study of the switching off the category in Tokamak T-10 method to injection admixture particles with big Z // Physics of the plasma, vol.27, №3, 2001, P.1101 - 1110.

[Betchelor, 1973] G.K. Betchelor. An introduction to fluid dynamics – Cambridge at the University Press, 1973.- 760 p.

[Gogish, 1990] Gogish L.V.,Stepanov G.YU. Detachable and cavitational flows -M.:Nauka, 1990,-384 s.

Authors' Information

Valery G. Pisarenko – V.M. Glushkov Institute of the Cybernetics of National Academy of Sciences of Ukraine; Ukraine, Kiev, Krasnoarmejskaya street, 43, 38: 775; e-mail: jvpisarenko@mail.ru

Ivan A. Varava – V.M. Glushkov Institute of the Cybernetics of National Academy of Sciences of Ukraine; Ukraine, Kiev, Polkovaya street, 74/76a, 76; e-mail: jvpisarenko@mail.ru

Julia V. Pisarenko – V.M. Glushkov Institute of the Cybernetics of National Academy of Sciences of Ukraine; Ukraine, Kiev, prospect 40-year Oktober, 82, 84 ; e-mail: jvpisarenko@mail.ru

Viktoriya I. Prokopchuk – V.M. Glushkov Institute of the Cybernetics of National Academy of Sciences of Ukraine; Ukraine, Kiev, akademika Zabolotnogo street, 22, 12; e-mail: jvpisarenko@mail.ru

4.5. Intelligent Systems

STATIC AND DYNAMIC INTEGRATED EXPERT SYSTEMS: STATE OF THE ART, PROBLEMS AND TRENDS

Galina Rybina, Victor Rybin

Abstract: *Systemized analysis of trends towards integration and hybridization in contemporary expert systems is conducted, and a particular class of applied expert systems, integrated expert systems, is considered. For this purpose, terminology, classification, and models, proposed by the author, are employed. As examples of integrated expert systems, Russian systems designed in this field and available to the majority of specialists are analyzed.*

Keywords: *integrated expert systems, real-time, simulation modelling, object-oriented model, rule, complex engineering systems, software tools, task-oriented methodology.*

Introduction

In the mid-1980s and early 1990s, the scientific and commercial success of expert systems (ES), one of the rapidly developing directions of artificial intelligence (AI), showed that, in actual practice, there is a sufficiently large class of problems that cannot be solved by the methods of conventional programming (for example, it is impossible to formulate the solution of the problem in mathematical terms in the form of a system of equations). This class is rather important, since the overwhelming majority of problems are very important in practice. In other words, if, 20 years ago, 95% of all problems solved by computer methods were problems that had an *algorithmic solution,* (hen, under modem conditions, especially in the so-called *describing application* domains (medicine, ecology, business, geology, design of unique equipment, etc.), the situation is completely different—*nonformalized problems* (NF-problems) make up a considerable amount of problems. These problems have one or several of the following characteristics: they cannot be given in a numerical form; the goals cannot be expressed in terms of a strictly defined goal function; and an algorithmic solution exists, but cannot be used due to the limited resources (time and/or memory).

However, NF-problems are not isolated from formalized problems (F-problems) as it is supposed in the basic concept of ESs (KNOWLEDGE + INFERENCE = EXPERT SYSTEM), but are a constituent of problems of real complexity and importance that are solved by conventional methods of programming. In practice, especially taking into account the constantly growing complexity of modern technological and management-technological systems and complexes, this results in combining in the framework of a 'joint programming system such diverse components as ESs and DBs comprising engineering, manufacturing, id management data; ESs and applied software packages (ASP) with developed computing, modelling, and graphical tools; ESs and learning systems including ring and learning systems; ESs and simulation sys-s for dynamical applications; ESs and developed hypertext systems; etc.

It is worth noting that *integration processes* that initiate important new classes of software systems are characteristic not only of ESs, but of all AI. They considerably expand the theoretical and technological tools are employed in designing *integrated intelligent stems* of various types and destinations, which was stressed in the papers of Pospelov, Popov, Erlikh, Tarasov, Kusiak and, later, Val'kman, Emel'yanov and Zafirov, Vittikh and Smirnov, Fominykh, Kolesnikov, Jackson, as well as other Russian and foreign specialists that have considerably contributed to the investigation and concept formulation of integrated intelligent systems.

It is worth noting that the trend towards integration of the AI paradigm with other scientific-technological paradigms when designing applications of intelligent systems has activated processes of so-called *hybridization* connected with designing *hybrid methods of knowledge representation*. The ideas of designing various hybrids and hybrid systems have been known for a fairly long time in AI. However, real practical applications have only

recently emerged. For example, the design of *neural expert systems,* that combine neural network methods and models for searching for solutions with mechanisms of ESs based on expert (logical-linguistic) models of knowledge representation and models of human reasoning became possible due to modern platforms and tools like Neur On-Line (Gensym Corp). However, despite successful individual theoretical investigations and implementations, there is no adequate theory and technique *of integration* and *hybridization* in AI, which partially explains the terminological confusion in using such terms as *integrated system* and *hybrid system* [Rybina,2002].

Thus, the trend towards integration of investigations in different fields in recent years has led to the necessity of combining objects, models, concepts, and techniques that are semantically different. This has inevitably generated both completely new classes of problems and new *architectures of software systems* that cannot be implemented by a straightforward usage of the methodology of simple ESs and tools like the *shell* of ESs that supports this methodology, which is inefficient, labor consuming, and expensive [Rybina,2004].

However, despite the obvious advantages of the design of IESs that can solve real practical problems in the framework of a unified architecture of a software system ensuring efficient data processing based on the interaction of logical-linguistic, mathematical, informational, simulation models, etc., this problem has not left the stage of problem statement.

Isolated attempts to solve the mentioned problems for IESs have been made both abroad and in Russia. However, most of the investigations and developments encompassed only part of these problems and cannot' pretend to propose a particular methodology for designing IESs. Among the Russian fundamental investigations in the field of producing elements of the IES concept, the papers of Erlikh and, from the standpoint of the methodology and technology of designing the software of conventional ES, the results of Khoroshevskii are the best known. Among the foreign works, we note the G2 tool system (Gensym Corp.), which can be used for designing dynamical IES. However, this system does not support a number of important tasks connected with computer knowledge acquisition and automatic generation of a knowledge base (KB). It also does not deal with the problems of designing applied IESs for various classes of tasks typically used in the IES concept.

For the authors of this paper, static IESs became an object of research and development in the late 1980s [Rybina,1997]. Later, in the mid-1990s, I started my research in the field of real-time dynamical IESs (RT IESs) [Rybina,Rybin,1999]. During this period, based on the experience of designing a number of IESs (from simple static IESs and others to dynamical RT IESs for such important and resource-consuming tasks as diagnostics of technological objects, management of industrial complexes and control of technological processes, ecological monitoring, etc.), a scientific direction that has its own object of investigation, conceptual, theoretical, and methodological basis has been formed. A set of methods and software tools (the AT-TECHNOLOGY system) that enables us to support the complete life cycle of automatic designing IESs for particular classes of tasks of solved problems and types of application domains on the basis of a particular task-oriented methodology of designing IESs in static and dynamical application domains has been developed [Rybina,2004].

Characterizing the proposed methodology in short [Rybina,1997], it is worth noting that it is based on a multilevel model of analysis of the integration process in IESs, modelling of particular types of tasks, the relevant techniques of conventional ESs (a task-oriented approach), and the methods for designing a model of the architecture of IESs and the corresponding software components at each integration level. This experience allows one to analyze rather systematically and visually the basic types of integrated intelligent systems (that comprise components similar to simple ESs) irrespective of the specific features of the particular application domain employing only a principle conditionally called "from the standpoint of the architecture."

However, the goal of this paper is not to attempt to make an *inventory of* all types of existing systems of the IES class. The survey of functional, structural, and, partially, programming specific features of various architectures of IESs (using examples of Russian systems) proposed below, in addition to presenting the state of the art for the field, is mainly directed towards finding ways for overcoming the technological lag in designing ESs described above as compared with modern technologies of designing and programming informational systems. This survey may also be useful in order to help to solve the problems of mutual semantic adaptation of scientific terms and notions that can be used by specialists in different fields in the development of applications on the basis of ESs.

1. Classification of IESs, the Interrelation of the Integration and Hybridization Processes in IESs

At present, we can observe two main processes in AI. One of them is connected with the expansion of architectures of intelligent systems by supplementing them with new tools from other scientific fields (the

integration process). The other process is associated with the initiation of new subfields from known fields and the development of special new tools for them (the division process). No doubt, *integration processes* are of the most practical importance, since they join different approaches and paradigms, and, as a consequence, totally new classes of software systems.

An IES is a software system whose architecture, together with a conventional component, an ES that, as a rule, uses the ideology of simple production ESs in order to solve NF-problems, comprises certain components "N" that *extend functional capabilities* of ESs (e.g., a database, packages of application programs, learning systems). Therefore, all IESs can be split into two subclasses: IESs with *shallow component integration* and IESs with *deep component integration*.

Then, in the case of shallow integration of the components of ES and N components, the interaction between, e.g., the DBMS and ES can be carried out at the level of sending messages, and the results of the operation of each component are the source data for the other. In a more complex variant of the shallow integration, one of the software components (ES or N) calls the other in order to refine data or to solve some tasks. In this case, a closer interface is implemented, e.g., with the help of special programs that provide a bridge between the DBMS and ES for messaging. Examples of IESs with shallow integration are presented in Section 2.

The deep integration of ES and N components means that these components are improved by functions that are not characteristic of the conventional types of these components. For example, if we solve the integration problem improving the shell of an ES by including certain functions of a DBMS, this may mean that we set a goal to incorporate a generalized DBMS inside a knowledge base (KB). This approach is the most interesting if the KB is too large to be stored in the computer memory and when efficient mechanisms of data access are required. On the other hand, in relation to an improvement of the DBMS by incorporating basic functions of ESs, we should keep in mind that the DBMS has a developed specialized mechanism for executing such tasks as data storage, maintenance of data integrity, knowledge base navigation, and efficient servicing of complex queries. In addition, this mechanism can also be supplemented by functions of the ES. In particular, one of the approaches to solving this problem is to extend the DBMS in order to transform it into a generalized inference mechanism. It is worth noting that, for designing an IES with a deep integration of components, the approach associated with improvement of tools for ESs by including unconventional functions seems to be more preferable, which, in particular, was employed in the development of the task-oriented methodology for designing IESs.

Complete integration is the highest level of integration for IESs and joins the best properties of the ES and N components. For example, the complete integration of a DBMS and an ES consists in selecting the best specific features and mechanisms for designing the DMBS and the advantages of the ES and in designing new data and knowledge models in order to develop completely new systems with new capabilities. The undoubted advantage of this approach to integrating DBMS and ES components is that all components belong to one system with a unified structure for modelling facts and rules and uniform data and knowledge processing.

The so-called *multilevel model of integration processes* proposed in the framework of the task-oriented methodology allows one to analyze the trends towards integration and hybridization in modern IESs. This methodology considers the integration processes from the standpoint of the following key aspects:

integration in the architecture of IESs of different components that implement both F-problems (the N component) and NF-problems {the ES component) and determine the specific features of functioning of the IES *(high level of integration);*

integration *(functional, structural, and conceptual)* connected with the concepts and methodologies of designing and developing particular classes of IESs and their components *(middle level of integration);*

integration (informational, programming, and technical) connected with the technologies, tools, and platforms (low level of integration).

For at least three levels of integration in IESs, this model allows us to trace the interrelation of particular methods of integration of some components of the IES (N or ES) with the necessity of internal integration, i.e., *hybridization* of some models, methods, procedures, algorithms, etc. In this connection, the following conclusions and suggestions have been made:

the notion of integrated intelligent systems differs from the notion of hybrid intelligent systems;

an integrated intelligent system must not be hybrid and vice versa;

an integrated intelligent system is always hybrid only in the case of complete integration of components.

Therefore, the term "integrated expert system" should be used for complex programming systems that join methods and facilities of ESs with the techniques of conventional programming. The term "hybrid expert system" is advisable to use both for ESs and tools for ESs with hybrid methods of knowledge representation (i.e., joining inside one tool different models of knowledge representation and mechanisms of functioning).

These recommendations are apropos and useful, since the lack of semantic unification leads to misunderstandings in the usage of terms by specialists. Some problems in using the terms "hybrid" and "integrated" may be explained by historic reasons connected with the early paper, where IESs were defined as "hybrid ESs."

It is worth noting that the consideration of problems of shallow and deep integration of the base component ES with a component N (where N is a DBMS, ASP, etc.) at the *high* level allows one to construct in general the so-called generalized model of the architecture of the designed system that reflects its structure. The *intermediate level* is of the greatest interest, since, in the framework concepts and methodologies of IESs, it provides an opportunity to consider the distribution of functions among the ES and N components when developing an application, as well as some design decisions over all components of the designed system. In particular, we can consider possible variants of the joint functioning of the ES and N components. There are several possibilities: the ES and N can be included in each other; they can be used in parallel to solve the same problem; they can be partially included in each other; and an intelligent interface between them can be organized. The integration at the *lower level* is provided by the technology and capabilities of the tools used in the course of development.

Let us explain the terms "static IES" and "dynamic IES" beginning from the notions "data domain" and "application domain." It is well known that "data domain" means a domain of human activity that is specially selected and described (i.e., a set of entities that describe the domain of expertise). The notion *application domain* includes the *data domain* plus a totality of problems solved in it. If an IES is based on the assumption that the source data about the data domain that provides a background for solving the particular problem does not change in the course of solving the problem or the problems solved by the IES do not directly change knowledge about the data domain in the course of their solution, then this domain is *static* (i.e., the domain has a static representation in the IES), while the tasks solved by the IES are static. In other words, a *static IES* is supposed to operate in a static *application domain* if a static representation is employed and static tasks are solved. Otherwise, we deal with a *dynamic IES* that operates in a *dynamic application domain.* This means that a dynamic representation is used and dynamic problems are solved.

Below, based on the introduced terms, notions, and classifications, examples and specific features (functional, structural, and, partially, programming) of the design of a number of Russian and foreign IESs are considered corresponding to the types of N components.

2. Analysis of the Specific Features of the Design of an IES

The analysis of foreign and Russian developments of IESs has shown that almost all of them are devoted to the description of applications of IESs that combine in their structure a set of components that implement both F-tasks and NF-tasks; i.e., they are systems with a *high integration level* (an integration within the scope of the general architecture of the IES). Below, we consider some examples.

2.1. Integration of an ES and a DB

It is worth noting that, despite the fact that, historically, ESs in artificial intelligence and DBs have been developed separately, nevertheless, there are many application domains where, simultaneously, both access to an industrial DB and the use of an ES for decision making on the basis of experience or expertise are required. In this connection, the problems of integrating *static* ESs and DBs have been investigated the best. Therefore, at present, there are many publications devoted to a specialized application of IESs that combine the techniques of KBs and DBs. However, most authors have concentrated their efforts on the description of ways of implementing peculiarities of one or another application domain and have paid substantially less attention to the principles, methods, and models of integration in the framework of ESs and DBs. In this connection, we place the main emphasis only on the papers that are of interest from the standpoint of methods of joining ESs and DBs (DBMSs).

There are two approaches to joining ESs and DBs—*weak and strong coupling*. Strong coupling is used when there is an opportunity to split the operation of an ES into a sequence of steps that require definite predictable and limited information from the DB (e.g., in tasks of medical diagnosis, the information about a particular patient

can be requested before the beginning of a consultation). As a result, the opportunity to efficiently arrange the operation of the ES is a merit of this approach, since, in the course of consultations of the ES, there is no need to query the DB.

However, in other application domains, this approach may be impossible and a strong coupling of the ES and DB may be required. This coupling allows one to multiply read and modify the content of the DB. It is quite obvious that a strong coupling imposes more serious requirements on the system; in particular, the rate of transactions between the ES and DB supposes an efficient implementation of this functional component. Among the first Russian systems of this class, the shell for the DI*GEN ES that has a mechanism of strong coupling of the ES and DB is worth mentioning. To implement this mechanism, DI*GEN contains a set of functional primitives that allow one to position the DB, to read and modify data from an arbitrary field of any record of the DB, to add/delete any records in the DB, retrieve data using a condition, etc.

A radically different approach to integrating ESs and DBs was implemented in one of the earliest Russian applied ESs, the integrated intelligent system of the molecular design of physiologically active substances, which was also *hybrid,* since it employed several reasoning strategies and models of knowledge representation. From the standpoint of the technique for joining an ES and a DB, the characteristic specific feature of this system is the use of an original relational DBMS for generating training samples and storing data, since the decision strategy in this IES is based on two complementary concepts of inductive learning by examples.

Among other developments in the field of static IESs of the early period, a large group was presented by the systems, in which the integration between the ES and DB, ES and ASP, and ES and other components was obtained by using integrated tools like the well-known GURU (INTER-EXPERT) system that contains facilities for designing static ESs, built-in DBMSs, electronic worksheets, and graphical packages. In GURU, the *premise* of each rule in the KB can contain working variables with one value, multivalued fuzzy variables, static variables, cells of electronic worksheets, fields of the KM AN relational DB, numerical functions, operators of relations, Boolean operators, operators over numbers, operators over strings, and symbols, thus providing the integration. Thus, using GURU, from the standpoint of the *lower level of integration,* numerous components N can be joined with components of the ES in the scope of one operation, which allows one to efficiently arrange processing of various data in applied IESs both together with the ES and irrespective of it. This also eliminates the necessity to write special bridges for the IES.

In later *dynamical* IESs that operate in real time (RT IESs), the integration of the ES and DB at the high level is based on designing special bridges, e.g., when using G2 (Gensym Corp.), the G2 Database Bridge is used. For example, when designing subsystems for data acquisition and storage for a system of on-line monitoring, the capabilities of integrating G2 and DBMS Oracle through the G2-Oracle (Oracle Bridge) interface were used. In addition, the tools of the DBMS Oracle and the system software support data backup and long-term storage. In G2, the transmission of online data from external sources was supported by the standard interface of G2 (GSI).

2.2. Integration of ESs with Learning Systems

The next large group is represented by IESs whose architecture integrates ESs with various kinds of learning systems. In general, in connection with the problem of designing computer learning systems, which emerged earlier than ESs and have gone a long way from laboratory programs to power commercial systems, we cannot avoid mentioning the following important specific feature. There are two main processes of learning—"learning" and "tutoring"—that substantially affect the approaches of the soft implementation of IESs.

The direction connected with "learning" process *(learning systems)* includes self-learning, learning with a teacher, adaptation, self-organization, etc. Therefore, when designing learning systems, sufficiently strict models (perhaps, hybrid) are investigated and constructed. These models show the ability to adapt to the external world by storing information. However, from the standpoint of this survey, the direction connected with "tutoring" process *(tutoring systems)* is the most interesting. In this direction, models of the transmission of information from a teacher with the help of a computer are investigated. This is especially interesting, since, in pedagogic, there is no commonly accepted learning theory or formal models of a trained person, tutoring, teaching effects, explanations, etc. Due lo these facts, specialists depend on expert models and tools of their implementation in the scope of IESs.

Tutoring ESs and expert-tutoring systems are the most popular, as well as the entire field called "intelligent tutoring systems" (ITS), among others. It is worth noting that the interpenetrating of integration processes in AI and pedagogic was reflected in ITSs, as well as in tutoring ESs (which can be considered as a subclass of ITSs). This interpenetrating stresses the need for additional components N that can support a *model of a trained person.*

Based on this model, the teacher specifies a current subgoal at the strategic level and the components that realize a particular *tutoring model* in the form of *tutoring actions* at the tactical level. These components must also provide an opportunity to the teacher to observe the actions of the trained person and to help him by using the developed interface facilities.

However, none of the existing systems is able to completely implement the ideas of deep and, especially, complete integration of ESs and N components that support the specified models. In the best case, successive functioning of autonomous components that implement the simplest model of the trained person and a number of learning actions (one of which was "training" based on an ES) was provided.

2.3. Integration of ESs with Hypertext Systems

A wide range of papers is connected with the description of various IESs whose architecture includes integration of components of ESs with not only DBs and tutoring components, but components with hypertext facilities (HT-facilities) as well. The appearance of hypermedia and multimedia systems and interactive leaching environments in recent years has led to the necessity of developing a new type of integrated system—*intelligent teaching environments*—that combine the capabilities of ITSs, hypermedia systems, and teaching environments.

Modern investigations devoted to the problems of integrating ESs with HT-systems provide another sufficiently large range of investigations. Their practical orientation becomes more obvious under the mass distribution of Internet technologies, electronic encyclopedias, Internet shops, etc. The deep integration of the components of ISs and HT-tools actually supplement the capabilities of ESs for hypertexting and allow one to make logical inferences in searching for relevant text fragments, especially, in the fields of activity recommended by standard documents and in electronic business. For future applications, IESs of this type can evolve to "intelligent texts" or "expert-texts," and adaptive models of a user can be designed.

2.4. Integration of an ES and an ASP

Let us consider the analysis of processes of integrating an ES and an ASP in an IES. The global problem of designing *intelligent computer-aided design systems* (ICAD) has been of interest for a long time for researchers and designers. The number of publications devoted to this topic is very large. It is the scope of these papers where the problems of deep integration of ESs with developed mathematical, computational, modelling, and graphical tools, implemented, as a rule, in the form of ASP, were first posed. In such complex software systems, a conventional ES should play the role of an *intelligent user interface,* while the decision maker of the ES should service not only the KB, but also be a *monitor* for the ASP, which can be supplemented not only from the external world, but as a consequence of the operation of the ES.

However, despite the large number of papers that describe applied IESs for various but rather narrow applications, the problems of integrating ESs and ASPs are mainly solved in the scope of IESs with *shallow integration,* i.e., at the level of message exchange between components. The PRIIUS system that joins in its architecture an ES, a data retrieval system, and programs in C, Pascal, and FORTRAN based on the principles of complicated shallow integration is a typical example.

A large number of Russian IESs of this type use foreign ASPs and shells of ESs. For example, in surveys, a large group of IESs applicable in problems of analysis, design, and optimization of the structure of industrial structures and objects of building was described in detail. Among these systems are SACON, designed based on a shallow integration of the well known shell EMYCIN and the MARC package of structural analysis; HI-RISE, in which an ES functions in the DICE environment of the graphical interface; PROTEI, in which the ES plays the role of a monitor that provides the organization of the interaction of all subsystems and components of the IES; and BTEXPERT, in which an interactive FORTRAN program is combined with a ES shell written in Pascal/VS and the processing of symbolic data and the interface are implemented by a pair of mutually complementing tools produced by IBM (ESCE and ESDE). For other classes of problems (control problems), such systems as CACE-III implemented with the help of DELPHI and having access to an ASP of numerical analysis (CLADP, SSDP, SIMMON, and SEPACK) and CASCADE implemented with the help of the DECIDE shell and working with an ASP library that comprises 78 programs of numerical analysis written in FORTRAN are well-known developments.

However, as was shown in this paper, to describe the set of *typical situations* for each type of aircraft, a totality of mathematical models is used in the OOAES. To support these models, an autonomous functional unit, connected with a rule base and a mechanism for making decisions in a rather simple way (like that of GURU) by including in the working memory (data base) of the OOAES elements of the corresponding mathematical models and using

them in the left-hand sides of rules-productions, was developed. In (his case, the valuated elements of the working memory correspond to the actual values of the output signals from the unit for supporting models; the mechanism of logical inference in the OOAES—processing rules—initiates the execution of autonomous units of the support of mathematical models and, then, continues the search for a solution in the rule base. Thus, mathematical models are called from the decision-maker at the level of external programs. In this connection, early OOAESs can be referred to as IESs with *complicated shallow integration* of components of ESs and ASPs (in this case, by mathematical models and the facilities for their support) with further perspectives for deep integration of components and, perhaps, hybridization on the basis of hybrid models of knowledge representation and reasoning.

It is worth noting that the ideas of *integration and hybridization of diverse models* in constructing applied intelligent systems (including IESs, stipulated by modelling real objects, setting and solving problems on certain models, integration of models, problems, and solutions obtained) were initially in the background of AI, which has been mentioned several times. The design of a unified informational environment and formulation of nine features of integration allow one to analyze the interaction of a user, complex product, and data resources and processes from the standpoint of design, theoretical, and experimental investigations on the basis of synthesis and analysis of models (and their junction). Nevertheless, despite the general principles proposed and particular examples of implementation, the problem of designing IESs with complete integration of components, i.e., hybrid systems, remains the most complex problem.

2.5. Integration of ESs with Simulation Systems

The developments that join conventional ESs and simulation systems (SSs) in the framework of *dynamic* IESs are the most widespread. This is the case in which the necessity and possibility of integrating methods for solving NF- and F-problems reveal themselves in the most complete way. In particular, these two techniques are considered, their similarities and differences are analyzed, the expediency of their mutual complementing and interaction is substantiated, and classifications of possible approaches and integration of the ES and N components are proposed from the standpoint of the *intermediate level of integration* (built-in systems, parallel systems, cooperating systems, and intelligent interfaces). A dynamic real time IES that comprises N components represented by a subsystem for modelling the external world is a typical example of an IES designed on the principles of integrating ESs and SSs.

It should be noted that dynamic IESs that are used to support solutions of NF-problems and AI problems are the most complex systems from the standpoint of implementation. As a rule, either *complicated shallow integration* of certain components of the IES or *deep integration* is used. The problems of complete integration are partially discussed only at the conceptual level.

The facilities of simulation are incorporated not only in the architecture of IESs. Very frequently, simulation systems are included in CAD systems, since, due to the complexity of the designed system, only simulation methods allow one to use relevant information of various kinds, including exact data and quantitative data, as well as expert, heuristic knowledge-based experience and assessments. In this case, the structure of the CAD system, together with SSs, contains a certain number of ESs and a common DB.

2.6. Integration of ESs with Knowledge Acquisition Systems

The advances in the ES field have led to the growing importance of special methods and software tools for knowledge acquisition (eliciting, extracting), especially in designing large knowledge bases. In other words, the direction of knowledge acquisition, which is conventional and key for AI, is separated out into a particular research field and attracts methods and approaches from other scientific fields such as psychology, linguistics, neural networks, machine learning, mathematical statistics, regression analysis, data visualization, etc.

An analysis of papers devoted to new lines of AI investigation shows that, beginning in 1994, many specialists specify software tools for knowledge acquisition into a separate category, directly connect it with data mining that rapidly develops knowledge acquisition in DBs, and predict great prospects for this field. In some cases, the systems that combine the capabilities of conventional ESs with a component for knowledge acquisition (and, sometimes, even with a DB) are proposed to be called "partner systems". However, there is no information about the methods and ways of integrating the ES and N components in the architectures similar to IESs in that paper.

2.7. Integration of ESs with Other Software Components

Consider other examples of integrating the ES and N components in the architecture of static and dynamic IESs that are not so characteristic. Under modem conditions, the integration processes manifest themselves most visibly in the architecture of systems aimed at reengineering business processes of enterprises. As is known, the main task of business process reengineering is to find a new way of constructing the existing business on the basis of employing the most modern data technologies, including the methods and tools of ESs. Here, it is important to take note of two aspects.

The first is associated with the general methodology of designing ESs, OOA, and business process reengineering. All of these methodologies are based on the same principles and the same life cycle and the models of processes of a new business that are designed in the course of reengineering correspond to analogous models that are constructed when designing software systems for supporting business and can serve as models of application domains developed at the stages of "identification" and "conceptualization" when designing ESs.

The other aspect is the close relation of the conventional paradigm of ESs of the paradigm oriented to the rules with the so-called business rules. The languages for knowledge representation of production types by and large have the required power for the representation of business rules, in contrast to, e.g., the tools of diagram techniques of structural analysis or object data models of OOA. We can describe all language constructions proposed, in particular, by Ross for describing business rules using the facilities of the language for knowledge representation used in the famous ReThink system. However, only preliminary studies have been undertaken in this area.

Conventional decision support systems (CDSS) have great prospects in the field of integrating the ES and N components, as well as in the hybridization of models of knowledge representation and methods of reasoning. This is especially true in connection with their evolution toward designing *intelligent* CDSS, which, as was mentioned by the authors, can be fully implemented only if the modem techniques for designing intelligent systems based on the concepts of distributed AI and dynamic knowledge models and the methods of plausible reasoning are harnessed. It is also necessary to employ powerful computational platforms and corresponding tool systems similar to G2.

Completing the survey, we should note that, in real practice, together with the considered cases, it is possible to find other examples of the use of models, methods, and tools of ESs in combination with conventional approaches and programs, e.g., integration of ESs with geoinformation systems oriented toward processing cartographic data. And, no doubt, the most modern agent-oriented paradigms in AI are based on the further evolution of intersystem integration processes.

2.8. New Architectures of Distributed IESs

From the standpoint of further prospects in the field of development and evolution of IESs, we should notice (he rapid progress in the field of Web-oriented IESs (Web-IESs), which, on the whole, fits the modem trends in designing intelligent systems connected with the transition from isolated autonomous systems with centralized management to distributed integrated systems with network control structures.

Up to now, the methods of interchange between a Web browser and a network has considerably complicated the design of Web-IESs, in particular, the functioning of the inference engine of an ES. However, currently, the technologies that allow one to solve these problems have begun emerging. At present, there are two main types of architectures of Web-IESs, whose specific feature is completely determined by the methods of activating the inference engine of the ES in the network. The first is the "server-master" architecture that stipulates the ES to be run on the server. The second architecture—"client-master"—stipulates the ES to be run on the client computer (local user computer). Let us evaluate the merits and demerits of each of these methods from the standpoint of designing the architecture of a distributed IES.

Despite the certain access rate connected with the availability of all programs required for operating an ES on the server (which allows us to eliminate the loading of (his program on the user computer), the server-master approach implies a number of inconveniences. First of all, this is associated with the impossibility of meeting a number of conditions of normal functioning of an ES such as the impossibility of saving the history of the system's states and of tracing the steps of inferences made. Moreover, when several clients simultaneously work with the ES, the load of the server and response time to the user queries increase. When using the client-master method, data are processed on the user computer and thus eliminate the problem of storing the states and trajectories of the inference in the ES.

Consequently, each of these approaches has it advantages and disadvantages, and the future will show which variant will be preferable. It is clear that the server approach requires considerable resources on the server, while

the design of Java-client ESs shifts all processing problems to the user computer. In addition, in order to prefer some of the considered variants of designing a Web-IES, it is necessary to know the amount and rate of data transmitted by the corresponding N component.

ES and N components are combined at the level of shallow integration by special bridges that can be implemented by standard techniques (COM, DCOM and CORBA), which makes the processes of interaction of an ES with other N components transparent and simple to implement.

At present, a number of foreign Web-IESs an known that are implemented by either the tools of conventional programming such as C++, Visual Basic, Active Server Pages, and Active X or using special tools like KROL (an object language for knowledge representation) or others.

Conclusion

The questions of to what extent modern expert systems and, in a broader sense, *knowledge-based system* are expert and what are the future trends and prospect: of the development of this class of intelligent systems the most popular from the beginning of the middle of the 1980s are of interest for researchers and designer; of ESs, as well as potential users. The euphoria from the first successes in designing conventional ESs for limited application domains was marred by the difficulties and technological problems when trying to solve problems of real practical complexity and importance.

This survey was devoted to one of the most important problems connected with the change and complication of the architecture of modern ESs because of the dominating processes of integration and hybridization (the emergence of integrated, hybrid, and Web-ESs). Naturally, tile limited length of this survey cannot give a complete representation of the variety of types of the existing architectures of IESs and approaches to their implementation, since it is the ES field where the greatest experience has been accumulated. This experience has become an essential part of almost all *integrated intelligent systems.* Note that we were primarily in the analysis of Russian developments in this area using sources available to any specialist that might be interested in the theory and techniques of the entire process of designing intelligent systems.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research, project no. 03-01-00924.

Bibliography

[Rybina,1997] G.V.Rybina. Task-Oriented Methodology for Automated Design of Integrated Expert Systems for Static Application Domains, Izv. Ross. Akad. Nauk, Tear. Sist. Upr., 1997, no. 5.

[Rybina,2002] G.V.Rybina. Integrated Expert Systems: State of the Art, Problems, and Trends, Izv.Ross.Akad.Nauk, Teor.Sist.Upr., 2002, no. 5.

[Rybina,2004] G.V.Rybina. The new generation of software tools for application intellectual systems development, Aviakosmicheskoe Priborostroenie, 2004, no. 10.

[Rybina,Rybin,1999] G.V.Rybina, V.M. Rybin. Real-Time DynamicJB Expert Systems: The Analysis of the Research Experience and Developments, Prib. Sist. Upr, 1999. no. 8.

Author's Information

Galina Rybina – Moscow Engineering Physics Institute (State University),Kashirskoe shosse, 31, 115409, Moscow, Russia,Email: galina@ailab.mephi.ru

Victor Rybin – Moscow Engineering Physics Institute (State University),Kashirskoe shosse, 31, 115409, Moscow, Russia,Email: rybin@aie.mephi.ru

ADAPTIVE ROUTING AND MULTI-AGENT CONTROL FOR INFORMATION FLOWS IN IP-NETWORKS

Adil Timofeev

Abstract: *The principles of adaptive routing and multi-agent control for information flows in IP-networks.*

Keywords*: telecommunication system, adaptive quality service, multi-agent control, IP-network.*

Introduction

Evolution of informational and telecommunication system requires on the modern stage a development of theoretical bases of design for integrated infotelecommunication computer networks of new generation, including telecommunication systems (TCS) and distributed information and computer resources (local and regional computer networks, data stores, GRID-systems etc.). At this global TCS play role of tools for providing for users as external agents (clients) quality of service (QoS) for their plural access to information and computing resources, distributed in all over the world.

Improvement of global TCS is connected firstly with further development of theoretical bases and realization of methods of automation, optimization and intellectualization of systems for network control of information flow. Reason of it is that today network control network of global TCS depends significantly on network administrators and operators. However, their possibilities and abilities are limited principally.

Alternative way for improvement of network flows control in global TCS is its automation on the base of dynamical models of TCS as complex network plants of control, methods of optimization of processes of routing of data flows and principles of adaptive and intelligent control for traffic with use of multi-agent technologies and protocols of new generation (for example, IPv6-protocols). On this new way there is a possibility of consideration either only real dynamics of TCS, i.e. real change of structure (topology) and parameters (weights of communication channels) of TCS in a real time or adaptation to different factors of uncertainty on the base of monitoring and functional diagnosis of TCS.

In the paper dynamical models of global TCS with changing structure, mathematical models, optimization algorithms and protocols of dynamical, adaptive, neural and multi-agent routing of data flows are described. These models, methods and protocols of new generation are important part of modern theory of adaptive and intelligent control for information flows in global TCS. They reflects experience and scientific reserve, obtained in a process of fulfillment in 2002-2004 a state contract №37.029.0027 with the title “Adaptive methods for control of data flows in telecommunication systems of new generation” of Russian Federal Agency on Science and Innovations in the framework of Federal Scientific Technical Program “Researches and development on priority directions for science and technical development”.

1. Evolution of TCS and Intellectualization of Network Control

Globalization and other modern tendencies of TCS development cause not only significant overview of basic telecommunication concepts but also significant technological drifts as follows [1-3]:

- from speech traffic to data traffic and multimedia traffic;
- from special TCS to global TCS of new generation;
- from local special service to multimedia universal service and applications with guaranteed quality in every time and everywhere.

With consideration of existing tendencies, two scenarios of TCS development are possible: revolutionary and evolutionary. Revolutionary scenario is in a fast substitution of existing electronic TCS development by optical systems with channel capacity about thousands Gb/s. Realization for this scenario requires very big investigations for development and mass production of standardized optical components for global TCS. Evolutionary scenario for TCS development is based on their gradual modernization by advancement of systems, protocols and technology of control for transferred information flows. Today realization of this scenario goes very fast and causes creation of corporative and global TCS of new generations.

Fast grow of real time traffic and its heterogeneous (multimedia) character cause network collisions and overloading, blocking normal TCS work. Rapid development of new kinds of service (electronic commercial, electronic games and entertainment etc.) increased sharply requirements for quality of service (QoS) and information security.

Appeared problems caused necessity for creation of new mathematical models, algorithms of routing and protocols, providing solution of the following tasks:

- development of scaled address system;
- optimization for processes or data flows routing;
- providing of guaranteed quality of service (QoS);
- support and realization of mobile service and Internet.

New tasks and requirements for IP-technologies caused wide use of fourth version of classical protocol (IPv4) and development of its sixth version (IPv6), providing the following [1-2]:

- increase of address field of working part of package till 128 bits, that increases quantity of IP-addresses till 1020 on every TCS node;
- increase of length of package title till 320 bits with information localization, necessary for router work;
- increase of effectiveness by aggregation of addresses and fragmentation of big packages;
- providing of security of information by authentication of nodes-sources and nodes-receivers of information, coding and keeping of wholeness of transferred data.

Architecture of global TCS consists of four basic (basis) subsystems [3]:

1. Distributed communication system (DCS);
2. Network control system (NCS);
3. Distributed information system (DIS);
4. Distributed transport system (DTS).

These subsystems are connected each other and destined for controlled transfer to users of global TCS distributed and computing resources, stored in CS [7-10].

NCS of global TCS of new generation should be adaptive and intelligent [3-11], i.e. have abilities to:

- adaptation (automated self-adjustment) with relation to changing quantity of users, their queries and requirements to quality of provided service, changing structure (topology) of TCS and parameters of nodes and communication channels etc.;
- learning and self-learning for new functions and rules for TCS work;
- self-organization of structure and functions of NCS with dependence on TCS changes;
- prediction for network collisions and other sequences of network control.

Thus, adaptivity and intelligence are the most important features of perspective NCS, destined for regional and global TCS of new generation.

2. Optimization of Traffic Control and Routing of Information Flows

Problem of traffic control in global TCS is decomposed on two interconnected tasks:

1) Planning, optimization and adaptation of routes for flows transfer between TCS nodes;
2) Control of transfer for data flows on defined route with adaptation to changing traffic, possible overloading or changes of TCS topology.

Simplest static setting of task for planning and optimization of data transfer routes is based on suggestion that TCS structure (number of nodes, topology and communication channel cost) is known and constant and TCS internal agent –user is played by one client, forming query to one of node network computers.

Dynamical setting of routing task by user query considers that TCS structure may change with current time but reminds known. At this network information automatically renewals in definite time intervals (Time to Live, TTL), that causes corresponding change of optimal routes.

At adaptive setting of task routing is made in uncertainty conditions, when topology of communication channels and TCS traffic may change unpredictably and accessible information has local character, monitoring and renewal of network information allows to correct routes, adapting them to changing conditions of TCS work (network overloading, failures etc.).

Necessity in dynamical and adaptive routing of data flows in global TCS appears in the following cases:

1) change of cost of TCS communication channel (for example, at their substitution);
2) failure of one or several communication channels in TCS;
3) addition to TCS new communication channels;
4) failure f one or several communication nodes of TCS;
5) addition to TCS new nodes;
6) overloading of TCS communication channels;
7) overloading of buffers of TCS nodes.

In the first case, it is necessary to renewal data about weights (costs) of TCS graph edges, and in the second and third ones – to eliminate or add corresponding edges in TCS graph. In the fourth and fifth cases, it is necessary to change data about TCS nodes by elimination or addition of corresponding nodes in TCS graph. In the sixth and seventh cases corresponding edges and nodes of TCS graph are shown as "prohibited" communication channels and nodes, playing role of unavoidable obstacles for routing and data flows transfer.

Multi-agent setting of task requires development for methods of static, dynamical and adaptive routing in conditions of multi-address transfer and collective use of TCS, when number of external agents-clients of TCS and quantity of their queries may change unpredictably with current time. In this case network overloading and collisions may appear, for their avoidance or resolution special algorithms and tools for their realization are necessary [4-6].

Suggested methods of solution for network control are based on development of adaptive, neural and multi-agent routers, using protocols of new generation (for example, IPv6-protocols).

3. Models and Methods of Adaptive and Multi-agent Routing

Necessity in adaptive routing of data flows appears at unpredictable structure changes (topology of nodes and communication channels) or parameters of global TCS and also at overloading of node buffers or TCS communication channels. Thus routers should plan and correct optimal routes of data package transfer, adapting them to possible TCS changes, appearing in real time.

For that, it is necessary to have feedback communication about current state of nodes and TCS communication channels, which may be organized with the help of monitoring and information exchange between TCS nodes.

Adaptive routing of data flows in global TCS has a series of advantages in relation with non-adaptive (static or dynamic) routing, as follows:

- provides workability and reliability of NCS at unpredictable changes of their structure or TCS parameters;
- causes to more uniform loading of nodes and TCS communication channels by "equalizing" of load;
- simplifies control for data flow transfer at network overloading;
- increases time of dependable work and TCS productivity at high level of provided service at unpredictable changes of network parameters and TCS structure.

Achievement of these advantages depends significantly on used principles, models and algorithms of adaptive routing of TCS data flows with unpredictably changing structure and beforehand unknown traffic.

Principle of adaptive routing with local feedback communication from one node is that data package is transferred into communication channel with the shortest queue or with the most probability of channel preference. At that local equalizing of load in output TCS channels may take place. However, in this case deviation from optimal route is possible. More effective principles of adaptive routing are based on transfer of local information (feedback communication) from neighbour nodes or global information from all TCS nodes. Such information may be presented by data about failures or delays in nodes or communication channels in TCS.

Models and principles of adaptive routing of data flows in global TCS may be divided on three classes [3,5-7]:

- centralized (hierarchical) routing;
- decentralized (distributed) routing;
- multi-agent routing.

Principle of centralized routing is that at first every TCS node transfers information about its state (delay or output channel capacity) to central node-router. Then this router calculates optimal route on the base of obtained global

information about current TCS state and transfer it back to all route nodes. After that controlled data package transfer from node-source to node-receiver on planned optimal route begins.

Principle of decentralized routing is based on change of local information between TCS nodes and use of this information about current state of nodes and TCS communication channels for calculation of local-optimal route. As subsequent plots of this route are calculated distributed controlled package transfer from node-source to node-receiver of TCS is realized.

Conclusion

Suggested in the paper mathematical models and optimization methods of dynamical, adaptive, neural and multi-agent (multi-address and multi-flow) routing of information flows for global TCS of new generation are important step towards creation of the theory of adaptive multi-agent (mass) service of global informational and telecommunication networks, which should change traditional theory of mass service. They may be useful for organization of adaptive multi-agent (mass) service of GRID-infrastructure of different scale and destination or for creation of new generation of scientific-educational networks.

Work has been done at partial support of the grant of RFBR № 03-01-00224a, the grant of RHSF № 03-06-12016b and the RAS Presidium Program “GRID”.

Bibliography

[1] Heleby S., Mac-Pherson D. The Routing Principles in Internet. 2-nd Edition: Translation from English. - Moscow, Publishing House “Williams”, Urbana, 2003. – 448p.

[2] Stallings V. Modern Computer Networks. – Saint-Petersburg: Piter, 2003, - 783 p.

[3] Timofeev A.V. Problems and Methods of Adaptive Control for Data Flows in Telecommunication Systems. – Informatization and Communication,№ 1, 2003,pp. 68-73.

[4] Timofeev A.V. Methods of High-Quality Control, Intellectualization and Functional Diagnostics of Automated Systems. – Mechatronics, Automation, Control, 2003, № 2, pp. 13-17.

[5] Syrtzev A.V., Timofeev A.V. Neural and Multi-Agent Routing in Telecommunicational Networks. – International Journal “Information Theories and Their Applications”, 2003 , № 2, pp 167-172.

[6] Timofeev A.V. Models for Multi-Agent Dialogue and Informational Control in Global Telecommunicational Networks. – International Journal “Information Theories and Their Applications”, 2003, № 1, pp. 54-60.

[7] Timofeev A.V. Architecture and Principles of Design of Multi-Agent Telecommunication Systems of New Generation. – Proceedings of 11-th All-Russian Scientific Methodical Conference “Telematika-2004” (Saint-Petersburg, 7-10 June 2004), vol. 1, pp. 172-174.

[8] Timofeev A.V., Ostyuchenko I.V. Multi-Agent Quality Control in Telecommunication Networks. – Proceedings of 11-th All-Russian Scientific Methodical Conference “Telematika-2004” (Saint-Petersburg, 7-10 June 2004), vol. 1, pp. 177-179.

[9] Timofeev A.V. Multi-Agent Information Processing and Adaptive Control in Global Telecommunication and Computer Networks. – International Journal “Information Theories and Their Applications”, 2003, № 10, pp. 54–60.

[10] Timofeev A.V. Intellectualization for Man-Machine Interface and Network Control in Multi-Agent Infotelecommunication Systems of New Generation. – Proceedings of 9-th International Conference “Speech and Computer”(20–22 September, 2004), Saint-Petersburg, Russia, pp. 694–700.

[11] Timofeev A.V., Syrtsev A.V., Kolotaev A.V. Network Analysis, Adaptive Control and Imitation Simulation for Multi-Agent Telecommunication Systems. – Proceedings of IFAC International Conference on Physics and Control 2005 (August 24-26, 2005, Saint-Petersburg, Russia).

[12] Timofeev Adil. Adaptive Control and Multi-Agent Interface for Infotelecommunication Systems of New Generation. – International Journal “Information Theories and Their Applications”, 2004, № 11, pp.329–336.

Author's Information

Adil Timofeev – Saint-Petersburg Institute for Informatics and Automation; P.O.Box: 199178, 39, 14-th Line; Saint-Petersburg, Russia, e-mail: tav@iias.spb.su

THE ON-BOARD OPERATIVE ADVISORY EXPERT SYSTEMS FOR ANTHROPOCENTRIC OBJECT

Boris E. Fedunov

Abstract: *A class of intelligent systems located on anthropocentric objects that provide a crew with recommendations on the anthropocentric object's rational behavior in typical situations of operation is considered. We refer to this class of intelligent systems as onboard real-time advisory expert systems. Here, we present a formal model of the object domain, procedures for obtaining knowledge about the object domain, and a semantic structure of basic functional units of the onboard real-time advisory expert systems of typical situations. The stages of the development and improvement of knowledge bases for onboard real-time advisory expert systems of typical situations that are important in practice are considered.*

Keywords*: expert systems, AI architectures, inference technique.*

Introduction

In up-to-date complex engineering of man-machine objects, the problems of setting the goals of object operation (general and current) are always solved by a crew (an operator or a group of operators governing the man-machine object). Such objects are called anthropocentric. An anthropocentric object is a functionally consistent set of onboard measuring and executive devices (OM&ED), an onboard digital computing system (ODCS) with implemented algorithms for the anthropocentric object system-forming kernel (onboard digital computer (ODC) algorithms), and a crew with algorithms of its activity (CAA) supported by the information-control field (ICF) of a crew compartment, which is the carrier of everything listed above.

The anthropocentric object operation is naturally divided into operation sessions for which the anthropocentric object and its crew is prepared in advance. As a result of this preparation, prior information and the main (general) goal of the forthcoming operation session are transmitted aboard the anthropocentric object (inserted into the ODCS algorithms and into the crew activity algorithms).

Onboard Real-Time Advisory Expert Systems of Typical Situations for an Anthropocentric Object

The external world in which a TS ORTAES operates is the information medium of an anthropocentric object (Fig. 1) formed by the output information of onboard measuring devices, the standard (not included in the TS ORTAES) onboard digital computer algorithms, and the crew compartment information-control field signals. Before starting a forthcoming operation session [Fedunov,1998], *a priori* information about this session is loaded into the TS ORTAES. This information is transmitted from the intellectual system, preparing the anthropocentric object for the operation session [Vasil'ev and others, 2000].

For every problem sub-situation significant for a TS, the ORTAES generates solution recommendations for the crew with brief explanatory notes. The recommendations and explanatory notes for them come to the crew compartment information-control field (the information part). The crew can disregard the ORTAES recommendation and resolve the problem sub-situation by other methods without reporting about this to the ORTAES. In this case, the ORTAES must develop the next recommendation taking into account the method implemented by the crew. Any disregard by the crew toward ORTAES recommendations is registered by the onboard objective control system, and when the operation session is completed, this information is submitted to the intelligent system analyzing the completion of the operation session.

The ORTAES recommendations must be in constant agreement with the activated model of the crew behavior at the conceptual and operative levels. The crew perceives the external world where the anthropocentric object operates, and the object technical state itself (the inboard world) through the sense organs and through the crew compartment information model of the outboard and inboard situation in the information-control field (the observable external world). With professional skills and prior information (the most general information about the external world and specific information about the anthropocentric object operation session), the crew (within its activated model of behavior in this TS) each time selects the current problem sub-situation (PSS) without reporting about this to the TS ORTAES.

Precisely with respect to this PSS, the TS ORTAES must immediately elaborate a reasonable and efficient recommendation for solving it.

Let us enumerate the main features of a TS ORTAES:

—it must solve all the problems of its own typical situation;

—it must have a restricted dialogue with the crew (restrictions on the time interval allowed by the external situation and on the information input through the information-control field of the anthropocentric object crew compartment performed by an operator);

—as for algorithms and rules, it must be oriented to the situational control structures;

—it must be in constant agreement with the activated conceptual model of operator behavior by generating recommendations for solving the current problem at the level of a professional operator with a significance that is sufficient for him (so as not to convey information the operator already probably knows);

—it must have a delayed self-education component.

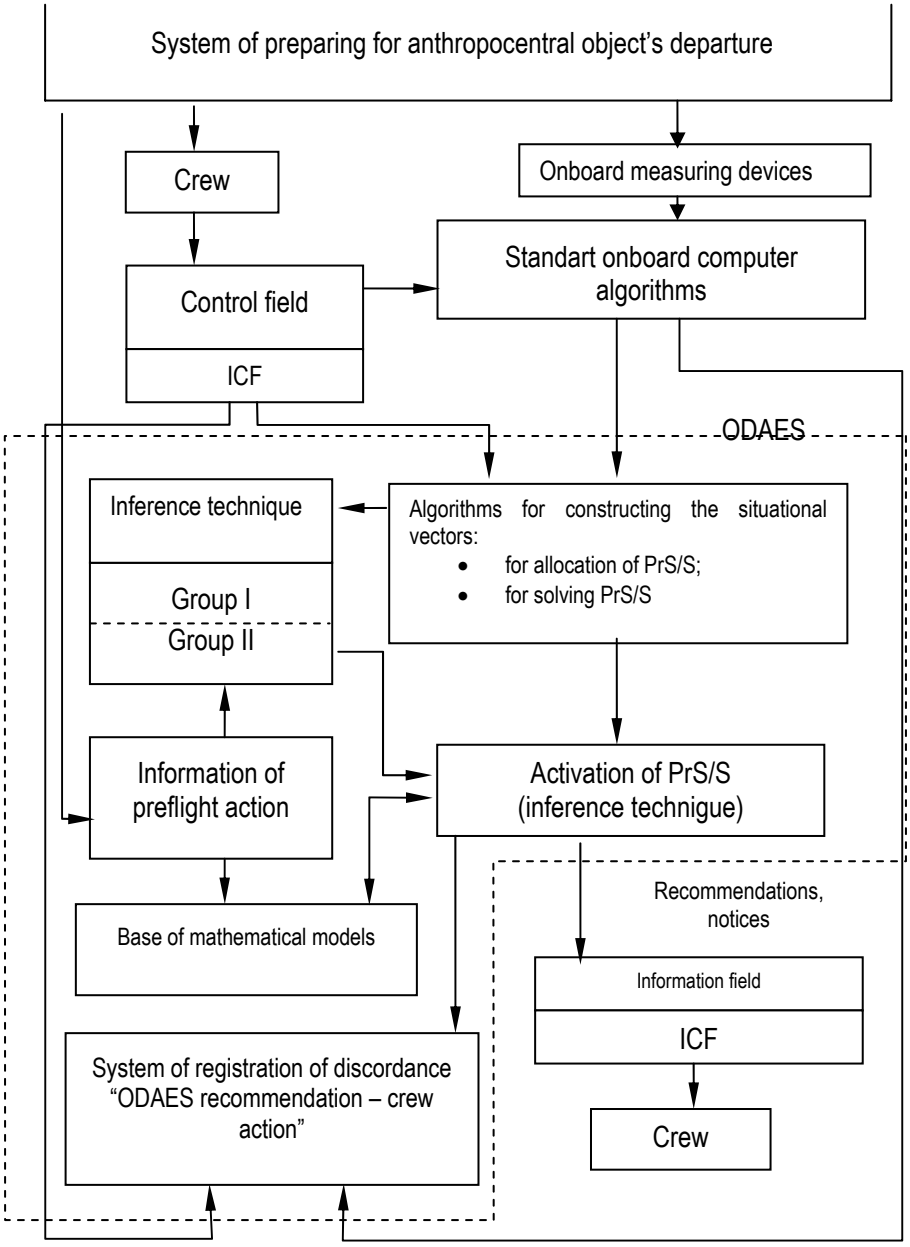


Fig. 1. The ORTAES of typical situation. Function units.

Structure of the Knowledge Base of the TS ORTAES

The TS ORTAES always operates (in the real-time mode) in cooperation with an operator [Fedunov, 1996]. It constantly produces and presents recommendations for the operator about a rational method for solving the current problem (PrS/S). In this context, the operator within the activated TS does not report to the ORTAES about the problem that has occurred and about the necessity to present recommendations for its solution to him. Furthermore, the operator can disregard (without reporting to the ORTAES) the recommendation elaborated by the ORTAES, and the ORTAES has to operate further by taking into account the operator's decision (made contrary to its recommendation).

We briefly dwell on the description of the destination and the form of the inference technique in the knowledge base of an ORTAES (Fig. 1).

Using the current information from onboard measuring devices, standard algorithms in onboard computer, signals from the information-control field (ICF) of the crew compartment, a situation vector SV(TS-PrS/S) is formed in the knowledge base of the ORTAES. This vector describes the state of the outboard and onboard environment for assigning (or identifying) current PrS/S. We call the technique of such an assignment the inference technique on the set of PrS/S. It is constructed on the basis of results of cooperating with experts who are specialists in the object domain considered. These mechanisms are implemented in the ORTAES in the form of the rules "if ..., then ..., else...". Their completeness and consistency is achieved by finalizing the ORTAES on systems of imitational modelling (SIM) [Fedunov, 1996] together with experts.

The inference technique used for (determining) a rational solution to the current PrS/S is represented by three types of mechanisms.

Inference Technique Taking into Account the Current Preferences of the Crew. Let a problem and some alternatives for its resolution be given. Suppose that, in each alternative, we are interested in some of its properties, which we will use for comparing the alternatives while choosing the most preferable one. We shall call these properties the comparison criteria. Suppose that we have several criteria which we use for comparing the alternatives. Also sup: that there is an expert (or experts) who has a suffitly definite opinion about the problem and the alternatives for solving this problem, which allows him to pairwise compare the alternatives according to each criterion.

The method of multicriteria choice of an alternative is a systematic procedure for hierarchical ordering of the elements of the problem. This method allows one to arrange the alternative according to their preferences with respect to a totality of specified comparison criteria. In order to constructively use this method in the inference technique, we present its justification based on studies by Saaty [Saaty,1991].

A Method of Pairwise Comparison of Alternatives with one Preference Criterion. Let us distinguish one of the comparison criteria and pairwise compare the alternatives with respect to this criterion.

To formalize the procedure for choosing a preferable alternative, first, we consider the simplest example, namely, the problem of choosing the heaviest object (the comparison criterion) in a given set of objects A_1,..., A_i ..., A_n (alternatives) whose absolute weights (physical) are known. The weight of the object A_1 is w_1, the weight of the object A_2 is w_2, and the weight of the object A_n is w_n. Since the object weights are known, to range them by weight, it is sufficient to arrange their weights in ascending order (to sort n numbers) and to choose the object with the greatest weight. Let us see in what ways one can choose the heaviest object.

We shall pairwise compare the weights of the objects A_i; i = 1 - n, recording the comparison results in the form of a table (matrix).

The properties of an ideal matrix of pairwise comparisons can also be used for experimental matrices.

Let a researcher have a sufficient amount of qualitative information about some instances A_1,..., A_n compared by a certain criterion. Suppose that the matrix of pairwise comparisons of order n (experimental matrix) is composed for these instances. Naturally, it differs from the ideal matrix (calculated for this case). To estimate this difference, Saaty proposes the following procedure [Saaty,1991].

Calculate the consistency index (CI) of the experimental matrix $CI = \frac{\lambda_{max} - n}{n - 1}$, where λ_{max} is the maximal eigenvalue of the experimental matrix of pairwise comparisons and n is the order of this matrix.

One can immediately see that, if the experimental matrix is ideal, then CI = 0.

In [Saaty,1991], an estimate (the random consistency index) is introduced for an arbitrary square matrix of order *n,* which is positive, inversely symmetric, and has a unit principal diagonal. In the same paper, a table of the random consistency indices (RCI) of such matrices is given. A fragment of the table from [Saaty,1991] is presented in Table 1.

Finally, Saaty proposes to calculate the consistency ratio (CR): CR = $\frac{CI}{RCI}$, where RCI is founded from table 1. For $0 \le CR \le 0.10\text{-}0.15$, he proposes to consider the experimental matrix as close to the ideal one and to use all useful properties of the latter for the experimental matrix. The main point of these properties is that, by the eigenvector of the experimental matrix corresponding to its maximal eigenvalue λ_{max}, one can judge the priorities of the instances compared according to the criterion considered.

Table 1

Order of the matrix, n	3	4	5	6	7	8	9	10
RCI	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

It is known that the determination of the eigenvalues and the corresponding eigenvectors for matrices of a high order n is a fairly complicated problem. We propose an approximate method for determining them. This method is based on the properties of the ideal matrix of pairwise comparisons.

Let us use these properties of the ideal matrix of pairwise comparisons for arranging the arbitrary alternatives (instances) A_1, ...,A_i, ...,A_n according to a certain criterion. Suppose that the matrix of pairwise comparisons of these alternatives is composed, and suppose that, checking its consistency, we find that this matrix is reasonably consistent; i.e., $0 \le CR < 0.10\text{-}0.15$ for this matrix.

Then, the following Rule 1is valid.

Rule 1**.** For a reasonably consistent experimental matrix of pairwise comparisons of the alternatives A_1, ...,A_i, ...,A_n with respect to the criterion *Kj,* the vector of priorities is determined in the an approximate method.

Rule 2. The vector of priorities can be used if the experimental matrix of pairwise comparisons is reasonably consistent.

Constructing the Matrices of Pairwise Comparisons (Experimental Matrices) in Practical Problems. We have considered the case where the "exact" weight ratios are presented in the matrix of pairwise comparisons.

However, in practical problems, it is often impossible to exactly measure the results of pairwise comparisons. One of the methods for quantitatively estimating the ratio of alternatives (for the ideal matrix, this is the ratio of the object weights) is the use of a numeral scale.

The method used most often is the Saaty 9-mark scale. If one uses the Saaty scale, then the matrix of pairwise comparisons is inversely symmetric.

It is convenient to represent the results of pairwise comparisons in the form of matrix. In the upper left corner of the matrix, we write the name of the criterion, according to which the objects are pairwise compared. In the example considered above, it is the weight. The rows and columns of the matrix correspond to the names of alternatives. The vertical ordering of the names (the first column of the matrix) and the horizontal one (the first row of the matrix) are the same. Any cell of the matrix contains the result of the pairwise comparison of the alternative in the row with the alternative in the column. This result is estimated according to the Saaty scale.

Multicriteria Choice of an Alternative. Let there be several alternatives A_1,..., A_i,.... A_n for the solution to a problem. These alternatives should be ordered according to criteria K_t,..., Kj,..., K_s.

For any criterion K_j, we estimate the weights of the alternatives A_1,..., A_i,.... A_n

$$S(Kj) = \{S_1(Kj),\ldots, Si(Kj),\ldots, Sn(Kj)\}$$

Using the method of pairwise comparisons and estimating the results by the Saaty scale, we determine the weights of the criterion significances $S = \{S_{lt} ...,Sj,...,S_s\}$ for the researcher.

Then, for any A_i, it is natural that its weight according to a criterion $S_i(K_j)$ is taken into account in the resulting weight for all criteria with the coefficient equal to the weight of this criterion's significance. The total weight (priority, rating) of the *i*-th object is determined by the formula

$$R_i = S_i(K_1)\,S_1 + \ldots + S_i(K_j)\,S_j + \ldots + S_i(K_s)\,S_s = \sum_{j=1}^{S} S_i(K_j)S_j$$

Finally, the alternatives $A_1,\ldots, A_i,\ldots. A_n$ in the problem of multicriteria choice are arranged in accordance with the total weights. The alternative with the greatest total weight is the most preferable according to the whole set of comparison criteria.

The Structure of the Inference Technique in ORTAES Constructed on the Basis of the Algorithm of Multicriteria Choice. The knowledge base of ORTAES contains a mathematical model for generating alternative versions for resolving problem sub-situations of admissible types which are fed into ORTAES at the stage of preparation for the operation session of the anthropocentric object (for piloted aircraft, when preparing for departure). The MM contains algorithms for determining the criterion values $K_j \in \{K_j\}$ for any alternative generated.

Current information characterizing the problem sub-situation and admissible types of alternatives for resolving this PrS/S is supplied at the input of the MM. On the basis of admissible types of alternatives and the existing conditions for the occurrence of the PrS/S, in the MM, a complete set of alternatives $\{A_i\}$ of admissible types is generated, and, for any alternative $A_i \in \{A_i\}$, the numerical value of each criterion $K_j \in \{K_j\}$ is calculated. Moreover, an operative correction (by the crew or onboard computer) of the values of some coordinates of the vector SV(PrS/S-solution) characterizing the PrS/S is possible.

Thus, any alternative (from the set generated by the MM) is characterized by a vector whose coordinates are the numerical values of the criteria Kj.

On the basis of these vectors, the matrices of pairwise comparisons of the alternatives are constructed for each criterion.

We separately dwell on the matrix of pairwise comparisons of criteria. When constructing this matrix, one should maximally take into account the crew preferences formed by analyzing the existing current (for the operation session) situation. Taking into account that the crew's possibility of inputting this information is unlikely, one should maximally use the transitivity property of the matrix of pairwise comparisons when constructing this matrix.

After this, the vector of total weights of alternatives is calculated by the algorithm presented in Subsection 2.1.4. An example of implementing the inference technique described is given in [Musarev and others,2001].

Inference Technique Based on Precedents. Such inference methods are used in problem subsituations, whose complexity does not allow one to constructively formalize them, but for which there is some experience (precedents) of their successful resolution.

One of difficulties of this approach is the correct choice of the coordinates $(x_1,\ldots,x_i,\ldots, x_n)$ of the situational vector SV(PrS/S-solution), both in their number and in the form of representation of each coordinate. The completeness of the description of the situational vector and the connection of a particular vector with a particular precedent is established by long-term cooperation with experts, who are actual bearers of this knowledge.

As a rule, the coordinates of the situational vector are linguistic variables.

Linguistic Variable as a Coordinate of the Situational Vector. A linguistic variable is defined by Zadeh in [Zadeh,1976] as a variable whose values belong to a specified set of terms or expressions of a natural language. The latter were also called terms.

To work with linguistic variables, one should represent each term via an appropriate fuzzy set [Zadeh,1976, Rotshtein,1999]. The latter, in turn, is represented via a universal set (universe) and the membership function of the elements of the universal set to the considered fuzzy set. The membership function takes the values in the interval [0, 1]. It quantitatively estimates the grade of membership of an element in a fuzzy set.

Note that both the universal sets and the membership functions on the set are specified on the basis of investigation results (together with experts) of the corresponding object domain.

For a large number of terms, their membership functions are usually specified in a unified form. Most often, this is a piecewise linear function.

Knowledge Matrices by Precedents. Let a state of a problem sub-situation be described by a situational vector with coordinates $(x_{1,\ldots,}\ x_i,\ ..,\ x_n)$ and each coordinate x_i be a linguistic variable with a set of terms $A_i = \left\{a_i^1,\ldots, a_i^j,\ldots, a_i^{K_i}\right\}$. For certain realizations of the situational vector, where each linguistic variable takes one of its possible values (a concrete term), there is a precedent of successful resolution of this PrS/S.

Suppose that a set d_j, $j = 1, ..., p$, of precedents is accumulated and each precedent is associated with a set of particular situational vectors, for which this precedent has been selected.

Table 2

Nos.	Coordinates of situational vector					Precedent
	x_1	…	x_i	…	x_n	
1.1 : $1\kappa_1$	a_1^{11} : $a_1^{1K_1}$	..... :	a_i^{11} : $a_i^{1K_1}$	..... :	a_n^{11} : $a_n^{1K_1}$	d_1
:	:	:	:	:	:	:
pl : $p\kappa_p$	a_1^{p1} : $a_1^{pK_p}$	..... :	a_i^{p1} : $a_{1i}^{pK_p}$	..... :	a_n^{p1} : $a_{1n}^{pK_p}$	d_p

Let us construct the matrix of this correspondence (Table 2). We select the rows of the matrix corresponding to a precedent (the block of the precedent). Any row of the matrix is a concrete situational vector for which the corresponding precedent has been successfully realized in the past.

We enumerate the rows of the block of precedent d_j, with two indices: the first index is the number of the precedent (here, it is the number of the block), and the second index is the serial number of the situational vector in this block.

This matrix determines a system of logical propositions of the form *"if..., then..., else..."* For instance, the row j_1 of the matrix encodes the following proposition:

$$\text{if } x_1 = a_1^{j1} \text{ and } x_2 = a_2^{j1} \text{ and ... and } x_i = a_i^{j1} \text{ and ... and } x_n = a_n^{j1}, \text{ then } d_j, \qquad (2.1)$$

else a similar proposition for the next row, etc. The obtained system of logical propositions ordered in this way is called a fuzzy knowledge matrix or, simply, a knowledge matrix.

Algorithm for Calculating the Membership Function of Precedent d_t. First of all, we present an algorithm [Zadeh,1976] for determining the membership function $\mu_{dj}(x_1, .. x_i, .. x_n)$ of the precedent d_j interpreted as a fuzzy set on a universal set $Ud = Ux_1 \times ... \times Ux_i \times ... \times Ux_n$, where Ux_i is a universal set on which the terms of the linguist variable x_i are defined, and U_d is the Cartesian product of the universal sets Ux_i .

Any logical proposition of the type (2.1) or, equivalently, any row of the knowledge matrix is a fuzzy relation of the corresponding fuzzy sets. For instance, for (2.1), this is $a_1^{j1} \times a_2^{j2} \times ... \times a_n^{jn}$.

In accordance with [6, 7], the membership function of a fuzzy set generated by this fuzzy relation is $\mu_{a_1^{j1}}(x_1) \wedge$ $... \wedge \mu_{a_i^{j1}}(x_i) \wedge ... \wedge \mu_{a_n^{j1}}(x_n)$, where "$\wedge$" we denote the "min" operation.

Analyzing the whole block of logical propositions related with precedent d_j (the block of the corresponding rows of the knowledge matrix), note that they form the union of the corresponding fuzzy sets generated while considering the rows of the selected block. In accordance with [6, 7], the membership function of this union, which is identified with the membership function of the precedent *dj,* is

$$\mu_{d_j}(x_1,...,x_i,...,x_n) = (\mu_{a_1^{j1}}(x_1) \wedge ... \wedge \mu_{a_i^{j1}}(x_i) \wedge ... \wedge \mu_{a_n^{j1}}(x_n)) \vee$$

$$... \vee (\mu_{a_1^{jK_j}}(x_1) \wedge ... \wedge \mu_{a_i^{jK_j}}(x_i) \wedge ... \wedge \mu_{a_n^{jK_j}}(x_n))$$

where by «$\vee$» we denote the "max" operation.

Table 3.

Nos.	Coordinates of the situational vector					min	max	d
	x_1		x_i		x_n			
:	:	:	:	:	:	:	:	:
j_1	$(a_1^{j1})^*$	.....	$(a_i^{j1})^*$	.....	$(a_n^{j1})^*$	$\min_i (a_i^{j1})^*$		
:	:	:	:	:	:	... ...		
j_s	$(a_1^{js})^*$	.....	$(a_i^{js})^*$	.....	$(a_n^{js})^*$	$\min_i (a_i^{is})^*$	$\max_{js} \min_i (a_i^{js})^*$	μ_{d_j}
:	:	:	:	:	:	... ...		
j_{K_j}	$(a_1^{jK_j})^*$	.....	$(a_i^{jK_j})^*$	.....	$(a_n^{jK_j})^*$	$\min_i (a_n^{jK_j})^*$		

Formally, this algorithm for determining the membership function of the precedent *dj* can be written in the following form:

(a). fix an arbitrary point $(x_1^*,..., x_i^*,..., x_n^*) \in U_{x_1} \times ... \times U_{x_i} \times ... \times U_{x_n}$;

(b). for any block of the knowledge matrix corresponding to d_j determine $\mu_{dj}(x_1, .. x_i, .. x_n)$ at this point according to the scheme of Table 3.

Note that, for any fixed point (x_1^*, x_i^*,..., x_n^*), the block of the matrix presented in Table 3 is numerical, because each term a_i^{js} from this block is replaced with the value of its membership function $(a_i^{js})^*$ calculated at the corresponding x_i^*. The operation $\min_i a_i^{js}$ is performed with the numbers located in rows "i" $1 \le i \le n$, and the minimal number in the corresponding row is placed in the column "min." The operation $\max_{jS} \min_i a_i^{js}$ selects the greatest of the row minima obtained for $1 \le j_s \le K_j$. This number is the value of the membership function $\mu_{dj}(x_1... x_i,...,x_n)$ at the fixed point $(x_1^*,...,x_i^*,...,x_n^*)$. Performing this calculation for every point of the universal set, we obtain the membership functions that interest us.

Algorithm for Choosing a Precedent when Observing a Situational Vector with Quantitative Coordinates. When observing a situational vector [Rotshtein,1999] with quantitative coordinates (all coordinates of the vector are measured by numerical scales), in order to select the most preferable precedent, it is not necessary to completely determine the membership functions $\mu_{dj}(x_1, .. x_i, .. x_n)$ on the whole set of points of the universal set. It is sufficient to calculate their values only for fixed numerical values of the coordinates of the vector, which is obtained by us as a result of the observation. For this purpose, we should use the algorithm from Subsection 3.3 once, taking the coordinates of the observed situational vector as $(x_1^*,..., x_i^*,...,x_n^*)$.

As a result, for any precedent *dj,* we obtain a number $d_j(x_1^*,..., x_i^*,..., x_n^*)$, which is the grade of membership of d_j to the point $(x_1^*,..., x_i^*,...,x_n^*)$.

Starting from this interpretation, the most preferable precedent for resolving the observed PrS/S is the precedent d_j^* such that

$$d_j^*(x_1^*,..., x_i^*,..., x_n^*) = \max_{1 \le j \le p} d_j(x_1^*,..., x_i^*,..., x_n^*).$$

Three types of inference techniques are presented. They are constructive for resolving problem substitutions of typical situations of anthropocentric object functioning (for instance, flights of piloted aircraft).

The first type of inference technique (on product rules) is based on results of the mathematical investigation of the PrS/S and, to a certain degree, on the knowledge of experts in the object domain under consideration. This technique is widely used by designers of the knowledge bases of the first Russian and foreign ORTAES.

The second type of inference technique is based on the use of the algorithm for multicriteria choice of an alternative. The technique is directed to the substantial use of the results of the preparation for the operation session of the anthropocentric object (prior information) and crew preferences formed while operatively analyzing the situation existing in the current operation session. The considered example of the use of this technique allows one to hope that it can be successfully employed in practice when designing the real knowledge bases of ORTAES.

The third type of inference technique has been studied only theoretically and has not been tested on practical examples. It is directed to be used in the inference technique of precedents and is based on the algorithms for choosing a solution on the basis of the knowledge matrix. These algorithms are successfully applied in diagnostic problems.

The choice of the type of inference technique for constructing in ORTAES a recommendation for resolving a problem sub-situation of a concrete type depends on its complexity and on the possibility of adequately formalizing it mathematically.

Conclusion

(1) The TS ORTAES of anthropocentric objects is a class of dynamic intelligent systems presenting, in the real-time mode, to the crew of an anthropocentric object the complete recommendations on its rational behavior under current operation conditions.

The TS ORTAES has the following specific features:

- working in the real-time mode of operation of the anthropocentric object that carries the TS ORTAES with the restricted dialogue "TS ORTAES-crew" under continuous coordination of ORTAES recommendations and activated (at the current instant) conceptual model of the crew behavior;
- solving all problems of the TS for which the TS ORTAES is developed;
- the procedure of the recommendation inference in the TS ORTAES is based on the principles of situational control;
- the delayed component of the TS ORTAES self-education.

(2) The semantic structure of the TS ORTAES knowledge base consists of the following:

- a two-level (with respect to semantics) rule base identifying, at the upper hierarchy level, the problem sub-situation (PSS) that occurred and determining methods for its solution at the lower hierarchy level;
- mathematical model (MM) bases consisting of MMs of tree types (the MMs predicting space-time occurrence of significant events, the MMs generating admissible solutions and their ranking, and the MMs for estimating unobservable phase coordinates;
- three types of inference techniques are presented. They are constructive for resolving problem sub situations of typical situations of anthropocentric object functioning.

Bibliography

[Fedunov,1998] Fedunov B.E. Constructive Semantics of Anthropocentric Systems for Development and Analysis of Specifications for Onboard Intelligent Systems, Izv. Ross. Akad. Nauk, Tear. Sist. Upr (Simultaneous English language translation of the journal is available from MAIK “Nauka/Interperiodica” (Russia). Journal of Computer and Systems Sciences International ISSN 1064-2307), 1998, №. 5.

[Fedunov,1996] Fedunov B.E., Problems of the Development of Onboard Operative and Advising Expert Systems for Anthropocentric Objects, Izv. Ross. Akad. Nauk, Tear. Sist. Upr., (Simultaneous English language translation of the journal is available from MAIK “Nauka/Interperiodica” (Russia). Journal of Computer and Systems Sciences International ISSN 1064-2307),1996, № 5.

[Vasil’ev and others,2000] Vasil’ev S.N., Gherlov A.K., Fedosov E.A., Fedunov B.E. Intelligent control of dynamic systems. Moscow, Fizmatlit, 2000, (by Russian).

[Musarev and others,2001] Musarev L.M. and Fedunov B.E., The Structure of Target Allocation Algorithms on Board the Commander of an Aircraft Group, *Izv.* Ross.Akad. Nauk, Teor. Sist. Upr. (Simultaneous English language translation of the journal is available from MAIK “Nauka/Interperiodica” (Russia). Journal of Computer and Systems Sciences International ISSN 1064-2307),,2001, № 6.

[Saaty,1991] Saaty T. The Analytic Hierarchy Process. Planning ,Priority Setting, Resource Allocation, New York: McGraw-Hill, 1980. Translated under the title Analiticheskoe planirovanie. Organizatsiya sistem, *M*ocow:Radio i Svyaz', 1991.

[Zadeh,1976] Zadeh L. The Concept of a Linguistic Variable and Its Application to Approximate Reasonong, New York: Elsevier, 1973. Translated under the title Ponyatie lingvisticheskoi peremennoi i ee primenenie k prinyatiyupriblizhennykh reshenii, Moscow: Mir, 1976.

Rotshtein,1999] Rotshtein A.P. Intellektual'nye tekhnologii identfikatsii (Intelligent Technologies of Verification), Vinnitsa: Universum, 1999.

Author's Information

Boris Evgenjevich Fedunov – State Scientific Research Institute for Aviation System (GosNIIAS), ul. Viktorenko 7, Moscow, 125319 Russia; e-mail boris_fed@gosniias.ru

ОПТИМИЗАЦИЯ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ С ТЕХНОЛОГИЕЙ ATM

Леонид Л. Гуляницкий, Андрей А. Баклан

***Abstract**: рассматриваются проблемы оптимизации сетей с технологией ATM, которые формулируются как задачи комбинаторной оптимизации. Для решения возникающих задач разработан ряд приближенных алгоритмов. Приведены результаты их сравнения при решении серии задач со случайными данными.*

***Keywords**: network, optimization,local search, simulated annealing, genetic algorithm*

О проблемах оптимизации сетей

Телекоммуникации и информационные технологии играют сегодня фундаментальную роль в развитии общества и экономики. Требования к телекоммуникационным сетям постоянно возрастают, в результате чего возникла концепция широкополосных цифровых сетей с комплексными услугами B-ISDN. Развитие этой концепции привело к возникновению новой сетевой технологии, универсальный характер которой делает ее чрезвычайно привлекательной, – это технология асинхронного режима передачи ATM (от англ. Asynchronous Transfer Mode – асинхронный режим передачи).

Возникновение и распространение этой технологии, а также ее широкие возможности делают актуальной разработку методов решения задач оптимизации сетей с технологией ATM. Одной из самых важных таких задач является задача оптимального выбора пропускных способностей каналов связи для разных видов трафика, для которой на сегодняшний день не существует эффективных подходов к решению.

Технология ATM – это высокоскоростная технология широкополосной передачи данных, которая основывается на технологиях коммутации пакетов (так называемых ячеек) и мультиплексирования и используется для передачи комплексной, разнородной информации, включающей данные, аудио- и видеоинформацию [1].

Основным требованием к сетям связи является требование увеличения пропускной способности и снижения ее стоимости. Доступность высокопроизводительных и относительно недорогих персональных компьютеров, рабочих станций, коммерческого программного обеспечения, распространение распределенного подхода к вычислениям – все это требует большей пропускной способности глобальных и региональных сетей по более низкой стоимости. Возникла острая необходимость в удобных для использования и приемлемых по цене широкополосных сервисах, предоставляемых по требованию.

Во время создания соединения ATM может требоваться определенная пропускная способность на всем пути между двумя конечными узлами этого соединения, а также устанавливаться требования к показателям качества сервиса. Это гарантирует, что приложение, которое делает запрос, будет способно выполнять свои функции во время соединения.

Основными показателями качества сервиса является:

– средняя задержка в передаче ячеек;
– вариация средней задержки;
– доля (или вероятность) потерянных ячеек;
– максимальный размер информационного пакета;
– максимальная скорость передачи;
– обеспеченная скорость передачи.

Для обеспечения качества сервиса технология ATM имеет несколько категорий сервиса (видов трафика):

а) CBR – передача с постоянной скоростью;
б) VBR – передача с переменной скоростью, которая может происходить как в реальном времени, так и нет;
в) ABR – передача с доступной скоростью;
г) UBR – передача с неустановленной скоростью.

Все эти категории сервиса были введены с целью достижения возможности передачи разнородного трафика, адекватного управления распределением ресурсов сети для каждой компоненты трафика, более высокой степени гибкости и использования сети. Введение категорий сервиса увеличивает преимущества ATM, делая эту технологию пригодной фактически для неограниченного спектра применений. Категории сервиса ATM делают возможным для пользователей выбирать специфические комбинации параметров трафика и производительности.

При проектировании новых и анализе существующих телекоммуникационных сетей возникают следующие задачи:

– задача оптимального выбора пропускных способностей существующих каналов связи;
– задача оптимального выбора маршрутов передачи и оптимального распределение потоков;
– комбинированная задача оптимального выбора пропускных способностей каналов связи, маршрутов передачи и распределения потоков;
– задача анализа показателей живучести;
– задача синтеза структуры сети.

Одной из самых важных задач является задача оптимального выбора пропускных способностей каналов связи, которую на содержательном уровне можно сформулировать следующим образом.

Задана структура сети, состоящая из коммутаторов, соединенных каналами связи. Для каждого канала связи определена его длина. Заданы также требования в передаче трафика для каждой упорядоченной пары коммутаторов. Кроме того, для каждого канала связи заданы величины общих потоков. Пропускная способность каждого канала связи пропорциональна пропускной способности базового канала. Известны также удельные стоимости каналов связи разной пропускной способности на единицу длины. Необходимо для всех каналов связи выбрать такие количества базовых каналов, при которых стоимость сети будет минимальной при выполнении ограничений на показатели качества обслуживания.

При решении задачи оптимального выбора пропускных способностей каналов связи сети ATM необходимо учитывать как разные виды трафика, так и разные показатели качества сервиса для них. Технология ATM определяет четыре вида трафика: CBR, VBR, ABR и UBR, причем для каждого устанавливаются соответствующие показатели качества сервиса. Учитывая очень высокие требования к показателям качества для трафика CBR, ему выделяется постоянная часть полосы пропускания в каждом канале связи. Если величина трафика CBR снижается, то часть полосы CBR, которая освободилась, не используется. Для трафиков VBR и ABR выделяется общая часть полосы пропускания, которая распределяется следующим образом: трафик VBR занимает большую часть полосы и обслуживается в коммутаторах с высшим относительным приоритетом по дисциплине FIFO, а если в очереди коммутатора нет ячеек VBR, то передаются ячейки трафика ABR. И, наконец, оставшаяся часть полосы канала используется для передачи трафика UBR, причем для него параметры качества сервиса не устанавливаются [1].

Учитывая сказанное, для сетей ATM содержательная постановка задачи имеет следующий вид. Задана структура сети ATM, состоящая из коммутаторов, соединенных каналами связи. Для каждого канала связи определена его длина. Заданы также требования в передаче трафиков CBR, VBR и ABR для каждой упорядоченной пары коммутаторов. Кроме того, для каждого канала связи заданы величины общих потоков трафиков CBR, VBR и ABR. Пропускная способность каждого канала связи пропорциональна пропускной способности базового канала. Известны также удельные стоимости каналов связи разной пропускной способности на единицу длины. Необходимо для всех каналов связи выбрать такие количества базовых каналов для трафика CBR, общего трафика VBR и ABR, а также часть общего трафика, занятую трафиком VBR, при которых стоимость сети будет минимальной. При этом должны выполняться следующие ограничения на показатели качества обслуживания: CLR и CTD для трафика CBR, CLR и CTD для трафика VBR, CTD для трафика ABR.

Задача оптимального выбора пропускных способностей каналов связи сети ATM является новой задачей дискретной оптимизации, для которой не существует эффективных методов решения. К этому времени предложен лишь один подход к решению задачи оптимального выбора пропускных способностей каналов связи – метод последовательного анализа и отсева вариантов [2]. Однако он не позволяет проводить улучшения найденных решений (находить несколько решений), а также имеет существенные

вычислительные сложности с ростом размерности задачи. Поэтому целесообразной является разработка приближенных алгоритмов решения.

Постановка задачи оптимального выбора пропускных способностей каналов связи для трафика CBR и трафиков VBR и ABR рассматривалась в работе [2]. Далее приведена модифицированная постановка задачи оптимального выбора пропускных способностей каналов связи в виде модели дискретной оптимизации.

Постановка задачи

Поскольку в связи со спецификой трафика CBR для него полоса пропускной способности в каждом канале связи выделяется постоянно, независимо от распределения потоков, то это позволяет решать задачу выбора пропускных способностей для трафика CBR независимо от потоков VBR и ABR, то есть выделить эту задачу в отдельную подзадачу.

Поэтому будем рассматривать две подзадачи: оптимальный выбор пропускных способностей каналов связи для трафика CBR и оптимальный выбор пропускных способностей каналов связи для трафиков VBR и ABR.

1) Оптимальный выбор пропускных способностей каналов связи для трафика CBR.

Задана структура сети в виде графа $G = (V, E)$, где $V = \{v_1, v_2, \ldots, v_n\}$ – множество узлов сети (коммутаторов), $E = \{e_1, e_2, \ldots, e_m\}$ – множество дуг (каналов связи). Каждый канал связи e_k имеет свою длину l_k, $k = \overline{1, m}$. Задана также матрица требований в передаче трафика CBR $H^{(0)} = \left\| h_{ij}^{(0)} \right\|_{i,j=\overline{1,n}}$, где $h_{ij}^{(0)}$ – интенсивность потока (Мбит/с), который необходимо передать из узла i в узел j. Кроме того, задан вектор потоков для трафика CBR $f^{(0)} = \left(f_1^{(0)} \quad f_2^{(0)} \quad \ldots \quad f_m^{(0)} \right)^T$, где $f_k^{(0)}$ – величина общего потока CBR в канале связи e_k, $k = \overline{1, m}$. Имеем прямую зависимость между пропускной способностью канала связи e_k и базовой пропускной способностью цифрового канала μ (Мбит/с): $\mu_k^{(0)} = x_k^{(0)} \mu$, где $x_k^{(0)} \in \{1,2,..., N\}$ – количество базовых каналов, $k = \overline{1, m}$.

Известны также удельные стоимости каналов связи разной пропускной способности на единицу длины $C = \{c_1, c_2, ..., c_N\}$. То есть стоимость канала связи e_k с количеством базовых каналов $x_k^{(0)}$ и длиной l_k равна $c_{x_k^{(0)}} l_k$, $k = \overline{1, m}$.

Необходимо выбрать такие количества базовых цифровых каналов для каждого канала связи $x^{(0)} = \left(x_1^{(0)} \quad x_2^{(0)} \quad \ldots \quad x_m^{(0)} \right)^T$, при которых стоимость сети CBR будет минимальной:

$$\sum_{k=1}^{m} c_{x_k^{(0)}} l_k \to \min \tag{1}$$

при выполнении ограничений на показатели качества обслуживания трафика CBR:

$$CLR^{(0)} \le CLR_{зад}^{(0)}, \tag{2}$$

$$CTD^{(0)} \le CTD_{зад}^{(0)}, \tag{3}$$

и условий

$$f_k^{(0)} < x_k^{(0)} \mu, \; k = \overline{1, m}, \tag{4}$$

$$x_k^{(0)} \in \{1,2,..., N\}, \; k = \overline{1, m}, \tag{5}$$

где $CLR^{(0)}$ – средняя вероятность потери (уровень потерь) ячеек CBR; $CTD^{(0)}$ – средняя задержка ячеек CBR; $CLR_{зад}^{(0)}$ – заданный уровень потерь ячеек CBR; $CTD_{зад}^{(0)}$ – заданная средняя задержка ячеек CBR в сети.

Аналитические выражения для $CTD^{(0)}$ и $CLR^{(0)}$, полученные на основе аппарата систем и сетей массового обслуживания в работе [2], в наших обозначениях имеют вид:

$$CTD^{(0)} = \frac{1}{H_{\Sigma}^{(0)}} \sum_{k=1}^{m} \frac{f_k^{(0)}}{x_k^{(0)}\mu - f_k^{(0)}}, \tag{6}$$

$$CLR^{(0)} = \frac{1}{m} \sum_{k=1}^{m} CLR_k^{(0)}, \tag{7}$$

$$CLR_k^{(0)} = P_0 \left(\frac{f_k^{(0)}}{\mu} \right)^{x_k^{(0)}} \frac{1}{x_k^{(0)}!} \left(\frac{f_k^{(0)}}{x_k^{(0)}\mu} \right)^{N_k^{(0)}}, \tag{8}$$

$$P_0 = \left[\sum_{t=0}^{x_k^{(0)}} \left(\frac{f_k^{(0)}}{\mu} \right)^t \frac{1}{t!} + \left(\frac{f_k^{(0)}}{\mu} \right)^{x_k^{(0)}} \frac{1}{x_k^{(0)}!} \sum_{t=1}^{N_k^{(0)}} \left(\frac{f_k^{(0)}}{x_k^{(0)}\mu} \right)^t \right]^{-1}, \tag{9}$$

где $H_{\Sigma}^{(0)} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{(0)}$ – общий размер внешнего потока типа CBR; $N_k^{(0)}$ – размер буфера коммутатора ATM для ячеек трафика CBR.

2) Оптимальный выбор пропускных способностей каналов связи для трафиков VBR и ABR.

Задана структура сети в виде графа $G = (V, E)$, где $V = \{v_1, v_2, \ldots, v_n\}$ – множество узлов сети (коммутаторов); $E = \{e_1, e_2, \ldots, e_m\}$ – множество дуг (каналов связи). Каждый канал связи e_k имеет свою длину l_k, $k = \overline{1,m}$. Заданы также матрицы требований в передаче трафика VBR $H^{(1)} = \left\| h_{ij}^{(1)} \right\|_{i,j=\overline{1,n}}$ и трафика ABR $H^{(2)} = \left\| h_{ij}^{(2)} \right\|_{i,j=\overline{1,n}}$, где $h_{ij}^{(1)}$ и $h_{ij}^{(2)}$ – интенсивности потоков (Мбит/с) VBR и ABR соответственно, которые необходимо передать из узла i в узел j. Кроме того, заданы векторы потоков для трафика VBR $f^{(1)} = \left(f_1^{(1)} \quad f_2^{(1)} \quad \ldots \quad f_m^{(1)} \right)^T$ и трафика ABR $f^{(2)} = \left(f_1^{(2)} \quad f_2^{(2)} \quad \ldots \quad f_m^{(2)} \right)^T$, где $f_k^{(1)}$ и $f_k^{(2)}$ – величины общего потока VBR и ABR в канале связи e_k соответственно, $k = \overline{1,m}$. Базовая пропускная способность канала связи e_k рассчитывается по формуле: $\mu_k = x_k \mu$, где $x_k \in \{1,2,...,N\}$ – количество базовых цифровых каналов, $k = \overline{1,m}$.

Известны также удельные стоимости каналов связи разной пропускной способности на единицу длины $C = \{c_1, c_2, ..., c_N\}$, тогда стоимость канала связи e_k с количеством базовых каналов x_k и длиной l_k равна $c_{x_k} l_k$, $k = \overline{1,m}$.

Необходимо для каждого канала связи выбрать такие количества базовых цифровых каналов, выделенных под трафик VBR и ABR $x = \left(x_1 \quad x_2 \quad \ldots \quad x_m \right)^T$, а также их часть, занятую трафиком VBR $x^{(1)} = \left(x_1^{(1)} \quad x_2^{(1)} \quad \ldots \quad x_m^{(1)} \right)^T$, при которых стоимость сети будет минимальной

$$\sum_{k=1}^{m} c_{x_k} l_k \to \min \tag{10}$$

при выполнении ограничений на показатели качества обслуживания трафиков VBR и ABR:

$$CLR^{(1)} \le CLR^{(1)}_{зад}, \quad (11)$$

$$CTD^{(1)} \le CTD^{(1)}_{зад}, \quad (12)$$

$$CTD^{(2)} \le CTD^{(2)}_{зад}, \quad (13)$$

и условий:

$$x_k^{(1)} < x_k,\ k = \overline{1,m}, \quad (14)$$

$$f_k^{(1)} + f_k^{(2)} < x_k \mu,\ k = \overline{1,m}, \quad (15)$$

$$f_k^{(1)} < x_k^{(1)} \mu,\ k = \overline{1,m}, \quad (16)$$

$$x_k \in \{1,2,...,N\},\ k = \overline{1,m}, \quad (17)$$

$$x_k^{(1)} \in \{1,2,...,N\},\ k = \overline{1,m}, \quad (18)$$

где $CLR^{(1)}$ – средняя вероятность потери (уровень потерь) ячеек VBR; $CTD^{(1)}$ и $CTD^{(2)}$ – средние задержки ячеек VBR и ABR соответственно; $CLR^{(1)}_{зад}$ – заданный уровень потерь ячеек VBR; $CTD^{(1)}_{зад}$ и $CTD^{(2)}_{зад}$ – заданные средние задержки в сети ячеек VBR и ABR соответственно.

Основываясь на [2], можем записать, что:

$$CTD^{(1)} = \frac{1}{H_\Sigma^{(1)}} \sum_{k=1}^{m} \frac{f_k^{(1)} \left(f_k^{(1)} + f_k^{(2)} \right)}{x_k \mu \left(x_k \mu - f_k^{(1)} \right)}, \quad (19)$$

$$CTD^{(2)} = \frac{1}{H_\Sigma^{(2)}} \sum_{k=1}^{m} \frac{f_k^{(2)} \left(f_k^{(1)} + f_k^{(2)} \right)}{\left(x_k \mu - f_k^{(1)} \right) \left(x_k \mu - f_k^{(1)} - f_k^{(2)} \right)}, \quad (20)$$

$$CLR^{(1)} = \frac{1}{m} \sum_{k=1}^{m} CLR_k^{(1)}, \quad (21)$$

$$CLR_k^{(1)} = P_0 \left(\frac{f_k^{(1)}}{\mu} \right)^{x_k^{(1)}} \frac{1}{x_k^{(1)}!} \left(\frac{f_k^{(1)}}{x_k^{(1)} \mu} \right)^{N_k^{(1)}}, \quad (22)$$

$$P_0 = \left[\sum_{t=0}^{x_k^{(1)}} \left(\frac{f_k^{(1)}}{\mu} \right)^t \frac{1}{t!} + \left(\frac{f_k^{(1)}}{\mu} \right)^{x_k^{(1)}} \frac{1}{x_k^{(1)}!} \sum_{t=1}^{N_k^{(1)}} \left(\frac{f_k^{(1)}}{x_k^{(1)} \mu} \right)^t \right]^{-1}, \quad (23)$$

где $H_\Sigma^{(1)} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{(1)}$ и $H_\Sigma^{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{(2)}$ – общие размеры внешних потоков типов VBR и ABR соответственно; $N_k^{(1)}$ – размер буфера коммутатора ATM для ячеек трафика VBR.

Результаты вычислительного эксперимента

С целью проведения экспериментальных исследований при решении задачи оптимального выбора пропускных способностей каналов связи был разработан программный комплекс, реализующий алгоритмы локального поиска (метода вектора спада – МВС [3]), повторяющегося локального поиска (ПЛП) [4], имитационного отжига (АИО) [5], *G*-алгоритма [6] и генетический алгоритм (ГА) [7]. Этот комплекс также позволяет производить генерацию тестовых задач со случайными данными.

На первом этапе эмпирическим путем были выбраны следующие значения параметров алгоритмов. Для алгоритма ПЛП: максимальное количество итераций $t_{\max}$=30, максимальное количество переходов $h_{\max}$=20. Для АИО: максимальное количество итераций $t_{\max}$=20; максимальное количество переходов $h_{\max}$=10000; начальное значение температуры T_0=50; коэффициент для температурного расписания

r =0,925; число для определения условия равновесия ε =0,01; количество прогонов k =2; количество переходов в прогоне v =35. Для G-алгоритма: максимальное количество итераций $t_{\max}$ =40; число для определения условия равновесия ε =10; количество прогонов k =2; количество переходов в прогоне v =35; начальное значение параметра μ_0 =0; параметр γ =0,05. Для ГА: максимальное количество поколений $t_{\max}$ =500; количество особей в начальной популяции *K*=*20*; количество отобранных для скрещивания особей Q =10; вероятность унаследования гена родителя P_c =0,1; вероятность изменения гена при мутации P_m =0,5. Радиус окрестностей в алгоритмах МВС, ПЛП и АИО выбирался равным 1.

На втором этапе проводились численные эксперименты с реализованными алгоритмами. Для этого был сформирован контрольный набор из 80 задач разной размерности, при этом для каждой размерности набор содержал 5 задач. Отметим, что размерность задачи определяется количеством ребер *m*.

Для каждой задачи из контрольного набора было проведено по пять запусков каждого алгоритма, в результате которых были найдены общее время их выполнения *t* (мс), наилучшее значение целевой функции f_* и величина улучшения *q* (%), которая определяется по формуле $q = (f_0 - f_*)/f_0 \cdot 100\%$, где f_0 – значение целевой функции для начального варианта решения задачи.

Усредненные результаты применения разработанных алгоритмов (времени вычисления и величины улучшения) для решения двух названных типов задач поданы в табл. 1 для трафика CBR – задача (1)-(9), а в табл. 2 для трафиков VBR и ABR – задача (10)-(23).

Разработка программного комплекса проводилась на языке программирования Object Pascal в интегрированной среде Borland Delphi 7.0. Экспериментальные исследования проводились на персональном компьютере с центральным процессором AMD Athlon XP 1700+ , тактовой частотой 1,47 ГГц и оперативной памятью DDR 333 Мгц объемом 256 Мбайт под управлением операционной системы Windows XP Professional.

Анализ результатов экспериментов показывает, что для задач с трафиком CBR наименьшее улучшение дает алгоритм локального поиска, другие находят примерно одинаковые решения, при этом лучшим является *G*-алгоритм. Для задач с трафиками VBR и ABR наилучшие решения находят алгоритм повторяющегося локального поиска и *G*-алгоритм, причем эти решения почти одинаковы.

Таблица 1 – Усредненные результаты применения разработанных алгоритмов для трафика CBR

m	МВС		ПЛП		АИО		G-алгоритм		ГА	
	q, %	t, мс	q, %	t, мс	q, %	t, мс	q, %	t, мс	q, %	t, мс
15	31,54	212,20	43,29	1540,20	45,80	12498,00	45,08	8860,60	42,62	4917,20
17	28,54	258,20	44,11	2069,20	46,55	13032,60	45,12	9756,00	43,32	5596,00
20	36,48	450,80	46,15	2889,80	50,00	17735,80	48,89	13120,80	46,07	6609,60
21	15,15	220,00	26,45	2665,80	32,23	14354,60	30,62	11011,80	28,82	6786,00
23	40,74	625,00	53,36	3759,20	57,91	19776,40	56,72	16207,20	53,08	7767,40
25	13,69	308,40	24,40	3623,20	32,65	15636,40	31,09	13481,40	27,21	8217,80
27	31,19	709,00	50,16	5542,20	56,99	24939,60	54,60	18220,20	51,78	8123,60
29	54,04	1382,20	65,86	6673,60	72,37	25462,40	70,14	20828,00	65,28	8632,60
30	28,62	817,00	42,61	6303,20	48,94	28242,80	46,93	20451,40	44,65	8908,80
32	6,83	362,60	27,03	6541,40	40,30	17787,60	38,53	18793,00	36,18	10002,60
33	31,65	1138,00	44,58	8101,60	52,77	31260,80	50,77	23273,40	46,78	9770,20
35	13,24	1019,80	27,83	8884,40	36,74	28689,20	34,62	20770,00	33,28	10152,60
37	40,92	1798,20	56,31	10307,00	64,93	34583,60	62,80	28122,20	60,08	10854,00
40	5,77	619,00	19,61	10086,80	30,52	24569,00	27,44	20215,20	26,03	11813,00
45	10,88	1120,00	29,03	14124,20	44,62	36861,00	41,67	29396,20	38,99	12818,60
50	8,62	1091,60	23,04	15654,60	41,36	36672,40	37,69	31221,40	35,81	14803,00

Таблица 2 – Усредненные результаты применения разработанных алгоритмов для трафиков VBR и ABR

M	МВС		ПЛП		АИО		G-алгоритм		ГА	
	q, %	t, мс	q, %	t, мс	q, %	t, мс	q, %	t, мс	q, %	t, мс
15	75,68	412,60	76,02	2251,20	75,88	9151,20	76,04	5874,60	75,93	4099,80
17	68,82	478,80	71,96	2457,80	72,22	11201,80	71,99	7410,60	71,86	4819,00
20	76,78	787,20	80,18	5013,00	80,28	12269,80	80,05	9711,80	79,14	5834,40
21	67,98	799,00	69,37	4474,80	68,82	13934,00	69,17	8910,80	68,44	5407,60
23	80,57	1021,40	82,93	4867,00	82,95	15246,00	83,13	10126,40	80,95	6060,80
25	59,20	983,40	68,92	8231,80	69,92	16930,40	68,90	12808,60	68,37	6291,00
27	76,12	1434,20	76,35	7036,00	75,42	19414,00	76,39	12283,80	73,61	7514,60
29	84,63	1732,40	84,48	6982,00	83,63	21020,20	84,64	15260,00	80,44	7885,20
30	73,95	1974,00	74,31	9041,20	73,31	22744,60	73,66	15125,80	70,96	8107,60
32	68,62	1946,60	75,68	12577,80	75,38	22801,00	75,53	17947,80	71,79	8530,40
33	74,65	2157,20	75,50	13861,60	74,27	25254,40	75,33	17607,40	71,04	8766,60
35	61,85	2303,40	66,22	14619,20	63,83	27962,20	66,10	17360,60	61,60	9077,20
37	81,30	2794,00	81,48	14088,20	79,92	26670,20	81,50	20229,40	75,54	9882,20
40	65,13	3052,20	66,33	17221,00	63,67	29652,60	65,84	20607,60	60,25	10288,60
45	69,07	3943,40	70,36	20485,80	67,34	34727,60	70,31	24819,80	61,97	11516,60
50	67,47	4708,80	72,81	27754,00	70,02	42056,20	72,59	28853,80	62,42	12770,20

Список литературы

1. Назаров А.Н., Симонов М.В. АТМ: технология высокоскоростных сетей. М.: ЭКО-ТРЕНДЗ, 1997.
2. Зайченко О.Ю. Оптимізація характеристик мереж з технологією АТМ // Системні дослідження та інформаційні технології. – 2002. – №3. – С.57– 73.
3. Сергиенко И.В. Математические модели и методы решения задач дискретной оптимизации. К.: Наук. думка, 1988.
4. Lourenço H. R., Martin O., Stützle T. Iterated local search // Handbook of Metaheuristics: International Series in Operations Research & Management Science, vol. 57 (Eds. F. Glover and G. Kochenberger). – Norwell: Kluwer Academic Publishers, MA, 2002. – P. 321-353.
5. Aarts E., Korst J. Simulated annealing and Boltzmann machines. – New York: John Willey and Sons, 1989.
6. Гуляницкий Л.Ф. Решение задач комбинаторной оптимизации алгоритмами ускоренного вероятностного моделирования // Компьютерная математика. – Киев: Институт кибернетики им. В.М. Глушкова НАН Украины, 2004. – №1. – С. 64-72.
7. Goldberg D.E. Genetic Algorithms in Search, Optimization and Machine Learning. – Boston: Addison-Wesley Longman Publishing Co., MA, 1989.

Информация об авторах

Леонид Гуляницкий – Институт кибернетики им. В.М.Глушкова НАН Украины, пр-т Глушкова, 40, Киев, 03680, Украина. e-mail: lh_dar@hotmail.com

Андрей Баклан – Институт кибернетики им. В.М.Глушкова НАН Украины, пр-т Глушкова, 40, Киев, 03680, Украина. e-mail: ascorpio@ukr.net

TESTING AI IN ONE ARTIFICIAL WORLD

Dimiter Dobrev

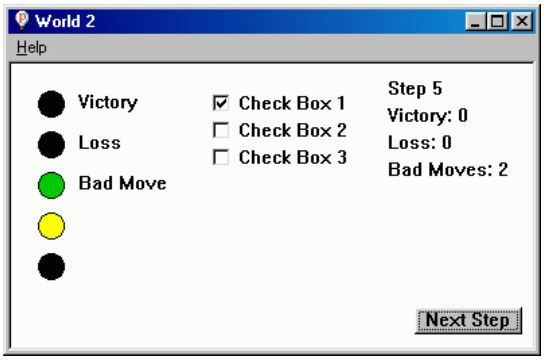
Abstract: *In order to build AI we have to create a program which copes well in an arbitrary world. In this paper we will restrict our attention on one concrete world, which represents the game Tick-Tack-Toe. This world is a very simple one but it is sufficiently complicated for our task because most people cannot manage with it. The main difficulty in this world is that the player cannot see the entire internal state of the world so he has to build a model in order to understand the world.*

Keywords: *AI Definition, Artificial Intelligence, Artificial World, Machine Learning.*

The World of the Tick-Tack-Toe Game

In this paper we will observe a concrete artificial world. You can find this world and test it in the examples of the compiler **Strawberry Prolog** [1]. You have to start the example program **World 2.spj**. After that you will see one panel with five lamps, three checkboxes and one button. You can observe this program as a test for intelligence. Really, it is made to distinguish AI but you can use it also to test a human being for its level of intelligence.

Here are the rules of the test. You are a step device which lives in an artificial world. To make a step you have to select your move in the checkboxes and to press the button "Next Step". What is your purpose? In order to cope well in this world you have to make more victories, and less losses and bad moves. You make victory when the lamp named "Victory" flashes. Respectively you make loss or bad move when the respective lamp flashes.



Of course, to cope well in this world you have to understand it first. This is extremely difficult even if you know that behind the panel is hidden the game Tick-Tack-Toe. The main problem is that you do not see the entire internal state of the game. This is obvious because you see only five lamps which do not carry enough information. Anyway, if you spare enough time for experiments in this world then you will understand it and will start to cope well in it. I am sure that you will do this because you manage to cope well in the real world, which is much more complicated. Really, you spare about 20 years for education before understanding the real world.

If you try to understand World 2 alone then you will understand the difficulties, which AI program has to overcome. It is better to try first the World 1. It is much easier because there is no hidden information in it. On every step you can observe the internal state of the World 1. In other words, the function **View** which gives the lamps status from the internal state of the world is inaction.

A Definition of AI

We are going to make a program which is able to cope in World 2 because we want to build AI. Our first question is what is AI. We will accept the definition for AI which is given in [2, 3, 4, 5]. Here is a short description of this definition.

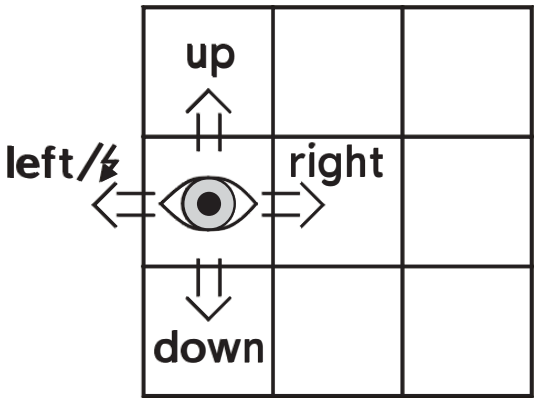
For us AI will be a step device which copes well in an arbitrary world. What is a world? For us this is a set **S** of internal states, one starting state $\mathbf{s_0}$ and two functions **World(s, d)** and **View(s)**. The function **World** will return the new state of the world from the current state and from the device's move (the state of the checkboxes). The function **View** will inform us what our device sees. That is, this function will return the state of the lamps from the current state of the world. One device copes better than another if it makes more victories and less losses.

This definition differs from the so called Turing's test [6, 7, 8] because it separates the intellect from the education. Alan Turing described a device which copes well in the natural world. Actually, even the human being cannot do this without education. So his device is already educated before the start of the test or his device is specially built for the natural world and includes within itself all information about the real world. In the case with World 2 it is not a problem to build a program which copes well in Tick-Tack-Toe world. There are many programs which play this game very successfully (one of them you can find in [1]). Our goal is not to build a program specialized for World 2 but for a class of worlds as big as possible. The best case is if the real world is inside this class.

So we will suppose that our program does not know anything for the world before it starts living (making steps). The AI program has to collect all information alone. This will make the task of building of AI program extremely difficult.

Detailed Description of World 2

In the World 2 you play Tick-Tack-Toe against an artificial partner which we will call Tom. The Tom's strategy is very simple. He plays every time randomly a correct move. You cannot see the whole board of the game but at every step you can see one cell by the state of the two yellow lamps. If the first lamp is on then this mean a cross. If the second lamp is on then this mean "O". If no lamp is on then there is nothing in the cell. Both lamps are never on together.



With the checkboxes you can select eight possible commands (moves). Four of them move your eye through the board. One command puts a cross at the position of the eye. Another one is the new game command. The last two commands are not used and they give bad move every time when you play them. When one Tick-Tack-Toe set is finished then one of the lamps Victory or Loss flashes depending on who won the set. If the set is equal then both lamps flash simultaneously. When after the end of the set you play the new game command then all cells are cleaned and become empty. The World 2 is made in this way because after the end of the set you will be able to continue to observe the Tick-Tack-Toe board in order to understand why you lost or won. Such observation is important for understanding of this world.

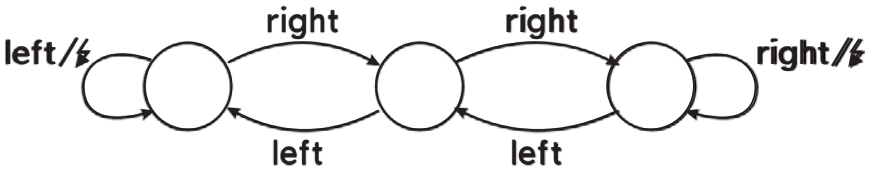
Every time when you try to do something which is not allowed the lamp "Bad Move" will flash red and the internal state of the world will stay unchanged.

Examples of bad moves: When you are in the first column and try to move left. When you try to put cross in the cell which is not empty or when the set is over. When you give the command new game before the end of the set.

Model of the World 2

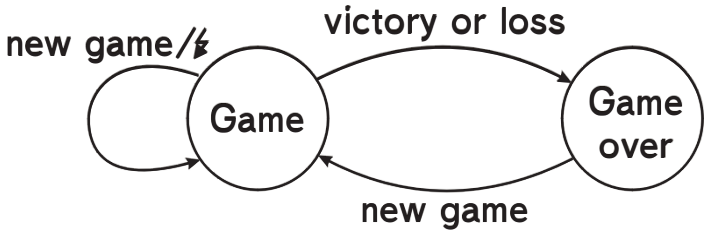
In order for AI to understand the World 2 it has to build a model of this world. Here we will build such a model, which will consist from very simple parts. Most of these parts will be final automata with maximum three states. For such small automata exists possibility AI to find them by brutal force or by more creative approach. For bigger automata there is not such possibility due to the combinatory explosion.

The automate (1) which will describe the position of the eye is the following one:



There are three states in this automate and they correspond to the left, middle and right column of the board. The change of the state occurs when AI makes the move "left" or "right". All other moves do not change the state of this automate. Here the flash stands for a "bad move". This means that if you try to play "left" from the left column then the lamp "bad move" will flash. Actually, this is the peculiarity of this automate, which will allow AI to find it. Actually, there is a huge number of final automata with three states but this one is special because it is connected with the rule that if the automate is in its first state and if AI plays "left" then the result is "bad move". From the other two states the move "left" every time is correct.

Of course, this automate is fundamental for the understanding of the World 2 because it gives to AI the information in which column the eye is. Analogically, AI will find the automate (2), which will say in which row the eye is and this two automata together will give the co-ordinates of the eye. Also analogically, AI will find the automate (3) with two states which will give information whether the game is over or not and when you can play the "new game" command. Interesting in this automate is that it can change its state due to the lamp status (if one or both of the lamps "victory" or "loss" flash).

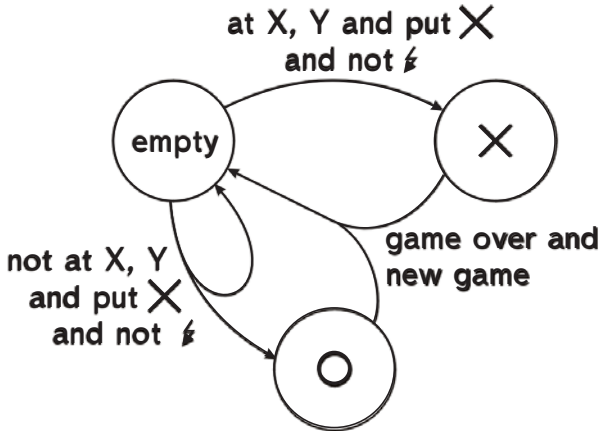


In order to begin to understand the rules and to play without "bad moves" we will need some easy rules like: If AI sees cross and plays "put cross" then the result is a "bad move". Simple rules like this can be represented like final automate with a single state. Of course, all single state automata are equivalent and that is why we can accept that we have only one automate of that kind.

In this way by constant rules and by deterministic automata we described the rules of World 2 and we can start playing correctly in it. Anyway this is not enough in order to understand it and to cope well in it. For this purpose we need to have an idea for the situation on the board (for the game status). Unfortunately, all possible situations on the board are too much. They are 3 on the power of 9. Of course, not all of these situations are really possible but they are too much anyway and it is impossible for AI to find an automate with so many states by brutal force.

That is why we will describe the situation on the board in a different way. We will introduce a special undeterministic final automate which will be of second level because it will change its states depending on the condition of other automata. The first level automata, which we will use for this, will be automata 1, 2 and 3. Actually, we will build 9 new automata - one for every possible co-ordinates (**X, Y**) of the eye. The condition "At X, Y" will be true if the automate (1) is at the state **X** and automate (2) is at the sate **Y**. The condition "Game over" will be true if the automate (3) is at the state "Game over".

Here is what the automate (**X, Y**) looks like:



What is the peculiarity of this automate, which will allow to AI to find it? This is the fact that when the automate (**X, Y**) is at the state "cross" and when the condition "At X, Y" is true then AI sees "cross". This peculiarity will also help to determine which state this undeterministic automate is in.

After all this we can understand the World 2 and the fact that if AI puts cross then Tom will answer immediately by putting "O" on the board. Actually we need also the rule that Tom cannot put more than one "O" as a response. For this we need first order formula which goes beyond the model from final automata. Also we need to understand why we win or lose and for this we need the concept of algorithm in order to understand "line" and "diagonal" on the board. We have "line" if we can three successive times see cross and go left (without "bad move" on the first two "left"). After this work which has to be done we will have the model of World 2 and we can start planning the future in this world. To plan its next move AI can use the algorithm Min-Max like the other programs, which play Tick-Tack-Toe. Anyway, the direct use of this algorithm will bring combinatory explosion because if we think ten steps in the future then we will make only few real moves in this ten steps. In order to win the game AI has to have an algorithm for moving of the eye from any position $\mathbf{X}_1$, $\mathbf{Y}_1$ to any new position $\mathbf{X}_2$, $\mathbf{Y}_2$. Also it needs to have an algorithm for checking the situation of the board to find where Tom responded with "O". With applying of such algorithms AI can reduce the Min-Max algorithm to calculating only the real moves instead of calculating of all possible steps.

Conclusion

This paper gives an idea of how we can make a program which will cope successfully in a Tick-Tack-Toe world without knowing apriory anything about this world. This program will be not that far from AI.

Bibliography

[1] Dobrev D. Strawberry Prolog. In: http://www.dobrev.com
[2] Dobrev D. AI Progect. In: http://www.dobrev.com/AI
[3] Dobrev D. AI - What is this. In: PC Magazine - Bulgaria, November'2000, 12-13.
[4] Dobrev D. AI - How does it cope in an arbitrary world. In: PC Magazine - Bulgaria, February'2001, 12-13.
[5] Dobrev D. A Definition of Artificial Intelligence. In: Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, 67-74.
[6] Turing, A. M. Intelligent machinery. In: report for National Physical Laboratory (1948). In Machine Intelligence 7, eds. B. Meltzer and D. Michie, 1969.
[7] Turing, A. M. Computing machinery and intelligence (1950) . In: Mind 49, 433-460.
[8] Turing, A. M. Can a Machine Think (1956) . In: The World of Mathematics, ed. James R. Newman, Simon & Schuster, Vol. 4, 2099-2123.

Author's Information

Dimiter Dobrev – Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; P.O.Box: 1274, Sofia-1000, Bulgaria; e-mail: d@dobrev.com

CONCURRENT ALGORITHM FOR FILTERING IMPULSE NOISE ON SATELLITE IMAGES

Nguyen Thanh Phuong

***Abstract**: A parallelized algorithm using improved "master/worker" architecture is applied to filtering noise on satellite images. Besides, being simple, natural and effective, paradigm of "data parallelism" is considered as a basis of parallel program. Finally, the most effective distribution is found basing on theoretical analyses and experimental results obtained from parallel–computing machines.*

***Keywords**: impulsive noise, noised streak, image distribution, parallelized algorithm.*

1. Introduction

Noise filtering is considered as an essential image pre-processing. Filtering must ensure that detailed information in image is maintained to the utmost. Operation of impulsive noises is grouped in "operation of low–level image processing" as every pixel needs to be operated separately [Nicolescu, Jonker, 2000]. It takes time for pixels on a big image to be operated one by one, thus it is not feasible in case real–time treatment required. This problem may be solved thanks to the system of parallel–computing machines [Seinstra, 2003].

This article introduces filtering impulsive noises on satellite images. These are images broadcasted from satellite Meteosat, their data is used for observations of the earth from space. For some technical reasons, impulsive noises can be recognized on the image with regular distribution and sometimes horizontal streaks. The common feature of these images is high frequency of broadcasting (every 30 minutes) and big size (100 MB/image from the 2nd generation Meteosat). Therefore, the system of parallel–computing machines plays an important role in processing such big volume of data.

This article presents not only parallelized algorithm for filtering but also the most effective method of image partition as well as task assignment between processes in system. Experimental results on parallel computer with 32 Xeon processors 2.66 GHz in the Institute of cybernetics Glushkov (Kiev, Ukraine) showed effectiveness of the proposed algorithm. Obtained results don't confine within filtering noises, they can be used in image processing at the next stages.

2. Serial Algorithm for Filtering Noises

Nowadays, Threshold Boolean Filter – TBF [Aizenberg, Astola, Butakoff, Egiazarian, Paliy, 2003] is one of the most effective algorithms for filtering impulsive noises. However, the appearance of noised streaks reduces its effectiveness. Thus, the combination of TBF and proposed method in [Phuong, 2004] improved filtering quality. In this article, we don't study these algorithms but introduce general steps by means of pseudo code (fig. 1) only.

```
<Image> = functionReadFileImage(<param.>)
// First step: Detect and Remove impulsive streak
if (functionDetectImpulsiveStreak(<Image>, <param.>) == TRUE)
   functionRemoveImpulsiveStreak(<Image>, <param.>);
// Second step: Detect and Remove isolated impulsive noise
for (count = 0; count < LOOP; count ++) {
   <Map> = functionDetectImpulsiveNoise(<Image>, <param.>);
   <RestoredImage> = functionRemoveImpulsiveNoise(<Image>,<Map>,<param.>);
   <Image> = <RestoredImage>;
}
functionWriteFileImage(<Image>, <param.>);
```

Figure 1. General serial algorithm of filtering noised image.

Removing the streaks of noises is always done at the first step of the algorithm. The result of this process is transferred to the second step. At the second step, restoration of isolated noised pixels happens only after detection of location of noises. Additionally, in order to enhance filtering quality, second step repeats **LOOP** = 4 times (the number of repetition is proved in experiments, the bigger is this number, the smoother image becomes). The result of the previous filtering is input data of the following one.

The most effective paradigms of parallel computation are data parallelism and task parallelism [Fissgus, Rauber, Rünger, 1999]. Data parallelism is only one program applied for all data at the same time. In contrast, task parallelism happens when different parts of the program works on independent processes or unrelated group of processes (processes in the same group can co-operate like in data parallelism). Sometimes the combination of these two paradigms in the same program will speed up processing compared with when being respectively operated [Nicolescu, Jonker, 2002]. However, to resolve our problem, let put it to the limit of data parallelism, which can be explained as follows:

(1) From the point of view of algorithm, first and second step (at a certain degree) may exist independently, so task parallelism can be used, but there are 2 problems:
   – Experiments in [Phuong, 2004] show that the proposed algorithm reaches the best quality only in case where first step will be done first and its result is transferred to second step.
   – A difficulty in task parallelism is equal distribution of tasks to different processes in order to improve effectiveness. It takes the most time to detect noised streaks (functionDetectImpulsiveStreak) and isolated noised pixels (functionDetectImpulsiveNoise) as it is operated on pixels one by one (including sorting neighboring pixels). Meanwhile, removing impulsive streaks (functionRemoveImpulsiveStreak) and isolated noised pixels (functionRemoveImpulsiveNoise) work with inappreciable number of pixels (less than 0.5% of total data). The stage of detecting and removing noised pixels (second step) belongs to one cycle of repetition **LOOP** times. Therefore, it is too complicated to balance the workload the workload of parallel processes.

(2) From the technical point of view, the algorithm of filtering uses 2 of 3 “operation of low-level image processing”. That is why it is a natural thing if we resolve the problem using data parallelism where every element of data being object of various operations.

Finally, the volume of data in our problem is not big enough to take full advantage of the combination of data and task parallelism.

3. Parallelized Algorithm for Filtering Noises

From the above analyses, the data parallelism is chosen to filter noise. During the processing, the whole surface of image is scanned by a window with size of $n \times n$, where n is odd integer. This processing can be done at the same time in different processes. However, one phenomenon happens that after every repetition, boundary data of neighboring processes will be exchanged. This is unexpected side effect for parallel computing. Fig. 2 shows pseudo code of the parallelized algorithm of filtering:

```
if (functionDetectImpulsiveStreak(<PartOfImage>, <param.>) == TRUE)
   functionRemoveImpulsiveStreak(<PartOfImage>, <param.>);
for (count = 0; count < LOOP; count ++)
{
   functionSendReceiveBorder(<param.>);
   <Map> = functionDetectImpulsiveNoise(<PartOfImage>, <param.>);
   <RestoredPartOfImage>=functionRemoveImpulsiveNoise(<PartOfImage>,<Map>,<param.>);
   <PartOfImage> = <RestoredPartOfImage>
}
```

Figure 2. The parallelized algorithm of filtering.

This parallelized algorithm is based on model SPMD [Aoyama, Nakano, 1999] (fig. 3a) with a small improvement in architecture "master/worker" of model MPMD [Aoyama, Nakano, 1999] (fig. 3b). Process "master" distributes parts of original image to processes "worker". In this case, process "master" also plays the role of "worker" and receive itself a part of image. After being operated by "workers", data will be gathered by "master" as the original size.

Following steps is required to design parallel processing program:

1. Increase the fraction of your program that can be parallelized.
2. Balance the workload of parallel processes.
3. Minimize the time spent for communication.

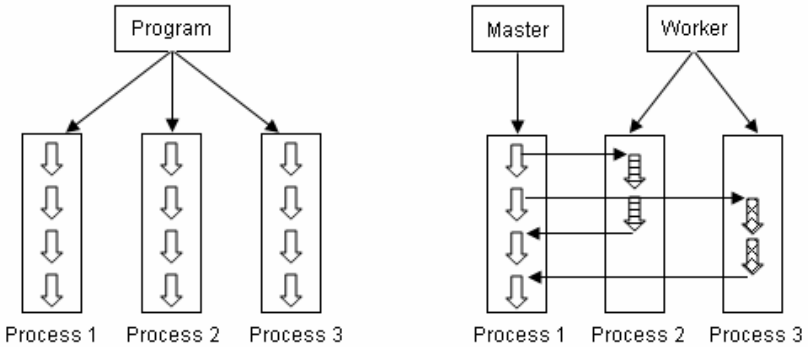


Figure 3. (a) Models SPMD and (b) MPMD: "master/worker" [Aoyama, Nakano, 1999].

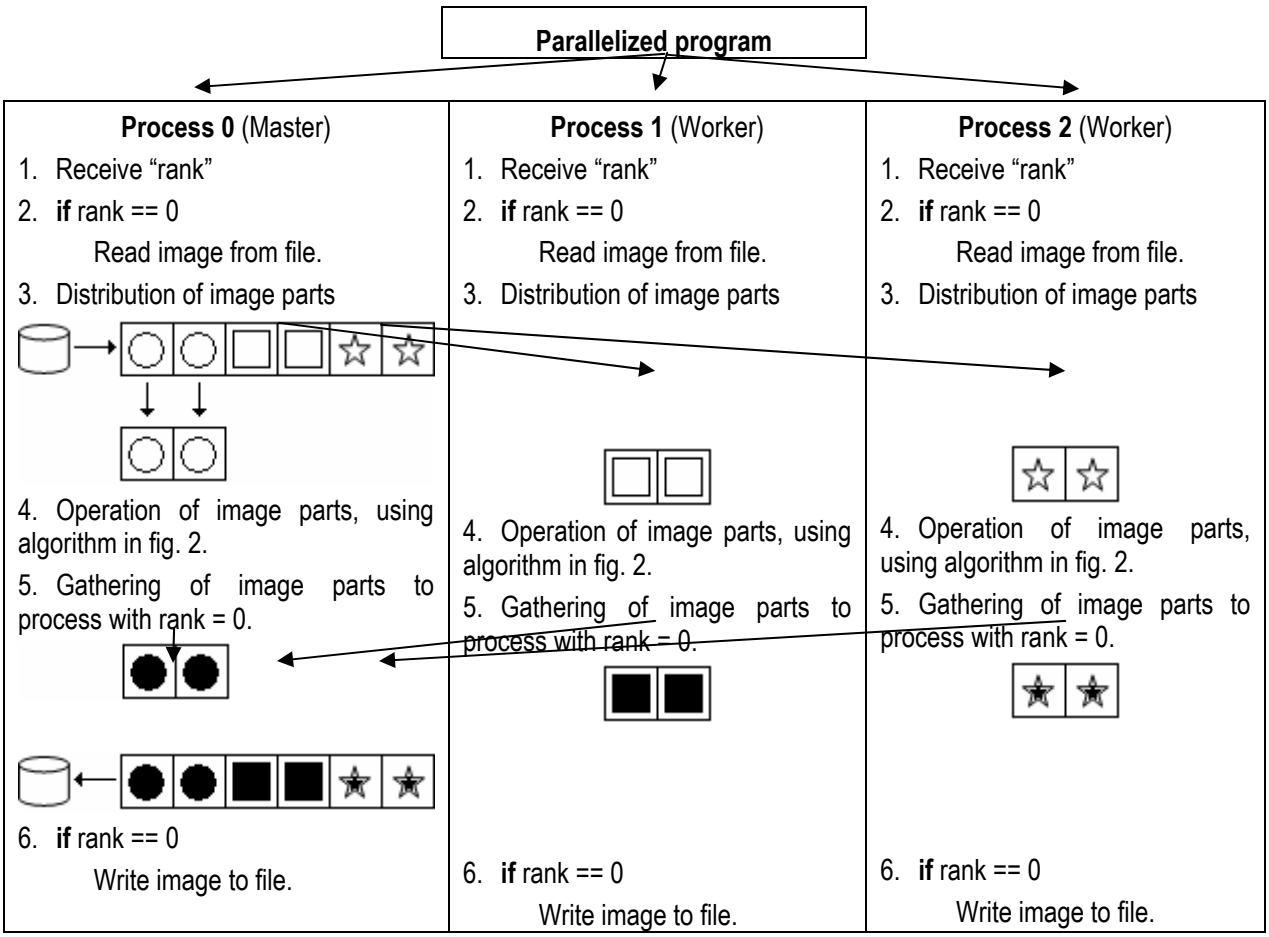


Figure 4 General diagram of parallel processing of filtering. Unfilled and black-painted circles, squares, stars mean data before and after processing respectively.

General diagram of parallel computation shows in fig. 4. At the first step, every process receives one “rank”. At the second one, the process with “rank” = 0 (“master”) reads image from file. At the next step, parts of original image are sent to all processes, including the one with “rank” = 0 (which is “worker” then). At the fourth step, processes operate their own data, using algorithm in fig. 2. After that, no process fully has initial image and at the fifth step, all parts will be transferred back to the process with “rank” = 0 (which plays the role of “master” then). The restored image will be saved into file at the sixth step.

The first requirement is obviously satisfied when we use the model of data parallelism. The feasibility of the second requirement depends on the existence of noised streaks on image parts. However, as mentioned above, the quantity of the noised pixels on streaks is unappreciable compared with total data (less than 0.5%), so it takes almost no time to remove streaks. Meanwhile, detection of streaks occurs on all processes. Therefore, the second requirement can be considered as satisfied if image parts have at least the same size.

The reduction of communication time between processes depends on 2 factors: the quantity of data and the number of exchange. Furthermore, access to data becomes quicker in case the data is contiguous in memory.

In practice, the width of satellite images is bigger than their height and the language usually used in developing parallelized software is C/C++ (the rows of image pixels are contiguous in memory). We will study 4 ways to distribute images.

3.1 Row-Wise Distribution

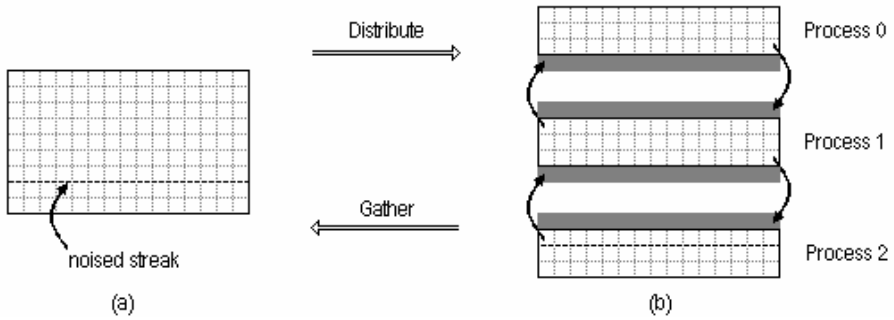


Figure 5. Row-Wise distribution. (a) Original image with noised streak, (b) Image parts.

This is the most natural distribution as data is stored in rows in memory. The grey-painted parts on fig. 5 contain exchange information. The arrows point out directions of transfer (these notations will be used in the remaining parts).

This way of access is restricted by following factors: firstly it fails to balance the required quantity of operations; secondly it takes the most data in exchange process.

3.2 Rotation and Row-Wise Distribution

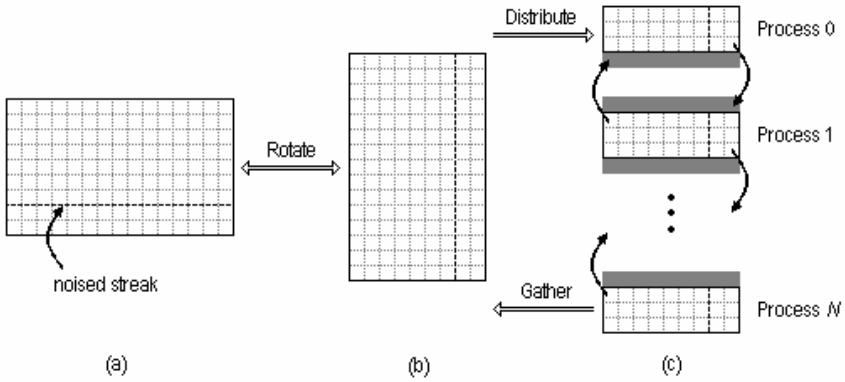


Figure 6. Rotation and Row-wise distribution.
(a) Original image with noised streak, (b) Rotated Image, (c) Image parts.

Image is rotated to 90° to minimize the above weak points. The balance of workload and the reduction of exchanged data are both satisfied. Furthermore, access to contiguous data remains unchanged. The only disadvantage of this method is time-taking for image rotation before and after operation (fig. 6).

3.3 Column-Wise Distribution

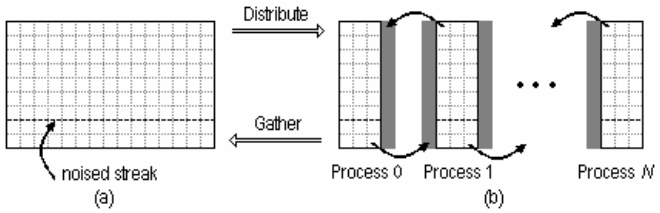


Figure 7. Column-wise distribution. (a) Original image with noised streak, (b) Image parts.

If it is proposed that it considerably takes time for image rotation, the return to Row-wise distribution seems logical then. This time image is vertically distributed (fig. 7). This Column-wise distribution meets requirement of balance of workload and reduction of exchanged data, however, elements necessary to be accessed are not contiguous in memory.

3.4 Rectangular-Block Distribution

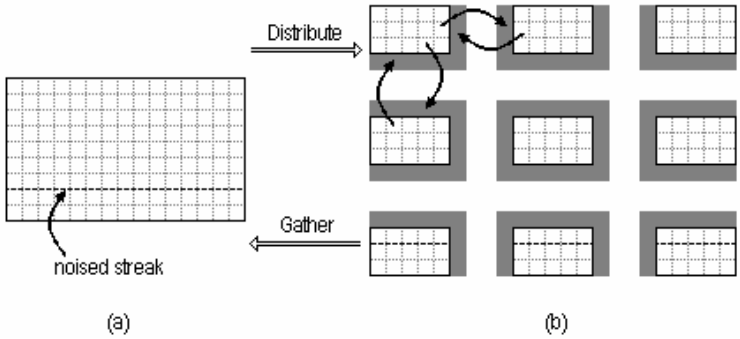


Figure 8. Rectangular-block distribution. (a) Original image with noised streak, (b) Image parts.

The advantage of this method is to minimize boundary elements in transfer because the perimeter of a rectangle is always bigger than the one of a square with the same area. However, from the point of view of program designing, it is the most complicated compared with others (fig. 8).

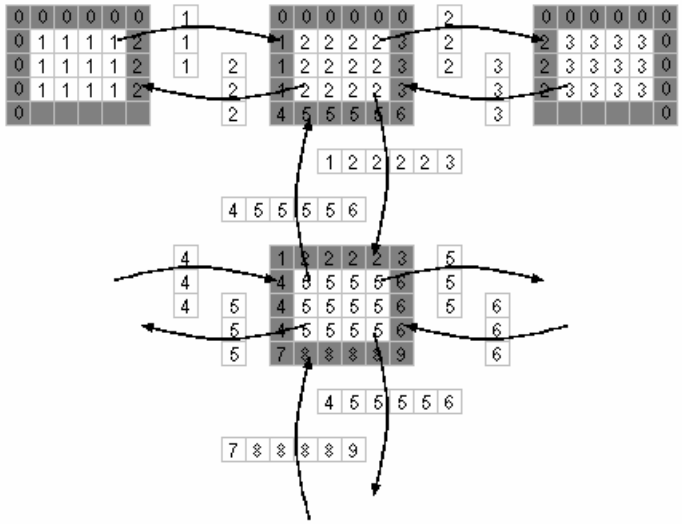


Figure 9. Sending/Receiving boundary elements. Firstly, “horizontal” boundary elements take in exchange process, and then “vertical” boundary elements.

Additionally, although this method allow us to reduce the quantity of exchanged data, it increases the number of transfers: 4 operations of sending/receiving “vertical” information + 4 operations of sending/receiving “horizontal” information + 4 operations of sending/receiving “angle” information. Thanks to a tip showed in [Aoyama, Nakano, 1999], 4 operations of sending/receiving “angle” information is excluded (fig. 9).

4. Experimental Results

The experiments were done on the system of 32 processors of Xeon 2.66 Ghz, using Linux operating system. The size of image in experiments is 100MB. Result of experiments is shown in fig. 10.

With rectangular block distribution, matrix logic of processors is automatically established in order to minimize the pixels taking part in data exchange between processes.

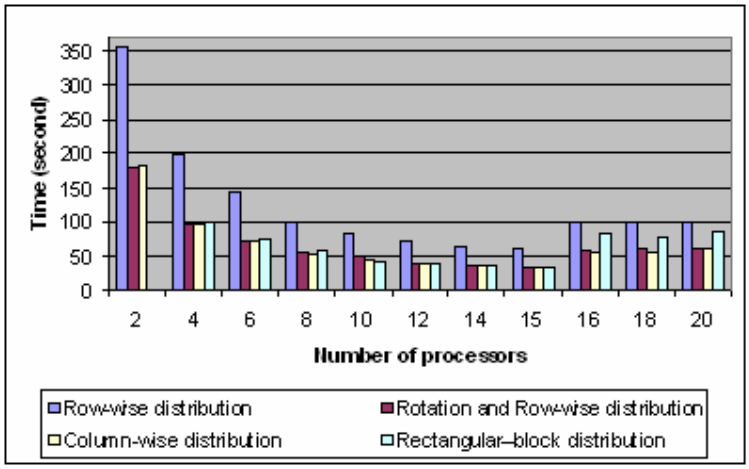


Figure 10. Comparison of result of experiments.

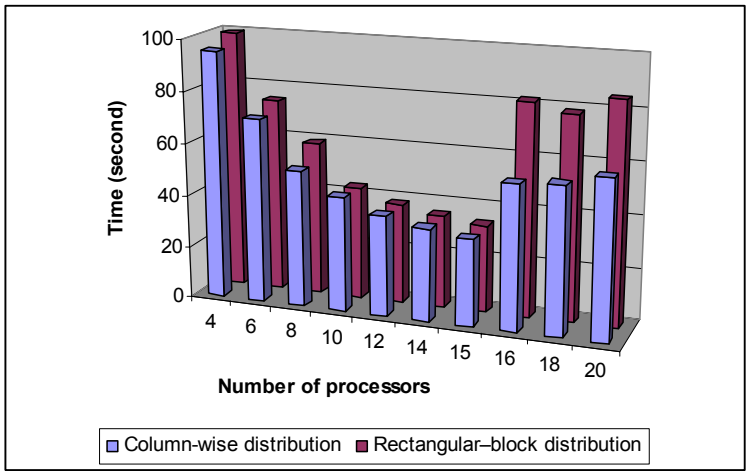


Figure 11. Comparison of result of operations when using Column-wise distribution and Rectangular-block distribution

Basing on obtained results, we have following comments:

– With Row-wise distribution, we have the worst result. In this case, the quantity of boundary elements in data exchange is biggest, which results in taking much time for processing.

– Rotation and Row-wise distribution and Column-wise distribution take nearly same time. Image rotation is done in memory and time for this operation is quite small compared with the whole process. In case image size is big enough, Column-wise distribution is reasonable choice.

– In most cases, Column-wise distribution need less time than Rectangular-block distribution (fig. 10). Its basic reason is purely technical matter. The interface between processors in the system is Gigabit Ethernet. It means that at a point of time, only one pair of neighboring processes takes part in data exchange. With Rectangular-block distribution, the number of transfers is 2 times more than Column-wise distribution, which leads to the increase in time spent for communication.
– Fig. 11 clearly proves that time for image operation reduces to minimum when we use 15 processors. Effectiveness of image distributions decreases as long as the number of processors increases afterwards. When the quantity of processes increases, the number of exchanged data also goes up. If the system uses professional interfaces, e.g. SCI ([IEEE, 1993], [Hellwagner, Reinefeld, 1998]) this phenomenon will not happen.

Satellite image from Meteosat before and after processing can be seen in fig. 12.

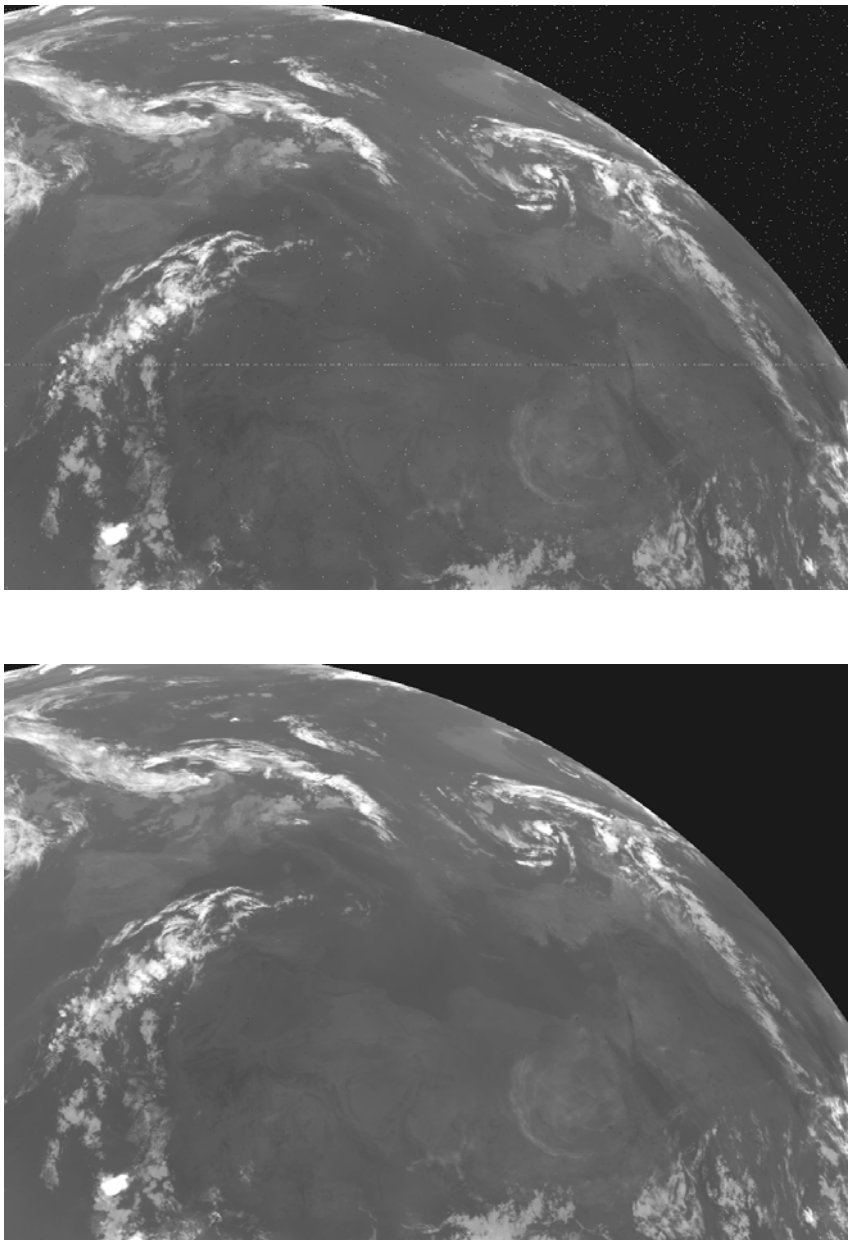


Figure 12. Satellite image from Meteosat before and after processing impulsive noises.

5. Conclusion

This article introduced parallelized algorithm for filtering impulsive noise. Its strong point is simple, natural and effective. If we count time for processing, column-wise distribution is the best solution. On the contrary, if the simplicity in program design is required, Rotation and Row-Wise Distribution is the reasonable choice. In this case, a pair of functions MPI_Scatter and MPI_Gather (MPI library – [Message Passing Interface Forum, 1994], [Pacheco, 1998], [Gropp, 2004], [Snir, Otto, Huss-Lederman, Walker, Dongarra, 1996], [Gropp, Lusk, Doss, Skjellum, 1996]) becomes very helpful for this task.

The algorithm represented herewith seems effective in processing images with big size and meets the requirement of real-time processing. It will be used as pre-processing module of a system for cloudiness dynamic identification based on the space images of Meteosat of Second Generation.

Bibliography

[Nicolescu, Jonker, 2000] C. Nicolescu, P. Jonker. Parallel low-level image processing on a distributed-memory system. The International Parallel and Distributed Processing Symposium Conference, 2000.

[Seinstra, 2003] F. J. Seinstra. User Transparent Parallel Image Processing. PhD Thesis. Intelligent Sensory Information System, University of Amsterdam, the Netherlands, 2003.

[Aoyama, Nakano, 1999] Y. Aoyama, J. Nakano. RS/6000 SP: Practical MPI Programming. RS/6000 Technical Support Center, IBM Japan, 1999.

[Message Passing Interface Forum, 1994] MPI: A Message-passing Interface Standard. Technical Report: UT-CS-94-230, 1994.

[Pacheco, 1998] P. S. Pacheco. A user's guide to MPI. Technical report, Department of Mathematics, University of San Fransisco, California, 1998.

[Gropp, 2004] W. D. Gropp. Issues in Accurate and Reliable Use of Parallel Computing in Numerical Programs. Preprint ANL/MCS-P1193-0804, Argonne National Laboratory, 2004.

[Snir, Otto, Huss-Lederman, Walker, Dongarra, 1996] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, J. Dongarra. MPI: The Complete Reference. The MIT Press, 1996.

[Gropp, Lusk, Doss, Skjellum, 1996] W. D. Gropp, E. Lusk, N. Doss, A. Skjellum. A High-Performance, Portable Implementation of the MPI Message Passing Interface. Parallel Computing, 1996.

[Aizenberg, Astola, Butakoff, Egiazarian, Paliy, 2003] I. Aizenberg, J. Astola, C. Butakoff, K. Egiazarian, D. Paliy. Effective Detection and Elimination of Impulse Noise with a Minimal Image Smoothing. Proceedings of IEEE International Conference on Image Processing, 2003.

[Phuong, 2004] N. T. Phuong. A Simple and Effective Method for Detection and Deletion of Impulsive Noised Streaks on Images. Problems of Controls and Informatics, 2004. (in Russian)

[Fissgus, Rauber, Rünger, 1999] U. Fissgus, T. Rauber, G. Rünger. A Framework for Generating Task Parallel Programs. The 7th Symposium on the Frontiers of Massively Parallel Computation, 1999.

[Nicolescu, Jonker, 2002] C. Nicolescu, P. A. Jonker. Data and Task Parallel Image Processing Environment. Parallel Computing, 2002.

[IEEE, 1993] IEEE Standard for Scalable Coherent Interface (SCI). IEEE Standard 1596–1992, IEEE Standards Board, 1993.

[Hellwagner, Reinefeld, 1998] H. Hellwagner, A. Reinefeld. Scalable Coherent Interface: Technology and Applications. Proceedings of SCI Europe '98, 1998.

Author's Information

Nguyen Thanh Phuong – The National Technical University of Ukraine "KPI", Vuborgska st. 1, (dormitory 17 of KPI, room 514), Kiev, Ukraine; e-mail: ntphuong72@yahoo.com.

4.6. Macro-economical Modelling

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЧЕТКОГО И НЕЧЕТКОГО МЕТОДОВ ИНДУКТИВНОГО МОДЕЛИРОВАНИЯ (МГУА) В ЗАДАЧАХ МАКРОЭКОНОМИЧЕСКОГО ПРОГНОЗИРОВАНИЯ

Юрий П. Зайченко

Abstract: *The fuzzy Group Method of Data Handling(FGMDH) is considered and investigated. The algorithm of fuzzy GMDH is suggested. The experimental investigations of the suggested algorithm of FGMDH in the problem of macroeconomic indexes forecasting are presented and the comparative analysis with the conventional algorithm GMDH are presented and discussed.*

Keywords: *fuzzy, group method of data handling, economy forecast.*

Введение.

Задача макроэкономического прогнозирования обладает рядом особенностей, а именно:

1) сложные неизвестные зависимости между различными макроэкономическими показателями;
2) нестационарность экономических процессов особенно для стран с переходной экономикой;
3) неопределенность и недостоверность исходных данных по ряду макроэкономических показателей.

Эти особенности не позволяют применить для задач макроэкономического прогнозирования традиционные методы регрессионного анализа и требуют разработки принципиально новых подходов и методов, свободных от ограничений классических методов.

К числу эффективных методов построения моделей сложных систем по экспериментальным наблюдениям относится метод индуктивного моделирования, известный как метод группового учета аргументов(МГУА), предложенный акад. А. Г. Ивахненко и развитый в его работах и работах учеников[1,2].

Метод базируется на принципах самоорганизации и внешнего дополнения. Этот метод обладает следующими пренципиальными преимуществами в сравнении с другими методами:

1) Возможность работы на «коротких» выборках данных, когда число точек наблюдения меньше, чем число неизвестных коэффициентов модели(регрессии);
2) Нет необходимости априорно задавать вид модели – модель конструируется в процессе работы алгоритма МГУА на основе идей самоорганизации.

Вместе с тем метод МГУА не лишен известных недостатков, одним из которых является возможность явления плохой обусловленности матрицы нормальных уравнений Гаусса, которая связана с использованием метода наименьших квадратов(МНК) для определения коэффициентов модели.

Кроме того, метод дает точечные оценки прогнозируемых величин, а желательно иметь интервальные, позволяющие судить о точности прогноза.

Указанные обстоятельства обусловили необходимость развития метода МГУА, который был бы свободен от вышеперечисленных недостатков. Таким методом является нечеткий МГУА, предложенный в [3], и исследованный и развитый в работах [4,6].

Целью настоящего доклада является исследование эффективности применения нечеткого МГУА в задачах макроэкономического прогнозирования и его сравнительный анализ с обычным(т. е. четким) МГУА.

1. Нечеткий МГУА. Основные идеи.

В работах [4,5] была рассмотрена линейная интервальная модель регрессии

$$Y = A_0Z_0+A_1Z_1+\ldots+A_nZ_n \quad (1)$$

где A_i – нечеткие числа треугольного вида, описываемые парой параметров

$$A_i = (\alpha_i, c_i)$$

где α_i – центр интервала, c_i – его ширина, $c_i \geq 0$.

Тогда Y – нечеткое число, параметры которого определяются следующим образом:

центр интервала $$\alpha_y = \sum \alpha_i z_i = \alpha_i^T \cdot z \quad (2)$$

ширина интервала $$c_y = \sum c_i \cdot |z_i| = c^T |z| \quad (3)$$

Для того, чтобы интервальная модель была корректной необходимо, чтобы действительное значение выходной величины Y принадлежало интервалу неопределенности, что описывается следующими ограничениями:

$$\begin{cases} \alpha^T z - c^T \cdot |z| \leq y \\ \alpha^T z + c^T \cdot |z| \geq y \end{cases} \quad (4)$$

Например, для частичного описания вида

$$f(x_i, x_j) = A_0 + A_1 x_i + A_2 x_j + A_3 x_i x_j + A_4 x_i^2 + A_5 x_j^2 \quad (5)$$

в общей модели (1) необходимо положить $z_0 = 1$, $z_1 = x_i$, $z_2 = x_j$, $z_3 = x_i x_j$, $z_4 = x_i^2$, $z_5 = x_j^2$.

Предположим, что мы наблюдаем обучающую выборку $\{z_1, z_2, \ldots, z_M\}$, $\{y_1, y_2, \ldots, y_M\}$. Тогда для адекватной модели вида (1) необходимо найти также $(\alpha_i, c_i)_{i=\overline{1,n}}$, для которых бы выполнялись соотношения вида:

$$\begin{cases} \alpha^T z_k - c^T \cdot |z_k| \leq y_k \\ \alpha^T z_k + c^T \cdot |z_k| \geq y_k \end{cases}, \quad k = \overline{1, M} \quad (6)$$

Сформулируем основные требования к оценочной линейной интервальной модели для частичного описания вида (5).

Найти такие значения параметров (α_i, c_i) нечетких коэффициентов, при которых:

а) наблюдаемые значения y_k попадали бы в оценочный интервал для Y_k;

б) суммарная ширина оценочного интервала была бы минимальна.

Эти требования можно свести к следующей задаче ЛП:

$$min(C_0 \cdot M + C_1 \sum_{k=1}^{M} |x_{ki}| + C_2 \sum_{k=1}^{M} |x_{kj}| + C_3 \sum_{k=1}^{M} |x_{ki} \cdot x_{kj}| + C_4 \sum_{k=1}^{M} |x_{ki}^2| + C_5 \sum_{k=1}^{M} |x_{kj}^2|) \quad (7)$$

при условиях:

$$\begin{aligned}
&\alpha_0+\alpha_1 x_{ki}+\alpha_2 x_{kj}+\alpha_3 x_{ki}x_{kj}+\alpha_4 x_{ki}^2+\alpha_5 x_{kj}^2-(C_0+C_1\left|x_{ki}\right|+C_2\left|x_{kj}\right|+\\
&+C_3\left|x_{ki}\cdot x_{kj}\right|+C_4\left|x_{ki}^2\right|+C_5\left|x_{kj}^2\right|)\le y_k \qquad k=\overline{1,M}\\
&\alpha_0+\alpha_1 x_{ki}+\alpha_2 x_{kj}+\alpha_3 x_{ki}x_{kj}+\alpha_4 x_{ki}^2+\alpha_5 x_{kj}^2+(C_0+C_1\left|x_{ki}\right|+C_2\left|x_{kj}\right|+\\
&+C_3\left|x_{ki}\cdot x_{kj}\right|+C_4\left|x_{ki}^2\right|+C_5\left|x_{kj}^2\right|)\ge y_k
\end{aligned} \tag{8}$$

$$C_p\ge 0,\ \ p=\overline{0,5} \tag{9}$$

где *к* – номер точки измерения.

Как видим, задача (7) – (9) является задачей ЛП. Однако неудобство формы (7) – (9) для применения стандартных методов ЛП состоит в том, что нет ограничений неотрицательности для переменных α_i. Поэтому для ее решения переходим к двойственной задаче, введя двойственные переменные $\{\delta_k\}$ и $\{\delta_{k+M}\}$. Решив двойственную задачу симплекс–методом и найдя оптимальные значения двойственных переменных $\{\delta_k\}$, $\{\delta_{k+M}\}$ найдем оптимальные значения искомых переменных c_i, α_i, $i=\overline{0,5}$, а также искомую нечеткую модель для частичного описания (5).

2. Описание алгоритма НМГУА.

Дадим краткое описание алгоритма.

1. Выбор общего вида модели, которым будет описываться искомая зависимость.
2. Выбор внешних критериев оптимальности (критерия регулярности $\bar{\delta}^2$ или несмещенности $N_{см}$).
3. Выбор общего вида опорной функции (вида частичных описаний), например, линейного или квадратичного.
4. Разбиение выборки на обучающую $N_{об}$ и проверочную $N_{пров}$.
5. Присваиваем нулевые значения счетчику числа моделей k и счетчику числа рядов r.
6. Генерируем новую частичную модель f_r вида (5) на обучающей выборке. Решаем задачу ЛП (10) - (12) и находим искомые значения α_i, c_i.
7. Определяем по проверочной выборке $N_{пров}$ значение внешнего критерия ($N_{см\,k}^{(r)}$) или ($\delta_k^{(2)}(r)$).
8. $k=k+1$. Если $k\ge C_F^2$, то $k=0$, $r=r+1$.
9. Вычисляем средний критерий для моделей *r*-й итерации ($N_{см}^{(r)}$ или $\delta^{(2)}(r)$). Если $r=1$, то переходим на шаг 6, иначе – на шаг 10.
10. Если $\left|N_{см}(r)-N_{см}(r-1)\right|\le\varepsilon$, то переходим на шаг 11, иначе – отбираем *F* лучших моделей и положив $r=r+1$, $k=1$, переходим на шаг 6 и выполняем следующую *(r+1)*-ю итерацию.
11. Из *F* моделей предыдущего ряда находим по критерию регуляризации наилучшую модель.

3. Описание экспериментальных исследований.

Алгоритм четкого МГУА был реализован в пакете GMDH, а предложенный алгоритм нечеткого МГУА реализован в пакете FGMDH. В качестве исходных данных для обеих программ выбирались данные по макроэкономическим показателям Украины за 2001-2003 гг. В качестве прогнозируемых показателей выбирался индекс потребительских цен IOC, а также ВВП. В процессе экспериментов варьировалось число оптимальных моделей передаваемых на следующий уровень F=5,6,7.(свобода выбора), а также % соотношение между обучающей и проверочной выборками $\left(\frac{20}{80}; \frac{30}{70}; \ldots \frac{80}{20}\right)$.

Некоторые из полученных экспериментальных результатов приводятся ниже(табл. 1-4)

Таблица 1. Результаты прогноза для соотношения обучающей и проверочной выборок 40%-60%

	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321055,36	321081,04	321096,45	321127,37	321527,67	321731,86
СКО	2637,4957	2310,1654	2584,4866	2165,1548	1369,6392	890,9118
%СКО	0,618251	0,516701	0,601806	0,507531	0,306356	0,207452

Таблица 2. Результаты прогноза для соотношения обучающей и проверочной выборок 20%-80%

	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321012,34	321081,04	321096,45	321127,37	321527,67	321731,86
СКО	2954,9378	2573,8456	2693,3674	2394,7324	1672,4382	1045,3724
%СКО	0,635622	0,5745323	0,6114884	0,603422	0,413345	0,378803

Таблица 3. Результаты прогноза для соотношения обучающей и проверочной выборок 30%-70%

	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321032,85	321073,81	321082,91	321102,43	321404,52	321594,83
СКО	2763,2425	2473,2324	2602,7313	2283,4662	1493,5241	943,4321
%СКО	0,625315	0,5426213	0,6053482	0,559243	0,359234	0,271042

Таблица 4.Результаты прогноза для соотношения обучающей и проверочной выборок 50%-50%

	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321089,23	321115,52	321123,48	321167,35	321672,83	321793,41
СКО	2601,6322	2263,5346	2492,4142	2087,3253	1295,3485	832,4143
%СКО	0,608532	0,5035522	0,601432	0,481414	0,286234	0,196434

Таблица 5. Результаты прогноза для соотношения обучающей и проверочной выборок 60%-40%

	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321103,52	321153,85	321173,74	321208,58	321694,52	321843,12
СКО	2544,3763	2134,2345	2405,1245	1975,3452	1192,4143	801,1334
%СКО	0,591452	0,482231	0,569241	0,458312	0,243535	0,167382

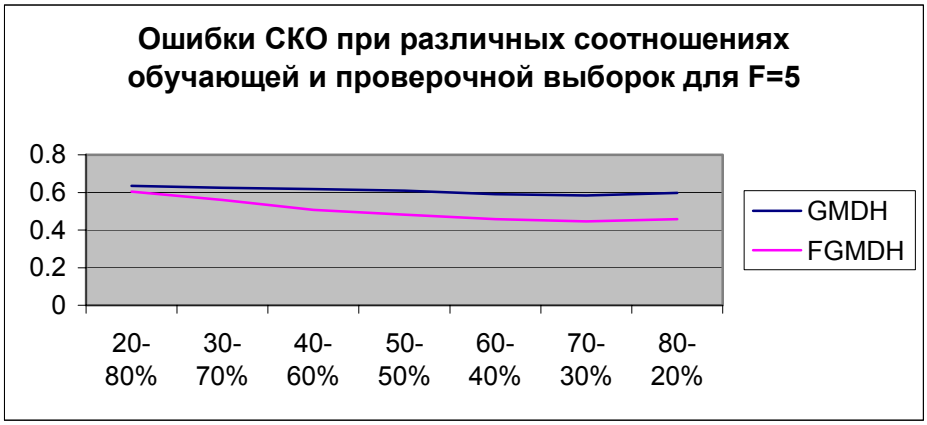
Таблица 6. Результаты прогноза для соотношения обучающей и проверочной выборок 70%-30%

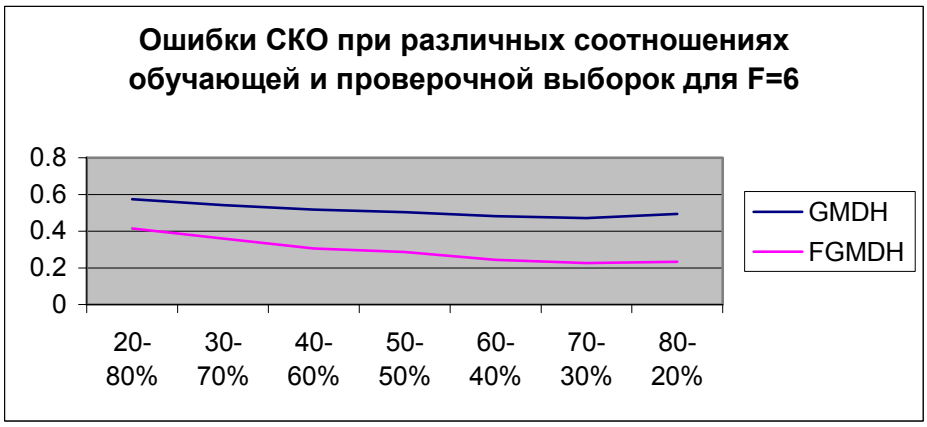
	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321108,62	321165,23	321191,65	321232,54	321707,75	321893,78
СКО	2532,4242	2104,7921	2483,4254	1952,5822	1174,5322	794,7325
%СКО	0,584325	0,471942	0,561432	0,445886	0,225544	0,161422

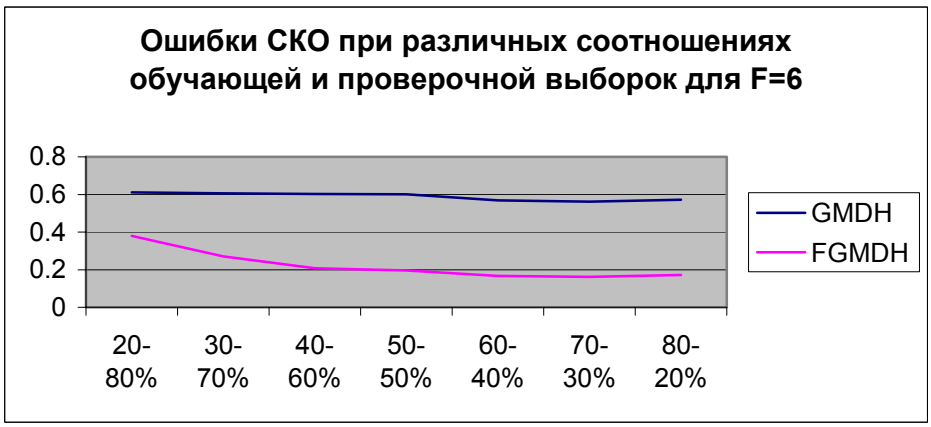
Таблица 7. Результаты прогноза для соотношения обучающей и проверочной выборок 80%-20%

	Пакет GMDH (четкий МГУА)			**Пакет FGMDH (нечеткий МГУА)**		
Кол-во передаваемых моделей	F=5	F=6	F=7	F=5	F=6	F=7
Прогнозное значение	321083,53	321112,53	321123,42	321102,54	321465,62	321645,59
СКО	2954,9378	2573,8456	2693,3674	2044,3323	1243,6531	820,2354
%СКО	0,597533	0,493543	0,572453	0,457932	0,233543	0,171343

Теперь приведем графики зависимости СКО прогноза от соотношения между обучающей и проверочной выборками для различных значений свободы выбора F.







Выводы.

Как следует из приведенных выше результатов, в целом нечеткий МГУА дает результаты лучше, чем четкий МГУА в задачах макроэкономического прогнозирования(меньше СКО).

Причем с ростом степени свободы выбора, этот выигрыш становится всё ощутимее. Это связано с тем, что, во-первых, отсутствует проблема плохой обусловленности матриц, а, во-вторых, в нечётком МГУА для спрогнозированных значений строятся доверительные интервалы, что позволяет проводить корректировку моделей путем их адаптации в случае явного непопадания значения в доверительный интервал. Еще одна интересная особенность состоит в том, что наилучший результат достигается при соотношении обучающей и проверочной выборок примерно 2:1.

Литература

1. Ивахненко А.Г., Мюллер И.А. Самоорганизация прогнозирующих моделей. Киев, Техника, 1985.
2. Mueller J.-A., Lemke F. Self-Organizing Data Mining. Berlin, Dresden 1999.
3. Ю.П. Зайченко, О.Г. Кебкал, В.Ф. Крачковський. Нечіткий метод групового врахування аргументів та його застосування в задачах прогнозування макроекономічних показників //Наукові вісті НТУУ КПІ, №2, 2000р., с. 18-26
4. Ю.П. Зайченко, І.О. Заєць. Синтез та адаптація нечітких прогнозуючих моделей на основі методу самоорганізації //Наукові вісті НТУУ КПІ, №2, 2001р.
5. Ю.П. Зайченко, І.О. Заєць. Застосування рекурсивних методів ідентифікації в задачах синтезу нечітких прогнозуючих моделей //Міжнародна конференція з індуктивного моделювання, Львів, 20-25 травня 2002: праці в 4-х томах.-Т.2.; стор. 59-65.
6. Ю.П. Зайченко, І.О. Заєць, О.В. Камоцький, О.В. Павлюк. Дослідженя різних видів функцій належності в нечіткому методі групового урахування аргументів //Управляющие системы и машины, №2, 2003г.

Информация об авторе

Zaychenko Yuri – Dr. of science, National technical University of Ukraine “KPI”, Institute for Applied System Analysis, dean, Address: Peremogy ave 37, Kiev, Ukraine, phone: 8044-241-86-93, e-mail: zaych@i.com.ua, zaych@mmsa.ntu-kpi.kiev.ua.

ИССЛЕДОВАНИЕ НЕЧЕТКОЙ НЕЙРОННОЙ СЕТИ ANFIS В ЗАДАЧАХ МАКРОЭКОНОМИЧЕСКОГО ПРОГНОЗИРОВАНИЯ

Юрий П. Зайченко, Фатма Севаее

Abstract: *The fuzzy neural network ANFIS with logical inference of Sugeno is considered. The training algorithm is described and its experimental investigations in the problem of macroeconomic indexes forecasting on the example of Ukrainian economy are presented.*

The influence of the number of linguistic variables and number of fuzzy rules on the efficiency of forecast is investigated and discussed.

Keywords *: fuzzy neural network, ANFIS, economy, forecast, training algorithm.*

Введение

Проблема макроэкономического прогнозирования в странах с переходной экономикой обладает рядом особенностей:

1) существенная нестационарность экономических процессов;
2) неопределенность и недостоверность исходных данных по ряду микроэкономических показателей;
3) ограниченность выборок данных (короткие выборки).

Указанные обстоятельства не позволяют применить для задач макроэкономического прогнозирования традиционные методы регрессионного и дисперсионного анализа и настоятельно требуют разработки принципиально новых подходов и методов, в частности использующих идеи искусственного интеллекта.

К числу перспективных направлений в области искусственного интеллекта относятся нечеткие нейронные сети(ННС). Нечеткие нейронные сети в отличие от обычных нейросетей, позволяют использовать априорную информацию экспертов в виде нечетких правил вывода: «ЕСЛИ-ТО».

Кроме того, они дают возможность работать в условиях неполной и недостоверной информации, когда значения ряда исходных показателей заданы интервально, а также когда некоторые из них являются качественными, и описываются как лингвистические переменные(малый, средний, большой и т.д.).

Для обучения ННС используется накопленный арсенал методов обучения, разработанных для обычных нейросетей, в частности, градиентный, сопряженных градиентов и др.[1,2]. В работе [3] было проведено исследование нечетких контроллеров с выводом Мамдани и Цукамото в задачах прогнозирования.

Цель настоящей работы – исследование и анализ эффективности применения ННС с логическим выводом Сугено в задачах макроэкономического прогнозирования на примере экономики Украины.

1.Архитектура и организация работы нечеткой нейросети ANFIS

Рассмотрим адаптивную нечёткую систему с механизмом логического вывода предложенного Сугено на базе правил ЕСЛИ-ТО [1, 3] , которая получила название сети ANFIS (*Adaptive Network Based Fuzzy Inference System*). Данная система может быть успешно использована для настройки функции принадлежности и настройки базы правил в нечёткой экспертной системе. Ниже представлена модель нечёткого вывода Сугено и структурная схема сети ANFIS.

ANFIS система использует следующую базу правил:

если $x = A_1$ **и** $y = B_1$ **то** $f_1 = a_1x + b_1y + r_1$

если $x = A_2$ **и** $y = B_2$ **то** $f_2 = a_2x + b_2y + r_2$,

где A_i и B_i являются лингвистическими переменными.

Слои данной нечёткой нейронной сети выполняют такие функции:

Слой 1. Каждый нейрон данного слоя является нейроном, который преобразует входной сигнал x или y с помощью функции принадлежности (фазификатор). Чаще всего используют колоколоподобную функцию

$$\mu_{Ai}(x) = \frac{1}{1 + \left[\left(\frac{x - c_i}{a_i}\right)^2\right]}, \tag{1}$$

или функцию Гаусса

$$\mu_{Ai}(x) = exp\left[-\left(\frac{x - c_i}{a_i}\right)^2\right] \tag{2}$$

Слой 2. Каждый нейрон в этом слое, отмеченный как Π, осуществляет пересечение множества входных сигналов, моделируя логическую операцию AND и посылает на выход

$$w_i = \mu_{Ai}(x) \times \mu_{Bi}(y), i = 1,2. \tag{3}$$

По сути, каждый нейрон представляет собой активирующую силу правила.

Слой 3. Каждый нейрон в этом слое вычисляет нормированную силу правила:

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2. \tag{4}$$

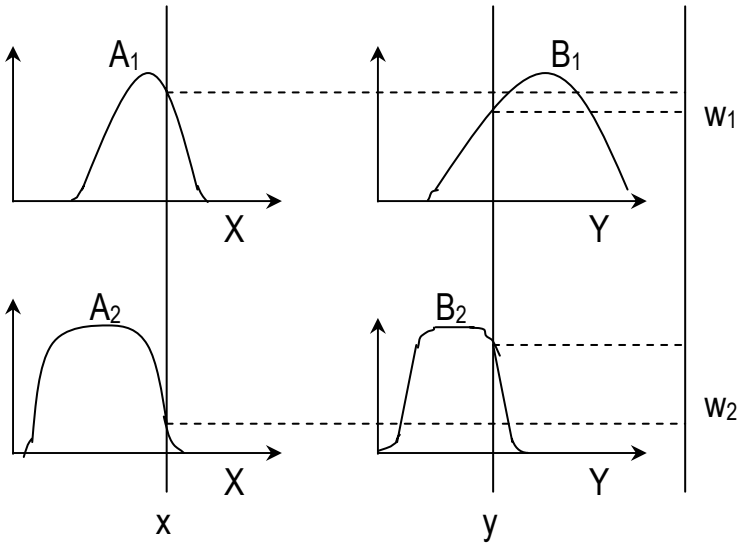
Слой 4. На данном слое в нейронах формируются значения выходных переменных:

$$O_i^4 = \overline{w_i} f_i = \overline{w_i}(a_i x + b_i y + r_i). \tag{5}$$

Слой 5. В последнем слое получаем выходной сигнал нейронной сети и выполняем дефазификацию результатов:

$$O_i^5 = overall\ output = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}. \tag{6}$$

Нейронная сеть архитектуры ANFIS обучается с помощью метода градиентного спуска, который в контексте нечетких нейронных сетей будет более детально рассмотрен в следующем разделе.



$f_1 = a_1 x + b_1 y + r_1$

$f_2 = a_2 x + b_2 y + r_2$

$$f = \frac{w_1 f_1 + w_2 f_2}{w_1 + w_2} = \overline{w_1} f_1 + \overline{w_2} f_2$$

Рис.1а: Схема логического вывода Сугено

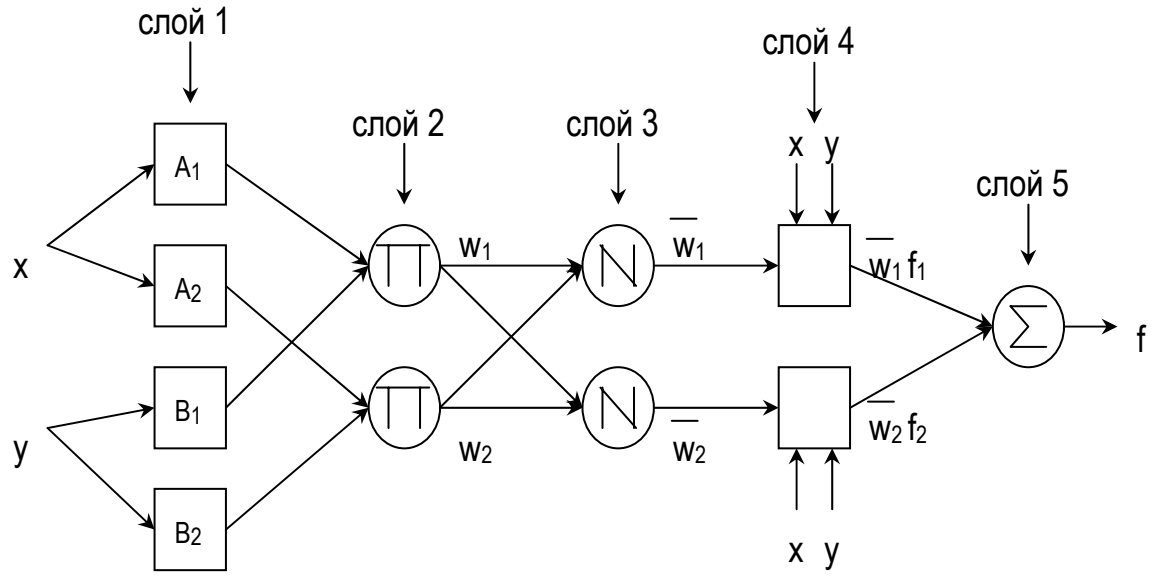


Рис.1б: эквивалентная структура нейронной сети ANFIS.

2. Воссоздание базы правил и настройка параметров функции принадлежности

В существующих системах с нечеткими нейронными сетями одним из важнейших вопросов является разработка оптимального метода настройки нечеткой базы правил, исходя из обучающей выборки, для получения конструктивных и оптимальных моделей нечетких систем с дальнейшим использованием с практических системах. В основном нечеткие правила описываются экспертами или операторами согласно их знаниям и опыту о соответствующих процессах. Но, в случае разработки нечетких систем иногда довольно тяжело или почти невозможно сразу получить четкие правила или функции принадлежности (membership functions) вследствие неясности, неполноты или сложности систем.

В таких случаях наиболее целесообразным считается генерирование и уточнение нечетких правил, используя специальные обучающие алгоритмы. На данный момент широко используется алгоритм

обратного распространения ошибки для нечетких сетей, который позволяет генерировать оптимальные модели нечетких систем и базы правил, а также градиентный метод обучения. Данный алгоритм был предложен независимо Ичиаши (Ichihashi), Номура (Nomura), Вангом и Менделом (Wang and Mendel) [1].

Без потери общности рассмотрим данный алгоритм на модели, которая содержит две входные лингвистические (x_1 , x_2) и одну выходную переменную y.

Пусть у нас есть база правил, которая содержит все возможные комбинации A_{1i} и A_{2j} (i=1… r ; j=1…k) такие, что:

Правило 1: $A_{11}, A_{21} \Rightarrow y_1$,
k: $A_{11}, A_{2k} \Rightarrow y_k$,
Правило …
$(i-1)k+j$: $A_{1i}, A_{2j} \Rightarrow y_{(i-1)k+j}$,
…
Правило $r \times k$: $A_{1r}, A_{2k} \Rightarrow y_{r \times k}$,

где A_{1i} та A_{2j} - нечеткие множества соответственно X_1 и X_2, а $y_{(i-1)\ k+j}$ - действительное число есть выход правила *(i-1)k+j.*

Исходя из вышесказанного, если нам дан набор величин (x_1 , x_2), то согласно нечеткой базе правил, выход *y* может быть получен на основе методов нечеткой логики.

Прежде всего, обозначим степень выполнения условий следующим образом:

$$h_{(i-1)k+j} = A_{1i}(x_1)A_{2j}(x_2). \tag{7}$$

Согласно методу расчета центра масс системы также запишем:

$$y = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k} h_{(i-1)k+j} y_{(i-1)k+j}}{\sum_{i=1}^{r}\sum_{j=1}^{k} h_{(i-1)k+j}} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k} A_{1i}(x_1)A_{2j}(x_2) y_{(i-1)k+j}}{\sum_{i=1}^{r}\sum_{j=1}^{k} A_{1i}(x_1)A_{2j}(x_2)}. \tag{8}$$

В случае обучения системы с помощью обучающей выборки (x_1 , x_2; y^*), ошибка системы может быть описана как $E = (y^* - y)^2/2$.

Исходя из описания нечетких величин для A_{1i} имеем a_{1i} – центр функции принадлежности, а b_{1i} - ширину для данной функции, аналогично для A_{2j} имеем a_{2j} и b_{2j} . Согласно методу *градиентного спуска* для минимизации ошибки выхода E можно записать формулы для расчета коефициентов a_{1i}, $b_{1i\ j}$ и $y_{(i-1)k+j}$ *(i=1,2,…, r ; j=1,2,…, k)* следующим образом:

$$\begin{aligned} a_{1i}(t+1) &= a_{1i}(t) - \alpha \partial E / \partial a_{1i}(t) = \\ &= a_{1i}(t) - \alpha(\partial E/\partial y)(\partial y/\partial h_{(i-1)k+j})(\partial h_{(i-1)k+j}/\partial A_{1i})(\partial A_{1i}/\partial a_{1i}(t)) = \\ &= a_{1i}(t) + \frac{\alpha(y^* - y)[\sum_{j=1}^{k}(y_{(i-1)k+j} - y)A_{2j}](\partial A_{1i}/\partial a_{1i}(t))}{\sum_{i=1}^{r}\sum_{j=1}^{k} h_{(i-1)k+j}}, \end{aligned} \tag{9}$$

$$\begin{aligned} b_{1i}(t+1) &= b_{1i}(t) - \beta \partial E / \partial b_{1i}(t) = \\ &= b_{1i}(t) - \beta(\partial E/\partial y)(\partial y/\partial h_{(i-1)k+j})(\partial h_{(i-1)k+j}/\partial A_{1i})(\partial A_{1i}/\partial b_{1i}(t)) = \\ &= b_{1i}(t) + \frac{\beta(y^* - y)[\sum_{j=1}^{k}(y_{(i-1)k+j} - y)A_{2j}](\partial A_{1i}/\partial b_{1i}(t))}{\sum_{i=1}^{r}\sum_{j=1}^{k} h_{(i-1)k+j}}, \end{aligned} \tag{10}$$

$$\begin{aligned} y_{(i-1)k+j}(t+1) &= y_{(i-1)k+j}(t) - \gamma \partial E / \partial y_{(i-1)k+j}(t) = \\ &= y_{(i-1)k+j}(t) - \gamma(\partial E/\partial y)(\partial y/\partial y_{(i-1)k+j}(t)) = \\ &= y_{(i-1)k+j}(t) + \frac{\gamma(y^* - y)h_{(i-1)k+j}}{\sum_{i=1}^{r}\sum_{j=1}^{k} h_{(i-1)k+j}}, \end{aligned} \tag{11}$$

где α, β, γ - скорость обучения, а t – означает итерацию в процессе обучения.

3. Постановка задачи прогнозирования

Исходные данные. Макроэкономические показатели экономики Украины представлены в виде статистических временных рядов(см. табл. 1).

ВВП - номинальный ВВП; ОПП – объем промышленной продукции, % к соответствующему периоду предыдущего года;ИРПП – индекс реальной промышленной продукции, % к соответствующему периоду предыдущего года. ИПЦ – индекс потребительских цен, % к соответствующему периоду предыдущего года; ИОЦ- индекс оптовых цен: РДН – реальные доходы населения; М2 – агрегат М2; ПБ – денежная база; ВК – всего кредиты, включая кредиты в иностранной валюте.

Таблица 1 Макроэкономические показатели Украины

Data	ВВП	ОПП	ИРПП	ИПЦ	ИОЦ	РДН	М2	ДБ	ВК
1	2358,3	102,2	100,3	103,1	99,6	113,1	1502	840,1	73,7
2	2308,5	102	99,7	101,2	100,3	111	1522,9	846,1	84,6
3	2267,7	103,7	99,9	101,1	99,2	108,1	1562,4	863,5	96,5
4	2428,5	104,3	102,2	101,2	100,7	118	1621,3	917,7	98,2
5	2535,6	102,8	102,5	101,7	102,2	107,8	1686	977,7	118,2
6	2522,8	104,4	103,1	100,5	104,4	105,8	1751,1	1020,7	138,6
7	2956,4	107,8	102,6	100,7	105,4	112,3	1776,1	1019,8	142,6
8	3025,9	103,4	101,7	100,1	105	109	1812,5	1065,6	157,8
9	3074,5	10,55	101,2	100,4	105,3	106,6	1846,6	1067,9	165.5
10	2854,3	103,9	102,1	101,1	105,3	108,5	1884,6	1078,6	158,9
11	2812,5	100,8	101,6	101,6	100,2	107,8	1930	1128,9	163,4
12	2998,4	103,2	99,8	101,5	105,7	106,9	2119,6	1232,6	262,5
13	2725,6	104,9	100,4	102,4	100,5	114,4	2026,5	1140,1	93,8
14	2853,4	1065	101,4	101,6	101,2	116,8	2108	1240,7	110,3
15	2893,1	106,7	101,3	101,1	103,3	115,4	2208,5	1284,5	125,9
16	3014,2	107,1	101,4	101	103,6	109,1	2311,2	1386,8	130,1
17	3102,6	108,5	99,8	100,8	103,9	119,7	2432,4	1505,7	158,8
18	3110,7	107	100,7	100,8	103,9	113,8	2604,5	1534	158,8
19	3192,4	107,1	102,2	100,7	104,9	112,7	2625,4	1510,8	181,9
20	3304,7	105,5	101,4	99,6	105,9	109,8	2683,2	1500,8	185
21	3205,8	108	101,4	100,3	106,9	112,6	2732,1	1484,5	205,8

Для построения базы правил необходимо определить значимые переменные и их лаги. В качестве степени взаимосвязи между входными переменными $x_1,x_2,\ldots x_n$ и выходной переменной Y используется коэффициент корреляции R, по значению которого отбирались существенные переменные. Далее по величине взаимно-корреляционной функции ВКФ $K_{yx_i}(\tau)$ определяются лаги(запаздывания)/

Была реализована программа , реализующая работу ННС ANFIS с градиентным алгоритмом обучения параметров ФП нечетких правил и проведены ее экспериментальные исследования в задачах прогнозирования макроэкономических показателей. В ходе экспериментов исследовалось влияние числа термов (значений лингвистических переменных) и количества нечетких правил на эффективность прогнозирования.

Определение максимального числа термов

Для определения максимального числа термов выполним прогноз ИПЦ. Значимые переменные и их лаги таковы: ИПЦ(2); ОПП(1); ИРПП(2).

Были проведены следующие эксперименты.

1. Допустим, что все переменные имеют одинаковое число термов(лингвистических значений) – 2. Используется полная база правил.

Результаты работы программы при прогнозировании показателя ИПЦ приведены в табл. 2 и на рис.2 приведен график результатов прогнозирования (среднеквадратичная ошибка прогноза в данном эксперименте СКО=0,53).

Таблица 2

Data	Реальное	Прогноз	Отклонение	Кв. откл.
12	101,5	101,23	0,27	0,0729
13	102,4	101,53	0,87	0,7569
14	101,6	101,39	0,21	0,0441
15	101,1	101,64	0,54	0,2916
16	101,0	100,90	0,10	0,0010
17	100,8	100,72	0,08	0,0064
18	100,8	99,61	1,19	1,4161
19	100,7	99,70	1,00	1,0000
20	99,6	98,69	0,91	0,8281
21	100,3	99,37	0,93	0,8649

2. Определим переменную с наибольшим диапазоном ОПП, увеличим число ее термов до 5, и используем полную систему правил. Результаты прогноза приведены в табл. 3.

Таблица 3

Data	Реальное	Прогноз	Отклонение	Кв. откл.
12	101,5	101,26	0,24	0,0576
13	102,4	101,62	0,78	0,6084
14	101,6	101,41	0,19	0,0361
15	101,1	101,58	0,48	0,2304
16	101,0	101,91	0,09	0,0081
17	100,8	100,73	0,07	0,0049
18	100,8	99,74	1,06	1,1236
19	100,7	99,80	0.90	0,8100
20	99,6	98,79	0,81	0,6561
21	100,3	99,47	0,83	0,6889

Значение СКО= 0,4224. Как видим, значение СКО уменьшилось.

Определение оптимального вида базы правил

1. Из предыдущих экспериментов следует, что для небольшого числа параметров целесообразно использовать полную базу правил. Определим оптимальный вид базы правил для большего числа параметров. Будем выполнять прогноз ИПЦ. Значимые переменные и их лаги таковы: ВВП(0); ИОЦ(3); ОПП(3); ИРПП(0).

Полная база правил состоит из 3*2*3*2*3=108 правил. Результаты работы программы приводятся в табл. 4. При этом значение СКО уменьшилось до 0.3025.

Таблица 4

Data	Реальное	Прогноз	Отклонение	Кв. откл.
12	101,5	101,49	0,01	0,0001
13	102,4	102,60	0,20	0,0400
14	101,6	102,28	0,68	0,4624
15	101,1	101,26	0,16	0,0256
16	101,0	100,56	0,44	0,1936
17	100,8	100,87	0,07	0,0049
18	100,8	100,86	0,06	0,0360
19	100,7	101,63	0.93	0.8649
20	99,6	100,39	0,79	0,6241
21	100,3	101,01	0,71	0,5041

2. Далее выполнялся прогноз, используя урезанную базу правил. База правил выбирается на основании знаний экспертов и содержит 60 правил, что составляет примерно 55% от полной базы. Результаты работы программы приведены в табл. 5.

Значение СКО составило 0.3524. Как видим, оно не значительно увеличилось в сравнении с полной базой правил.

Заключение

Проведенные эксперименты позволяют сделать следующие **выводы**.

1. Нечеткая нейронная сеть архитектуры ANFIS является весьма эффективным инструментом для прогнозирования макроэкономических показателей.
2. Число термов(значений) необходимо выбирать в зависимости от диапазона изменения переменной; чем больше диапазон, тем больше число термов. Для макроэкономических показателей, измеряемых в % относительно предыдущего периода(года), установлено следующее: если диапазон изменения показателя меньше 5, то целесообразно использовать 2 терма, если – больше, то 3. Дальнейшее увеличение термов приводит к сильному росту числа правил, а качество прогноза увеличивается незначительно.
3. Полную систему правил целесообразно использовать для не более, чем трех переменных. Для четырех и более переменных целесообразно использовать неполную базу правил, что существенно упрощает структуру ННС

Литература

1. Осовский С. Нейронные сети для обработки информации.(Пер. с польского И.Д. Рудинского – М.: Финансы и статистика, 2002.- 344с.)
2. Зайченко Ю.П. Основы проектирования интеллектуальных систем. Научное пособие. – К.: Издательский дом «Слово», 2004. – 352 с.
3. Зайченко Ю.П., Севаее Фатма, Титаренко К.М., Титаренко Н.В. Исследование нечетких нейронных сетей в макроэкономическом прогнозировании. Системные исследования и информационные технологии. – 2004.- №2.- с.70-86.
4. Jyh-Shing, Roger Jang. ANFIS: Adaptive-Network Based Fuzzy Inference System. Department of Electrical Engineering and Computer Science. University of California, Berkely, 1995.

Информация об авторах

Zaychenko Yuri – Dr. of science, professor, National technical University of Ukraine “KPI”, Institute for Applied System Analysis ,dean, Address: Peremogy ave. 37, Kiev , Ukraine, phone: 8044- 241-86-93, e-mail : zaych@i.com.ua, zaych@mmsa.ntu-kpi.kiev.ua

Sevaee Fatma – (Lybia), post- graduate student, NTUU “KPI”, address: Peremogy ave. 37, Kiev, Ukraine, phone: 8-044- 2272797

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РЕСТРУКТУРИЗАЦИИ СЛОЖНЫХ ТЕХНИКО-ЭКОНОМИЧЕСКИХ СТРУКТУР

Май Корнийчук, **Инна Совтус**, **Евгений Цареградский**

Резюме: *Излагается исследование и разработка математической модели оптимального распределения ресурсов (в основном финансовых) для обеспечения нового (повышенного) качества (надежности) сложной системы, относительно которой принято решение о ее реструктуризации. Итоговая модель дает ответы (алгоритм расчета) на вопросы: сколько элементов системы выделить на модернизацию, какие именно элементы, до какого уровня глубины нужна модернизация каждого из выделенных, причем оптимальные ответы по критерию минимизации финансовых расходов.*

Ключевые слова: *система, реструктуризация, качество, надежность.*

Основной текст

При разработке новых сложных систем, а также повышении их эффективности в процессе эксплуатации важным фактором повышения адекватности и достоверности математических моделей оценивания уровня их надежности является способность описания, формализации и учета в этих моделях возможности управления надежностью [1]. Приращение надежности *и* за счет рационального управления надежностью достигается совершенствованием алгоритма вариаций режима эксплуатации системы, варьированием мероприятий по техническому и профилактическому обслуживанию, ибо уменьшение интенсивности отказов после рациональной процедуры регламентных работ зависит от уровня оптимизации этой процедуры. Уменьшение интенсивности потока отказов, изменение его вероятностной структуры ограниченого последействия может достигаться также специальными режимами внешних воздействий. Так, например, отдельные виды интегральных схем при радиоактивном облучении резко повышают точность нахождения параметров в необходимых границах [1], хотя при этом время их жизни существенно уменьшается. Тем не менее вполне может быть оправдано проведение такой процедуры, когда системой выполняется важная и ответственная задача, значение которой позволяет пренебречь уменьшением общего времени жизни элемента за счет строгого сохранения параметров в определенных пределах, хотя в меньшем временном промежутке. В таких ситуациях возникает задача об управлении надежностью с целью оптимизации системы по определенному критерию.

Будем исследовать такой вариант сложной системы, который характеризуется линейной блок-схемой ее процесса функционирования (ПФ). Это может быть сложная экономическая структура, или сложная техническая система, или такая система, которая имеет и экономические, и технические свойства, то есть является сложной технико-экономической структурой (СТЭС). Пусть такая сложная система содержит n подсистем, функционально соединеных последовательно, и пусть она имеет определенный уровень качества функционирования, который мы для удобства будем интерпретировать его же уровнем надежности P. На определенной стадии эксплуатации системы (например, в силу развития рыночных отношений страны) наступает такой момент, когда состояние эффективности ПФ системы уровня P становится недостаточным, и возникает проблема его повышения до уровня $P^* = P + + \Delta P$. На этом этапе жизнедеятельности многих систем возникают проблемы по их реорганизации и реструктуризации, причем весомая часть общих инвестиций должна быть выделена на инновации.

Активные инвестиции в инновационный процесс элемента могут привести к существенному приросту надежности элемента или вероятности пребывания параметров в нужных пределах [2, 3]. Этим самым повышается надежность системы вцелом. Увеличивать ее можно было бы сколь угодно долго, если бы это не было связано с материальными затратами. Расширение количества модернизированных элементов и их параметров, увеличение глубины инновационного процесса каждого элемента приводит к соответствующему росту расходов на осуществление всей реструктуризации. Объективно существует

некоторая точка (значение уровня доработки), после которой увеличивать расходы на реструктуризацию из практических соображений становится нецелесообразным. Возникает задание по отысканию такой точки, а именно: где необходимо остановиться при выборе уровня глубины доработки отдельного элемента. Для системы в целом проблема состоит в том, какое количество элементов подвергнуть модернизации, чтобы не утратилась экономическая целесообразность выполняемой работы. И далее, когда вопрос для системы в целом о количестве выбранных для модернизации элементов, решен, то тут же возникнет следующая проблема: какие именно среди n элементов нужно выбирать, и до какого именно уровня глубины каждый из выбранных элементов нужно доработать? Понятно, что есть смысл решать эти задачи только с учетом затрат на доработку. При этом уровень доработки элемента с точки зрения надежности системы будет определяться новым необходимым уровнем надежности этого элемента по сравнению с надежностью p, заложенной в процессе производства элемента. Глубину доработки, модернизацию i-го элемента системы количественно можна охарактеризовать приростом x_i надежности этого элемента за счет инноваций. При этом, если i-й элемент или узел не дорабатывается (не модернизируется), то логично положить значение $x_i = 0$. Тогда глубину доработки или реконструкции всей системы можно охарактеризовать n-мерным вектором $\vec{x} = \{x_1, x_2, ..., x_n\}$, где x_i, $i = \overline{1,n}$ – суть величины, введенные выше. Если же обозначить через p_i надежность i-го елемента системы до начала доработки, то после реконструкции надежность элемента будет равна $p_i + x_i$. Надежность системы до проведения модернизации была функцией $P = P(\vec{p}),$ после модернизации она увеличится и приобретет новое значение:

$$P = P(\vec{p}, \vec{x}).$$

Необходимо исследовать зависимость расходов на доработку элемента от его качества, приобретенного за счет доработки, то есть от дополнительного прироста вероятности x_i, учитывая при этом приращение надежности за счет управления. В дальнейшем будем называть эту зависимость функцией стоимости доработки элемента [4] и обозначим ее через $K(x)$. Как и при решении любой другой практической задачи, мы не можем воспользоваться тут реальной (эмпирической) зависимостью расходов на доработку от прироста надежности. Для использования этой зависимости ее необходимо формализовать, построив аналитическую функцию, которая достаточно полно аппроксимировала бы данную эмпирическую зависимость, то есть необходимо построить математическую модель зависимости. Воспользуемся ее математической моделью, исследованной и построенной в работе [4], обоснованной экономическими аналогами [5], а именно:

$$K(x) = \frac{Ax}{q - u - x},$$

где $q = 1 - p$, а коэффициент A определяется из экспериментальных данных для конкретной системы. Поскольку нами рассматривается сложная система с последовательным соединением подсистем (элементов), то считаем, что система перестает быть надежной (наступает отказ системы), если хотя бы один из n параметров (элементов) системы вышел за допустимые пределы. В предположении последовательного соединения элементов и учета приращения надежности за счет модернизации и управления уравнение надежности системы будет иметь форму:

$$P = \prod_{i=1}^{n} (p_i + u_i + x_i). \tag{1}$$

и при этом суммарная функция расходов на модернизацию системы

$$K(\vec{x}) = \sum_{i=1}^{n} \frac{A_i x_i}{q_i - u_i - x_i}. \tag{2}$$

Заметим, что функция расходов на модернизацию (2) и уравнение надежности (1) имеют смысл только при определенных ограничениях, которые являются следствием их физического содержания, а именно:

1. $K(x) \geq 0$;
2. $K(0) = 0$;
3. $\lim\limits_{x \uparrow 1-p} K(x) = \infty.$ (3)

$x_i \geq 0, \quad x_i < q_i - u_i, \quad i = \overline{1,n}$.

Теперь задача свелась к задаче нелинейного математического программирования, а точнее к нахождению такого решения уравнения (1), которое обеспечивало бы условный, при условиях (3), минимум функции (2). Как мы уже отмечали, подобные задачи теории нелинейного программирования еще требуют исследований, и поэтому мы очень ограничены в выборе методов для нахождения решений. Для определения оптимальных требований к доработке каждого элемента, то есть для нахождения условного минимума функции (2) с учетом ограничения на переменные (1) и граничных условий воспользуемся модифицированным [6] методом неопределенных множителей Лагранжа.

Сущность модифицированного метода состоит в том, что непосредственное использование его невозможно из-за открытой (3) области ограничений, а потому используем одно содержательное равенство-ограничение, а именно (1), которое информационно поглощает в этом смысле остальные.

Тогда строим функцию типа функции Лагранжа, имеем:

$$f(\vec{x},\lambda) = \sum_{i=1}^{n} \frac{A_i x_i}{q_i - u_i - x_i} + \lambda \left[\prod_{i=1}^{n} (p_i + u_i + x_i) - P^* \right]. \tag{4}$$

С построения функции (4) видно, что в ней использовано только ограничение в виде уравнения (1) и не учтены граничные условия в виде неравенств. Такой шаг, вообще говоря, является не совсем верным, но в данном случае это оправдано информационной емкостью ограничения (1). Ниже будет доказано, что предложенное использование ограничения (1) приводит к обязательному выполнению граничных условий (3), а процедура отыскания решения это обеспечивает. Вместе с тем предложенный способ освобождает ход решения задачи выбранным методом от аналитической громоздкости введения огромного количества неизвестных дополнительных множителей типа множителей Лагранжа.

Для определения составляющих вектора $\vec{x}$, то есть оптимального набора x_i, i = 1, 2, ..., n построим систему из n уравнений, вычислив частные производные от функции Лагранжа $f(\vec{x},\lambda)$ по всем переменным x_i и приравняв их к нулю. Получим:

$$\frac{f}{x_i} = \frac{A_i q_i}{(q_i - u_i - x_i)^2} - \frac{\lambda}{p_i + u_i + x_i} \prod_{j=1}^{n} (p_j + u_j + x_j) = 0, \quad i = 1, 2, ..., n. \tag{5}$$

Система уравнений (5) содержит (n+1) неизвестное: n неизвестных x_i и (n+1)-е неизвестное λ. Если к этой системе присовокупить уравнение (1), то получим полную систему n + 1 уравнений с n + 1 неизвестными. Полученная система уравнений (5), (1) имеет решение, и к тому же единственное, что будет обосновано ниже.

Для удобства преобразуем уравнения (5) к виду:

$$-\lambda \prod_{j=1}^{n} (p_j + u_j + x_j) = \frac{A_i q_i (p_i + u_i + x_i)}{(q_i - u_i - x_i)^2}, \quad i = 1, 2, ..., n, \tag{6}$$

выделив влево их одинаковые части. Поскольку выражение, которое находится в левой части каждого уравнения (6) является константой по отношению к индексу i, то и правые части уравнений должны совпадать. Допустим из некоторых физических или практических условий можно придти к выводу, что для k-го элемента системы обязательно необходима доработка, что означает $x_k > 0$. Воспользуемся k-м уравнением системы (6) и вместо выражения в левой части подставим равное ему i-е выражение из текущего уравнения системы. Тогда система уравнений (6) перепишется следующим образом:

$$\frac{A_i q_i (p_i + u_i + x_i)}{(q_i - u_i - x_i)^2} = \frac{A_k q_k (p_k + u_k + x_k)}{(q_k - u_k - x_k)^2}, \quad i = 1, 2, ..., k-1, k+1, ... n. \tag{7}$$

Поскольку мы зафиксировали выбор *k*-го элемента, то выражение справа имеет функцию, которая не зависит от *i*, поэтому обозначим общую правую часть соотношения (7) через B_k, а именно:

$$B_k = \frac{A_k q_k (p_k + u_k + x_k)}{(q_k - u_k - x_k)^2}. \tag{8}$$

Далее переходим от системы (7) к классическому виду квадратичного уравнения:

$$B_k (x_i + u_i - q_i)^2 - A_i q_i (x_i + u_i - q_i) - A_i q_i = 0, \quad i = 1, 2, ..., k-1, k+1, ... n. \tag{9}$$

Получена система уравнений второй степени (9), которая является эквивалентной к системе уравнений (5). Исследуем ее на существование и единственность решения. С этой целью введем функции

$$_i(x_i) = A_i q_i (p_i + u_i + x_i),$$
$$\psi_i(x_i) = B_k (x_i + u_i - q_i)^2,$$
$$i = 1, 2, ..., k-1, k+1, ... n.$$

соответственно правой и левой частей уравнения $\varphi(x_i) = \psi(x_i)$ системы, эквивалентной (9).

Поскольку для *i*-го уравнения системы при значении $x_i = q_i - u_i$ функции приобретают значения $_i(q_i - u_i) = A_i q_i > 0$ і $\psi_i(q_i - u_i) = 0$, то имеет место соотношение $_i(q_i - u_i) > \psi_i(q_i - u_i)$. Графически это означает, что прямая $\varphi_i(x_i)$ пересечет параболу $\psi_i(x_i)$ в двух точках x_i: одна из точек x_i расположена левее $q_i - u_i$, то есть $x_i < q_i - u_i$, а другая – справа от $q_i - u_i$, то есть $x_i > q_i - u_i$. Таким образом, вторые корни системы уравнений (9) не имеют смысла, то есть не представляют для нас интереса, поскольку они лежат вне области физически дозволенных значений x_i, вне области существования решения поставленной задачи и не удовлетворяют граничным условиям (3). Из свойств монотонности и непрерывности функций $\varphi_i(x_i)$ и $\psi_i(x_i)$ вытекает, что точка пересечения этих кривых левее значения $x_i = q_i - u_i$ является единственной. Тогда единственные левые корни системы квадратных уравнений (9) имеют вид:

$$x_i = q_i - u_i + \frac{1 - \sqrt{1 + 2\alpha_i}}{\alpha_i}, \quad \text{где } \alpha_i = 2B_k / A_i q_i, \quad i = 1, 2, ..., k-1, k+1, ..., n, \tag{10}$$

Вычислим значения введенных функций в левых концах промежутка [0; $q_i - u_i$) – области определения переменных x_i, поскольку граничные значения их в правых концах мы уже сравнили. При $x_i = 0$ функция $\varphi_i(x_i)$ принимает значение $\varphi_i(0) = A_i q_i (p_i + u_i)$, а функция $\psi_i(x_i)$ превращается в число $\psi_i(0) = B_k (q_i - u_i)^2$.

Для тех *i*-х элементов, для которых $\varphi_i(0) > \psi_i(0)$, точка пересечения кривых $\varphi_i(x_i)$ и $\psi_i(x_i)$ лежит левее нуля, а значения корня получаются $x_i < 0$, что противоречит физической природе переменной x_i. Это означает, что дорабатывать такой элемент не нужно, и следует положить $x_i = 0$, при этом условие (3) будет обеспечено. Если для определенного *i*-го элемента значения введенных функций сравняются $\varphi_i(0) = \psi_i(0)$, то есть $\frac{A_i q_i (p_i + u_i)}{(q_i - u_i)^2} = B_k$, то точка пересечения графиков $\varphi_i(x_i)$ и $\psi_i(x_i)$ попадет на ось ординат. При этом получится корень $x_i = 0$, что также означает: дорабатывать элемент не нужно. Таким образом граничное условие (3) обеспечилось автоматически, что и обосновывает высказанное выше утверждение.

Как итог имеем: по выбранному значению уровня прироста надежности x_k для k-го элемента системы однозначно определяем уровень доработки i-го (i = 1, 2, ..., k – 1, k + 1, ..., n) элемента x_i при помощи соотношений:

$$x_i = \begin{cases} 0, & \dfrac{A_i q_i (p_i + u_i)}{(q_i - u_i)^2} \geq B_k, \\ q_i - u_i + \dfrac{1 - \sqrt{1 + 2\alpha_i}}{\alpha_i}, & \dfrac{A_i q_i (p_i + u_i)}{(q_i - u_i)^2} < B_k, \end{cases} \quad (11)$$
$$i = 1, 2, ..., k-1, k+1, ..., n.$$

Построенная математическая модель (11) решает поставленную задачу распределения ресурсов (финансовых) на оптимальную реструктуризацию сложной системы с целью повышения ее качества (надежности) с учетом управления надежностью от уровня P до уровня $P^* = P + \Delta P$.

Заключение

Используя исходную информацию о системе (надежность, стоимость элементов и узлов, имеющийся неудовлетворительный уровень надежности P, новый P^* требуемый уровень качества системы в достижении которого состоит смысл реструктуризации), построена математическая модель (11), которая дает ответы на все вопросы достижения конкретных решений, которые сформулированы в постановке проблемы и текстуально сформулированы в резюме в начале работы. Отметим простоту вычислений, прикладную и прозрачную их интерпретацию при практическом использовании разработанной модели.

Библиография

1. *Северцев Н.А.* Надежность сложных систем в эксплуатации и отработке. – М.: Высшая школа, 1989. – 432 с.
2. *Валтер Я.* Стохастические модели в экономике. – М.: Статистика, 1976. – 232 с.
3. *Саркисян С.А., Каспин В.И., Лисичкин В.А.* Теория прогнозирования и принятия решений. – М.: Высшая школа, 1977. – 352 с.
4. *Корнийчук М.Т.* Математические модели оптимизации и оценивания надежности и эффективности функционирования сложных РТС. – К.: КВИРТУ ПВО, 1980.–280 с.
5. *Мескон Н.Х., Альберт М., Хедоурн Ф.* Основы менеджмента. Пер. с англ. – М.: «Дело», 1992. – 702 с.
6. *Корнійчук М.Т., Совтус І.К.* Стохастичні моделі інформаційних технологій оптимізації надійності складних систем. – К.: КВІУЗ, 2000. – 316 с.

Информация об авторах

Май Корнийчук – д.т.н., проф. Киевский национальный экономический университет, пр.Победы, 54/1, Киев-03680, Украина, e-mail: sovtus@bigmir.net.

Инна Совтус – д.т.н., проф. Киевский национальный экономический университет, пр.Победы, 54/1, Киев-03680, Украина, e-mail: sovtus@bigmir.net.

Евгений Цареградский – соискатель, Киевский национальный экономический университет, пр.Победы, 54/1, Киев-03680, Украина, YTsaregradskyy@bmw.ua.

Section 5. Mathematical Foundations of AI

5.1. Algorithms

RAISING EFFICIENCY OF COMBINATORIAL ALGORITHMS BY RANDOMIZED PARALLELIZATION

Arkadij D. Zakrevskij

***Abstract**: A new approach is proposed to deal with some hard combinatorial optimization problems, which admit a certain reformulation. Considering such a problem, several similar problems are prepared differing in initial data but having the same set of solutions. They are solved in parallel until one of them will be solved, and that solution is accepted. Notwithstanding the evident overhead, the whole run-time could be significantly reduced due to dispersion of velocities of combinatorial search in regarded cases. The efficiency of this approach is investigated on the concrete problem of finding short solutions of non-deterministic system of linear logical equations.*

***Keywords**: combinatorial problems, combinatorial search, parallel computations, randomization, run-time, acceleration.*

Introduction

There exist a variety of various hard combinatorial optimization problems, which could be more quickly solved when changing each of them for a selection of other problems solved in parallel. These new problems could seem quite different, but nevertheless they can be made equivalent to the initial problem, that means they will have the same set of solutions. The overhead expenses for constructing these sets of new problems and dealing with them can be compensated with interest by the essential acceleration of the search for solution of the considered problem.

That idea is demonstrated below by the problem of finding a short solution of an undefined system of linear logical equations.

Let us regard a system of m *linear* Boolean equations with n *variables*

$$\begin{aligned} &a_1^1x_1 \oplus a_1^2x_2 \oplus \ldots \oplus a_1^nx_n = y_1 ,\\ &a_2^1x_1 \oplus a_2^2x_2 \oplus \ldots \oplus a_2^nx_n = y_2 ,\\ &\qquad\qquad \ldots \\ &a_m^1x_1 \oplus a_m^2x_2 \oplus \ldots \oplus a_m^nx_n = y_m , \end{aligned} \qquad (1)$$

where a_i^j are the coefficients of the system, x_j – unknown variables ($x_j \in \{0, 1\}$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$), $\oplus$ is the EXOR operator.

This system can be represented in a compact matrix form:

$$\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{y} , \qquad (2)$$

where $\boldsymbol{A}$ is a Boolean $m \times n$ coefficient matrix, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ – Boolean vectors. AND operator is used as an internal one, and EXOR as the external one for the matrix multiplication.

$$\bigoplus_{j=1}^{n} a_i^j x_j = y_i . \quad (3)$$

Solutions (roots) of this system are the sets of variable values, for which all equations will be satisfied. In this case EXOR sum of the columns of matrix ***A***, corresponding to the component-wise sum by modulo 2 of the columns of ***A***, which correspond to the variables having value 1, should be equal to the column ***y***. The system can be deterministic (having the single solution), non-deterministic (several solutions exist) or contradictory, where there are no solutions [1, 2]. Let us assume that $n > m$ and all rows of ***A*** are linearly independent. Then the number of possible solutions is equal to 2^{n-m}.

A great property of the linear equations is that they can be solved in regard of any variable. An effective Gauss method of variable exclusion is based on this property [3] and allows finding the only solution for a deterministic system or the set of all possible solutions for a non-deterministic one. But it is not sufficient when some optimal solution should be selected from that set which could be very large.

Looking for Short Solutions of Systems of Linear Logical Equations

Solving a non-deterministic system of linear logical equations, it is necessary sometimes to find a solution with minimum number of variables taking value 1 – such a solution is called *shortest*. That is important in many practical cases, for instance, when logic circuits in AND/EXOR basic are synthesized [4-6] or when some problems of information security are considered.

Evidently, a shortest solution could be found by means of selecting from matrix ***A*** one by one all different combinations of columns, consisting first of 1 column, then of 2 columns, etc. and examining each of them to see if their sum equals vector ***y***. As soon as it happens, the current combination is accepted as the sought-for solution. That moment could be forecasted. If the weight of the shortest solution (the number of 1s in vector ***x***) is w, the number N of checked combinations is defined approximately by the formula (where C_n^i is the number of i-element subsets taken from a set of n elements).

$$N = \sum_{I=0}^{w} C_n^i \quad (4)$$

and could be very large, as is demonstrated below.

It was shown [7], that the expected weight γ of the shortest solution of an SLLE with parameters m and n can be estimated before finding the solution itself. We represent this weight as a function $\gamma(m, n)$. First we find the mathematical expectation $\alpha(m, n, k)$ of the number of solutions with weight k. We assume that the considered system was randomly generated, which means that each element of ***A*** takes value 1 with the probability 0.5 and any two elements are independent of each other. Then the probability that a randomly selected column subset in matrix ***A*** is a solution equals 2^{-m} (probability that two randomly generated Boolean vectors of size m are equal). Since the number of all such subsets having k elements equals $C_n^k = n! / ((n-k)!\, k!)$, we get:

$$\alpha(m, n, k) = C_n^k 2^{-m}. \quad (5)$$

Similarly, we denote as $\beta(m, n, k)$ the expected number of the solutions with weight not greater than k:

$$\beta(m, n, k) = \sum_{i=0}^{k} C_n^i 2^{-m}. \quad (6)$$

Now, the expected weight γ of the shortest solution can be estimated well enough by the maximal value of k, for which $\beta(m, n, k) < 1$:

$$\gamma(m, n) = k, \quad \text{where } \beta(m, n, k) < 1 \le \beta(m, n, k+1). \quad (7)$$

For example, for a system of 40 equations with 70 variables the expected weight γ equals 10, and reaches 31 when $m = 100$ and $n = 130$.

In the last case, the described above simple algorithm should check about 10^{30} combinations, according to formula (2) in which $w = \gamma$. Examining them with the speed of one million combinations per second we need about **30 000 000 000 000 000 years** to find the solution. Practically impossible!

A great acceleration can be achieved by using Gaussian method of variables exclusion [3] developed for solving systems of linear equations with real variables and adjusted for Boolean variables [5]. It enables to avoid checking all subsets of columns from ***A*** which have up to *w* columns, when only one of 2^{n-m} regarded combinations presents some roots of the system.

Its main idea consists in transforming the extended matrix of the system (matrix ***A*** with the added column ***y***) to the *canonical form*. A maximal subset of *m* linear independent columns (does not matter which one) is selected from ***A*** and by means of equivalent matrix transformations (adding one row to another) is transformed to **I**-matrix, with 1s only on the main diagonal. That part is called a basic, the rest $n - m$ columns of the transformed matrix ***A*** constitute a *remainder*. The column y is changed by that, too.

According to this method the subsets of the remainder are regarded, i.e. combinations selected from the set which has only $n - m$ columns (not all *n* columns!). It is easy to show that every of these combinations enables to get a solution of the considered system (any sum of its elements can be supplemented with some columns from matrix **I** to make it equal to ***y***). When we are looking for a shortest solution using this method, we have to consider different subsets of columns from the remainder, find the solution for each such subset and select a subset, which generates the shortest solution. If it is known that the weight of the shortest solution is not greater than *w*, then the level of search (the cardinality of inspected subsets) is restricted by *w*. Note that if $w \geq n - m$, then all 2^{n-m} subsets must be searched through.

Now, for the same example ($m = 100$ and $n = 130$), $N \cong 10^9$, which means that the run-time of Gaussian method is about only **17 minutes**.

Unfortunately (for practical application), it rises very quickly when increasing parameters *m* and *n*. For example, when $m = 625$ and $n = 700$, the number of checked combinations equals 2^{75} and the run-time surpasses one milliard years.

It could be significantly reduced when we know that there exists a short solution with weight *w* less than γ. For example, the following situation could be imagined: a random matrix ***A*** is generated, then *w* columns are selected by random, and their component-wise modulo 2 sum is defined as vector ***y***. In that case the searching (checking different solutions one by one in order to find the shortest) could be interrupted as soon as the weight *v* of current solution satisfies the inequality $v < \gamma$ and we conclude that the sought-for solution is found. This idea lies in the base of the recognition method [8, 9].

The Level of Search and its Dispersion

Suppose, a short solution exists presented by *w* columns of matrix ***A*** (their sum equals vector ***y***). Let *p* of them belong to the remainder. Then this solution will be found when such subsets of columns from the remainder are checked which have exactly *p* elements. In other words, the solution will be found on the *level of search* L when this one equals *p*.

That quantity depends, first, on the regarded example (how matrix ***A*** was generated) and, second, on the way of getting the remainder (how *m* linearly independent columns were selected from initial matrix ***A***, which determines splitting that matrix into two parts, basic and remainder).

An experiment was conducted, where 10 different random matrices ***A*** were generated with parameters $n = 500$ and $m = 430$. In each of them 30 different subsets of linearly independent columns were selected and corresponding canonical forms were constructed, which produced 300 variants of the remainder. A set of 75 columns was chosen from ***A*** by random constituting a short solution - their sum was accepted as vector ***y***.

All intersections of that set with remainders were found. The cardinalities of these intersections (numbers of columns in them) define the level of search *L* on which the considered short solution will be found. The values of *L* for all 300 variants are shown in Table 1. Its columns correspond to different examples (*i*) of system and the rows correspond to different remainders in them (*j*) selected at random. As one can see, this quantity is subjected to a great dispersion.

The values of run-time (using PC COMPAC Presario, 1000 MH) corresponding to different levels are presented in Table 2, measured in seconds (*s*), minutes (*m*), hours (*h*), days (*d*) and years (*y*). The distribution of levels is shown on the right side of the table, *N* indicating the number of entries with value *L* in Table 1.

Table 1. Dispersion of level L in space of examples and remainders

j \ *i*	1	2	3	4	5	6	7	8	9	10
1	10	13	12	12	16	11	13	10	12	13
2	8	9	13	9	9	10	11	10	7	9
3	10	10	11	9	11	13	11	10	12	14
4	10	8	11	11	10	8	12	10	13	11
5	10	11	14	6	11	8	17	15	10	6
6	12	6	7	10	12	11	11	12	8	16
7	9	11	8	9	10	12	8	9	8	17
8	8	15	6	13	14	16	15	12	12	12
9	10	13	9	5	14	11	13	6	16	10
10	14	15	9	8	11	10	8	6	13	12
11	14	12	10	12	6	10	12	15	13	9
12	8	9	12	16	16	7	18	11	9	12
13	9	11	6	13	15	9	9	10	10	9
14	14	12	9	14	6	9	7	13	11	13
15	9	5	13	12	11	10	7	13	9	16
16	13	16	6	15	10	7	4	10	11	13
17	6	10	15	13	11	11	13	11	9	12
18	9	9	10	11	10	6	14	8	10	12
19	5	15	7	11	10	16	9	12	9	13
20	11	11	7	8	7	7	11	9	13	10
21	8	12	10	8	8	13	9	6	12	8
22	15	8	11	8	16	6	12	11	15	5
23	10	12	7	14	11	12	13	13	9	9
24	10	9	10	12	4	10	12	11	10	10
25	9	11	10	10	11	11	14	9	7	19
26	11	14	8	10	11	5	14	8	12	10
27	11	12	14	9	10	7	14	11	11	8
28	18	12	13	12	6	8	8	10	12	11
29	12	12	7	9	14	12	7	8	9	12
30	10	7	6	11	13	8	11	15	11	18

Table 2. Dependence of run-time *T* on level *L*

L	*T*	*N*	
4	1.6s	2	11
5	22s	5	11111
6	4m	16	1111111111111111
7	38m	16	1111111111111111
8	5h	27	111111111111111111111111111
9	1.5d	36	111111111111111111111111111111111111
10	10d	44	11111111111111111111111111111111111111111111
11	55d	44	11111111111111111111111111111111111111111111
12	280d	39	111111111111111111111111111111111111111
13	3.6y	27	111111111111111111111111111
14	15y	16	1111111111111111
15	59y	12	111111111111
16	214y	10	1111111111
17	713y	2	11
18	2213y	3	111
19	6394y	1	1

Note that fulfilled experiments were virtual ones: the run-time was not measured directly but was forecasted by the method suggested in [10] and ensuring a rather good approximation.

It logically follows from these tables that efficient algorithms for finding short solutions may be constructed which solve the problem in quasi-parallel mode using a set of q canonical forms of the system ($\boldsymbol{A}$, $\boldsymbol{z}$) with different basics selected at random. Using these forms one by one, such an algorithm consecutively increases the cardinality of checked combinations of columns taken from the remainders and terminates when finding the short solution.

Experiments

Some results of the program virtual implementation of that algorithm for different values of q are shown in Table 3. Thirty examples of random systems with parameters n = 1000, m = 900 and the weight of the short solution w = 100 have been generated and solved. The following denotation is used in the table: **No** is the number of the regarded example, **N** is the number of form where its solution was found, **L** is the level at which it was found, and **T** is the time spent for it.

Table 3. Results of solving non-deterministic systems of linear logical equations with parameters n = 1000, m = 900, w = 100.

	q = 1			q = 10			q = 30			q = 300		
No	**N**	**L**	**T**	**N**	**L**	**T**	**N**	**L**	**T**	**N**	**L**	**T**
1	1	8	7d	9	6	16h	9	6	18h	33	4	19m
2	1	9	69d	8	6	14h	16	5	2h	254	2	4m
3	1	10	2y	7	7	7d	28	6	2d	234	2	4m
4	1	13	776y	3	6	5h	3	6	8h	117	3	5m
5	1	3	4s	1	3	4s	1	3	5s	1	3	4m
6	1	10	2y	5	7	5d	26	5	3h	56	2	4m
7	1	12	112y	9	4	3m	9	4	3m	57	3	5m
8	1	8	7d	10	5	1h	10	5	1h	106	3	5m
9	1	12	112y	10	5	1h	28	4	10m	141	3	6m
10	1	9	69d	2	6	4h	2	6	6h	95	2	4m
11	1	13	776y	7	8	82d	15	4	5m	50	2	4m
12	1	13	776y	9	8	104d	14	5	2h	117	3	5m
13	1	10	2y	8	6	14h	8	6	16h	35	3	4m
14	1	6	58m	1	6	2h	1	6	4h	39	3	4m
15	1	6	58m	1	6	2h	11	5	1h	134	3	6m
16	1	14	4954y	2	8	27d	28	4	10m	205	3	7m
17	1	6	58m	1	6	2h	14	5	2h	49	3	4m
18	1	10	2y	2	7	2d	27	5	3h	285	2	4m
19	1	13	776y	8	7	8d	8	7	9d	84	3	5m
20	1	10	2y	2	6	4h	2	6	6h	203	4	1,3h
21	1	8	7d	7	5	45m	7	5	52m	93	4	39m
22	1	7	13h	10	6	17h	27	4	9m	190	3	6m
23	1	12	112y	2	5	13m	16	2	4s	16	2	4m
24	1	15	29185y	4	7	4d	24	6	2d	226	2	4m
25	1	10	2y	2	6	4h	2	6	6h	46	3	4m
26	1	13	776y	9	7	9d	16	5	2h	112	4	45m
27	1	10	2y	6	8	71d	16	4	6m	281	3	8m
28	1	12	112y	7	8	82d	18	6	1d	87	3	5m
29	1	12	112y	6	4	2m	6	4	2m	254	2	4m
30	1	12	112y	7	8	82d	22	6	2d	31	4	18m
The sum:			38704y			1,3y			20d			5,3h

Note that value 1 of parameter q corresponds to the pure Gaussian method dealing with one canonical form of the system. Changing it for pseudo-parallel algorithm on the base of 300 forms we accelerate finding the short solution on an average in 46 million times. The immense acceleration!

Conclusion

A new approach to solving hard combinatorial optimization problems is suggested, demonstrated on the problem of finding a short solution of a non-deterministic system of linear logical equations. Its idea is in changing the regarded problem for a set of other similar problems equivalent to the given one and solving them in parallel. The run-time could be considerably reduced by that, possibly in many millions times, as computer experiments show.

Acknowledgements

The work had been performed thanks to the partly support of ISTC (Project B-986) and the Fond of Fundamental Researches of Belarus.

Bibliography

1. Kostrikin A., Manin Y. Linear algebra and geometry. Translated from the second Russian (1986) edition by M. E. Alferieff. Revised reprint of the 1989 English edition. Gordon and Breach, 1997.
2. Lankaster P. Theory of matrices, Academic Press, New York – London, 1969.
3. Gauss C.F.Beitrage zur Theorie der algebraischen Gleichungen, Gött, 1849.
4. Sasao T. Representations of logic functions using EXOR operators. – In "Representations of discrete functions" (ed. by T. Sasao and M. Fujita) – Kluwer Academic Publishers, Boston/London/Dordrecht, 1996, pp. 29-54.
5. Zakrevskij A.D. Looking for shortest solutions of systems of linear logical equations: theory and applications in logic design. - 2.Workshop "Boolesche Probleme", 19./20. September 1996, Freiberg/Sachsen, pp. 63-69.
6. Zakrevskij A.D., Toropov N.R. Polynomial representation of partial Boolean functions and systems. – Minsk, Institute of technical cybernetics, Belarusian NAS, 2001. (In Russian).
7. Zakrevskij A.D., Vasilkova I.V. Fast algorithm to find the shortest solution of the system of linear logical equations. - Problems of logical design. - Minsk, 2002. – Pp. 5-12. (In Russian).
8. Zakrevskij A.D. Randomization of a parallel algorithm for solving undefined systems of linear logical equations. – Proceedings of the International Workshop on Discrete-Event System Design – DESDes'04. – University of Zielona Gora Press, Poland, 2004, pp. 97-102.
9. Zakrevskij A.D. Efficient methods for finding shortest solutions of systems of linear logical equations. – Control Sciences, 2003, No 4, pp. 16-22. (In Russian).
10. Zakrevskij A.D., Vasilkova I.V. (2003). Forecasting the run-time of combinatorial algorithms implementation. -Methods of logical design, issue 2. Minsk: UIIP of NAS of Belarus, pp. 26-32. (In Russian).

Author's Information

Arkadij Zakrevskij – United Institute of Informatics Problems of the NAS of Belarus, Surganov Str. 6, 220012 Minsk, Belarus; e-mail: zakr@newman.bas-net.by

SPECIFYING AGENT INTERACTION PROTOCOLS WITH PARALLEL CONTROL ALGORITHMS

Dmitry Cheremisinov, Liudmila Cheremisinova

Abstract. *The purpose of the paper is to explore the possibility of applying existing formal theories of description of distributed and concurrent systems to interaction protocols for real-time multi-agent systems. In particular it is shown how the language, proposed for description of parallel logical control algorithms and rooted in the Petri net formalism, can be used for the modelling of complex concurrent conversations between agents in a multi-agent system. It is demonstrated with a known example of English auction on how to specify an agent interaction protocol using considered means.*

Keywords: *multi-agent system, interaction protocols, parallel control algorithm*

Introduction

Agents are becoming one of the most important topics in distributed and autonomous decentralized systems, and there are increasing attempts to use agent technologies to develop large-scale commercial and industrial software systems. Over the last decade, the specification, design, verification and application of a particular type of agents, called BDI (belief, desire, intention) agents [1], have received a great deal of attention. BDI agents are systems that are situated in changing environment receive continuous perceptual input and take actions to affect

their environment based on their internal state. In particular, there are many efforts aimed at developing agent-oriented designs that are typically structured as multi-agent systems (MASs). MAS is a computational system in which two or more agents interact or work together to perform a set of tasks or to achieve a set of goals [2]. Agents of a MAS interact with others toward their common specific objective or individual benefit. Agent interactions are established through exchanging messages that specify the desired performatives of other agents (such as notice, request) and declarative representations of the content of messages.

MAS is usually specified as a concurrent system based on the notion of autonomous, reactive and internally-motivated agents acting in a decentralized environment. One key reason of the growth of interest in MAS is that the idea of an agent as an autonomous system, capable of interacting with other agents in order to satisfy its design objectives, is a naturally appealing one for software designers. Agents are used in an increasingly wide variety of applications such as distributed and autonomous decentralized systems. The complexity of MASs suggests a pressing need for system modelling techniques to support reliable, maintainable and extensible design. Although there are many efforts aimed at developing such MASs, there is sparse research on formal specification and design of such systems.

Agent system can operate if the agents are able to exchange information in the form of messages and if they have a common understanding of the possible types of messages that are connected with message “content”. This shared understanding is referred to as an ontology [3]. A great deal of agent-based research has devoted to the development of techniques for representing ontology’s. Multi-agent conversations are built upon two components: agent communication language and interaction protocol. There are a number of agent communication languages, such as Knowledge Query and Manipulation Language (KQML) [4] and the Foundation for Intelligent Physical Agents (FIPA) ACL [5] and others designed for special purposes and that are like mentioned ones. These agent communication languages specify a domain specific vocabulary (ontology) and the individual messages that can be exchanged between agents.

Interaction protocols [5] specify the sequences in which these messages should be arranged in agent interactions. A group of rational agents complies with an interaction protocol in order to engage in task-oriented sequences of message exchange. Thus, when an agent sends a message, it can expect a receiver's response to be among a set of messages indicated by the protocol and the interaction history. With a common interpretation of the protocol, each member of the group can also use the rules of the interaction in order to satisfy its own goals. In other words protocol constrains the sequences of allowed messages for each agent at any stage during a communicative interaction (dialogue), i.e. it describes some standard pattern messages exchanged between agents need to follow. Protocol plays a central role in agent communication. It specifies the rules of interaction between communicating agents and the first thing should be done by the designer developing any particular real-time system is to impose interaction protocol.

This paper explores the possibility of applying existing formal theories of description of distributed and concurrent systems to interaction protocols for real-time multi-agent systems. In particular it is shown how the language PRALU [6, 7], proposed for description of parallel logical control algorithms and rooted in the Petri net formalism, can be used to describe agent interaction protocols. The described approach can be used for the modelling of complex, concurrent conversations between agents in a multi-agent system. It can be used to define protocols for complex conversations composed of a great number of simpler conversations. With the language PRALU it is possible to express graphically the concurrent characteristics of a conversation, to capture the state of a complex conversation during runtime, and to reuse described conversation structure for processing multiple concurrent messages. It is demonstrated with a known example of English auction [5] on how to specify an agent interaction protocol using considered means. Finally, using PRALU language we can verify some key behavioral properties of our protocol description that is facilitated by the use of existing software for the language PRALU [6, 7, 8].

Agent Interaction Protocols

The application domains of MASs are getting more and more complex. Firstly, many current application domains of MASs require agents to work in changing environment (or world) that acts on or is acted on by the system. A closed system is one that has no environment; it is completely self-contained in contrast to an open (uncertain) system, which interacts with its environment. Any real system is open. The MAS must decide what to do and develop a strategy in order to achieve its assigned goals. For this, the MAS must have a representation or a

model of the environment in which it evolves. The environment is composed of situations. A situation is the complete state of the world at an instant of time.

The application of multi-agent systems to real-time environments can provide new solutions to very complex and restrictive systems such as real-time systems. A suitable method for real-time multi-agent system development must take into account the intrinsic characteristics of systems of this type. As a rule they are distributed, concurrent systems with adaptive and intelligent behaviour. For agent-based systems to operate effectively, they must understand messages that have a common ontology underlying them. Understanding messages that refer to ontology can require a considerable amount of reasoning on the part of the agents, and this can affect system performance.

BDI agents are systems that are situated in a changing environment, receive continuous perceptual input, and take actions to affect their environment, all based on their internal mental state. Within the BDI architecture agents are associated with beliefs (typically about the environment and other agents), desires or goals to achieve, and intentions or plans to act upon to achieve its desires. In practical terms, beliefs represent the information an agent has about the state of the environment. It is updated appropriately after each action. The desires denote the objectives to be accomplished, including what priorities are associated with the various objectives. Intentions reflect the actions that must be fulfilled to achieve the goal (the rules to be fired).

To reduce the search space of possible responses to agent messages, interaction protocols can be employed. They specify a limited range of responses that are appropriate to specific message types for a given protocol. When an agent is involved in a conversation that uses an interaction protocol, it maintains a representation of the protocol that keeps track of the current state of the conversation. After a message is received or sent, it updates the state of the conversation in this representation.

By the very nature of protocols as public conventions, it is desirable to use a formal language to represent them. When agents are involved in interactions where no concurrency is allowed, conversation protocols are traditionally specified as deterministic finite automata (DFA) of which there are numerous examples in the literature. DFA consists of a set of states, an input alphabet and a transition function, which maps every pair of state and input to next state. In the context of interaction protocols [9], the transitions specify the communicative actions to be used by the various agents involved in a conversation. A protocol based on such a DFA representation determines a class of well-formed conversations. Conversations that are defined in this way have a fixed structure that can be laid down using some kind of graphical representation.

Protocols can be represented as well in a variety of other ways. The simplest is a message flow diagram, as used by FIPA [5]. More complex protocols will be better represented using a UML sequence (Uniffied Modelling Language) [10] and AUML [11], interaction diagram, statechart [12] and Colored Petri Net (CPN) [13]. UML is one of the currently most popular graphical design languages that is de facto standard for the description of software systems. AUML extends UML sequence diagrams to represent asynchronous exchange of messages between agents. The advantage of AUML is its visual representation. Statecharts [12] are an extension of conventional DFAs. However, expressing protocols of realistic complexity using statecharts or AUML requires substantial efforts for developing, debugging and understanding.

A CPN model of a system describes the states, which the system may be in, and the transitions between these states. CPNs provide an appropriate mathematical formalism for the description, construction and analysis of distributed and concurrent systems. CPNs can express a great range of interactions in graphical representations and well-defined semantics, and allow formal analysis and transformations [14]. By using CPNs, an agent interaction protocol can be modeled as a net of components, which are carriers of the protocol structure. Using CPNs to model agent interaction protocol, the states of an agent interaction are represented by CPN places. Each place has an associated type determining the kind of data that the place may contain. Data exchanged between agents are represented by tokens, and the colors of tokens indicate the data value of the tokens. The interaction policies of a protocol are carried by CPN transitions and their associated arcs. A transition is enabled if all of its input places have tokens, and the colors of these tokens can satisfy constraints that are specified on the arcs. A transition can be *fire*d, which means the actions of this transition can occur, when this transition is enabled. When a transition occurs, it consumes all the input tokens as computing parameters, conducts conversation policy and adds new tokens into all of its output places. After a transition occurs, the state (marking) of a protocol has been changed and a protocol will be in terminal state when there is no enabled or fired transition.

There are a number of works of using Petri Nets or CPNs to model agent interaction protocols [15, 16], there have been also some works on the investigation of protocols' flexibility, robustness and extensibility [17]. It is becoming consensus [14] that a CPN is one of the best ways to model agent interaction protocols. However the notion of an agent executing an action with Petri Net is not explicit in the notation [18]. Different PNs can be assigned to each agent role, raising questions about how the entire protocol is inferred and the reachability and consistency of shared places. If a single Petri net is partitioned for each role, this leads to a complex diagram where a partition is required for each agent identified. Furthermore, alternative actions and states such as either agree or reject but not both, cannot be expressed in standard Petri nets.

Keeping in mind complex systems characterized by complex interaction, asynchronism and concurrency we propose to use for the purposes of their description a special language PRALU [6, 7] having its background in the Petri net theory (expanded nets of free choice – EFC–nets investigated by Hack [19]) but possessing special means for keeping track of the current states of the conversation, receiving messages and initiating responses. The formal language PRALU combines properties of "cause-effect" models with Petri nets. It is intended for a wide application in engineering practice and is well suited for representation of the interactions involved in concurrent system, synchronization among them and then it is simple enough for understanding. The language PRALU supports hierarchical description of the algorithms that is especially important in the case of complex systems. At last, powerful software has been developed that provides correctness verifying, simulation, hardware and software PRALU-description's implementation [7, 8]. The review of obtained results it can be found in [8].

PRALU Language

Any algorithm in PRALU consists of sequences of operations to be executed in some pre-determined order. There exist two basic kinds of operations used in PRALU: acting operations "$\rightarrow A$" and waiting operations "$- p$". Action operation changes the state of the object under control, whereas a waiting operation is passive waiting for some event without affecting anything. In simple case A and p are conjunctive terms, so acting and waiting operations can be interpreted as waiting for event $p = 1$ and producing event $A = 1$. But A can be understood too as a formulae defining operations to be performed and p as a predicate defining condition to be verified [20]. For example, acting and waiting operations could be specified by the expressions such as

"$- (a > b + c)$" and "$\rightarrow (a := b + c)$".

The sequences consisting of action and waiting operations are considered to be linear algorithms. For instance, the following expression means: wait for p and execute A, execute B, then wait for q and execute C:

$$-p \rightarrow A \rightarrow B - q \rightarrow C$$

In general, a logical control algorithm can be presented as an unordered set of chains α_j in the form

$$\mu_j : - p_j L_j \rightarrow \nu_j ,$$

where L_j is a linear algorithm, μ_j and ν_j denote the initial and the terminal chain labels represented by some subsets of integers from the set $M = \{1, 2, ..., m\}$: $\mu_j, \nu_j \subset M$ and the expression "$\rightarrow \nu_i$" presents the transition operation: to the chains with labels from ν_j.

Chains can be fulfilled both serially and in parallel. The order in which they should be fulfilled is determined by the variable starting set $N_t \subseteq M$ (its initial value $N_0 = \{1\}$ as a rule): a chain $\alpha_j = \mu_j : - p_j L_j \rightarrow \nu_j$ (that was passive) is activated if $\mu_j \subseteq N_t$ and $p_j = 1$. After executing the operations of the algorithm L_j, N_t gets a new value $N_{t+1} = (N_t \setminus \mu_j) \cup \nu_j$. The algorithm can finish when some terminal value of N is reached (one-element as a rule), at which time all chains became passive. But the algorithms can also be cyclic; they are widely used when describing production processes.

When the conditions $\mu_j \subseteq N_t$ and $p_j = 1$ are satisfied for several chains simultaneously these chains will be fulfilled concurrently. On the contrary chains with the same initial labels are alternative (only one of them can be fulfilled at a time), they are united in a sentence with the same label as will be shown below.

Thus PRALU allows concurrent and alternative branching, as well as merging concurrent and converging alternative branches. These possibilities are illustrated with the following examples of simplified fragments [20]:

Concurrent branching	Merging concurrent branching	Alternative branching	Converging alternative branching
1: ...→2.3	2: ...→4	1: – *a*...→2	2: ...→4
2: ...	3: ...→5	– $\bar{a}$ ...→3	3: ...→4
3: ..	4.5: ...		4: ...

In PRALU there are two syntactic constraints on chains that restrict concurrent and alternative brunching. If some chains are united in the same sentence (they have equal initial labels) they should have orthogonal predicates in the waiting operations opening the chains:

$$(i \neq j)\ \&\ (\mu_i \cap \mu_j \neq \varnothing) \rightarrow (p_i\ \&\ p_j = 0).$$

The other constraint is similar to the corresponding condition specific for extended nets of free choice (Hack [19]):

$$(i \neq j)\ \&\ (\mu_i \cap \mu_j \neq \varnothing) \rightarrow (\mu_i = \mu_j).$$

PRALU language has some more useful properties that come in handy for description of complex interaction protocols.

1. PRALU algorithms can be expressed both in graphical and symbolic forms.

2. PRALU language permits hierarchical descriptions. The two terminal algorithms (having the only terminal label) may be used as blocks (invoked as complex acting operations) in hierarchical algorithms.

3. In PRALU there exist some additional interesting operations that can be useful for description of interaction protocols. Those are suppression operations and some arithmetic operations. The first ones provide response on special events that can take place outside or within control system. Suppression operation ("→∗", "→∗γ", "→′γ") interrupts the execution all concurrently executed chains of the algorithm or those ones mentioned in γ, (or do not mentioned in γ – in the case of "→′γ"). This operation breaks the normal algorithm flow. Among arithmetic operations it ought to be mention timeout operations (waiting for *n* unit times "– *n*") and counting operation that counts event occurrences.

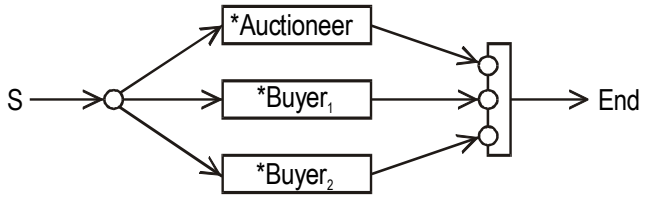
Specifying Protocols in PRALU

For an example of an interaction protocol, consider an English auction [5]. The auctioneer seeks to find the market price of a good by initially proposing a price below that of the supposed market value and then gradually raising the price. Each time the price is announced, the auctioneer waits to see if any buyers will signal their willingness to pay the proposed price. As soon as one buyer indicates that it will accept the price, the auctioneer issues a new call for bids with an incremented price. The auction continues until no buyers are prepared to pay the proposed price, at which point the auction ends. If the last price that was accepted by a buyer exceeds the auctioneer's (privately known) reservation price, the good is sold to that buyer for the agreed price. If the last accepted price is less than the reservation price, the good is not sold.

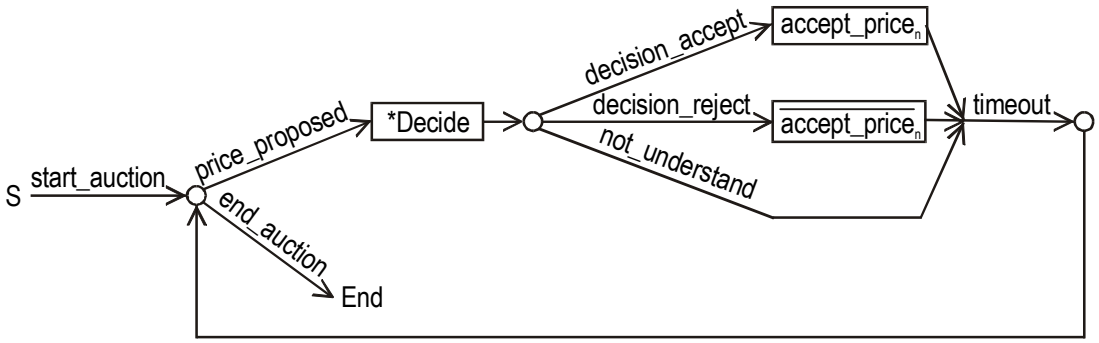
In the case of the auction there are participants of two types: the initiator of the conversation – Auctioneer, and others – Buyers. So, we have two kinds of interaction protocols – those of Auctioneer and of Buyers. The last participants are peer and should be described with identical interaction protocols modeled in our case by the operation "Buyer".

Interaction protocol as a whole can be represented in PRALU as three complex acting operations – blocks. PRALU-blocks are exchanging with values of logical (binary) variables, only such variables are mentioned in them. Each block has some sets of input and output that are enumerated in brackets following the block name (the other variables of a block are its internal). Initialization of a complex acting operation is depicted by the fragment such as "→∗Buyer". The operation Buyer exists in as many copies as the number of participants of the auction, so the copies of the operation differ in their indexes only.

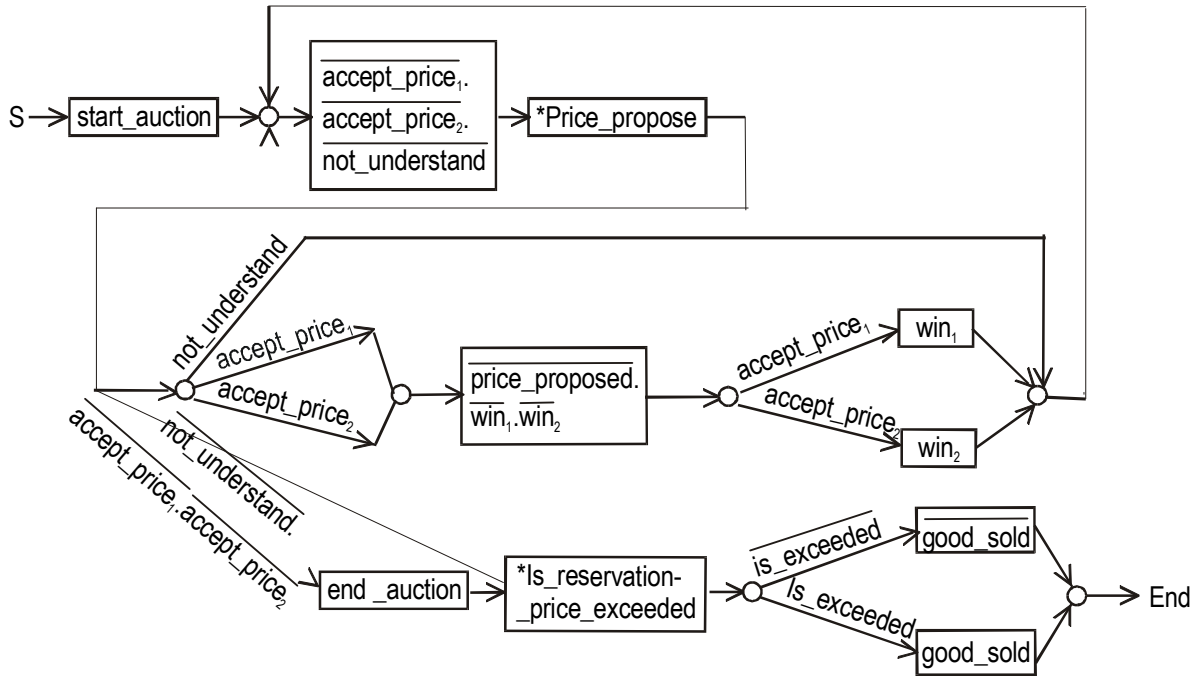
The modelling of the process of auction begins with the execution of "Main_process" triggering event that initiates the interaction protocol execution. Here the processes Auctioneer and Buyer_ns are executed concurrently. For the sake of simplicity we limit the number of buyers to two (in principle it can be simply increased). The process Auctioneer starts with sending the first message (start_auction) that is waited by others participants to continue communication.



Main_process ()



Buyer*n* (start_auction, price_proposed, end_auction / accept_price*n*, not_understand)



Auctioneer (accept_price1, accept_price2, not_understand / start_auction, price_proposed, end_auction)

Fig. 1. English auction interaction protocol in PRALU

Below PRALU description of the auction interaction protocol is shown. Here we cite the only block Buyer*n*, but for real application (intending to simulate the process of auction, for example) we should have as many proper copies as it has been used (in our case – two). Fig. 1 depicts graphical schemes of three mentioned PRALU-blocks: Main_process, Auctioneer and Buyer*n*.

Main_process ()

1: →2.3.4
2: →∗Auctioneer →5
3: →∗Buyer$_1$ →6
4: →∗Buyer$_2$ →7
5.6.7: →**.**

Buyer$_n$ (start_auction, price_proposed, end_auction / accept_price$_n$, not_understand)

1: –start_auction →2
2: –price_proposed →∗Decide(/decision_accept,decision_reject) →3
–end_auction →**.**
3: –decision_accept → accept_price$_n$ →4
–decision_reject →′accept_price$_n$ →4
–not_understand →4
4: –timeout →2

Auctioneer (accept_price$_1$, accept_price$_2$, not_understand / start_auction, price_proposed, end_auction)

1: →start_auction →2
2: →′accept_price$_1$.′accept_price$_1$.′not_understand →∗Price_propose(/price_proposed) →3
3: –not_understand →2
– accept_ price$_1$ →4
– accept_ price$_2$ →4
–′not_understand.′accept_ price$_1$.′accept_ price$_2$ →end_auction
→∗Is_reservation_price_exceeded(/is_exceeded) →6
4: →′price_proposed.′win$_1$.′win$_2$ →5
5: – accept_ price$_1$ →win1 →2
– accept_ price$_2$ →win2 →2
6: –is_exceeded →good_sold →7
–′ is_exceeded →′good_sold →7
7: →**.**

It is assumed that all unformalized operations are referred to as acting operations that set values of logical variables concerned with them. For example, Buyer's operation "Decide" decides for accepting or rejecting the announced price. Depending on adopted decision, it outputs true value of logical variable "decision_accept" or "decision_reject". In a similar, Auctioneers operation "Price_propose" proposes an initial price or increments the charged price outputting true value of logical variable "price_proposed"; the operation "Is_reservation_price_exceeded" verifies if the price accepted by a buyer exceeds the auctioneer's reservation price outputting true or false value of logical variable "is_exceeded".

The operation "–timeout" (where timeout is integer number) means waiting for timeout unit times before doing something followed it. The operation "→**."** is interpreted as the transition to an end of a process described by the block. When the processes of Auctioneer and all Buyers reach their end in the Main_process the transition to its end is executed.

Conclusion

This paper has addressed the need for formalized and more expressive logical and graphical methodologies for specifying (and then validating) interaction protocols in multi-agent systems. Towards this, it was proposed to use the formal language PRALU intended for the representation of complex interactions involved in concurrent system, being in need of synchronization among these interactions. It was demonstrated as well how PRALU algorithms could be used for the specification of multi-agent interaction protocols by the example of English auction. In favour of using the language PRALU is the existence of a great deal of methods and software developed for simulation and logical design of PRALU algorithms as well as for their hardware and software implementation.

References

1. B. Burmeister and K. Sundermeyer, "Cooperative problem-solving guided by intensions and perception", edited by E. Werner and Y. Demazean, Decentralized A.I. 3, Amsterdam, The Netherlands, North Holland, 1992.
2. V. Lesser, "Cooperative Multiagent Systems: A Personal View of the State of the Art", IEEE Trans. Knowledge and Data Engineering, vol. 11, no 1, pp. 133–142, 1999.
3. T. R. Gruber, "A Translation Approach to Portable Ontologies", Knowledge Acquisition, 1993, vol. 5, no 2, pp. 199-220, 1993
4. Finin, Y. Labrou and J. Mayfield, "KQML as an agent communication language", edited by J.M. Bradshaw, Software Agents, MIT Press, pp. 291–316, 1997.
4. ARPA Knowledge Sharing Initiative. Specification of the KQML agent-communication language. ARPA Knowledge Sharing Initiative, External Interfaces Working Group, July 1993.
5. Foundation for Intelligent Physical Agents (FIPA), http://www.fipa.org.
6. A.D. Zakrevskij, V.K. Vasiljonok, "The formal description of logical control algorithms for designing discrete systems", Electronnoje modelirovanie, vol. 6, no 4, pp. 79–84, 1984 (in Russian).
7. A.D. Zakrevskij, "Parallel algorithms for logical control", Minsk, Institute of Engineering Cybernetics of NAS of Belarus, 202 p., 1999 (in Russian).
8. L.D. Cheremisinova, "Realization of parallel algorithms for logical control", Minsk, Institute of Engineering Cybernetics of NAS of Belarus, 246 p., 2002 (in Russian).
9. M. Greaves, and J. Bradshaw J., "Specifying and Implementing Conversation Policies", Autonomous Agents '99 Workshop, Seattle, WA, May 1999.
10. G. Booch, J. Rumbaugh, and I. Jacobson, "*The Unified Modelling Language User Guide*", Addison Wesley, 1999.
11. B. Bauer, J.P. Müller, J. Odell, "Agent UML: A Formalism for Specifying Multiagent Interaction," *Agent-Oriented Software Engineering* , edited by P. Ciancarini and M. Wooldridge, Springer-Verlag, Berlin, pp. 91–103, 2001.
12. D. Harel and M. Politi, "Modelling reactive systems with statecharts", McGraw-Hill, 1998.
13. K. Jensen, "Colored Petri Nets – Basic Concepts, Analysis Methods and Practical Use", vol. 1: Basic Concepts, Springer-Verlag, Berlin, 1992.
14. Quan Bai, Minjie Zhang and Khin Than Win, "A Colored Petri Net Based Approach for Multi-agent Interactions", 2nd International Conference on Autonomous Robots and Agents, Palmerston North, New Zealand, December 13-15, pp. 152–157, 2004
15. R. Cost, "Modelling Agent Conversations with Coloured Petri Nets", Working Notes of the Workshop on Specifying and Implementing Conversation Policies, Seattle,Washington, pp. 59–66, 1999.
16. D. Poutakidis, L. Padgham, and M. Winikoff, "Debugging Multi-Agent Systems Using Design Artefacts: The Case of Interaction Protocols", Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi Agent Systems, Bologna, Italy, pp. 960–967, 2002.
17. J. Hutchison and M. Winikoff, "Flexibility and Robustness in Agent Interaction Protocols", Proceedings of the 1st International Workshop on Challenges in Open Agent Systems, Bologna, Italy, 2002.
18. S. Paurobally, J. Cunningham, "Achieving Common Interaction Protocols in Open Agent Environments", *AAMAS* '02, Melbourne, Australia, 2002.
19. M. Hack, "Analysis of production schemata by Petri nets", Project MAC-94, Cambridge, 1972.
20. A. Zakrevskij, V. Sklyarov, "The Specification and Design of Parallel Logical Control Devices", Proceedings of PDPTA'2000, June, Las Vegas, USA, pp. 1635–1641, 2000.

Authors' Information

Dmitry Cheremisinov, **Liudmila Cheremisinova** – The United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganov str., 6, Minsk, 220012, Belarus, Tel.: (10-375-17) 284-20-76, e-mail: cher@newman.bas-net.by, cld@newman.bas-net.by

ОБ ОДНОЙ МОДИФИКАЦИИ TSS-АЛГОРИТМА

Руслан А. Багрий

Аннотация: *В работе рассматривается одна модификация TSS–алгоритма, который генерирует минимальное порождающее множество решений системы однородных линейных диофантовых уравнений над множеством натуральных чисел.*

Ключевые слова: *Система линейных диофантовых уравнений, базис решений, минимальное порождающее множество.*

Введение

Системы линейных диофантовых уравнений используются во многих разделах современной науки. Важное место среди таких систем занимают системы линейных однородных диофантовых уравнений (СЛОДУ), поскольку методы решения неоднородных систем уравнений или неравенств сводятся к решению СЛОДУ. Методы решения СЛОДУ и систем линейных неоднородных диофантовых уравнений (СЛНДУ) в области натуральных чисел используются в таких областях как искусственный интеллект, логическое программирование[Lloyd J., 1987], компьютерная алгебра[Buchberger B, Kollins J, Loos R., 1986], автоматизация доказательств теорем, унификация, сети Петри [Murata, 1989] и т. п.

Методи решения системы линейных диофантовых уравнений условно можно разделить на методы построения базиса множества всех решений СЛОДУ и методы построения минимального порождающего множества решений СЛОДУ. Известно несколько методов и алгоритмов решения задачи построения базиса и наиболее часто используемым является метод Контенжан-Дэви. В некоторых случаях СЛОДУ может быть охарактеризована с помощью минимального порождающего множества решений, которое в работе [Krivoi S., 1999] было названо TSS (от английского ***Truncated Set of Solutions***).

В данной статье рассматривается одна модификация TSS-метода и соответствующего алгоритма и ее сопоставление с алгоритмом Контенжан-Вэви.

1 Необходимые сведения и определения

Системой линейных диофантовых уравнений (СЛДУ) называется система вида

$$S = \begin{cases} a_{11}x_1 + ... + a_{1q}x_q = b_1; \\ a_{21}x_1 + ... + a_{2q}x_q = b_2; \\ ... \quad ... \quad ... \quad ... \quad ... \\ a_{p1}x_1 + ... + a_{pq}x_q = b_p; \end{cases} \tag{1}$$

где $a_{ij}, b_j \in Z$, $x_j \in N$ и i = 1,2,...,p, j = 1,2,...q.

Единичные вектора из множества N^q $e_1 = (1,0,...,0)$, $e_2 = (0,1,...,0)$,..., $e_q = (0,0,...,1)$ называются векторами канонического базиса множества N^q. Решением СЛДУ (1) называется вектор $c=(c_1,c_2, \dots , c_q)$, где $c_i \in N$, $I = 1,2,...,q$, если выполняется условие $\forall I \in [1,p]$ $a_{i1}c_1 + ... + a_{iq}c_q \equiv b_i$.

Если $\forall I \in [1,p]$ $b_I = 0$, то система называется однородной. СЛОДУ всегда имеет решение вида (0,0,...,0), называемое нулевым или тривиальным. Всякое решение СЛОДУ, отличное от нулевого, называется нетривиальным. СЛОДУ называется совместной, если она имеет хотя бы одно нетривиальное решение. Введем на множестве N^q отношение порядка $\leq$, которое определяется таким образом: если $x=(x_1,x_2, \dots , x_q)$, $y=(y_1,y_2, \dots , y_q) \in N^q$, то $x \leq y$ тогда и только тогда, когда для всех I = 1,2,...,p $x_i \leq y_i$.

Это отношение является частичным порядком и относительно этого порядка множество минимальных решений СЛОДУ является базисом. Известно, что это множество всегда конечно и задача сводится к его построению.

2.1 Краткая характеристика алгоритма Контенжан-Дэви

Метод Контенжан-Дэви[Contenjean E., Devie H., 1994] был первым из алгоритмов построения базиса множества всех решений СЛДОДУ, который не использовал идею сокращения числа неизвестных и уравнений. Этот алгоритм использовал идею метода Фортенбахера[Clausen M. Fortenbacher A., 1989], который строился для построения базиса множества решений единственного диофантового уравнения (ЛОДУ). Идея Фортенбахера состоит в том, чтобы искать минимальные базисные решения, начиная с канонических векторов N^q. При этом если некоторый текущий вектор $x=(x_1,x_2, \dots , x_q)$ еще не является решением, то условие наращивания компонент текущего вектора выполняется с соблюдением такого условия (условие Фортенбахера):

(C_1): увеличивать на 1 то x_j, для которого $a(x)\cdot a(e_j)<0$, где $x = (x_1,x_2, \dots , x_q)$, $a(e_j) = a_j$, где e_j – вектор канонического базиса.

Метод Контенжан-Дави обобщает условие Фортенбахера для СЛОДУ до такого:

(C_p): увеличивать на 1 то x_j, для которого $a(x)\cdot a(e_j)<0$, где $\cdot$ означает скалярное произведение векторов-столбцов $a(x)$ и $a(e_j)$.

Преимуществами алгоритма Контенжан-Дэви есть свойство инкрементальности, возможность применения не только к СЛОДУ, но и СЛНДУ и отсутствие необходимости предварительных расчетов.

Недостатками алгоритма есть его малая эффективность в случае СЛОДУ с большими величинами коэффициентов a_{ij}, а количество уравнений экспоненциально влияет на количество необходимых итераций для поиска решения и как следствие на время решения.

2.2 Краткая характеристика TSS-алгоритма

В ряде задач нет необходимости искать весь базис решений системы, а достаточно убедиться в том, что система совместна. Совместность СЛОДУ можно охарактеризовать с помощью построения ее минимального порождающего множества решений. Минимальное порождающее множество решений генерируется TSS-алгоритмом (описание этого алгоритма можно найти в [Krivoi S., 1999, Krivoi S., 2003]) и по этому множеству осуществляется проверка СЛОДУ на совместность. Работа TSS-алгоритма сводится к выполнению следующих шагов:

- для заданной СЛОДУ $L_1(x) = 0$ & $L_2(x) = 0$ & … & $L_j(x) = 0$, множество канонических векторов N^q разбиваются на три группы M^0, M^+, M^- таким образом, что $M^0 = \{e|L_1(e)=0\}$, $M^+ = \{e|L_1(e)>0\}$, $M^- = \{e|L_1(e)<0\}$; если множество $M^0 \cup M^+$ или $M^0 \cup M^-$ пустое, то уравнение не имеет нетривиальных решений в множестве натуральных чисел, иначе множество решений уравнения $L_1(x) = 0$, находять по следующей формуле:

$$M_1 = M^0 \cup \{y_{ij} \mid y_{ij} = -L_1(e_i)\cdot e_j + L_1(e_j)\cdot e_i, e_j \in M^+, e_i \in M^-\} \quad (2)$$

- полученное множество M_1 с помощью функции $L_2(x)$ тоже разбивается на три группы – $M^0 = \{e|L_2(e)=0\}$, $M^+ = \{e|L_2(e)>0\}$, $M^- = \{e|L_2(e)<0\}$;
- далее повторяются все действия, аналогичные предыдущему уравнению.
- алгоритм заканчивает работу, когда все уравнения данной СЛОДУ будут обработаны и элементы множества M_j составляют TSS этой СЛОДУ.

Основное свойство TSS выражается такой теоремой [6, 7].

Теорема 1. Пусть M и M_j - множество всех решений СЛОДУ и ее TSS, соответственно. Тогда для любого $x \in M$ и $x \notin M_j$ имеет место представление

$$tx = a_1e_1 + a_2e_2 + ... + a_ke_k,$$

где a_i, $t > 1 \in N$, $e_i \in M$.

Из этой теоремы и способа построения TSS вытекает, что TSS совпадает с базисом всего множества решений данной СЛОДУ, если |TSS| = 1.

Из способа построения элементов множества M_j вытекает такое утверждение

Теорема 2. TSS СЛОДУ состоит из единственного решения, если СЛОДУ включает p линейно независимых уравнения, $q - p \leq 1$ и множество $M_i^0 = \varnothing$ для всех $i \in [1,p]$, где q – количество неизвестных.

На эффективность работы данного метода мало влияют значения коэффициентов при неизвестных. Данный метод лишен тех недостатков свойственных предыдущему методу, но в отличие от него не дает полного базиса. Это ограничение позволяет использовать его только в задачах, где поиск всего базиса не требуется.

Таким образом, каждый из алгоритмов, существующих на данный момент времени, может использоваться только в каких-то определенных задачах обусловленных ограничениями самого алгоритма. Следовательно, расширив границы, при которых усеченное множество TSS алгоритма будет совпадать с полным минимальным базисом, мы тем самым, увеличим круг задач, в которых он может использоваться. Рассмотрим одну простую модификацию TSS-алгоритма, который основывается на теореме 2.

3 Модификация TSS-алгоритма

Как следует из вышесказанного, TSS-алгоритм в процессе своей работы оперирует с каждым из уравнений СЛОДУ и тремя множествами M^0, M^+, M (первое множество M^0 уже включает решения очередного уравнения). Комбинирование элементов из последних двух множеств в объединении с первым множеством и дает TSS данной СЛОДУ. Рассмотрим на примере алгоритм работы и проанализируем полученный результат.

Пусть система уравнений будет следующая:

$$\begin{cases} L_1(x) = 4x_1 + 2x_2 - 3x_3 - 2x_4 - x_5 = 0 \\ L_2(x) = 2x_1 - x_2 - 4x_3 + x_4 - 5x_5 = 0 \\ L_3(x) = 0x_1 + 2x_2 + 4x_3 + 0x_4 - 9x_5 = 0 \end{cases} \quad (3)$$

Рассмотрим уравнение $L_1(x)$. Множество канонических векторов N^q в данном случае разбиваются на три группы: $M^0 = \varnothing$, $M^+ = \{(1,0,0,0,0), (0,1,0,0,0)\}$, $M^- = \{(0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1)\}$. После комбинирования элементов этих множества по формуле (2) получаем такое TSS, состоящее из 6 векторов:

$$e_1 = (3,0,4,0,0), e_2 = (0,3,2,0,0), e_3 = (1,0,0,2,0),$$
$$e_4 = (0,1,0,1,0), e_5 = (1,0,0,0,4), e_6 = (0,1,0,0,2).$$

При анализе полученных результатов очевиден, тот факт, что количество компонент в каждом векторе не равных нулю будет равно двум, поскольку только два вектора из множества канонических векторов комбинировались и формировали решение.

Идея данной модификации состоит в комбинировании не двух векторов принадлежащие к разным множествам M^+ и M^-, а к комбинированию трех векторов, два из которых принадлежат одному множеству, а третий другому. Тогда будем получать решения, в которых количество ненулевых компонент будет больше двух. Это равносильно решению линейного диофантового уравнения с тремя неизвестными, а, следовательно, необходим алгоритм нахождения базиса множества решений такого уравнения, который для данного случая был бы наиболее эффективным.

Для нахождения базиса уравнений вида $ax + by + cz = 0$ был использован расширенный алгоритм Евклида. Этот алгоритм вычисляет наибольший общий делитель двух чисел, а также позволяет решать

уравнения вида $ax+by=c$, где a,b,c – целые, и получить частные решения x_0 и y_0. Общий вид решений имеет вид $x = x_0+kb$, $y = y_0-ka$, где k – произвольное любое целое число.

Исходя из алгоритма Евклида, опишем порядок нахождения минимального базиса для уравнения вида $ax + by + cz = 0$:

1. Преобразуем уравнение по правилу:
   - если $\mathrm{sgn}(a) = \mathrm{sgn}(b)$, то уравнение примет вид $by + cz = -ax$;
   - если $\mathrm{sgn}(a) = \mathrm{sgn}(c)$, то уравнение примет вид $ax + by = -cz$;
   - если $\mathrm{sgn}(b) = \mathrm{sgn}(c)$, то уравнение примет вид $ax + cz = -by$;
2. Примем значение неизвестного в правой части уравнения равной 0.
3. Решим полученное уравнение при помощи алгоритма Евклида.
4. Увеличим неизвестную в правой части уравнения на 1.
5. Будем повторять пункты 3-4 до тех пор, пока значение одного из неизвестных в левой части не будет равно нулю.

Рассмотрим работу алгоритма на примере.

Возьмем два вектора из множества M^+ предыдущего примера – (1,0,0,0,0) и (0,1,0,0,0), один из множества M^- (0,0,1,0,0), то необходимо решить уравнение $4x_1 + 2x_2 - 3x_3 = 0$. Следуя правилу преобразования, приведем уравнения к виду $2x_2 - 3x_3 = -4x_1$. Порядок получения базиса сведен в таблицу:

Уравнение	Вектор-решение (x_1,x_2,x_3)
$2x_2 - 3x_3 = -4 \cdot 0$	(0,3,2)
$2x_2 - 3x_3 = -4 \cdot 1$	(1,1,2)
$2x_2 - 3x_3 = -4 \cdot 2$	(2,2,6)
$2x_2 - 3x_3 = -4 \cdot 3$	(3,0,4)

Третий вектор-решение (2,2,6) не относится к базису, так как все его компоненты больше или равны второго вектора (1,1,2).

Использую расширенный алгоритм TSS, к усеченному множеству решений системы (3) добавляется еще 5 векторов.

$$e_7 = (1,1,2,0,0), e_8 = (2,0,2,1,0), e_9 = (1,0,0,1,2), e_{10} = (0,2,1,0,1), e_{11} = (1,0,1,0,1).$$

В данном случае расширенное усеченное множество векторов совпало с полным базисом решений, но в большинстве случаев такое совпадение редкость.

Дальнейшее решение системы полностью соответствует алгоритму TSS за исключением того, что векторы-решения образуются за новым принципом. Суть его заключается в использовании трех векторов-решений предыдущего уравнения для получения расширенного усеченного базиса. Решение системы уравнений (3) при помощи оригинального алгоритма TSS дает 2 решения:

$$s_1 = (0,63,18,25,22),\ s_2 = (63,0,72,2,32).$$

Используя новый предложенный алгоритм, к ним добавляются еще 4 решения:

$$s_3 = (19,1,22,1,10),\ s_4 = (7,7,10,3,6), s_5 = (13,4,16,2,8),\ s_6 = (1,10,4,4,4).$$

4 Эксперименты

В таблице приведены экспериментальные данные позволяющие оценить временные характеристики одной из версий TSS алгоритма, расширенного TSS алгоритма и метода Контежан-Дави. Также приведено количество полученных ответов решенных этими алгоритмами.

СЛОДУ	Метод TSS		Расширенный метод TSS		Метод Контежан-Дави	
	Кол-во реше-ний	Время мс	Кол-во реше-ний	Время мс	Кол-во реше-ний	Время мс
$\begin{cases} L_1(x) = 4x_1 + 2x_2 - 3x_3 - 2x_4 - x_5 = 0 \\ L_2(x) = 2x_1 - x_2 - 4x_3 + x_4 - 5x_5 = 0 \\ L_3(x) = 0x_1 + 2x_2 + 4x_3 + 0x_4 - 9x_5 = 0 \end{cases}$	2	0	6	15	6	11400
$\begin{cases} L_1(x) = x_1 - 3x_2 + 9x_3 - 1x_4 + 81x_5 = 0 \\ L_2(x) = -9x_1 + 27x_2 + 81x_3 - 9x_4 + x_5 = 0 \\ L_3(x) = 1x_1 - 3x_2 + 9x_3 + 1x_4 - 81x_5 = 0 \end{cases}$	2	0	2	1300	2	2600
$\begin{cases} L_1(x) = -1x_1 + 6x_2 - 3x_3 + 2x_4 + 1x_5 = 0 \\ L_2(x) = 1x_1 + 2x_2 - 5x_3 + 3x_4 + 6x_5 = 0 \\ L_3(x) = 7x_1 + 4x_2 - 2x_3 + 7x_4 - 5x_5 = 0 \end{cases}$	2	0	3	16	3	125
$\begin{cases} L_1(x) = -3x_1 - 14x_2 + 1x_3 + 12x_4 - 3x_5 = 0 \\ L_2(x) = 4x_1 - 5x_2 + 16x_3 + 1x_4 - 4x_5 = 0 \\ L_3(x) = -2x_1 - 1x_2 - 5x_3 - 2x_4 + 14x_5 = 0 \end{cases}$	2	0	66	4700	66	26000

Заключение

Предложенный алгоритм генерирует базис множества всех решений СЛОДУ при условии выполнения требований, приведенных в теореме 2. Кроме того, из экспериментальных результатов следует, что временные характеристики предлагаемого алгоритма выглядят лучше по сравнению с экспериментальными данными алгоритма Контенжан-Дэви.

Библиография

[Lloyd J., 1987] Lloyd J., Foundations of Logic Programming (2-d addition). Springer Verlag. - 1987. -276 p.

[Buchberger B, Kollins J, Loos R., 1986] Компьютерная алгебра (Символьные и алгебраические вычисления). Под ред. Бухберга Б., Коллинза Дж., Лооса Р., М.:Мир, 1986, 386с.

[Contenjean E., Devie H., 1994] Contenjean E., Devie H. An efficient algorithm for solving system of Diophantine equations. Inform. Comput., 1994, 113. v.1, PP.143-173.

[Clausen M. Fortenbacher A., 1989]Clausen M. Fortenbacher A. Efficient solution of linear Diophantine equations. Journ. Symbolic Computation, 1989, v.8, PP. 201-216.

[Murata, 1989] T. Murata. Petri Nets: Properties, Analysis and Applications. In Proceedings of the IEEE, 1989, - vol. 77, -N4, -P. 541-580.

[Krivoi S., 1999] Крывый С.Л. О некоторых методах и критериях совместности систем линейных диофантовых уравнений в области натуральных чисел //Кибернетика и системный анализ., 1999, N 4, с. 12-36.

[Krivoi S., 2003] Кривой С.Л. Об алгоритмах решения систем линейных диофантовых констрейнтов в области {0,1}, .ж. Кибернетика и системный анализ. – 2003, - № 5. – С. 58-69.

Информация об авторе

Руслан А. Багрий (Ruslan A. Bagriy) – Хмельницкий национальный университет, Хмельницкий, Украина (National University of Kchmelnitsk, Kchmelnitsk, Ukraine), e-mail: ruslan@beta.tup.km.ua

THE DEVELOPMENT OF PARALLEL RESOLUTION ALGORITHMS USING THE GRAPH REPRESENTATION

Andrey Averin, Vadim Vagin

Abstract. *The parallel resolution procedures based on graph structures method are presented. OR-, AND- and DCDP- parallel inference on connection graph representation is explored and modifications to these algorithms using heuristic estimation are proposed. The principles for designing these heuristic functions are thoroughly discussed. The colored clause graphs resolution principle is presented. The comparison of efficiency (on the Steamroller problem) is carried out and the results are presented. The parallel unification algorithm used in the parallel inference procedure is briefly outlined in the final part of the paper.*

Keywords: *Automated Reasoning, Logical inference*

1. Introduction

The deductive inference procedures are broadly used in variety of fields, such as expert systems, decision support systems, deductive databases and intelligent information systems. Due to high amount of data in practical problems and the exponential growth of the search space, the efficiency of deductive procedures becomes the key factor in the development of deductive inference systems. One of the ways to improve the efficiency is the parallelization of inference procedures. We present a parallel inference procedure based on resolution principle. The connection graph representation is chosen as the basis for designing the parallel resolution procedures. Using the graph representation simplifies the parallelization of the inference process and allows to apply the different parallelization techniques such as OR, AND and DCDP parallelism. We study and implement OR-, AND- and DCDP- parallel resolution procedures, develop useful heuristics which can be used in parallel resolution procedures on connection graphs and make a comparison of the obtained results with the results of algorithms developed by other researchers. Also we describe and implement the clause graphs inference procedure. As the test task, the Schubert's Steamroller problem is examined [1]. The problem of parallelism on the term level is also investigated. The data structure for the term representation and the parallel unification algorithm using this data structure are presented.

2. Connection Graph

The connection graph method was designed by R. Kowalski [2]. A connection graph is a scheme for representing the proper first-order formulas in disjunctive normal form. Each literal is associated with a node in the connection graph. Literals in a clause are combined into a group. If the literals in two clauses form a contrary pair (P and $\neg$P) then there is an edge between the respective nodes of the connection graph.

Example 1.

The initial set of clauses:

1.Q(c)
2. Q(b)
3.R(x) $\vee \neg$Q(y) $\vee \neg$P(x)
4.P(b)
5.$\neg$R(x) $\vee$ S(x,y) $\vee \neg$T(x)
6.T(y) $\vee \neg$B(y)
7.B(a)
8.$\neg$F(y) $\vee \neg$S(y,z) $\vee \neg$B(z)
9.B(x) $\vee \neg$C(x) $\vee \neg$D(y)
10.D(c)
11.F(b)
12.F(c)
13.C(b)

The corresponding connection graph is shown in fig. 1.

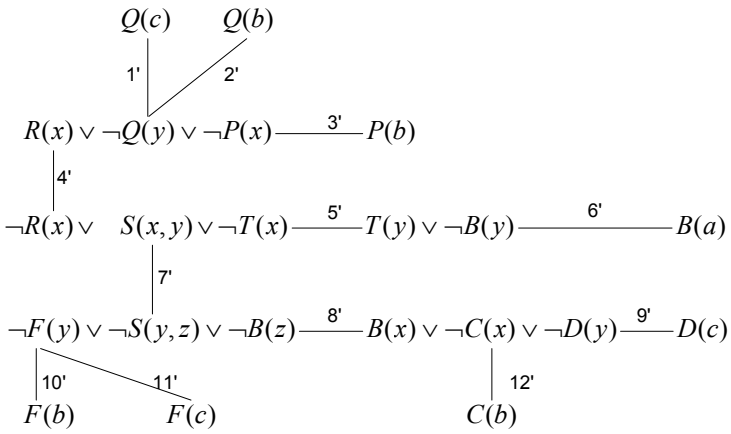


Fig.1

3. Methods of Inference on Connection Graphs

3.1 The Sequential Proof Procedure

To prove the unsatisfiability of a clause set we must generate and resolve the initial connection graph, i.e. derive an empty clause.

The main operation in connection graph refutation is the link resolution, when the resolvent is computed and added to the graph. The corresponding link is deleted and the links of the added resolvent are inserted. A pure clause is a literal group containing a node with no links. Pure clause with all its links can be easily removed from the graph without losing the completeness of the connection graph refutation procedure. Similarly, if we have a tautology clause, it also can be removed from a graph. If a resolvent on some step is a pure clause or a tautology, there is no need to insert this clause into a graph.

The refutation algorithm consists of the following steps:

1. the verification whether there is any clause in a graph or not. If there are no clauses, the algorithm terminates unsuccessfully. If there is the empty clause, then the algorithm is successfully terminated, else go to step 2;
2. if a graph does not contain any link, then the algorithm is unsuccessfully terminated, else go to step 3;
3. a link selection. The link is resolved and the resolvent is generated;
4. if an empty resolvent is obtained, then the algorithm terminates successfully, else the resolvent is inserted into the graph, its links are added, and the algorithm goes to step 2.

The fundamental problem in the connection graph refutation is the choice of suitable links by some criteria at each step of an algorithmic operation. Links are usually selected by using heuristics.

3.2 Parallel Inference on the Kowalski Connection Graph

The Kowalski connection graph can easily be used as the basis for designing parallel resolution algorithms [3]. Since the search space is complete, there is a possibility of using parallel computation strategies for enhancing the inference procedure efficiency. Parallel resolution algorithms differ from the sequential algorithm in step 3 at which a set of links satisfying certain criteria (not a single link as in the sequential procedure) is chosen and parallel resolution of all the links in this set is carried out.

3.2.1. OR-parallel Resolution on a Connection Graph

In case of OR-parallelism, the inference system associates some goal clause with the heads of clauses – possible candidates for resolution. Literals are unified and new clauses are generated. Admissible OR-links sets for the connection graph of Fig.1 are {1', 2'} and {10', 11'}.

3.2.2. DCDP-parallel Resolution on a Connection Graph

One modification of the parallel inference on the Kowalski connection graph is called DCDP parallelism (parallelism for distinct clauses) [4]. The correctness of the DCDP parallel resolution is proved in [5].

Definition 1. Clauses are said to be adjacent if there exist one or several links joining the literals of one clause with the literals of another clause.

Definition 2. A set of links joining pairs of distinct clauses is called a DCDP-link set if the clauses of every pair are not adjacent to any clause of other pairs.

To illustrate these definitions let us study the DCDP-link set for the connection graph of Fig.1. Adjacent pairs of clauses for this set are {(1), (3)}, {(2), (3)}, {(3), (4)}, etc... Thus, one of the DCDP-link sets is {1', 6', 9'}. Other examples of DCDP-link sets are: {2', 6', 12'}, {4', 12'}, {1', 6', 10'}.

3.2.3 AND-parallel Inference

Definition 3. A clause where all its links are resolved in parallel is called a SUN-clause.

Definition 4. Clauses joined with the literals of a SUN-clause are called the satellite clauses.

In AND-parallelism all links of literals of the SUN clause are resolved simultaneously. All resolvents are inserted in the graph along with all inherited links of a satellite clause. A SUN clause with all its links is removed. There is proved the correctness of the AND-parallel resolution [5]. The correct unification of separable variables under AND-parallel resolution is studied in [3].

Let us consider the choice of an AND-link set for the connection graph of fig.1. Admissible AND-link sets are, for example, {5', 4', 7'} (SUN-clause –clause (5)) and {5',6'} (SUN-clause – clause (6)). Detailed description of methods and algorithms can be found in [3,6].

4. Modification of Parallel Inference Procedures

Different heuristics can be used for choosing a link in resolving upon a connection graph. In the parallel resolution we must choose a set of links satisfying certain conditions. Note that the inference procedure becomes unsuitable if links are chosen unsuccessfully. The main principles underlying the design of heuristics are:

1. the number of literals in resolved clauses must be minimal,
2. the number of links in resolved clauses must be minimal,
3. the number of links in a literal for which the clauses are resolved must be minimal,
4. the unifiers of a resolved link must have a substitution of the type {*c*/*x*}, where *c* is a constant or a functional term, and *x* is a variable.

Further we describe the meaning of each principle in more detail.

Principle (1) simplifies a resulting connection graph, because a clause with a small number of literals, usually has a fewer number of links.

Principle (2) also simplifies a resulting connection graph, because the resolution of clauses with a small number of links yields clauses also with a small number of links.

Principle (3) prefers those links, the resolution of which yields "pure" clauses that on removing, can considerably simplify a connection graph.

The same is true for principle (4): as a result of the resolution by links containing a substitution of a variable for a constant, we obtain a clause containing constant terms. Such a clause has, at the first, a small number of links and, at the second, can be effectively used in the resolution. Principles (I)-(4) are taken into consideration in the heuristic function described below.

4.1. The Heuristic Function HI

In the heuristic function *H1* the link estimation is represented as a linear combination of estimations of the objects in the link (unifiers, clauses, and predicate literals in the link).

4.1.1 Computation of the Value of the Heuristic Function

Let *WeightLink* denote the heuristic estimation of a link. Then:

$$WeightLink = k_{clause} \cdot (Clause_{1Heur} + Clause_{2Heur}) + Uni_{Heur} \cdot k_{uni} + k_{pred} \cdot (Pred_{1Heur} + Pred_{2Heur}),$$

where $Clause_{1Heur}$, $Clause_{2Heur}$, Uni_{Heur}, $Pred_{1Heur}$ and $Pred_{2Heur}$ are the heuristic estimations for the first clause, the second clause, the link unifier, the predicate literal in the first clause of a link, and the predicate literal in the second clause of a link, respectively, and k_{uni}, k_{clause} and $k_{pred} \in [1; 100]$ are coefficients. Let us describe these symbols in more detail.

4.1.2 Heuristic Estimation of a Clause

The heuristic estimation must take into account the changes taking place in the characteristics of a connection graph during the inference (for example, changes in the number of links and the number of literals in a clause). Let us examine the heuristic estimation based on principles (1) and (2):

$$Clause_{Heur} = k_1 \cdot ClauseNumberOfLinks + k_2 \cdot ClauseNumberOfClauses,$$

where *ClauseNumberOfLinks* is the number of links in a clause, *ClauseNumberOfClauses* is the number of predicate literals in a clause; k_1, k_2 are arbitrary coefficients, chosen a priori on the basis of graph characteristics such as an average number of literals and links for the clause.

Let for example the average number of literals and links for the clause be 3.2 and 4.8, respectively. Then we can take k_1 = 4.8/3.2 = 1.5 and k_2= 1. In this case, both principles (1) and (2) have the same weight. Characteristics may change their values during the inference. In this case, the initially chosen values of coefficients become "obsolete," and one principle gains a greater weight over the other. To avoid such a situation, the values of coefficients must be changed during the inference. Let *AverageLinkCount* and *AverageClauseLength* denote an average number of links and literals in a clause, respectively.

We can take k_1 = *AverageLinkCount/AverageClauseLength* and k_2=*1*. In this case, principles (1) and (2) both gain the same weight in the course of the whole inference process. The heuristic clause estimation takes the final form:

$$Clause_{Heur}=(AverageLinkCount/AverageClauseLength)\times ClauseNumberOfLinks+ ClauseNumberOfClauses.$$

4.1.3 Heuristic Estimation of the Unifier

The heuristic unifier estimation must be based on principle 4 (the unifier of a resolved link must have a substitution of the type *c/x*, where *c* is a constant or a functional term with constants and *x* is a variable) and must take into account the changes in values of the graph characteristics. The heuristic estimation of the unifier Uni_{Heur} is computed by the formula:

$$Uni_{heur}=AverageLinkCount/(1+NumberOfConstantSubst +NumberOfFuncSubst),$$

where *NumberOfConstantsSubst* is the number of substitutions of the type *{c/x}* in the unifier, *c* is a constant term and *x* is a variable.

NumberOfFuncSubst is the number of substitutions of the type {*f/x*} m the unifier, where *f* is a functional term with constants and *x* is a variable.

4.1.4 Heuristic Estimation of a Predicate Literal

The heuristic estimation of a predicate literal must be based on principle (3) (the number of links in a literal, for which clauses are resolved upon must be minimal). The heuristic function must also take into account the changes in the graph characteristics. The heuristic estimation $Pred_{Heur}$ of a predicate literal is computed by the formula:

$$Pred_{Heur}=PredNumberOfLinks\cdot AverageClauseLength,$$

where *PredNumberOf Links* is the number of links in a predicate literal.

4.1.5 Selection of Coefficients

The coefficients k_{uni}, k_{clause} and k_{pred} are chosen experimentally. We have chosen the following ratio $k_{uni}=k_{pred}=(1/10)\cdot k_{clause}$ for coefficients. In this relationship, the greater weight is attached to principles (1) and (2). As a rule, k_{clause} is taken from the interval [10; 100] so that k_{uni} and k_{pred} lie in the interval [1; 10]. Thus, the heuristic function *H1* takes account of all principles (I)-(4). The weight of principles can be changed. The changes in the graph characteristics are also taken into account. The function *H1* enhances the efficiency of parallel inference algorithms for problems of practical complexity.

5. Deductive Inference on Clause Graphs

The deduction algorithm transforms a clause graph via two special operators – a predicate node elimination operator and a predicate node splitting operator [6,7]. They are applied to a predicate vertex depending on whether the node has multiarcs or not. A predicate node I said to be joined to a clause with multiarcs if the clause contains more than one literal with predicate symbol of the predicate node. For example, the node P in Fig.2 is joined to clause 1,2 and 3 (clause 4 is joined with the predicate node P with an ordinary arc) by the multiarcs.

1. $P(x,y) \rightarrow P(f(x),y)$ $(or \quad \neg P(x,y) \vee P(f(x),y))$
2. $P(u,f(x)) \,\&\, P(v,g(w)) \rightarrow$ $(or \quad \neg P(u,f(x)) \vee \neg P(v,g(w)))$
3. $\rightarrow P(g(x),y) \vee P(a,z)$ $(\text{or } P(g(x),y) \vee P(a,z))$
4. $\rightarrow P(a,x)$ $(or \quad P(a,x))$

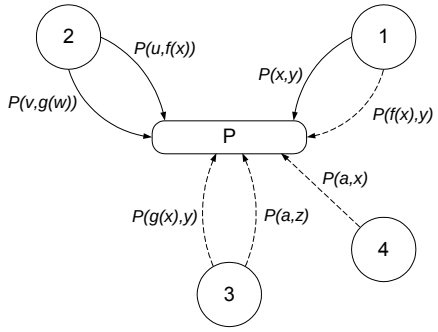


Fig. 2

In this method clauses are expressed in implicative form.

Here 1-4 are the nodes corresponding to the clauses of the logical model. The oval vertex represents the predicate symbol P. Continuous arcs are weighted with conditional literals of clauses, whereas dotted ars are weighted with inference literals of the clause.

An operation similar to colouring of clause graphs is introduced. Clause condition is «network is coloured» by color C1 (the corresponding arcs in figures are shown by a continuous line) and the condition for inference is colored by color C2 (a dotted line in figures). If clause graphs are represented in colored form, then it is easy to search for information and new assertions can be inferred easily and thus the effectiveness of the inference system is enhanced.

The general inference scheme for an empty clause via transformation of clause graphs consists of the following:

1. if the network contains predicate nodes to which the node elimination operator can be applied, then such nodes are removed by this elimination operator;
2. if there are nodes with multiarcs, then the splitting operator is applied to generate nodes without multiarcs. Then the node elimination operator is applied, etc. until a contradiction is obtained in network.

If the node elimination operator consists of a set of usual operations of resolution of specially chosen pairs of clauses, then the splitting operator contains a distinct feature of this algorithm, i.e., a feature that has no analog in other logical inference mechanisms.

The predicate node elimination operator is applied to nodes having no multiarcs. It resolves in all possible ways those clauses that contain contrary pairs of literals. Resolution is implemented by the predicate of the chosen predicate node. Upon completion of all resolutions, parent clauses and the predicate node are removed from the network and the new clauses found via resolution are included in the network. Figures 3a and 3b show a clause graph before and after the application of the elimination operator to the vertex M for the following disjunct set:

1. $Q \,\&\, M \rightarrow$
2. $M \,\&\, H \rightarrow Q$
3. $F \,\&\, M \rightarrow H$
4. $\rightarrow M$
5. $\rightarrow F$

The node M and clauses (1)-(4) were removed and the clauses

6. $Q \rightarrow .$ (1,4)
7. $F \rightarrow .H$ (3,4)
8. $H \rightarrow Q$ (2,4)

were added to the network.

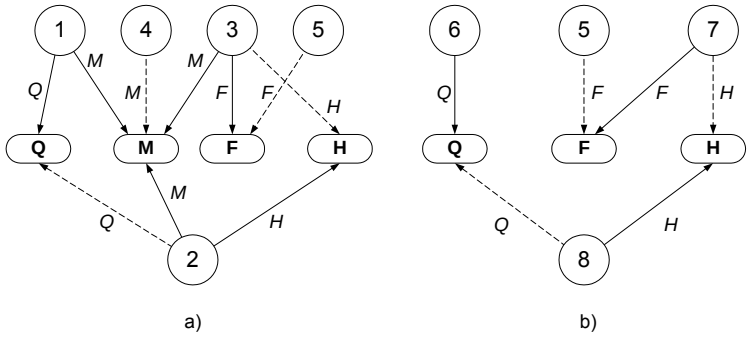


Fig. 3

The predicate splitting operator is introduced in order to remove multiarcs from nodes and thereby create conditions for the application of the elimination operator. The splitting operator generates copies of clauses and a few new vertices having the same predicate symbol as the split vertex, but with new indexes 1,2,3, … . The clause copies and new nodes are interrelated with each other. First, the elimination operator removes those nodes that have higher indexes.

Figure 4 shows the action of a splitting operator for the case of three multiarcs starting from clauses 1,2 and 3 (Fig. 2).

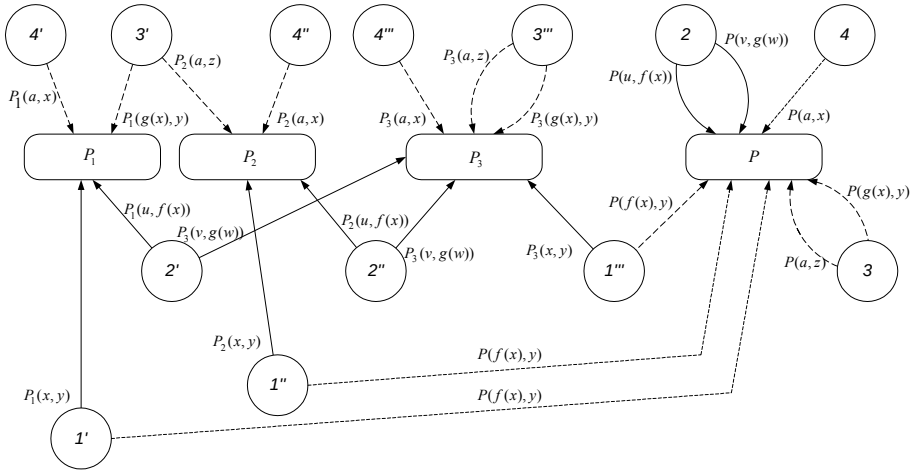


Fig.4

Splitting the three multiarcs (“petals”) of the node P by this operator 1we obtain four nodes P_1, P_2, P_3 and P (their amount is one more than the amount of «petals»), four copies of clause 4, three copies of the «split» clause 1, two copies of the «split» clause 2 with a duplicate at vertex P, and a «split» clause 3 with a duplicate at nodes P and P_3. Splitting occurs in the following order: first the multiarc (1 – P), then the multiarc (2 – P), and finally the multiarc (3 – P) are split. The priority elimination strategies for removing nodes having no multiarcs and nodes with minimal number of arcs are reasonable for forming a sequence of processed nodes in a clause graph. They yield a sequence of eliminations and generate a lesser number of intermediate clauses than the variant in which these strategies are not used. Both these operations are correct i.e. unsatisfiable clauses remain unsatisfiable after the application of these operators[1]. A parallel algorithm for deductive inference on colored clause graphs is described in [7,8].

5. Efficiency

As the test problem, we have researched the "steamroller" problem formulated by L. Schubert in 1978 to test automatic proof systems [1]. This problem requires the generation of an exponential number of intermediate clauses and unifications during the inference process. Below we describe the results obtained by our and other algorithms of deductive inference for the «steamroller» problem.

CG is the inference strategy with a connection graph. SOS is the inference on a connection graph with a goal statement as a support set. TR is the inference within Theory Links - the extension of the standard resolution method. UR is inference with Unit Resolution - a modification of the resolution method. LUR is the inference by Linked Unit Resolution - a modification of the UR resolution method.

According to Figs. 5 and 6, the best results are obtained by the McCune/LUR procedures. Parallel inference procedures have also shown their efficiency for deductive inference on connection graphs.

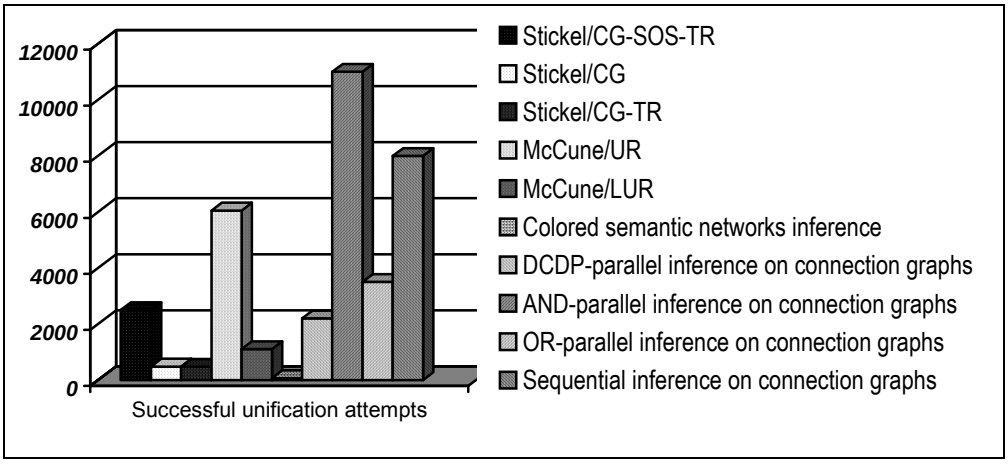


Fig. 5

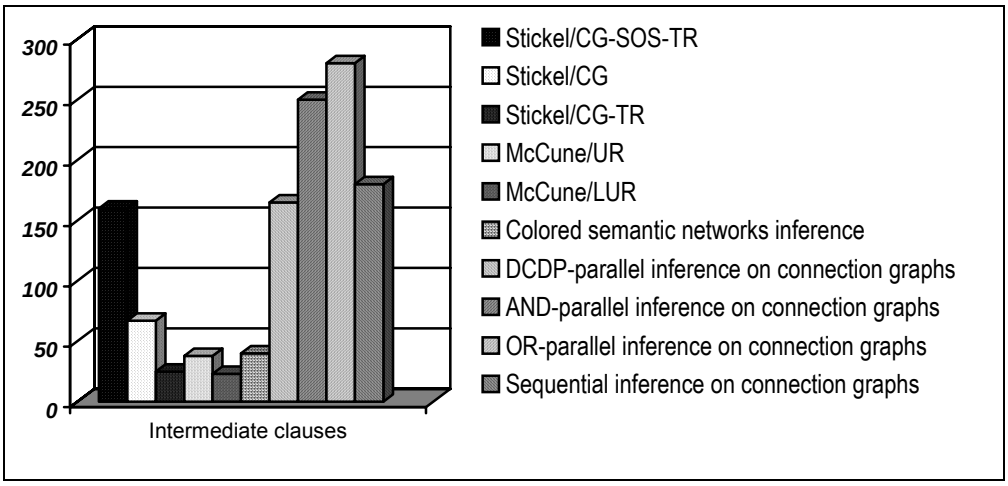


Fig. 6.

6. Parallel Unification in Connection Graph Inference

Due to a high amount of the unification tasks in deductive inference, the efficiency of the implementation of the unification procedure is one of the main factors in designing deductive inference procedures. There is a variety of effective unification procedures, but all of them have their own disadvantages. The main disadvantage is the use of complex, tightly coupled data structures, which are very difficult to parallelize.

We develop rather simple data structure for the term representation based on the notion of path strings. Terms (and clauses) are stored in tables, which are connected with links. Some tables and parts of the tables can be processed in parallel (with having links in mind). As we have no possibility to describe the algorithm and the data structure for term representation in detail, we just briefly outline the main principles used in this approach.

The term is stored as the sequence of strings, where every string presents one symbol in the term. Also such information as the type of a symbol (a variable or a constant symbol), the index number of an argument in a function symbol and the depth of a symbol (i.e. the number of function symbols which this symbol belongs to) is stored. The terms that can be unified are connected with the links of two types.

The first type corresponds to the terms (and strings) of different depth. The second type corresponds to the terms of the same depth. Let us illustrate these notions by the simple example.

Consider the unification task $\{t_1=t_2\}$, where t_1= *f(a,g(h(x)))* and t_2=*f(x,g(y))*. The representation of the term *f(a,g(h(x)))* is *f1.a.1.constant*(1) and *f2.g1.h1.x.3.variable* (2).

The representation of the term *f(x,g(y))* is *f1.x.1.variable*(3) and *f2.g1.x.2.variable*(4) (the index number is shown in brackets). The links are created between the strings (1) and (3) and between the strings (2) and (4). The first link has the second type and the second link has the first type.

The task of link establishing is one of the main tasks in the proposed approach. Two techniques can be used. The simplest one is the special sorting procedure on tables containing strings from the unification task. The main drawback of this approach is the deficit of efficiency caused by the nature of the sorting problem (the complexity of the sorting problem is *nlog(n)*). The second approach is based on automata representation of strings and is similar to discrimination trees. Using automata increases the efficiency of the procedure, but requires higher memory consumption.

The main idea lying in the parallelization of the unification procedure is the use of dependencies graph, where strings connected with arcs cannot be proceeded in parallel. The complexity of the task of determining maximum sets for parallel proceeding is equal to the complexity of the graph coloring task, though heuristics can be used to find non-maximum but satisfactory sets. This unification procedure has been combined with the parallel inference procedure on connection graphs. The structure for term representation and unification are thoroughly described in [9].

7. Conclusions

The Kowalski connection graph and the sequential inference procedure on connection graphs are investigated. The structure of the OR-, AND- and DCDP-parallel inference methods is described. Methods of modifying parallel inference procedures are analyzed and the main principles underlying the design of heuristic link estimations are stated. The heuristic function for link selection is designed. A procedure of inference on colored clause graphs is also described. The developed deductive inference procedures are compared with other procedures on the "steamroller" problem. The procedures of parallel inference on connection graphs and clause graphs are the effective ones for the deductive inference. They can be applied in expert and decision making systems.

Bibliography

[1] M.F. Stickel, Schubert's Steamroller Problem: Formulations and Solutions // *J. Autom. Reasoning,* 1986, vol. 2, pp. 89-100.

[2] R. Kowalski, A Proof Procedure using Connection Graphs // J. ACM, 1975, 22(4), pp. 595-599.

[3] V.N. Vagin and N.O. Salapina, A System of Parallel Inference with the Use of a Connection Graph // *Journal of Computer and System Sciences International,* Vol. 37, No. 5, 1998, pp. 699-708.

[4] G. Hornung, A. Knapp and U. Knapp, A Parallel Connection Graph Proof Procedure // *German Workshop on Artificial Intelligence. Lecture Notes in Computer Science,* Berlin: Springer-Verlag, 1981. pp. 160-167.

[5] R. Loganantharaj, Theoretical and Implemetational Aspects of Parallel Link Resolution in Connection Graphs, *Ph.D. Thesis,* Dept. of Computer Science, Colorado State Univ., 1985.

[6] Vagin V.N. Deduktsiya i obobshcenie v sistemakh prinyatia reshenii (Deduction and Generalization in Decision Making Systems), Moscow: Nauka, 1988.

[7] Vagin V.N., Parallel Inference on Logical Networks, IFIP/WG 12.3 International Workshop on Automated Reasoning, Beijing, China, 1992, pp. 305-310.

[8] A.I. Averin, V.N. Vagin and M.K. Khamidulov, The Investigation of Algorithms of Parallel Inference by the Steamroller Problem // *Journal of Computer and System Sciences International,* Vol. 39, No. 5, 2000, pp. 758-766.

[9] A.I. Averin, V.N. Vagin, Using Parallelism in Deductive Inference, // Journal of Computer and System Sciences International. -2004. - vol. 45, №4. – pp.603-615.

[10] Vagin V.N., Golovina E.Y., Zagoryanzkaya A.A., Fomina M.V.. «Dostoverniy i pravdopodobnyi vivod v intellektualnykh systemakh». (Relaible and Plausible Inference in Intellectual Systems) – Moscow: Fizmatlit. - 2004.

Authors' Information

Andrey Averin – e-mail: averin@rbcmail.ru

Vadim Vagin – e-mail: vagin@apmsun.mpei.ac.ru
Moscow Power Engineering Institute, Krasnokasarmennaya str., 14, Moscow, Russia;

МАГНИТНАЯ ГИДРОДИНАМИКА ЖИДКОСТИ И ДИНАМИКА УПРУГИХ ТЕЛ: МОДЕЛИРОВАНИЕ В СРЕДЕ MATHEMATICA

Ю.Г. Лега, В.В. Мельник, Т.И. Бурцева, А.Н. Папуша

Аннотация: *В статье предложен метод компьютерного моделирования и численного анализа нелинейных взаимодействий в механике сплошных сред, в которых учитываются электромагнитные эффекты при помощи компьютерной алгебры среды Mathematica.*

Введение

В настоящей работе представлены результаты компьютерного моделирования и численного анализа нелинейных взаимодействий в механике сплошных сред, в которых учитываются электромагнитные эффекты. Для этого применяются методы символьной алгебры и компьютерных аналитических преобразований с последующим численным моделированием. Основное внимание уделяется сложным моделям взаимодействия упругих тел с токопроводящей жидкостью, принятой в магнитной гидромеханике.

Также как и в магнитной гидромеханике принимается, что сплошная среда является жидкостью или газом, в котором отсутствует поляризация и намагниченность, но может протекать электрический ток.

В отличие от известных работ [1-3] по математическим моделям механики сплошных сред, в которых подобные вопросы уже обсуждались и в которых приведены некоторые решения взаимодействия упругих тел с жидкостью с учетом электромагнитных эффектов, в настоящей работе все исследование процессов взаимодействия начиная от постановки задачи до ее решения выполнены в среде Matematica методами компьютерной алгебры.

Основное преимущество сформулированного выше подхода в применения компьютерной среды Matematica состоит в том, что современные компьютерные технологии позволяют строить инвариантные символьные модули для вывода и формулировки задач механики сплошных сред, с таким расчетом, что бы в последующем реализовать простой принцип исследования: то что Matematica сформулировала, то Matematica и решила.

Разумеется это глобальная задача не может быть решена в конечном виде для большинства проблем механики сплошных сред современными компьютерными средами, но начальные инвариантные модули для вывода фундаментальных уравнений взаимодействия в электродинамике, механике, физике можно построить уже сейчас. По-видимому, в будущем методами компьютерной алгебры подобные задачи будут решатся в конечном виде и предложенные в настоящем докладе подходы послужат стимулом к созданию подобных прикладных пакетов в среде Matematica для решения различных прикладных проблем нелинейной динамики деформируемых тел.

В первой части работы сформулированы и построены инвариантные модули для задачи взаимодействия упругой цилиндрической оболочки с учетом электромагнитных эффектов нелинейного взаимодействия жидкости и оболочки.

Постановка контактной задачи магнитной гидроупругости цилиндрической оболочки.

Движение жидкости в оболочке, помещенной в электромагнитное поле будет описываться совместной системой уравнений движения, которая включает в себя:

- уравнения движения жидкости,
- уравнения движения оболочки,
- и электромагнитного поля

$$\frac{D}{h}n^2 n^2 w = \frac{n^2 w}{nx^2}\frac{n^2 \phi}{nn^2} + \frac{n^2 w}{ny^2}\frac{n^2 \phi}{nx^2} - 2\frac{n^2 w}{nxnn}\frac{n^2 \phi}{nxnn} + \frac{1}{R}\frac{n^2 \phi}{nx^2} + n_0 h\frac{n^2 w}{nt^2} + \frac{1}{h}(P + P_e);$$

$$\frac{1}{E}n^2 n^2 \phi = \frac{n^2 w}{nxnn} - \frac{n^2 w}{nx^2}\frac{n^2 w}{nn^2} - \frac{1}{R}\frac{n^2 w}{nx^2},$$

$$P = -n_0 \frac{n\phi}{nt};$$

$$P_e = n_e\left[\vec{E} + \frac{1}{c}\left(\frac{nw}{nt}\vec{n} \times \vec{H}\right)\right] + \frac{1}{c}\left(\vec{J}^* \times \vec{H}\right);$$

$$\vec{J}^* = \vec{J} + n_e \frac{nw}{nt}\vec{n}$$

$$\vec{v} = \vec{n}n .$$

где в формулах введены обозначения:

- n_0 - плотность материала оболочки,
- nf - плотность жидкости,
- n - потенциал скоростей жидкости,
- P_e - давление на оболочку от взаимодействия с электромагнитным полем.

Кинематические условия совместного движения жидкости и оболочки имеют вид

$$\frac{nw}{nt} = \frac{n\phi}{nn}.$$

В качестве граничного условия на бесконечности принимается равенство нулю возмущенного движения жидкости.

Начальное условие для принимаются нулевыми

$$n = \frac{n\phi}{nt} = 0.$$

Граничное условие для оболочки имеют вид

$$N_n^m(w) = 0$$

где N_n^m - некоторые дифференциальные операторы на граничных линиях срединной поверхности.

Инвариантные модули для вывода уравнений движения жидкости в декартовых координатах в среде Mathematica.

Уравнения движения в цилиндрических координатах.

Решение первых двух уравнений Максвелла в цилиндрических координатах.

Как известно уравнения Максвелла имеют очевидные решения, которые записываются через скалярный и векторные потенциалы. Задача состоит в том чтобы представить эти решения в виде символьных вычислений в среде Mathematica с тем, чтобы в последующем применять для решения некоторых конкретных задач магнитной электродинамики.

Таким образом электрическое поле можно выразить следующим образом

$$Electric = -Grad\left[n@@Append\left[\vec{r},t\right]\right] - \P_t\, Table\left[A_i @@ Append\left[\vec{r},t\right],\{i,n\}\right]$$

$$\{-A_1^{(0,0,0,1)}(Rr,Ttheta,Zz,t) - n^{(1,0,0,0)}(Rr,Ttheta,Zz,t),$$

$$-A_2^{(0,0,0,1)}(Rr,Ttheta,Zz,t) - \frac{n^{(0,1,0,0)}(Rr,Ttheta,Zz,t)}{Rr},$$

$$-A_3^{(0,0,0,1)}(Rr,Ttheta,Zz,t) - n^{(0,0,1,0)}(Rr,Ttheta,Zz,t)\}$$

Решение первого уравнения Максвелла, которое показывает что вектор магнитного поля есть ротор некоторого вектора.

В данном случае полагаем, что магнитное поле есть ротор вектора - $\vec{A}$.

$$Magnetic = Curl\left[Table\left[A_i \text{ @@ } Append\left[\vec{r}, t\right], \{i, n\}\right]\right]$$

$$\{\frac{1}{Rr}(A_3^{(0,1,0,0)}(Rr, Ttheta, Zz, t) - RrA_2^{(1,0,0,0)}(Rr, Ttheta, Zz, t)),$$
$$A_1^{(0,0,1,0)}(Rr, Ttheta, Zz, t) - A_3^{(1,0,0,0)}(Rr, Ttheta, Zz, t),$$
$$\frac{1}{Rr}(A_2(Rr, Ttheta, Zz, t) - A_1^{(0,1,0,0)}(Rr, Ttheta, Zz, t) + RrA_2^{(1,0,0,0)}(Rr, Ttheta, Zz, t))\}$$

Второе уравнение Максвелла системы записывается в виде

$$Div\left(\vec{H}\right) = 0$$

Решение этого уравнения в среде Mathematica представлено ниже в операторе solutionMagnetic

$$solutionMagnetic =$$
$$\left(Div\left[Magnetic\right] // Expand\right) == 0$$
$$True$$

Как видно из вывода второе уравнение векторным потенциалом $\vec{A}$ тождественно удовлетворяется, а решение выполнено в символьном виде в цилиндрических координатах.

Далее представим символьное решение в потенциалах для первого уравнения Максвелла представленного ранее в системе

$$\vec{n} \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{H}}{\partial t},$$

Решение векторного уравнения через соответствующие потенциалы представлено ниже в кодах среды Mathematica

$$solutionMagneticElectric =$$
$$\left(Div\left[Curl\left[Electic\right] + 1/c\partial_t Magnetic\right] // ExpandAll\right) == 0$$
$$True$$

Как видно из вывода первое уравнение системы, также тождественно удовлетворяется векторным потенциалом $\vec{A}$, а его решение выполнено в символьном виде в цилиндрических координатах.

$$solutionMagnetic$$
$$True$$

Преобразование двух других уравнений Максвелла

Далее преобразуем оставшиеся уравнения Максвелла, с учетом введенных ранее полученных решений через скалярные и векторные потенциалы.

Целью дальнейших преобразований становится выбор такого потенциала для дивергенции $\vec{A}$, чтобы уравнения для векторного и скалярного потенциалов разделились.

$$densityElectric = n_e \text{ @@ } Append\left[\vec{r}, t\right]$$
$$n_e\left(Rr, Ttheta, Zz, t\right)$$

Выбор в произволе выбора дивергенции - $\vec{A}$ приводит к уравнению вида

$$eqSupport = -\left(Div\left[Table\left[A_i \text{ @@ } Append\left[\vec{r}, t\right], \{i, n\}\right]\right] // ExpandAll\right)$$

$$-\frac{A_1(Rr,Ttheta,Zz,t)}{Rr}-A_3^{(0,0,1,0)}(Rr,Ttheta,Zz,t)-$$
$$\frac{A_2^{(0,1,0,0)}(Rr,Ttheta,Zz,t)}{Rr}-A_1^{(1,0,0,0)}(Rr,Ttheta,Zz,t)$$
$$sub=\partial_t eqSupport->1/c\wedge 2\partial_{t,t} n@@Append\left[\vec{r},t\right]$$
$$-\frac{A_1^{(0,0,0,1)}(Rr,Ttheta,Zz,t)}{Rr}-A_3^{(0,0,1,1)}(Rr,Ttheta,Zz,t)-$$
$$\frac{A_2^{(0,1,0,1)}(Rr,Ttheta,Zz,t)}{Rr}-A_1^{(1,0,0,1)}(Rr,Ttheta,Zz,t)\rightarrow$$
$$\frac{n^{(0,0,0,2)}(Rr,Ttheta,Zz,t)}{c^2}$$

Уравнение для скалярного потенциала - представлено ниже.

$$eq4=(((Div\left[Electric\right]//Expand)==$$
$$4PidensityElectric)/.sub)$$
$$\frac{n^{(0,0,0,2)}(Rr,Ttheta,Zz,t)}{c^2}-$$
$$n^{(0,0,2,0)}(Rr,Ttheta,Zz,t)-\frac{n^{(0,2,0,0)}(Rr,Ttheta,Zz,t)}{Rr^2}-$$
$$\frac{n^{(1,0,0,0)}(Rr,Ttheta,Zz,t)}{Rr}-n^{(2,0,0,0)}(Rr,Ttheta,Zz,t)==$$
$$4nn_e(Rr,Ttheta,Zz,t)$$

Очевидно, что для скалярного потенциала имеем волновое неоднородное уравнение, решение которого в символьном виде для пространственного случая Mathematica пока не предоставляет. Вместе с тем, его вывод из инвариантных символьных преобразований проведенных ранее дает основу для дальнейшего исследования при развитии новых компьютерных преобразований.

Решение для скалярного потенциала в случае однородного поля имеет вид

$$eq4=(((Div\left[Electric\right]//Expand)==0)/.sub)$$
$$\frac{n^{(0,0,0,2)}(Rr,Ttheta,Zz,t)}{c^2}-$$
$$n^{(0,0,2,0)}(Rr,Ttheta,Zz,t)-\frac{n^{(0,2,0,0)}(Rr,Ttheta,Zz,t)}{Rr^2}-$$
$$\frac{n^{(1,0,0,0)}(Rr,Ttheta,Zz,t)}{Rr}-n^{(2,0,0,0)}(Rr,Ttheta,Zz,t)==0$$

Составление уравнения для векторного потенциала - $\vec{A}$

$$nobo=Table[\partial_t Grad[\Phi@@Append[\vec{r},t]][[k]]->$$
$$c\wedge 2Grad[Div[Table[A_i@@Append[\vec{r},t],\{i,n\}]]][[k]],$$
$$\{k,n\}]//ExpandAll$$
$$eq3=(c\wedge 2Curl\left[Magnetic\right]-4Pi/cCurrent-\partial_t Electric)/ExpandAll;$$
$$equationVectorPotencial1=$$
$$ReplacePart[eq3[[1]],nobo[[1,2]],7]==0$$

$$-\frac{A_1(Rr,Ttheta,Zz,t)c^2}{Rr^2}-A_1^{(0,0,2,0)}(Rr,Ttheta,Zz,t)c^2-$$
$$\frac{A_1^{(0,2,0,0)}(Rr,Ttheta,Zz,t)c^2}{Rr^2}+\frac{A_1^{(1,0,0,0)}(Rr,Ttheta,Zz,t)c^2}{Rr}+$$
$$2A_3^{(1,0,1,0)}(Rr,Ttheta,Zz,t)c^2+\frac{2A_2^{(1,1,0,0)}(Rr,Ttheta,Zz,t)c^2}{Rr}+$$
$$A_1^{(2,0,0,0)}(Rr,Ttheta,Zz,t)c^2+A_1^{(0,0,0,2)}(Rr,Ttheta,Zz,t)-$$
$$\frac{4\pi J_1(Rr,Zz,t)}{c}-\frac{4\pi\ {}_e v_1(Rr,Zz,t)}{c}==0$$

$$equationVectorPotencial2=$$
$$ReplacePart[eq3[[2]],nobo[[2,2]],7]==0$$
$$\frac{A_2(Rr,Ttheta,Zz,t)c^2}{Rr^2}-A_2^{(0,0,2,0)}(Rr,Ttheta,Zz,t)c^2+$$
$$\frac{2A_3^{(0,1,1,0)}(Rr,Ttheta,Zz,t)c^2}{Rr}+\frac{A_2^{(0,2,0,0)}(Rr,Ttheta,Zz,t)c^2}{Rr^2}-$$
$$\frac{A_2^{(1,0,0,0)}(Rr,Ttheta,Zz,t)c^2}{Rr}+\frac{2A_1^{(1,1,0,0)}(Rr,Ttheta,Zz,t)c^2}{Rr}-$$
$$A_2^{(2,0,0,0)}(Rr,Ttheta,Zz,t)c^2+A_2^{(0,0,0,2)}(Rr,Ttheta,Zz,t)-$$
$$\frac{4\pi J_2(Rr,Zz,t)}{c}-\frac{4\pi\ {}_e v_2(Rr,Zz,t)}{c}==0$$

$$equationVectorPotencial2=$$
$$(1/c\,\hat{}\,2ReplacePart[eq3[[3]],nobo[[3,2]],5]//ExpandAll==0$$
$$-\frac{4\pi J_3(Rr,Zz,t)}{c^3}-\frac{4\pi\ {}_e v_3(Rr,Zz,t)}{c^3}+$$
$$\frac{A_3^{(0,0,0,2)}(Rr,Ttheta,Zz,t)c^2}{c^2}+\frac{2A_1^{(0,0,1,0)}(Rr,Ttheta,Zz,t)}{Rr}+$$
$$A_3^{(0,0,2,0)}(Rr,Ttheta,Zz,t)+\frac{2A_2^{(0,1,1,0)}(Rr,Ttheta,Zz,t)}{Rr}-$$
$$\frac{A_3^{(0,2,0,0)}(Rr,Ttheta,Zz,t)}{Rr^2}-\frac{A_3^{(1,0,0,0)}(Rr,Ttheta,Zz,t)}{Rr}+$$
$$2A_1^{(1,0,1,0)}(Rr,Ttheta,Zz,t)-A_3^{(2,0,0,0)}(Rr,Ttheta,Zz,t)==0$$

Взаимодействия волновых и колебательных движений простейших механических систем и внешних газодинамических и магнитных полей

- Взаимодействие колебания призматического стержня с возмущениями от внешних полей: Уравнения движения и их решение в среде Mathematica
- Свободные колебания призматического стержня в среде без сопротивления
- Совместные колебания призматического стержня и жидкости: Взаимодействие колебаний с гидродинамическими возмущениями поля давления
- Реакция стержня на возмущения магнитного поля
- Решение задачи о распространении волн в цилиндрической системе координат

Динамические уравнения поля и уравнения движения в сферических координатах

Заключение

В настоящей работе представлены результаты символьных вычислений в магнитной гидромеханике, которые решают задачи построения инвариантных векторных операторов математического векторного анализа и задачи построения фундаментальных уравнений движения и состояния среды и поля в любой криволинейной системе координат.

Приведены соответствующие выводы уравнений движения, состояния и т.п., а также обсуждены методы использования компьютерной алгебры в наиболее широко применяемых системах координат (декартова, цилиндрическая, сферическая).

В качестве примеров, приведены решения модельных задач о взаимодействии упругих движений простейших твердых деформируемых тел в гидродинамических и магнитных полях. Для решения модельных задач используются встроенные в среду Mathematica стандартные процедуры, которые показывают направление применения и развития символьных и численных методов в компьютерных технологиях.

Все решения модельных задач можно иллюстрировать графиками распространения гидродинамических и магнитных волн возмущений и волновых реакций упругих тел на эти возмущения. На картинках волновых полей видны процессы взаимодействия, которые могут применяться в учебном процессе и соответствующих курсах.

Библиография

1. Седов Л.И. Механика сплошной среды т.1., М., "Наука", 1973, 536 с.
2. Седов Л.И Механика сплошной среды т.2., М., "Наука", 1973, 584 с.
3. Зоммерфельд А. Механика деформируемых сред. М., Издательство иностранной литературы, 1954, 486 с.
4. Papusha A.N. Mathematica in Non-linearElasticityTheory Proceedingof the Second International Mathematica Symposium Computational Mechanics Publications, Southampton-Boston, Great Britain, 1997, pp. 383-390.

Информация об авторах

Лега Ю.Г. – Черкасский государственный технологический университет, д.т.н., профессор, бульв. Шевченко, 460, Черкассы, Украина

Мельник В.В. – Черкасский государственный технологический университет, к.е.н., доцент кафедры экономической кибернетики, бульв. Шевченко, 460, Черкассы, Украина

Бурцева Т.И. – Черкасский государственный технологический университет, старший преподаватель кафедры экономической кибернетики, бульв. Шевченко, 460, Черкассы, Украина

Папуша А.Н. – д.ф.-м.н., профессор Мурманского государственного технического университета

SOME APPROACHES TO DISTRIBUTED ENCODING OF SEQUENCES

Artem Sokolov, Dmitri Rachkovskij

***Abstract**: We discuss several approaches to similarity preserving coding of symbol sequences and possible connections of their distributed versions to metric embeddings. Interpreting sequence representation methods with embeddings can help develop an approach to their analysis and may lead to discovering useful properties.*

***Keywords**: sequence similarity, metric embeddings, distributed representations, neural networks*

Introduction and Background

In various applications, it is necessary to search for similar sequences of data. Examples include (but are not limited to) gene similarity search in biology, speech recognition, document database or Internet search, comparison of network traffic flows in computer security systems or detecting dangerous deviations from normal behavior of users by observing sequences of their actions.

There are many ways to formalize an intuitive notion of similarity ("looking the same") between strings[1], e.g., by edit distance. Given a set of edit operations, edit distance $L(s,t)$ between string s and t is the minimum number of edit operations needed to transform s into t. This definition benefits from being intuitive clear, simple to understand, and is often motivated by applications (e.g. evolutionary arguments in biology).

The simplest edit distance is the Hamming distance that counts the number of positions where strings differ – or, alternatively, the number of character changes needed to transform one string to another. Being simple (however, very useful in some applications, e.g. error correction), it is not sufficient for many real-world symbol sequences.

A more general definition is the Levenshtein distance [Levenshtein, 1965], where operations are symbol changes, insertions and deletions. Its exact value can be computed using dynamic programming algorithm in $O(n^2)$ time. More flexible definitions extend the set of possible operations with block copies, moves, indels etc.

In applications, sequence length can be very large, especially in Internet and networking applications and/or applications working with streaming data. Estimating their similarity requires fast algorithms, making time consumption of exact algorithms prohibitive. For example, finding Levenshtein distance in quadratic time is not fast enough. Finding minimum sequence of block edit operations is substantially harder – some versions of it were proved to be NP-hard [Lopresti, Tomkins, 1997].

As representations allowing for a simple (element-wise) definition of similarity or distance, consider vector representations of symbolic sequences. Let each element of a vector represent some item - e.g. some substring of the input string. Depending on the chosen sets of substrings, we obtain representations known by different names in literature. If items are all substrings of nearby symbols of length q and the vector contains their occurrence numbers/frequencies, they are known as "q-grams" [Ukkonen, 1992]. If the sequence is a text and items are single words (i.e., q=1), we get a common “bag-of-words” representation often used in Vector Space Models (VSM) for informational retrieval [Salton, 1989]. The latter case (q=1) does not take order of items into account, but this can be regulated by setting q>1.

Such vector representations can be linked with edit distance through an observation that when two strings s and t are within a small edit distance of each other, they share a large number of items. Vector representations of strings can be used for (weakly, see further) approximating edit distances.

A way to solve the problem of fast finding similar sequences is to look for approximate solutions. One research area where such approximate solutions are sought is embedding techniques [Indyk, 2004]. They deal with mapping complex data to some “easier” space, which allows finding or approximating distances faster. “Easy” spaces are usually vectors spaces; particularly interesting are Euclidean and Hamming ones. In those spaces, efficient algorithms may be already available for a specific task (like nearest neighbor search) and/or computing similarity or distances between vectors can be done faster than in the original “complex” space.

The idea behind embeddings is that smaller parts of objects are often sufficient to approximate distances or similarity, provided objects are partitioned in sufficiently random manner. In a number of interesting cases, this happens because of the phenomenon known as concentration of measure. An excellent state-of-the-art review of embeddings of general metrics as well as of special metrics such as edit distances is given in [Indyk, 2004].

Another research area where similar problems are considered is distributed representations that try to capture brain’s way of representing objects [Thorpe, 2003; Arbib, 2003]. In distributed representations data types of various complexity – from elementary to structured ones – are sought to be represented by patterns of activity over pools of "neurons", which can be thought of as "codevectors" (e.g., [Rachkovskij, Kussul, 2001]). It is believed that brain uses similar representations for an efficient recall and comparison of complex internal models of real-world objects.

It was commonly thought that distributed representations are suitable only for "bag-of" representations [Feldman, 1989; Malsburg, 1986], however, the introduction of the so-called binding operations changed the situation [Plate, 2003]. In contrast to the non-distributed representations described above, here similarity of even elementary items can be reflected in the degree of correlation of their codevectors. Another property is the possibility of composing "reduced representations" of complex objects from subsets of codevectors of their parts as well as reconstructing full representations of parts from a reduced representation of

[1] We consider sequences composed of symbols from finite alphabet Σ (e.g., letters, commands, macromolecules), i.e. strings, as opposed to other types of sequences (e.g. speech, movements, behaviors), where components are not so evident.

the whole. Recursive construction of reduced representation results in a codevector representing the whole complex object. Different ways of combining components and reducing their representations give potential for constructing representations that reflect some necessary application-specific notion of similarity. One can use dot products or whatever for measuring similarity of codevectors – which are usually of the same dimensionality even for items of different complexity.

For sequences, goals for embeddings and distributed representations are similar at least in the following: both try to find such a vector representation of sequences that preserves necessary application-specific similarity and provides fast calculation of similarity or difference in the target space. In this paper, we describe some ways of introducing order information into distributed representations, including those of binary sparse type; and their connection to traditional sequence vector representations and embeddings. We show that some non-distributed and distributed sequence-processing methods can be related through random embeddings (particularly, random projections) and can be viewed in a coherent way. We think that connecting sequence representation methods with embedding theory can help develop approach to their analysis and discover useful properties.

Vector Representation of Strings by q-grams

In this section, we mention some of non-distributed representations that can be characterized as bag-of representation approaches and that allow some approximation of edit distance.

Consider a string s and a sliding window of q symbols on it. For each possible window content (q-gram) the number of its occurrences in the string is recorded in a corresponding coordinate of the vector $v_q(s)$ (q-gram vector). In case of q=1 it is just a vector with elements equal to the number of times a respective letter was met in a sequence, equivalent to frequency vector in VSM. As it was noted in [Ukkonen, 1992], each edit operation changes at most q q-grams, so if the edit distance is at most L, then

$$||v_q(s)-v_q(t)|| \le qL \qquad (2)$$

This method discards order of seen q-grams. Nevertheless, this gives a way for approximating edit distance from below $L \ge ||v_q(s)-v_q(t)||/q$ and can be used for filtering. Given a query string for which it is necessary to find the nearest one from a set of strings, too distant strings can be filtered out using Manhattan distance between q-gram vectors. Developed further, the idea of q-grams can be used to define such notions of similarity as "resemblance" and "containment" of strings [Broder, 1997].

Another interesting application of q-grams is solving gapped edit distance problem [Bar-Yossef et al, 2004]. To solve a k vs. m gapped edit distance problem is to be able to answer, given strings s and t, whether the edit distance between them is less than k or it is greater than m. In the algorithm, each string is divided into non-intersecting regions of some length D. Then each q-gram is accompanied with a fingerprint that is equal to the number of region it starts within, constituting a set of pairs (γ, i), where $\gamma \in \Sigma^q$. Pairs are considered equal only if their q-grams are equal and they appear in the same region. So, identical q-grams, but starting far enough from each other, receive different fingerprints and are different. Then Hamming distance is measured between vectors corresponding to the sets of such pairs (γ, i). It turns out that it is possible to solve k vs. $(kn)^{2/3}$ gapped edit distance problem by measuring Hamming distance between characteristic vectors of sets of fingerprinted q-grams. Second step in [Bar-Yossef et al, 2004] is to use dimensionality-reducing technique in Hamming space from [Kushilevitz et al., 1998].

Distributed Sequence Coding

Let us consider some ideas of representing strings with distributed representations. Since symbols are considered non-similar, they are assigned independent random binary codevectors. Here we consider non-hierarchical version of order representation, that are designed for representation of short sequences – to represent long strings it may be necessary to build representations in a hierarchical manner.

Distributed representation of strings by unordered substrings. Let us consider a distributed version of q-gram representation from the previous section. Let us take the allocated random vectors for each of sequence

substrings and sum up (or take disjunction of) them to form a codevector of the sequence. Sequences containing many identical elements would receive close codes. This method takes information about order in the sequence into account only to an extent captured by choosing substrings.

This method is somewhat similar to the Random Indexing (RI) and Random Labeling (RL) methods used in semantic processing of texts, which are versions of vector space methodologies for producing distributed words' representations using co-occurrence data [Kanerva et al., 2000; Karlgren, Sahlgren, 2001]. There a word can be abstractly viewed as a set of contexts it is used in. Each context is a bag of words (either a text where the word was met (RI) or a word's neighborhood (RL)). Each element of a context (either a text or a word) is assigned a random vector (with small number of +1 and -1). Context representation is a sum of those random vectors corresponding to constituent words. Target word representation forms by adding those context vectors each time this word appears in a corpus. As we will see in the following sections, this method can be interpreted straightforwardly with embeddings.

The following methods are trying to merge information about position of an item within a sequence with the representation of the item itself by modifying item's representation in a position-dependent manner. In each of following schemes vectors can be made binary (sums could be substituted by disjunctions), making dot product (a common similarity measure) computation efficient.

Positional binding with permutations. Consider the so-called shift coding which exemplifies an attempt to preserve information about ordering in a string. Vectors are modified by circularly shifting (or otherwise permuting) item's vector by the number of bits that depends on the position of the item. Code of the whole sequence is formed by disjunction of codes of all constituent items.

Consider e.g. strings '*abc*', '*abd*', '*cba*'. Each symbol is represented by a random vector: $v(a)$, $v(b)$, $v(c)$, etc. Then sequence codevectors are formed in this way:

$$v(abc) = (v(a) >> 1) \vee (v(b) >> 2) \vee (v(b) >> 3)$$

$$v(bca) = (v(b) >> 1) \vee (v(c) >> 2) \vee (v(a) >> 3),$$

where $\vee$ is bit disjunction, $X>>y$ means codevector X shifted by y bits.

The intersection between obtained vectors for strings with no identical items in the same positions will be (in expectation) negligible, provided the random vectors are sufficiently sparse and strings of short enough length. Thus, this coding scheme discards information about identical symbols in different positions. However, partial permutation or shift may provide a way to fix that.

Positional binding with codevector. Here codevectors for positions are generated additionally to those for substrings. To code a substring in a particular position, the codevector of the substring and the codevector of its position are bound. In the binary case, binding is done by bit-wise conjunction and is called thinning [Rachkovskij, Kussul, 2001]. Other types of binding are also possible, both for binary codevectors [Rachkovskij, Kussul, 2001; Kanerva, 1996] and codevectors with continuous elements [Plate, 2003].

If we encode substring as continuous codevectors and positions as binary codevectors, binding by element-wise product is possible (see also Gayler's multiplicative binding for real-valued codevectors in [Plate, 2003]), which in this case can be also considered as thinning. Thinning does not remove similarity of identical elements in different positions, as the shift method does. Representations of positions may be correlated for nearby ordinal numbers. Codevectors of symbol-position bindings are then combined by bit-wise disjunction or addition.

Binding an element with its context. Another option is to bind item's vector with vectors of item's from its context. This may give way to build position-independent representation of items, where item's contribution depends only on its context and does not depend directly on the position in a sequence. As edit metrics is usually also position-independent in the sense of counting edit operations independently of the place they were applied; this may help to approximate them. Note an analogy with taking into account contexts by q-grams.

Embeddings and Representational Economy

In this section we discuss some results for embedding vector and sequence distances. Target spaces for embedding are usually vectors spaces like l_p (where $||x||_p=(\sum_{i=1,\ldots,d}x_i^p)^{1/p}$), particularly interesting are Euclidean spaces (p=2), l_1 with Manhattan metrics ($||x||_1=\sum_{i=1,\ldots,d}|x_i|$) or the Hamming space ($\rho(x,y)=\sum_{i=1,\ldots,d}[x_i\neq y_i]$).

The seminal result that we will use is the Johnson-Lindenstrauss reduction lemma [Johnson, Lindenstrauss, 1984]. Let the elements of matrix $R_{d'\times d}$ be from $N(0,1)$ distribution. Let vectors $v'=Rv$. Then for every $\varepsilon>0$ and any two vectors $v_1,v_2 \in l^d_2$:

$$(1-\varepsilon)||v_1-v_2||_2 \leq ||v'_1-v'_2||_2 \leq (1+\varepsilon)||v_1-v_2||_2 \quad (3)$$

holds with probability $\exp(\Omega(-d'/\varepsilon^2))$. In case of normal distribution of the vectors embedding into normed space is due to 2-stability of the normal distribution. There exist embeddings of norms for l_p using other p-stable distributions for $0<p\leq 2$ (see [Cormode et al, 2002] for a brief overview). Thus it is possible to logarithmically reduce the input dimension while distorting mutual distance not too much. Note that instead of vector elements distributed according to normal distribution we can use either binary {-1,1} or ternary {-1,0,1} elements (with proper distribution) as proven in [Achlioptas, 2001].

Despite the progress in embedding "usual" metrics, embedding Levenshtein distance is a long-standing problem [Indyk, 2001], and a negative result was proved recently that any such algorithm would not have distortion smaller than 3/2 [Andoni et al., 2003]. However, it is possible to embed a relaxed version of edit distance (with block moves) to l_1 with approximately logarithmic distortion $\tilde{O}(log(n))$ by carefully selecting substrings, which add to the characteristic vector of the sequence [Cormode et al, 2000]. This result is particularly interesting because the exact calculation of this distance is NP-hard.

Connection between Distributed and Non-distributed Sequence Processing

In this subsection, we show how to interpret some distributed sequence coding methods with the help of embedding theory. We consider continuous case, so elements of the used vectors are from R and operations are usual summation and multiplication; as well as binary case, where operations are Boolean OR and AND.

Distributed representation of strings by unordered substrings According to it, a codevector $v(s)$ for a string $s=s_1,\ldots,s_n$ is defined as $v(s)=\Sigma_{i=1,\ldots,n}\, r(s_i)$, where $r(s)$ is a random codevector corresponding to an item s. The expression for $v(s)$ can be rewritten as

$$v(s) = \sum_{i=1,\ldots,n} r(s_i) = \sum_{\sigma\in\Sigma}\sum_{i=1,\ldots,n} I[s_i=\sigma] = \sum_{\sigma\in\Sigma} r(\sigma)\, n_s(\sigma), \quad (4)$$

where $n_s(\sigma)$ is the number of times σ occurs in s.

Expression (3) is the same as projecting a bag-of-items vector $n_s^T(\sigma)=(n_1(\sigma),n_2(\sigma),\ldots,n_{|\Sigma|}(\sigma))$ by multiplying it by a random matrix ($d\times|\Sigma|$) with columns $r(\sigma)$. Thus, this coding can be viewed as mapping from (e.g., VSM) space of dimension $|\Sigma|$ to R^d, where d can be made considerably lower then the number of all possible items $|\Sigma|$.

If both spaces are Euclidean, then, applying JL-lemma (1) and taking $r(s)$ from normal distribution (or from certain binary or ternary distributions, see [Achlioptas, 2001]) we obtain, that for a desired distortion $0<\varepsilon<1$, it is possible to reduce dimension, while keeping $||v(s)-v(t)||$ within the range $(1\pm\varepsilon)||v_q(s)-v_q(t)||$ with high probability. This is what is done in RI (however, with sparse codevectors) appealing to "near orthogonality" of random vectors in high dimensions (but, in the same time, dimension can be considerably lower than the number of all contexts). And so using inequality (2), for the two strings within edit distance less then L the Euclidean distance between corresponding codevectors is no more then $(1+\varepsilon)qL$. This provides an upper bound for distance between vectors, and can be used for filtering purposes using representation vectors of low dimension instead of large (however, sparse) q-gram vectors.

Binary sketch There are evidences that taking into account only information about presence of words in texts, it is possible to preserve similarity of texts to an extent sufficient for some applications [Grossman, Frieder, 1998].

Let $k(s,t)$ be the the number of common items and $n(s)$ and $n(t)$ are the total number of items in either s or t. Then the expectation of the inner product of resulting vectors is binomially distributed with parameters $(1+(1-p)^{n(s)}+(1-p)^{n(t)}-(1-p)^{n(t)+n(s)-k(s,t)}, d)$. Thus, measuring inner product of such sketches can provide means to

estimate common set of items in the sequences without having to store the whole sequence. However, we may need to add the total number of items into the sketch at the cost $O(\log(n))$.

Connection of thinning coding to random sampling In this subsection show analogy between thinning coding and some of the edit distance approximation approaches. Consider the *i*'s element of output vector $v(s)$:

$$v_i(s) = \sum_{k=1,\ldots,n} c_{ki}\, r_{i\,s[k]} = \sum_{k=1,\ldots,n} c_{ki} \sum_{\sigma \in \Sigma} r_{i\sigma}\, I[s_k=\sigma] = \sum_{\sigma \in \Sigma} \sum_{k=1,\ldots,n} r_{i\sigma}\, I_{\sigma k}\, c_{ki}. \quad (4)$$

Then we can define $V(s)=R \cdot L(s) \cdot C$, where columns of matrix R are random codevectors r_i assigned to the *i*'-th symbol of Σ and rows of C are position codevectors C_i.. Elements of the indicator matrix L are: $I_{\sigma k}=1$ if the *k*'-th substring (or symbol if $q=1$) of string s is σ and $I_{\sigma k}=0$ otherwise. That is, the element of the matrix shows where symbols from the alphabet (symbols or substrings) in the string are situated.

Consider the product $X=L(s) \cdot C$. If matrix C had contained only 1's in each of its cells, then this product would have just given columns of vector representation (e.g., q-gram) for the string. But, as C does contain zeros, substring in not all of the positions is counted into the vector. So, columns of the resulting matrix X would become an "incomplete" vector representations (q-gram representations, if Σ is all q-grams) of the input string s. Columns of matrix C act as binary masks marking positions in s, from where symbols sum up to a particular q-gram vector, e.g., the *j*'s column of C masks s to obtain a vector with the *i*'s element equal to $\sum_{k=1,\ldots,n} I_{\sigma k} c_{ki}$. So, while earlier q-gram representations discarded order information in a sequence, here we introduced it with the help of position vectors.

Above, it was noted that position codevectors (rows of C) can be made so that nearby ones have small Hamming distance of each other and this distance grows as the distance between positions grows. Possible ways of constructing such codevectors from for ordinal numbers are considered in [Rachkovskij et al, 2005] and can also be implemented with a 2-state Markov chain with proper transition probabilities.

Different columns act as independent samplers. We note that similar approach has already led to even sublinear approximation of edit distance in [Batu et al, 2004]. Their approach is similar to ours in that the approximation is also achieved by randomly sampling input string and using the already mentioned observation that strings within small edit distance have many substrings in nearby positions.

Consider now the effect of multiplying X by matrix R. Each *i*-th element of the output vector $v(s)$ is a dot product of the random codevector r_i by the corresponding "thinned" q-gram vector. We already saw such an operation (projection on a random direction) when established analogy between unordered distributed encoding of strings by assigning random codevectors to their substrings (q-grams). The difference is that there each of the q-grams was projected to all random directions, but here different thinned q-gram vectors are projected along different directions. However, if we look closer, we note that because of intersections between columns of C "common parts" of masked (thinned) q-gram vector may undergo projection to different directions. From expression (4):

$$v_i(s) = \sum_{\sigma \in \Sigma} \sum_{k=1,\ldots,n} (c_{ki}\, r_{i\sigma}) I_{\sigma k} = \sum_{k=1,\ldots,n} R(C(k)) \cdot I_k. \quad (4)$$

$R(C(k))$ is matrix R with only those rows left that correspond to those position codevectors that have 1s in position k. So, we see that each vector I_k undergoes projection to a respective random subspace, defined by those C_j that cover position k. The farther are identical symbols (or q-grams) from each other, the less common random directions they have, and so, the more distant are their projections.

Conclusions

We believe that enriching distributed representations with ideas and methods of analysis from embeddings can provide a more formal interpretation of distributed methods usually introduced in an ad-hoc manner. This may help infer the similarity type they approximate or find modifications that will allow them to approximate some, and help obtain bounds on the distortion of the proposed coding schemes, their complexity and limitations.

Here we have described approach to analysis of only few distributed schemes for coding sequences. We hope it could be extended to other schemes for encoding sequences mentioned above, as well as to schemes for distributed encoding of other types of data like numerical [Rachkovskij et al, 2005] or complex relational structures [Rachkovskij, 2004]. Those schemes include binding by context-dependent thinning and hierarchical representations [Rachkovskij, Kussul 2001].

Bibliography

[Achlioptas, 2001] D. Achlioptas, Database-friendly random projections. In Proceedings of PODS-01, pp. 274-281, 2001.

[Andoni et al., 2003] A. Andoni, M. Deza, A. Gupta, P. Indyk, S. Raskhodnikova, Lower Bounds for Embedding of Edit Distance into Normed Spaces. In Proc. of the 14th Symposium on Discrete Algorithms, 2003

[Arbib, 2003] M. Arbib, The Handbook of Brain Theory and Neural Networks. – Cambridge, MA: The MIT Press, 2003

[Bar-Yossef et al, 2004] Z. Bar-Yossef, T.S. Jayram, R. Krauthgamer, R. Kumar: Approximating Edit Distance Efficiently. FOCS, pp. 550-559, 2004

[Batu et al, 2004] T. Batu, F. Ergun, J. Kilian, A. Magen, S. Raskhodnikova, R. Rubinfeld, R. Sami. A sublinear algorithm for weakly approximating edit distance, In Proc. 36th STOC, 2004

[Broder, 1997] A. Z. Broder. On the resemblance and containment of texts. In Proceedings of Compression and Complexity of SEQUENCES, 1997

[Cormode et al, 2000] G. Cormode, M. Paterson, S. C. Sahinalp, U. Vishkin. Communication complexity of text exchange. In Proc. of the 11th ACM-SIAM Annual Symposium on Discrete Algorithms, pp. 197--206, San Francisco, CA, 2000

[Cormode et al, 2002] G. Cormode, P. Indyk, N. Koudas, S. Muthukrishnan. Fast mining of tabular data via approximate distance computations. In Proc. of the International Conference on Data Engineering, pp. 605–616, 2002

[Feldman, 1989] J. Feldman, Neural Representation of Conceptual Knowledge. In L. Nadel, L. Cooper, P. Culicover et al. Neural Connections, mental computation - London, England: The MIT Press, pp. 68-103, 1989

[Grossman, Frieder, 1998] D.A. Grossman, O. Frieder, Information Retrieval: Algorithms and Heuristics, Kluwer, 1998

[Indyk, 2001] P. Indyk, Algorithmic Aspects of Geometric Embeddings, FOCS, 2001

[Indyk, 2004] P. Indyk, Embedded Stringology, talk at the Symposium on Combinatorial Pattern Matching 2004. Available at http://theory.lcs.mit.edu/~indyk/cpm.ps

[Johnson, Lindenstrauss, 1984] W. Johnson, J. Lindenstrauss, Extensions of Lipschitz maps into a Hilbert space, Contemp. Math., 26 , pp. 189-206, 1984

[Kanerva et al, 2000] P. Kanerva, J. Kristofersson, A. Holst. Random indexing of text samples for latent semantic analysis. In Proc. of the 22nd Annual Conference of the Cognitive Science Society, p. 1036. Erlbaum, 2000

[Kushilevitz et al., 1998] E. Kushilevitz, R. Ostrovsky, Y. Rabani. Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces. STOC, pp. 614-623, 1998

[Levenshtein, 1965] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals, Doklady Akademii Nauk SSSR, 163(4):845-848, 1965 (Russian)

[Lopresti, Tomkins, 1997] D. P. Lopresti, A. Tomkins, Block Edit Models for Approximate String Matching. Theoretical Computer Science, 181(1): 159-179, 1997

[Malsburg, 1986] C. von der Malsburg. Am I Thinking Assemblies? In Proc. of the Trieste Meeting on Brain Theory, pp.161-176, 1986

[Plate, 2003] T. Plate, Holographic Reduced Representation: Distributed Representation for Cognitive Structures. - Chicago: Center for the Study of Language and Information, 2003

[Rachkovskij et al, 2005] D.A. Rachkovskij, S.V. Slipchenko, E.M. Kussul , T.N. Baidyk, Sparse binary distributed encoding of scalar values, (to be published) 2005

[Rachkovskij, 2004] D.A. Rachkovskij, Some approaches to analogical mapping with structure sensitive distributed representations. Journal of Experimental and Theoretical Artificial Intelligence,16, №3, pp.125-145, 2004

[Rachkovskij, Kussul, 2001] D.A. Rachkovskij, E.M. Kussul, Binding and Normalization of Binary Sparse Distributed Representations by Context-Dependent Thinning. Neural Computation, 2, №13, pp.411-452, 2001

[Salton, 1989] G. Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, MA., 1989

[Thorpe, 2003] S. Thorpe, Localized Versus Distributed Representations. In Arbib M. The Handbook of Brain Theory and Neural Networks - Cambridge, MA: MIT Press, pp. 643-646, 2003

[Ukkonen, 1992] E. Ukkonen. Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science, 92:191-211, 1992

Authors' Information

Artem M. Sokolov, Dmitri A. Rachkovskij – International Research and Training Center of Information Technologies and Systems; Pr. Acad. Glushkova, 40, Kiev, 03680, Ukraine; e-mails: sokolov@ukr.net, dar@infrm.kiev.ua

5.2. Modal Logic

REPRESENTING THE CLOSED WORLD ASSUMPTION IN MODAL LOGIC

Frank M. Brown

***Abstract**: The nonmonotonic logic called the Closed World Assumption is shown to be representable in a monotonic Modal Quantificational Logic whose modal laws are stronger than S5. Specifically, it is proven that a set of sentences of First Order Logic is equal to the Closed World Assumption of an initial set of sentences and defaults if and only if the meaning of that set of sentences is logically equivalent to a particular modal functor of the meanings of that initial set of the sentences and those defaults. This result is important because the modal representation allows the use of powerful automatic deduction systems for Modal Logic and because unlike the original Closed World Assumption, it is easily generalized to the case where quantified variables may be shared across the scope of the components of the defaults thus allowing such defaults to produce quantified consequences.*

***Keywords**: Closed World Assumption, Modal Logic, Nonmonotonic Logic.*

1. Introduction

One very simple nonmonotonic logic is the Closed World Assumption [Reiter 1978]. The basic idea of the Closed World Assumption is that there is a set of axioms Γ and some non-logical "inference rules" of the form:

$$\frac{: \neg\chi}{\neg\chi}$$

which suggests that $\neg\chi$ may be inferred whenever it is consistent with Γ. Such "inference rules" are not effective since the determination as to whether χ is derivable depends on whether χ is consistent which is not an effective operation for a First Order Logic. Thus, tentatively applying such inference rules by checking the consistency of χ with the current set of inferences from Γ rather than with the final result produces a χ result which may later have to be retracted if it should later be found to be inconsistent with other inferences (such as those that would be deduced when another axiom were examined). For this reason valid inferences in a nonmonotonic logic such as the Closed World Assumption are essentially carried out not in the original nonmonotonic language, but rather in some (monotonic) metatheory in which that nonmonotonic logic is defined.

This intuition may be explicated by defining the Closed World Assumption in terms of the set theoretic proof theory metalanguage of First Order Logic (i.e. FOL) with the following sentence:

$$\kappa=(\text{cwa } '\Gamma\ '\chi_i)$$

where cwa is defined as:

$(\text{cwa } '\Gamma\ '\chi_i) =_{df} (\text{fol}('\Gamma\cup\{'(\neg\chi_i): ('\chi_i\notin(\text{fol } '\Gamma))\}))$[1,2]

where $'\chi_i$ is the closed sentence of FOL occurring in the ith "inference rule". and $'\Gamma$ is a set of closed sentences of FOL. A closed sentence is a sentence without any free variables. fol is a function which produces the set of theorems derivable in FOL from the set of sentences to which it is applied. The quotations appended to the front of these Greek letters indicate references in the metalanguage to the sentences of the FOL object language. Interpreted doxastically this fixed point equation states:

[1] [Reiter 1978] defined cwa as: $(\text{cwa } '\Gamma\ '\chi_i) =_{df} ('\Gamma\cup\{'(\neg\chi_i): ('\chi_i\notin(\text{fol } '\Gamma))\})$. If that definition were used then all the theorems in this paper should have $(\text{cwa } '\Gamma\ '\chi_i)$ replaced by $(\text{fol } (\text{cwa } '\Gamma\ '\chi_i))$.

[2] The Closed World Assumption is often presented less generally by restricting $'\chi_i$ to begin with a particular predicate symbol or one or more such predicate symbols and by requiring that the arguments to those predicates range over all possible variable free terms of particular types [Reiter 1978].

the set of closed sentences which are believed is equal to:
the set of closed sentences derived in FOL from
the union of the initial beliefs consisting of a set of closed sentences: 'Γ
and the set of closed sentences of the form '($\neg\chi_i$)
such that for each i, the closed sentence 'χ_i is not initially believed.

The purpose of this paper is to show that all this metatheoretic machinery including the formalized syntax of FOL, the proof theory of FOL, the axioms of a strong set theory, and the set theoretic equation is not needed and that the essence of the Closed World Assumption is representable as a necessary equivalence in a simple (monotonic) Modal Quantificational Logic. Interpreted as a doxastic logic this necessary equivalence states:

that which is believed is logically equivalent to
the initial beliefs Γ and for each i, if $\neg\chi_i$ is initially believable then $\neg\chi_i$

thereby eliminating all mention of any metatheoretic machinery.

The remainder of this paper proves that this modal representation is equivalent to the Closed World Assumption. Section 2 describes a formalized syntax for a FOL object language. Section 3 describes the part of the proof theory of FOL needed herein (i.e. theorems FOL1-FOL4). Section 4 describes the Intensional Semantics of FOL which includes laws giving the meaning of FOL sentences: M0-M7, theorems giving the meaning of sets of sentences: MS1, MS2, MS3, and laws specifying the relationship of meaning and modality to the proof theory of FOL (i.e. the laws R0, A1, A2, and A3 and the theorems: C1, C2, C3, and C4). The modal version of the Closed World Assumption, called CWA, is defined in section 5 and explicated with theorems MC1-MC6 and SS1-SS2. In section 6, this modal version is shown by theorems CWA1 and CWA2 to be equivalent to the set theoretic fixed point equation for the Closed World Assumption. Finally, in section 7, some consequences are discussed.

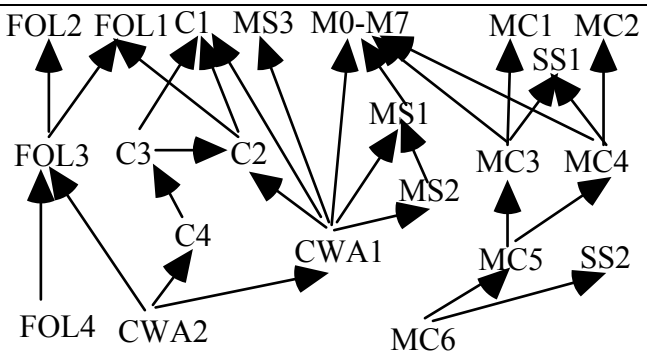


Figure 1: Dependencies among the Theorems

2. Formal Syntax of First Order Logic

We use a First Order Logic (i.e. FOL) defined as the six tuple: ($\rightarrow$, #f, $\forall$, *vars*, *predicates*, *functions*) where $\rightarrow$, #f, and $\forall$ are logical symbols, *vars* is a set of variable symbols, *predicates* is a set of predicate symbols each of which has an implicit arity specifying the number of associated terms, and *functions* is a set of function symbols each of which has an implicit arity specifying the number of associated terms. The sets of logical symbols, variables, predicate symbols, and function symbols are pairwise disjoint. Lower case Roman letters possibly indexed with digits are used as variables. Greek letters possibly indexed with digits or lowercase roman letters are used as syntactic metavariables. γ, γ_1,...γ_n, range over the variables, ξ, ξ_1...ξ_n range over sequences of variables of an appropriate arity, π,π_1...π_n range over the predicate symbols, ϕ, ϕ_1...ϕ_n range over function symbols, δ, δ_1...δ_n, σ range over terms, and α, α_1...α_n, β, β_1...β_n,χ, χ_1...χ_n,Γ,Γ_1,...Γ_n,φ range over sentences. The terms are of the forms γ and (ϕ δ1...δn), and the sentences are of the forms ($\alpha\rightarrow\beta$), #f, ($\forall\gamma$ α), and (π δ_1...δ_n). A nullary predicate π or function ϕ is written as a sentence or a term without parenthesis.

$\varphi\{\pi/\lambda\xi\alpha\}$ represents the replacement of all unmodalized occurrences of π in φ by $\lambda\xi\alpha$ followed by lambda conversion. The primitive symbols are shown in Figure 2 with their intuitive interpretations.

Symbol	Meaning
$\alpha \rightarrow \beta$	if α then β.
#f	falsity
$\forall\gamma\ \alpha$	for all γ, α.

Figure 2: Primitive Symbols of First Order Logic

The defined symbols are listed in Figure 3 with their definitions and intuitive interpretations.

Symbol	Definition	Meaning		Symbol	Definition	Meaning
$\neg\alpha$	$\alpha \rightarrow$ #f	not α		$\alpha\wedge\beta$	$\neg(\alpha \rightarrow \neg\beta)$	α and β
#t	$\neg$ #f	truth		$\alpha\leftrightarrow\beta$	$(\alpha\rightarrow\beta)$ $\wedge(\beta\rightarrow\alpha)$	α if and only if β
$\alpha\vee\beta$	$(\neg\alpha)\rightarrow\beta$	α or β		$\exists\gamma\ \alpha$	$\neg\forall\gamma\ \neg\alpha$	for some γ , α

Figure 3: Defined Symbols of First Order Logic

The FOL object language expressions are referred in the metalanguage (which also includes a FOL syntax) by inserting a quote sign in front of the object language entity thereby making a structural descriptive name of that entity. Generally, a set of sentences is represented as: $\{'\Gamma_i\}$ which is defined as: $\{'\Gamma_i: \#t\}$ which in turn is defined as: $\{s: \exists i(s='\Gamma_i)\}$ where i ranges over some range of numbers (which may be finite or infinite). With a slight abuse of notation we also write $'\kappa$, $'\Gamma$ to refer to such sets.

3. Proof Theory of First Order Logic

FOL is axiomatized with a recursively enumerable set of theorems as the set of axioms is itself recursively enumerable and its inference rules are recursive. The axioms and inference rules of FOL [Mendelson 1964] are given in Figure 4:

MA1: $\alpha \rightarrow (\beta\rightarrow \alpha)$ MR1: from α and $(\alpha\rightarrow \beta)$ infer β

MA2: $(\alpha\rightarrow (\beta\rightarrow \rho)) \rightarrow ((\alpha\rightarrow \beta)\rightarrow (\alpha\rightarrow \rho))$ MR2: from α infer $(\forall\gamma\ \alpha)$

MA3: $((\neg\alpha)\rightarrow (\neg\beta))\rightarrow (((\neg\alpha)\rightarrow \beta)\rightarrow\alpha)$

MA4: $(\forall\gamma\ \alpha)\rightarrow \beta$ where β is the result of substituting an expression (which is free for the free positions of γ in α) for all the free occurrences of γ in α.

MA5: $(\forall\gamma(\alpha \rightarrow \beta)) \rightarrow (\alpha\rightarrow(\forall\gamma\ \beta))$ where γ does not occur in α.

Figure 4: Inferences Rules and Axioms of FOL

In order to talk about sets of sentences we include in the metatheory set theory symbols This set theory includes the symbols ε, $\notin$, $\supseteq$, $=$, $\cup$ as defined in [Quine 1969].

The derivation operation (i.e. fol) of any FOL obeys the Inclusion (i.e. FOL1), and Idempotence (i.e. FOL2) properties:

FOL1: $(\text{fol } '\kappa)\supseteq'\kappa$ Inclusion

FOL2: $(\text{fol } '\kappa)\supseteq(\text{fol}(\text{fol } '\kappa))$ Idempotence

From these two properties we prove:

FOL3 $(\text{cwa}(\text{fol } '\Gamma)'\chi_i)=(\text{fol}(\text{cwa}(\text{fol } '\Gamma)'\chi_i))$

proof: FOL1 and FOL2 imply that $(\text{fol}(\text{fol } '\kappa))=(\text{fol } '\kappa)$. Since cwa begins with fol this implies: $'k=(\text{fol}(\text{cwa } 'k))$ QED.

FOL4: $('\kappa=(\text{cwa } '\Gamma\ '\chi_i))\rightarrow('\kappa=(\text{fol } '\kappa))$

proof: From the hypothesis and FOL3 $'\kappa=(\text{fol}(\text{cwa } '\Gamma\ '\chi_i))$ is derived. Using the hypothesis to replace $(\text{cwa } '\Gamma\ '\chi_i)$ by $'\kappa$ in this result gives: $('\kappa=(\text{fol } '\kappa))$ QED.

4. Intensional Semantics of FOL

The meaning (i.e. mg) [Brown 1977, Boyer & Moore 1981] or rather disquotation of a sentence of FOL is defined in Figure 5 below[1]. mg is defined in terms of mgs which maps each FOL object language sentence and an association list into a meaning. mgn maps each FOL object language term and an association list into a meaning. An association list is a list of pairs consisting of an object language variable and the meaning to which it is bound.

M0: (mg 'α) =df (mgs '($\forall\gamma 1...\gamma n\ \alpha$)'()) where '$\gamma 1...'\gamma n$ are all the free variables in 'α

M1: (mgs '($\alpha \rightarrow \beta$)a) $\leftrightarrow$ ((mgs 'α a)$\rightarrow$(mgs 'β a))

M2: (mgs '#f a) $\leftrightarrow$ #f

M3: (mgs '($\forall\ \gamma\ \alpha$)a) $\leftrightarrow$ $\forall$x(mgs 'α(cons(cons 'γ x)a))

M4: (mgs '($\pi\ \delta 1...\delta n$)a) $\leftrightarrow$ (π(mgn '$\delta 1$ a)...(mgn 'δn a)) for each predicate symbol 'π.

M5: (mgn '($\phi\ \ \delta 1...\delta n$)a) = ($\phi$(mgn '$\delta 1$ a)...(mgn 'δn a)) for each function symbol 'ϕ.

M6: (mgn 'γ a) = (cdr(assoc 'γ a))

M7: (assoc v L)= (if(eq? v(car(car L)))(car L)(assoc v(cdr L))) where: cons, car, cdr, eq?, and if are as in Scheme.

Figure 5: The Meaning of FOL Sentences

The meaning operator (i.e. mg) is a disquotation operation: (mg 'α)$\leftrightarrow\alpha$

The meaning of a set of sentences is defined in terms of the meanings of the sentences in the set as:

(ms 'κ) =df $\forall$s((sε'κ)$\rightarrow$(mg s)).

MS1: (ms{'α: Γ}) $\leftrightarrow$ $\forall\xi$((Γ{s/'α})$\rightarrow\alpha$) where ξ is the sequence of all the free variables in 'α and where Γ is any sentence of the intensional semantics. proof: (ms{'α:Γ}) Unfolding ms and the set pattern abstraction symbol gives: $\forall$s((sε{s: $\exists\xi$((s='α)$\wedge\Gamma$)})$\rightarrow$(mg s)) where ξ is a sequence of the free variables in 'a. This is equivalent to: $\forall$s(($\exists\xi$((s='α)$\wedge\Gamma$))$\rightarrow$(mg s)) which is: $\forall$s$\forall\xi$ (((s='α)$\wedge\Gamma$)$\rightarrow$(mg s)) which is: $\forall\xi$(Γ{s/'α}$\rightarrow$(mg 'α)). Unfolding mg using M0-M7 then gives: $\forall\xi$((Γ{s/'α})$\rightarrow\alpha$) QED

The meaning of a set is the meaning of all the sentences in the set (i.e. MS2):

MS2: (ms{'Γ_i}) $\leftrightarrow$ $\forall i\forall\xi_i\Gamma_i$ proof: (ms{'Γ_i}) Unfolding the set notation gives: (ms{'Γ_i: #t}). By MS1 this is equivalent to: $\forall_i\forall\xi_i$(($\#t\{s/'\alpha\}$)$\rightarrow\Gamma_i$) which is equivalent to: $\forall i\forall\xi_i\Gamma_i$ QED.

The meaning of the union of two sets of FOL sentences is the conjunction of their meanings (i.e. MS3):

MS3: (ms('$\kappa\cup\Gamma$)) $\leftrightarrow$ ((ms 'κ)$\wedge$(ms 'Γ)) proof: Unfolding ms and union in: (ms('$\kappa\cup\Gamma$)) gives:

$\forall$s((sε{s: (sε'κ)$\vee$(sε'Γ)})$\rightarrow$(mg s)) or rather: $\forall$s(((sε'κ)$\vee$(sε'Γ))$\rightarrow$(mg s)) which is logically equivalent to: ($\forall$s((sε'κ)$\rightarrow$(mg s)))$\wedge$($\forall$s((sε'Γ)$\rightarrow$(mg s))). Folding ms twice then gives:((ms 'κ)$\wedge$(ms 'Γ)) QED.

The meaning operation may be used to develop an Intensional Semantics for a FOL object language by axiomatizing the modal concept of necessity so that it satisfies the theorem:

C1: ('$\alpha\varepsilon$(fol 'κ)) $\leftrightarrow$ ([] ((ms 'κ)$\rightarrow$(mg 'α)))

for every sentence 'α and every set of sentences 'κ of that FOL object language. The necessity symbol is represented by a box: []. C1 states that a sentence of FOL is a FOL-theorem (i.e. fol) of a set of sentences of FOL if and only if the meaning of that set of sentences necessarily implies the meaning of that sentence.

One modal logic which satisfies C1 for FOL is the Z Modal Quantificational Logic described in [Brown 1987; Brown 1989] whose theorems are recursively enumerable. Z has the metatheorem: (<>Γ{$\pi/\lambda\xi\alpha$})$\rightarrow$ (<>Γ) where Γ is a sentence of FOL . Z includes all the laws of S5 Modal Logic [Hughes & Cresswell 1968] whose modal axioms and inference rules are in Figure 6. Therein, κ and Γ are sentences of the intentional semantics.

[1] The laws M0-M7 are analogous to Tarski's definition of truth except that finite association lists are used to bind variables to values rather than infinite sequences. M4 is different since mg is interpreted as being meaning rather than truth.

R0: from κ infer ([] κ)	A2: $([](\kappa \rightarrow \Gamma)) \rightarrow (([]\kappa) \rightarrow ([]\Gamma))$
A1: $([]\kappa) \rightarrow \kappa$	A3: $([]\kappa) \vee ([]\neg[]\kappa)$

Figure 6: The Laws of S5 Modal Logic

These S5 modal laws and the laws of FOL given in Figure 6 constitute an S5 Modal Quantificational Logic similar to [Carnap 1946; Carnap 1956], and a FOL version [Parks 1976] of [Bressan 1972] in which the Barcan formula: $(\forall\gamma([]\kappa))\rightarrow([]\forall\gamma\kappa)$ and its converse hold. The defined Modal symbols are in Figure 7 with their definitions and interpretations.

Symbol	Definition	Meaning		Symbol	Definition	Meaning
$<>\kappa$	$\neg [] \neg\kappa$	α is logically possible		$[\kappa]\ \Gamma$	$[] (\kappa\rightarrow\Gamma)$	β entails α
$\kappa\equiv \Gamma$	$[] (\kappa\leftrightarrow\Gamma)$	α is logically equivalent to β		$<\kappa>\ \Gamma$	$<> (\kappa\wedge\Gamma)$	α and β is logically possible

Figure 7: Defined Symbols of Modal Logic

From the laws of the Intensional Semantics we prove that the meaning of the set of FOL consequences of a set of sentences is the meaning of that set of sentences (C2), the FOL consequences of a set of sentences contain the FOL consequences of another set if and only if the meaning of the first set entails the meaning of the second set (C3), and the sets of FOL consequences of two sets of sentences are equal if and only if the meanings of the two sets are logically equivalent (C4):

C2: $(ms(fol\ '\kappa))\equiv(ms\ '\kappa)$ proof: The proof divides into two cases:

(1) $[(ms\ '\kappa)](ms(fol\ '\kappa))$ Unfolding the second ms gives: $[(ms\ '\kappa)]\forall s((s\varepsilon(fol\ '\kappa))\rightarrow(mg\ s))$

By the soundness part of C1 this is equivalent to: $[(ms\ '\kappa)]\forall s(([(ms\ '\kappa)](mg\ s))\rightarrow(mg\ s))$

By the S5 laws this is equivalent to: $\forall s(([(ms\ '\kappa)](mg\ s))\rightarrow [(ms\ '\kappa)](mg\ s))$ which is a tautology.

(2) $[(ms(fol\ '\kappa))](ms\ '\kappa)$ Unfolding ms twice gives: $[\forall s((s\varepsilon(fol\ '\kappa))\rightarrow(mg\ s))]\forall s((s\varepsilon'\kappa)\rightarrow(mg\ s))$

which is: $[\forall s((s\varepsilon(fol\ '\kappa))\rightarrow(mg\ s))](s\varepsilon'\kappa)\rightarrow(mg\ s))$ Backchaining on the hypothesis and then dropping it gives: $(s\varepsilon'\kappa)\rightarrow(s\varepsilon(fol\ '\kappa))$. Folding $\supseteq$ gives an instance of FOL1. QED.

C3: $(fol\ '\kappa)\supseteq(fol\ '\Gamma) \leftrightarrow ([(ms\ '\kappa)](ms\ '\Gamma))$ proof: Unfolding $\supseteq$ gives: $\forall s((s\varepsilon(fol\ '\Gamma))\rightarrow(s\varepsilon(fol\ '\kappa)))$

By C1 twice this is equivalent to: $\forall s(([(ms\ '\Gamma)](mg\ s))\rightarrow([(ms\ '\kappa)](mg\ s)))$

By the laws of S5 modal logic this is equivalent to: $([(ms\ '\kappa)]\forall s(([(ms\ '\Gamma)](mg\ s))\rightarrow(mg\ s)))$

By C1 this is equivalent to: $[(ms\ '\kappa)]\forall s((s\varepsilon(fol\ '\Gamma))\rightarrow(mg\ s))$. Folding ms then gives: $[(ms\ '\kappa)](ms(fol\ '\Gamma))$

By C2 this is equivalent to: $[(ms\ '\kappa)](ms\ '\Gamma)$. QED.

C4: $((fol\ '\kappa)=(fol\ '\Gamma)) \leftrightarrow ((ms\ '\kappa)\equiv(ms\ '\Gamma))$ proof: This is equivalent to $(((fol\ '\kappa)\supseteq(fol\ '\Gamma))\wedge((fol\ '\Gamma)\supseteq(fol\ '\kappa))) \leftrightarrow ([(ms\ '\kappa)](ms\ '\Gamma))\wedge([(ms\ '\Gamma)](ms\ '\kappa))$ which follows by using C3 twice.

5. The Closed World Assumption in Modal Logic

The equation for the Closed World Assumption may be expressed as a necessary equivalence in an S5 Modal Quantificational Logic [Brown 1989] as follows:

$$k\equiv(CWA\ \Gamma\ \chi_i)$$

where CWA is defined in Modal Logic as follows: $(CWA\ \Gamma\ \chi_i) =df\ \Gamma\wedge\forall i((<\Gamma>\neg\chi_i)\rightarrow\neg\chi_i)$ where Γ and χ_i are propositions of FOL. Given below are some simple properties of CWA:

MC1: $[(CWA\ \Gamma\ \chi_i)]\Gamma$

proof: By R0 it suffices to prove: $(CWA\ \Gamma\ \chi_i)\rightarrow\Gamma$. Unfolding CWA gives: $(\Gamma\wedge\forall_i((<\Gamma>\neg\chi_i)\rightarrow\neg\chi_i))\rightarrow\Gamma$

which is a tautology. QED.

MC2: $(<\Gamma>\neg\chi_i)\rightarrow([(CWA\ \Gamma\ \chi_i)]\neg\chi_i)$

proof: Unfolding CWA gives: $(<\Gamma>\neg\chi_i)\rightarrow([\Gamma\wedge\forall_i((<\Gamma>\neg\chi_i)\rightarrow\neg\chi_i)]\neg\chi_i)$. Using the hypotheses on the ith instance gives: $(<\Gamma>\neg\chi_i)\rightarrow([\Gamma\wedge\forall_i((<\Gamma>\neg\chi_i)\rightarrow\neg\chi_i)\wedge\neg\chi_i]\neg\chi_i)$ which is a tautology. QED.

The concept (i.e. ss) of the combined meaning of all the sentences of the FOL object language whose meanings are entailed by a proposition is defined as follows:

$$(ss\ \kappa) =df \forall s(([\kappa](mg\ s))\rightarrow(mg\ s))$$

SS1 shows that a proposition entails the combined meaning of the FOL object language sentences that it entails. SS2 shows that if a proposition is necessarily equivalent to the combined meaning of the FOL object language sentences that it entails, then there exists a set of FOL object language sentences whose meaning is necessarily equivalent to that proposition:

SS1: $[\kappa](ss\ \kappa)$ proof: By R0 it suffices to prove: $\kappa\rightarrow(ss\ \kappa)$. Unfolding ss gives: $\kappa\rightarrow\forall s(([\kappa](mg\ s))\rightarrow(mg\ s))$ which is equivalent to:$\forall s(([\kappa](mg\ s))\rightarrow(\kappa\rightarrow(mg\ s)))$ which is an instance of A1. QED.

SS2: $(\kappa\equiv(ss\ \kappa))\rightarrow \exists s(\kappa\equiv(ms\ s))$ proof: Letting s be $\{s: ([\kappa](mg\ s))\}$ gives: $(\kappa\equiv(ss\ \kappa))\rightarrow (\kappa\equiv(ms\{s: ([\kappa](mg\ s))\}))$. Unfolding ms and lambda conversion gives: $(\kappa\equiv(ss\ \kappa))\leftrightarrow (\kappa\equiv\forall s(([\kappa](mg\ s))\rightarrow(mg\ s)))$. Folding ss gives a tautology. QED.

The theorems MC3 and MC4 are analogous to MC1 and MC2 except that CWA is replaced by the combined meanings of the sentences entailed by CWA.

MC3: $[ss(CWA\ \forall i\Gamma_i\ \chi_i)]\forall i\Gamma_i$

proof: By R0 it suffices to prove: $(ss(CWA\ \forall i\Gamma_i\ \chi_i))\rightarrow\forall i\Gamma_i$ which is equivalent to: $(ss(CWA\ \forall i\Gamma_i\ \chi_i))\rightarrow\Gamma_i$

Unfolding ss gives: $(\forall s(([(CWA\ \forall i\Gamma_i\ \chi_i)](mg\ s))\rightarrow(mg\ s)))\rightarrow\Gamma_i$ which by the meaning laws M0-M8 is equivalent to: $(\forall s(([(CWA\ \forall i\Gamma_i\ \chi_i)](mg\ s))\rightarrow(mg\ s)))\rightarrow(mg\ '\Gamma_i)$. Backchaining on $(mg\ '\Gamma_i)$ with s in the hypothesis assigned to be $'\Gamma_i$ in the conclusion shows that it suffices to prove: $([(CWA\ \forall i\Gamma_i\ \chi_i)](mg\ '\Gamma_i))$ which by M0-M8 is equivalent to: $([(CWA\ \forall i\Gamma_i\ \chi_i)]\Gamma_i)$ which by the laws of S5 Modal Logic is equivalent to: $([(CWA\ \forall i\Gamma_i\ /\chi_i)]\forall i\Gamma_i)$ which is an instance of MC1. QED.

MC4: $(<k>\neg\chi_i)\rightarrow ([ss(CWA\ \Gamma\ \chi_i)]\neg\chi_i)$

proof: Unfolding the last ss gives: $(<k>\neg\chi_i)\rightarrow([\forall s(([(CWA\ \Gamma(mg\ '\chi_i))](mg\ s))\rightarrow(mg\ s))]\neg\chi_i)$

Instantiating s in the hypothesis to $'(\neg\chi_i)$ and then dropping the hypothesis gives:

$(<k>\neg\chi_i)\rightarrow([(([(CWA\ \Gamma(mg\ '\chi_i))](mg\ '(\neg\chi_i)))\rightarrow(mg\ '(\neg\chi_i)))]\neg\chi_i)$. Using the meaning laws M0-M7 gives: $(<k>\neg\chi_i)\rightarrow([(([(CWA\ \Gamma\ \chi_i)]\neg\chi_i)\rightarrow\neg\chi_i)]\neg\chi_i)$. Backchaining on $\neg\chi_i$ shows that it suffices to prove: $(<k>\neg\chi_i)\rightarrow([(CWA\ \Gamma\ \chi_i)]\neg\chi_i)$ which is an instance of theorem MC2. QED.

MC5 and MC6 show that talking about the meanings of sets of FOL sentences in the modal representation of the Closed World Assumption is equivalent to talking about propositions-

MC5: $(ss(CWA(\forall i\Gamma_i)\chi_i)) \leftrightarrow (CWA(\forall i\Gamma_i)\chi_i)$

proof: In view of SS1, it suffices to prove: $[(ss(CWA(\forall i\Gamma_i)\chi_i))](CWA(\forall i\Gamma_i)\chi_i)$

Unfolding the second occurrence of CWA gives: $[(ss(CWA(\forall i\Gamma_i)\chi_i))]((\forall i\Gamma_i)\wedge\forall_i((<(\forall i\Gamma_i)>\neg\chi_i)\rightarrow\neg\chi_i))$

which holds by theorems MC3 and MC4. QED.

MC6: $(\kappa\equiv(CWA(\forall i\Gamma_i)\chi_i))\rightarrow\exists s(\kappa\equiv(ms\ s))$

proof: From the hypothesis and MC5 $\kappa\equiv(ss(CWA(\forall i\Gamma_i)\chi_i))$ is derived. Using the hypothesis to replace $(CWA(\forall i\Gamma_i)\chi_i)$ by κ in this result gives: $\kappa\equiv(ss(CWA\ \kappa))$, By SS2 this implies the conclusion. QED.

6. The Relationship Between The Closed World Assumption and Modal Logic

The relationship between the proof theoretic definition of the Closed World Assumption [Brown 1989] and the modal representation is proven in two steps. First theorem CWA1 shows that the meaning of the set cwa is the proposition CWA and then theorem CWA2 shows that a set of FOL sentences which contains its FOL theorems is equal to cwa of an initial set of axioms and defaults if and only if the meaning (or rather disquotation) of that set of sentences is logically equivalent to CWA of the meaning of that initial set of sentences and those defaults.

CWA1: $(ms(cwa\{'\Gamma_i\}'\chi_i))\equiv(CWA(\forall i\Gamma_i)\chi_i)$

proof: $(ms(cwa\{'\Gamma_i\}'\chi_i))$.

Unfolding the definition of cwa gives: $ms(fol(\{'\Gamma_i\}\cup\{'(\neg\chi_i): ('\chi_i\notin(fol\{'\Gamma_i\}))\}))$

By C2 this is equivalent to: $ms(\{'\Gamma_i\}\cup\{'(\neg\chi_i): ('\chi_i\notin(fol\{'\Gamma_i\}))\})$.

Using C1 gives: $ms(\{'\Gamma_i\}\cup\{'(\neg\chi_i): \neg([(ms\{'\Gamma_i\})](mg\ '\chi_i))\})$.

Using MS3 gives: $(ms\{'\Gamma_i\})\wedge(ms\{'(\neg\chi_i):(\neg([(ms\{'\Gamma_i\})](mg\ '\chi_i)))\})$

Using MS2 twice gives: $(\forall i\Gamma_i)\wedge(ms\{'(\neg\chi_i):(\neg([(\forall i\Gamma_i)](mg\ '\chi_i)))\})$.

By MS1 this is equivalent to: $(\forall i\Gamma_i)\wedge\forall i((\neg([(\forall i\Gamma_i)](mg\ '\chi_i)))\rightarrow(mg'(\neg\chi_i)))$.

Using M0-M7 to gives: $(\forall i\Gamma_i)\wedge\forall i((\neg([(\forall i\Gamma_i)]\chi_i))\rightarrow\neg\chi_i)$ which is equivalent to: $(\forall i\Gamma_i)\wedge\forall i((<(\forall i\Gamma_i)>\neg\chi_i)\rightarrow\neg\chi_i)$.
Folding the definition of CWA gives: $(CWA\ (\forall i\Gamma_i)\chi_i)$ QED.

CWA2: $((fol\ '\kappa)=(cwa\{'\Gamma_i\}'\chi_i))\leftrightarrow((ms\ '\kappa)\equiv(CWA(\forall i\Gamma_i)\chi_i))$

proof: $(fol\ '\kappa)=(cwa\{'\Gamma_i\}'\chi_i)$.

By FOL3 this is equivalent to: $(fol\ '\kappa)=(fol(cwa\{'\Gamma_i\}'\chi_i))$.

By C4 this is equivalent to: $(ms\ '\kappa)\equiv(ms(cwa\{'\Gamma_i\}'\chi_i))$.

By CWA1 this is equivalent to: $(ms\ '\kappa)\equiv(CWA(\forall i\Gamma_i)\chi_i)$ QED.

Theorem CWA2 shows that the set of theorems: $(fol\ '\kappa)$ of a set $'\kappa$ equals cwa if and only if the meaning $(ms\ '\kappa)$ of $'\kappa$: is equivalent to CWA. Furthermore, by FOL4 it cannot be anything else (such as a set not containing all its theorems) and by MC6 there are no other solutions (such as a proposition not representable as a sentence in the FOL object language). Therefore, the Modal representation of the Closed World Assumption (i.e. CWA), faithfully represents the set theoretic description of the Closed World Assumption (i.e. cwa). Finally, we note that $(\forall i\Gamma_i)$ and $(ms\ '\kappa)$ may be generalized to be arbitrary propositions Γ and κ giving the more general modal representation: $\kappa\equiv(CWA\ \Gamma\ \chi_i)$.

7. Conclusion

Theorems CWA2, FOL4, and MC6 prove that The Closed World Assumption can be represented in a monotonic modal quantificational logic which is a slight extension of S5. Since the modal representation:

$$(CWA\ \Gamma\ \chi_i)=df\ \Gamma\wedge\forall i((<\Gamma>\neg\chi_i)\rightarrow\neg\chi_i)$$

is equivalent to the Closed World Assumption and since it an instance of the more general case where free variables may occur in the χ_i sentences:

$$(QCWA\ \Gamma\ \chi_i)\ =df\ \Gamma\wedge\forall i\forall\xi_i((<\Gamma>\neg\chi_i)\rightarrow\neg\chi_i)$$

we now have a logic for inferring consequences from that more general case where universal quantifiers may cross modal scopes. For example,

$$(QCWA\ (P\ 1)\wedge\forall x((P\ x)\rightarrow(P(add1\ x)))\ \ \neg(P\ x))$$

is equivalent to: $\forall x((P\ x)\leftrightarrow([(\pi\ 1)\wedge\forall x((\pi\ x)\rightarrow(\pi(add1\ x)))](\pi\ x)))$

which states that the natural numbers and only the natural numbers bear the property P.

Finally, we note that the techniques used herein to develop a modal representation of the Closed World Assumption and then to generalize it to handle the case where bound variables cross modal scopes, may be applied to other nonmonotonic systems such as those defined by a fixed-point equation [Brown 2003a], thereby allowing those systems to be extended to analogous cases where variables are allowed to cross the scope of nonmonotonic operations. An example of using the Modal Representation of the Closed World Assumption on multiple knowledgebases involving actions and epistemic reasoning is given in [Brown 1986].

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers: MIP-9526532, EIA-9818341, and EIA-9972843.

Bibliography

[Bressan 1972] Bressan, Aldo 1972. *A General Interpreted Modal Calculus,* Yale University Press.

[Boyer&Moore 1981] R. S. Boyer and J Strother Moore, "Metafunctions: proving them correct and using them efficiently as new proof procedures," *The Correctness Problem in Computer Science,* R. S. Boyer and J. Strother Moore, eds., Academic Press, New York, 1981.

[Brown 1977] Brown, F. M., "The Theory of Meaning", Department of Artificial Intelligence Research Report 35, University of Edinburgh, June 1977, University of Edinburgh, *Edinburgh Artificial Intelligence Research Papers 1973-1986*, Scientific Datalink microfische, 1989.

[Brown 1986] Brown, Frank M. 1986. "Reasoning in a Hierarchy of Deontic Defaults", *Proceedings of the Canadian Artificial Intelligence Conference CSCI 86*, Montreal, Canada, Morgan-Kaufmann, Los Altos.

[Brown 1987] Brown, Frank M. 1987. "The Modal Logic Z", In *The Frame Problem in AI"; Proc. of the 1987 AAAI Workshop*, Morgan Kaufmann, Los Altos, CA .

[Brown 1989] Brown, Frank M. 1989. "The Modal Quantificational Logic Z Applied to the Frame Problem", advance paper *First International Workshop on Human & Machine Cognition*, May 1989 Pensacola, Florida. Abbreviated version in *International Journal of Expert Systems Research and Applications, Special Issue: The Frame Problem. Part A.* eds. Kenneth Ford and Patrick Hayes, vol. 3 number 3, pp169-206 JAI Press 1990. Reprinted in *Reasoning Agents in a Dynamic World: The Frame problem*, editors: Kenneth M. Ford, Patrick J. Hayes, JAI Press 1991.

Brown, Frank M., 2003a, "Representing Autoepistemic Logic in Modal Logic", *Information Theories and Applications*, Vol. 10, no. 4, pages 455-462, December 2003.

[Carnap 1946] Carnap, Rudolf 1946. "Modalities and Quantification" *Journal of Symbolic Logic*, vol. 11, no. 2.

[Carnap 1956] Carnap, Rudolf 1956. *Meaning and Necessity: A Study in the Semantics of Modal Logic*, The University of Chicago Press.

[Hughes & Cresswell 1968] Hughes, G. E. and Creswell, M. J., 1968. *An Introduction to Modal Logic*, Methuen & Co. Ltd., London.

[Mendelson 1964] Mendelson, E. 1964. *Introduction to Mathematical Logic*, Van Norstrand, Reinhold Co. New York.

[Parks 1976] Parks, Z. 1976. "An Investigation into Quantified Modal Logic", Studia Logica 35, p109-125.

[Quine 1969] Quine, W.V.O., Set Theory and Its Logic, revised edition, Oxford University Press, London, 1969.

[Reiter 1978] Reiter, R. 1978. "On Closed World Databases", in, *Logic and Databases*, eds. Gallaire and Minker, J., Plenum Press, Ney York.

Author's Information

Frank M. Brown – University of Kansas, Lawrence, Kansas, 66045, e-mail: brown@ku.edu.

REPRESENTING SKEPTICAL LOGICS IN MODAL LOGIC

Frank M. Brown

***Abstract**: Several skeptical nonmonotonic logics are shown to be representable in a monotonic Modal Quantificational Logic whose modal laws are stronger than S5. Specifically, it is proven that under certain conditions a set of sentences of First Order Logic is the intersection of the possibly infinite number of fixed-points of the fixed-point equation of a base nonmonotonic logic with an initial set of axioms and defaults if and only if the meaning of that set of sentences is logically equivalent to the meaning of the set of sentences entailed by the disjunction of the possibly infinite number of solutions to a necessary equivalence formed from a particular modal functor of the meanings of that initial set of sentences and of the sentences in those defaults. This result is important because the modal representation allows the use of powerful automatic deduction systems for Modal Logic and because unlike the set theoretic definition of a skeptical logic, the modal representation is easily generalized to the case quantified variables may be shared across the scope of the components of the defaults thus allowing the such defaults to produce quantified consequences.*

***Keywords**: Skeptical Fixed-point Logics, Modal Logic, Nonmonotonic Logic.*

1. Introduction

In general, a fixed-point equation defining a nonmonotonic logic may have any number of solutions including an infinite number or even none. When an equation has more than one fixed-point, the question arises as to what is actually the case, since a different answer follows for each distinct fixed-point. This question could be answered in the "skeptical" sense of not believing anything that does not hold in all the fixed-points[1] by computing the intersection of all the fixed-points: $\cap\{k: k=(\text{nml } k)\}$ where nml is a set theoretic functor defining the theorems of some nonmonotonic logic. Thus, the set of closed sentences which are skeptically believed is:

the intersection of all sets of closed sentences such that: each such set is equal to nml of itself.

The purpose of this paper is to show that all this metatheoretic machinery including the formalized syntax and the proof theory of First Order Logic (i.e. FOL), the axioms of a strong set theory, the intersection of the fixed-points of the set theoretic fixed-point equation is not needed and that the essence of a Skeptical Logic may be representable in a simple (monotonic) Modal Quantificational Logic as an expression involving a modal functor: NML which is analogous to the set theoretic functor: nml. In this case, that which is skeptically believed is:

the disjunction[2] of all propositions such that each is necessarily equivalent to NML of itself.

thereby eliminating all mention of any metatheoretic machinery.

The remainder of this paper proves that this modal representation is equivalent to the set theoretic description. Section 2 describes a formalized syntax for a FOL object language. Section 3 describes the part of the proof theory of FOL needed herein (i.e. theorems FOL1-FOL3). Section 4 describes the Intensional Semantics of FOL including the meaning operator (i.e. the laws M0-M7) and the relationship of meaning and modality to the proof theory of FOL (i.e. the laws R0, A1, A2 and A3 and the theorems C1, C2, C3, and C4). The modal version of Skepticity, called S, is defined in section 5 and explicated with theorems MSK1-MSK3 and SS1. In section 6, this modal version is shown by theorems SL1 to be equivalent to the set theoretic representation of a Skeptical Logic. Figure 1 outlines the relationship of all these theorems in producing the final theorems SL1 and MSK3. In section 7 these theorems are applied to producing the modal representations of the Skeptical Logics constructed from FOL, the Closed World Assumption [Reiter 1978], Reflective Logic [Brown 1989], Default Logic [Reiter 1980], the recursive definition of Default Logic [Reiter 1980], and Autoepistemic Logic [Moore 1985]. Finally, in section 8, some conclusions are made.

[1] An alternative answer is the "credulous" sense of believing anything that holds in any of the fixed-points.
[2] Just as the word "intersection" in the set theoretic description encompasses an infinite and even a nondenumerable number of sets, our usage of the word "disjunction" encompasses an infinite and even a nondenumerable number of propositions.

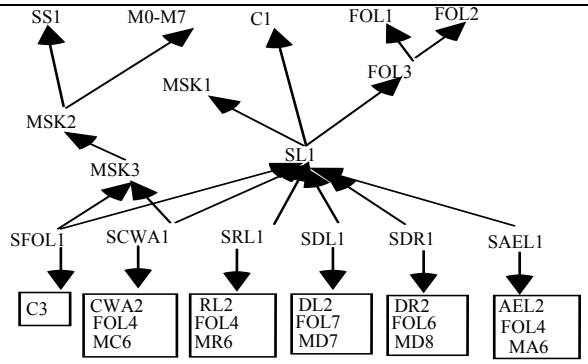


Figure 1: Dependences among the Theorems
(The theorems in the boxes are proven in other papers as noted in section 7.)

2. Formal Syntax of First Order Logic

We use a First Order Logic (i.e. FOL) defined as the six tuple: ($\rightarrow$, #f, $\forall$, *vars*, *predicates*, *functions*) where $\rightarrow$, #f, and $\forall$ are logical symbols, *vars* is a set of variable symbols, *predicates* is a set of predicate symbols each of which has an implicit arity specifying the number of associated terms, and *functions* is a set of function symbols each of which has an implicit arity specifying the number of associated terms. The sets of logical symbols, variables, predicate symbols, and function symbols are pairwise disjoint. Lower case Roman letters possibly indexed with digits are used as variables. Greek letters possibly indexed with digits or lowercase roman letters are used as syntactic metavariables. γ, γ_1,...γ_n, range over the variables, ξ, ξ_1...ξ_n range over sequences of variables of an appropriate arity, π,π_1...π_n range over the predicate symbols, ϕ, ϕ_1...ϕ_n range over function symbols, δ, δ_1...δ_n, σ range over terms, and α, α_1...α_n, β, β_1...β_n,χ, χ_1...χ_n,Γ,Γ_1,...Γ_n,φ range over sentences. The terms are of the forms γ and (ϕ δ_1...δ_n), and the sentences are of the forms ($\alpha\rightarrow\beta$), #f, ($\forall\gamma$ α), and (π δ_1...δ_n). A nullary predicate π or function ϕ is written as a sentence or a term without parenthesis. $\varphi\{\pi/\lambda\xi\alpha\}$ represents the replacement of all unmodalized occurrences of π in φ by $\lambda\xi\alpha$ followed by lambda conversion. The primitive symbols are shown in Figure 2 with their intuitive interpretations.

Symbol	Meaning
$\alpha \rightarrow \beta$	if α then β.
#f	falsity
$\forall\gamma\ \alpha$	for all γ, α.

Figure 2: Primitive Symbols of First Order Logic

The defined symbols are listed in Figure 3 with their definitions and intuitive interpretations.

Symbol	Definition	Meaning		Symbol	Definition	Meaning
$\neg\alpha$	$\alpha \rightarrow$ #f	not α		$\alpha\wedge\beta$	$\neg(\alpha \rightarrow \neg\beta)$	α and β
#t	$\neg$ #f	truth		$\alpha\leftrightarrow\beta$	$(\alpha\rightarrow\beta) \wedge (\beta\rightarrow\alpha)$	α if and only if β
$\alpha\vee\beta$	$(\neg\alpha)\rightarrow\beta$	α or β		$\exists\gamma\ \alpha$	$\neg\forall\gamma\ \neg\alpha$	for some γ , α

Figure 3: Defined Symbols of First Order Logic

The FOL object language expressions are referred in the metalanguage (which also includes a FOL syntax) by inserting a quote sign in front of the object language entity thereby making a structural descriptive name of that entity. Generally, a set of sentences is represented as: $\{'\Gamma_i\}$ which is defined as: $\{'\Gamma_i: \#t\}$ which in turn is defined as: $\{s: \exists i(s='\Gamma_i)\}$ where i ranges over some range of numbers (which may be finite or infinite). With a slight abuse of notation we also write $'\kappa$, $'\Gamma$ to refer to such sets.

3. Proof Theory of First Order Logic

The axioms and inference rules of FOL [Mendelson 1964] are given in Figure 4:

MA1: $\alpha \rightarrow (\beta \rightarrow \alpha)$ MR1: from α and $(\alpha \rightarrow \beta)$ infer β

MA2: $(\alpha \rightarrow (\beta \rightarrow \rho)) \rightarrow ((\alpha \rightarrow \beta) \rightarrow (\alpha \rightarrow \rho))$ MR2: from α infer $(\forall \gamma\ \alpha)$

MA3: $((\neg \alpha) \rightarrow (\neg \beta)) \rightarrow (((\neg \alpha) \rightarrow \beta) \rightarrow \alpha)$

MA4: $(\forall \gamma\ \alpha) \rightarrow \beta$ where β is the result of substituting an expression (which is free for the free positions of γ in α) for all the free occurrences of γ in α.

MA5: $(\forall \gamma(\alpha \rightarrow \beta)) \rightarrow (\alpha \rightarrow (\forall \gamma\ \beta))$ where γ does not occur in α.

Figure 4: Inferences Rules and Axioms of FOL

In order to talk about sets of sentences we include in the metatheory set theory symbols ε, $\notin$, $\supseteq$, $=$, $\cap$ as defined in [Quine 1969]. The derivation operation (i.e. fol) of FOL obeys the Inclusion and Idempotence properties:

FOL1: (fol 'κ)$\supseteq$'κ Inclusion

FOL2: (fol 'κ)$\supseteq$(fol(fol 'κ)) Idempotence

From these two properties we prove:

FOL3: $\forall p((p=(\text{fol } p)) \rightarrow \alpha) \leftrightarrow \forall p(\alpha\{p/(\text{fol } p)\})$ and $\exists p((p=(\text{fol } p)) \wedge \alpha) \leftrightarrow \exists p(\alpha\{p/(\text{fol } p)\})$

proof: The universal quantifier version follows from the existential quantifier version by running negation through both sides of the bi-implication. The existential version is proven as follows. There are two cases:

(1) $((p=(\text{fol } p)) \wedge \alpha) \rightarrow \exists\ p(\alpha\{p/(\text{fol } p)\})$. The existentially quantified p is replaced by p giving: $((p=(\text{fol } p)) \wedge \alpha) \rightarrow (\alpha\{p/(\text{fol } p)\})$. The equation in the hypothesis is used to replace p in α by (fol p) giving the conclusion.

(2) $(\alpha\{p/(\text{fol } p)\}) \rightarrow \exists p((p=(\text{fol } p)) \wedge \alpha)$. Letting p in the conclusion be (fol p) gives: $(\alpha\{p/(\text{fol } p)\}) \rightarrow (((\text{fol } p)=(\text{fol}(\text{fol } p))) \wedge (\alpha\{p/(\text{fol } p)\}))$ which holds by FOL1 and FOL2.

4. Intensional Semantics of FOL

The meaning (i.e. mg) [Brown 1977] or rather disquotation of a sentence of FOL is defined in Figure 5 below. mg is defined in terms of mgs which maps each FOL object language sentence and an association list into a meaning. mgn maps each FOL object language term and an association list into a meaning. An association list is a list of pairs consisting of an object language variable and the meaning to which it is bound.

M0: (mg 'α) =df (mgs '$(\forall \gamma 1 \ldots \gamma n\ \alpha)$'()) where '$\gamma 1 \ldots$'$\gamma n$ are all the free variables in 'α

M1: (mgs '$(\alpha \rightarrow \beta)$a) $\leftrightarrow$ ((mgs 'α a)$\rightarrow$(mgs 'β a))

M2: (mgs '#f a) $\leftrightarrow$ #f

M3: (mgs '$(\forall\ \gamma\ \alpha)$a) $\leftrightarrow \forall x$(mgs '$\alpha$(cons(cons '$\gamma$ x)a))

M4: (mgs '$(\pi\ \delta_1 \ldots \delta_n)$a) $\leftrightarrow$ (π(mgn 'δ_1 a)...(mgn 'δ_n a)) for each predicate symbol 'π.

M5: (mgn '$(\phi\ \delta_1 \ldots \delta_n)$a) = ($\phi$(mgn '$\delta_1$ a)...(mgn 'δ_n a)) for each function symbol 'ϕ.

M6: (mgn 'γ a) = (cdr(assoc 'γ a))

M7: (assoc v L)= (if(eq? v(car(car L)))(car L)(assoc v(cdr L))) where: cons, car, cdr, eq?, and if are as in Scheme.

Figure 5: The Meaning of FOL Sentences

The meaning operator (i.e. mg) is a disquotation operation: (mg 'α)$\leftrightarrow\alpha$. It may be used as an Intensional Semantics for a FOL object language by axiomatizing the modal concept of necessity to satisfy the theorem:

C1: ('$\alpha\varepsilon$(fol 'κ)) $\leftrightarrow$ ([] ((ms 'κ)$\rightarrow$(mg 'α)))

for every sentence 'α and every set of sentences 'κ of that FOL object language. The necessity symbol is represented by a box: []. C1 states that a sentence of FOL is a FOL-theorem (i.e. fol) of a set of sentences of FOL if and only if the meaning of that set of sentences necessarily implies the meaning of that sentence.

One modal logic which satisfies C1 for FOL is the Z Modal Quantificational Logic described in [Brown 1987; Brown 1989] whose theorems are recursively enumerable. Z has the metatheorem: $(<>\Gamma\{\pi/\lambda\xi\alpha\})\rightarrow(<>\Gamma)$ where Γ is a sentence of FOL . Z includes all the laws of S5 Modal Logic [Hughes & Cresswell 1968] whose modal axioms and inference rules are in Figure 6. Therein, κ and Γ are sentences of the intentional semantics.

R0: from κ infer $([]\kappa)$	A2: $([](\kappa\rightarrow\Gamma))\rightarrow(([]\kappa)\rightarrow([]\Gamma))$
A1: $([]\kappa)\rightarrow\kappa$	A3: $([]\kappa)\vee([]\neg[]\kappa)$

Figure 6: The Laws of S5 Modal Logic

These S5 modal laws and the laws of FOL given in Figure 6 constitute an S5 Modal Quantificational Logic similar to [Carnap 1946; Carnap 1956], and a FOL version [Parks 1976] of [Bressan 1972] in which the Barcan formula: $(\forall\gamma([]\kappa))\rightarrow([]\forall\gamma\kappa)$ and its converse hold. The defined Modal symbols are given in Figure 7.

Symbol	Definition	Meaning		Symbol	Definition	Meaning
$<>\kappa$	$\neg[]\neg\kappa$	α is logically possible		$[\kappa]\Gamma$	$[](\kappa\rightarrow\Gamma)$	β entails α
$\kappa\equiv\Gamma$	$[](\kappa\leftrightarrow\Gamma)$	α is logically equivalent to β		$<\kappa>\Gamma$	$<>(\kappa\wedge\Gamma)$	α and β is logically possible

Figure 7: Defined Symbols of Modal Logic

5. Skepticity Represented in Modal Logic

The set theoretic intersection defining a Skeptical Logic in terms of a base nonmonotonic logic called nml:

$$\cap\{k: k=(\text{nml } k)\}$$

may be expressed in an S5 Modal Quantificational Logic supplemented with propositional quantifiers [Fine 1970; Bressan 1972] which obey the normal laws of Second Order Logic (i.e. laws analogous to MR2, MA4, and MA5 given in Figure 4 where γ is now a propositional variable), as follows:

$$\exists k(k\wedge(k\equiv(\text{NML } k)))$$

where NML is the modal functor corresponding to nml. The idiom $\exists k(k\wedge(k\equiv(\varphi\ k)))$ is intuitively read as a nominal as the (possibly infinite) disjunction of all propositions such that κ_i. When the necessary equivalence has a finite number of solutions: $\kappa_1,...,\kappa_\eta$ then the idiom is equivalent to: $\kappa_1\vee...\vee\kappa_\eta$, but there is in general no requirement that the necessary equivalence holds for only a finite or even only a denumerable number of solutions.

Some theorems are now proven which show that reasoning about meanings of sets of sentences is equivalent to reasoning about propositions in general. MSK1 shows (in the cases of interest) that quantifying over the meanings of sets of FOL sentences is equivalent to quantifying propositions in general:

MSK1 if $(\kappa\equiv(\text{NML }\kappa))\rightarrow\exists s(\kappa\equiv(\text{ms } s))$ then $(\exists k((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k)))))\equiv(\exists k(k\wedge(k\equiv(\text{NML } k))))$

proof: By R0 it suffices to prove: $(\exists k((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k)))))\leftrightarrow\exists k(k\wedge(k\equiv(\text{NML } k)))$. There are two cases: (1) $\exists k((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k))))\rightarrow\exists k(k\wedge(k\equiv(\text{NML } k)))$. By FOL this is equivalent to:

$((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k))))\rightarrow\exists k(k\wedge(k\equiv(\text{NML } k)))$. Letting k in the conclusion be (ms k) gives a tautology.

(2) $\exists k(k\wedge(k\equiv(\text{NML } k)))\rightarrow\exists k((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k)))))$. By FOL this is :$(k\wedge(k\equiv(\text{NML } k)))$

$\rightarrow(\exists k((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k)))))$. From $(k\equiv(\text{NML } k))$ and the hypothesis to the theorem we infer $\exists s(\kappa\equiv(\text{ms } s))$ and then $\kappa\equiv(\text{ms } s)$. Using this to replace k by (ms s) gives: $((\text{ms } s)\wedge((\text{ms } s)\equiv(\text{NML}(\text{ms } s))))\rightarrow(\exists k((\text{ms } k)\wedge((\text{ms } k)\equiv(\text{NML}(\text{ms } k)))))$. Letting k in the conclusion be s then gives a tautology. QED.

The concept (i.e. ss) of the combined meaning of all the sentences of the FOL object language whose meanings are entailed by a proposition is defined as follows: $(\text{ss }\kappa)=_{df}\forall s(([\kappa](\text{mg } s))\rightarrow(\text{mg } s))$. SS1 shows that a proposition entails the combined meaning of the FOL object language sentences that it entails.

SS1: $[\kappa](\text{ss }\kappa)$ proof: By R0 it suffices to prove: $\kappa\rightarrow(\text{ss }\kappa)$. Unfolding ss gives: $\kappa\rightarrow\forall s(([\kappa](\text{mg } s))\rightarrow(\text{mg } s))$ which is equivalent to:$\forall s(([\kappa](\text{mg } s))\rightarrow(\kappa\rightarrow(\text{mg } s)))$ which is an instance of A1. QED.

One might think that: $(\text{ss}(\exists k(k\wedge(k\equiv(\text{NML } k))))\leftrightarrow(\exists k(k\wedge(k\equiv(\text{NML } k)))$, but in general $\text{ss}(\exists k(k\wedge(k\equiv(\text{NML } k)))$ does not imply: $\exists k(k\wedge(k\equiv(\text{NML } k))$ because if there were an infinite number of solutions to: $k\equiv(\text{NML } k))$ none of which

was entailed by the others, then that disjunction of all the solutions would be infinite and therefore might not be representable as a sentence of FOL. However, if there are only a finite number of solutions then this relation will hold as is proven by MSK2 below:

MSK2: If k does not occur in any 'Γi then: $(ss(\exists k(k\wedge(\vee_{i=1,n}(k\equiv(ms\ '\Gamma i))))))\equiv(\vee_{i=1,n}(ms\ '\Gamma_i))$

proof: By SS1 it suffices to prove: $[(ss(\exists k(k\wedge(\vee_{i=1,n}(k\equiv(ms\ '\Gamma_i))))))](\vee_{i=1,n}(ms\ '\Gamma_i))$

By the laws of FOL this is equivalent to: $[(ss(\vee_{i=1,n}(\exists k(k\wedge(k\equiv(ms\ '\Gamma_i))))))](\vee_{i=1,n}(ms\ '\Gamma_i))$

Since k does not occur free in Γ_i by the laws of S5 this is equivalent to: $[(ss(\vee_{i=1,n}(ms\ '\Gamma_i)))](\vee_{i=1,n}(ms\ '\Gamma_i))$. The last ms is unfolded giving: $[ss(\vee_{i=1,n}(ms\ '\Gamma_i))](\vee_{i=1,n}(\forall s((s\varepsilon'\Gamma_i)\rightarrow(mg\ s))))$. By C1 this is equivalent to: $[ss(\vee_{i=1,n}(ms\ '\Gamma_i))](\vee_{i=1,n}(\forall s(([(ms\ '\Gamma_i)](mg\ s))\rightarrow(mg\ s))))$. Renaming and pulling out the universal quantifiers gives: $[ss(\vee_{i=1,n}(ms\ \Gamma_i))]\forall s_1...\forall s_n(\vee_{i=1,n}(([(ms\ '\Gamma_i)](mg\ s_i))\rightarrow(mg\ s_i)))$. This is equivalent to:

$[ss(\vee_{i=1,n}(ms\ \Gamma_i))]\forall s_1...\forall s_n((\wedge_{i=1,n}([(ms\ '\Gamma_i)](mg\ s_i)))\rightarrow(\vee_{i=1,n}(mg\ s_i)))$ By the mg laws this is equivalent to: $[ss(\vee_{i=1,n}(ms\ \Gamma_i))]\forall s_1...\forall s_n((\wedge_{i=1,n}([(ms\ '\Gamma_i)](mg\ s_i)))\rightarrow(mg\ `(\vee_{i=1,n}\ ,s_i)))$. Unfolding ss gives:

$[(\forall s(([(\vee_{i=1,n}(ms\ \Gamma_i))](mg\ s))\rightarrow(mg\ s))]\ \forall s_1...\forall s_n((\wedge_{i=1,n}([(ms\ \Gamma_i)](mg\ s_i)))\rightarrow(mg'(\vee_{i=1,n}\ ,s_i)))$

Backchaining letting $s:=(\vee_{i=1,n}\ ,s_i)$ shows that it suffices to prove:

$(\forall s_1...\forall s_n((\wedge_{i=1,n}([(ms\ \Gamma_i)](mg\ s_i)))\rightarrow([(\vee_{i=1,n}(ms\ \Gamma_i))](mg\ `(\vee_{i=1,n}\ ,s_i))))$.

In view of the hypothesis it suffices to prove: $([(\vee_{i=1,n}(mg\ s_i))](mg\ `(\vee_{i=1,n}\ ,s_i)))$.

By the mg laws this is equivalent to: $([(\vee_{i=1,n}(mg\ s_i))](\vee_{i=1,n}(mgs_i)))$ which is a tautology. QED.

MSK3 is the instance of MSK2 with only one solution:

MSK3 $(ss(\exists k(k\wedge(k\equiv(ms\ '\Gamma)))))\equiv(ms\ '\Gamma)$. and $(ss(ms\ '\Gamma))\equiv(ms\ '\Gamma)$ proof: Let n=1 in theorem MSK2.

6. The Relationship between Skeptical Logics and Modal Logic

The relationship between the proof theoretic definition of Skepticity and the modal representation is proven with theorem SL1. Theorem SL1 shows that the meaning of the intersection of all the fixed-points of the set theoretic functor nml is the meaning of the sentences entailed by the "disjunction" of the solutions of the Modal functor NML provided that hypotheses H1, H2, and H3 (listed therein) hold.

SL1: Let nml be a set theoretic functor and let NML be a modal functor. If the following three hypotheses hold:

H1: $('\kappa=(nml\ '\kappa))\rightarrow('\kappa=(fol\ '\kappa))$

H2: $((fol\ '\kappa)=(nml(fol\ '\kappa)))\leftrightarrow((ms\ '\kappa)\equiv(NML(ms\ '\kappa)))$

H3: $(\kappa\equiv(NML\ \kappa))\rightarrow\exists s(\kappa\equiv(ms\ s))$

then: $(ms(\cap\{k:\ k=(nml\ k)\}))\equiv(ss(\exists k(k\wedge k\equiv(NML\ k))))$

proof: $ms(\cap\{k:\ k=(nml\ k)\})$. By hypothesis H1 this is equivalent to: $ms(\cap\{k:\ (k=(fol\ k))\wedge(k=(nml\ k))\})$

Unfolding the definition of intersection gives: $ms\{s:\forall k((k\varepsilon\{k:\ (k=(fol\ k))\wedge(k=(nml\ k))\})\rightarrow(s\varepsilon k))\}$ which is equivalent to: $ms\{s:\forall'k(((k=(fol\ k))\wedge(k=(nml\ k)))\rightarrow(s\varepsilon k))\}$. By FOL3 this is equivalent to: $ms\{s:\forall'k(((fol\ k)=(nml(fol\ k))))\rightarrow(s\varepsilon(fol\ k)))\}$. By the Hypothesis H2, this is equivalent to: $ms\{s:\forall k(((ms\ k)\equiv(NML(ms\ k)))\rightarrow(s\varepsilon(fol\ k))))\}$

By C1 this is equivalent to: $ms\{s:\forall k(((ms\ k)\equiv(NML(ms\ k)))\rightarrow([ms\ k](mg\ s)))\}$. By the S5 laws pushing $\forall k$ to lowest scope gives: $ms\{s:\ ([\exists k((ms\ k)\wedge((ms\ k)\equiv(NML(ms\ k)))))](mg\ s))\}$

Hypothesis H3 is equivalent to the hypothesis of theorem MSK1 therefore by the conclusion of MSK1 the above expression is equivalent to: $ms\{s:\ ([\exists k(k\wedge(k\equiv(NML\ k)))](mg\ s))\}$. Unfolding ms and lambda conversion gives: $\forall s(([\exists k(k\wedge(k\equiv(NML\ k)))](mg\ s))\rightarrow(mg\ s))$. Folding ss gives $ss(\exists k(k\wedge(k\equiv(NML\ k))))$ QED.

Theorem SL1 shows that the modal representation of Skepticity using ss is logically equivalent to the set theoretic representation. $ss(\exists k(k\wedge(k\equiv(NML\ k))))$ may be generalized to: $(\exists k(k\wedge(k\equiv(NML\ k))))$ since the latter entails the same FOL object language sentences.

7. Example Relationships between Skeptical Logics and Modal Logics

Figure 8 gives the set theoretic representations of six Skeptical Logics in terms of their base logics. The six base logics are FOL: (i.e. fol), The Closed World Assumption (i.e. cwa) [Reiter 1978], Reflective Logic (i.e. rl) [Brown 1989], Default Logic (i.e. dl) [Reiter 1980], the "recursive" definition of Default Logic [Reiter 1980], and Autoepistemic Logic (i.e. ael) [Moore 1986]. The first two Skeptical Logics are defined with degenerate equations of the form k=(nml) where nml does not contain k, and thus are identical to their base Logics. These exemplify the theorems in section 6 relating set theoretic descriptions of Skeptical Logics to their corresponding modal descriptions. The last four Skeptical Logics involve the more interesting nondegenerate cases.

Skeptical Logic	Skeptical Logic constructed from the Base Logic
sfol	$\cap\{k: k=(\text{fol } '\Gamma)\}$ or $(\text{fol } '\Gamma)$
scwa	$\cap\{k: k=(\text{cwa } '\Gamma\ '\chi_i)\}$ or $(\text{cwa } '\Gamma\ '\chi_i)$where: $(\text{cwa}'\Gamma'\chi_i)=\text{df fol}('\Gamma\cup\{'\chi_i:'(\neg\chi_i)\notin(\text{fol } '\Gamma)\})$
srl	$\cap\{k: k=(\text{rl k } '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)\}$where: $(\text{rl } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)=\text{df fol}('\Gamma\cup\{'\chi_i:('\alpha_i\varepsilon'\kappa)\wedge\wedge_{j=1,mi}'(\neg\beta_{ij})\notin'\kappa\})$
sdl	$\cap\{k: k=(\text{dl k } '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)$ where: $(\text{dl } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)=\text{df}\cap\{p:(p\supseteq(\text{fol } p))\wedge(p\supseteq'\Gamma)\wedge\forall_i((('\alpha_i\varepsilon p)\wedge\wedge_{j=1,mi}'(\neg\beta_{ij})\notin'\kappa)\rightarrow('\chi_i\varepsilon p))\}$
sdr	$\cap\{k: '\kappa=(\text{dr } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)\}$ where: $(\text{dr } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)=\text{df}\cup_{t=1,\omega}(\text{r t } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)$ $(\text{r } 0\ '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)=\text{df (fol } '\Gamma)$ $(\text{r t+1 } 'k)=\text{df (fol}((\text{r t } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i)\cup\{'\chi_i:('\alpha_i\varepsilon(\text{r t } '\kappa\ '\Gamma\ '\alpha_i:'\beta_{ij}/'\chi_i))\wedge\wedge_{j=1,mi}('(\neg\beta_{ij})\notin'\kappa)\}))$
sael	$\cap\{k: k=(\text{ael k } \Gamma)\}$ where: $(\text{ael } '\kappa\ '\Gamma)=\text{df (fol}('\Gamma\cup\{'(L'\chi_i):'\chi_i\varepsilon'\kappa\}\cup\{'(\neg(L'\chi_i)):'\chi_i\notin'\kappa\}))$

Figure 8: Set Theoretic Representations of Skeptical Logics

Figure 9 gives the Modal Logic representations of six Skeptical Logics in terms of their base logic. The six base logics are the modal representations of FOL (i.e. the identity operator), The Closed World Assumption (i.e. CWA) [Brown 2005], Reflective Logic (i.e. RL) [Brown 2003a], Default Logic (i.e. DL) [Brown 2003b], "recursive" definition of Default Logic [Brown 2004], and Autoepistemic Logic (i.e. AEL) [Moore 2003c].

Skeptical Logic	Skeptical Logic constructed from the Base Logic
SFOL	$\exists k(k\wedge(k\equiv(\text{FOL } \Gamma)$ or Γ where: $(\text{FOL } \Gamma)=\text{df } \Gamma$
SCWA	$\exists k(k\wedge(k\equiv(\text{CWA } \Gamma\ \chi_i)))$ or $(\text{CWA } \Gamma\ \chi_i)$ where: $(\text{CWA } \Gamma\ \chi_i)=\text{df } \Gamma\wedge\forall_i((<\Gamma>\chi_i)\rightarrow\chi_i)$
SRL	$\exists k(k\wedge(k\equiv(\text{RL } \kappa\ \Gamma\ \alpha_i:\beta_{ij}/\chi_i)$ where:$(\text{RL k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i)=\text{df } \Gamma\wedge\forall_i((([\kappa]\alpha_i)\wedge\wedge_{j=1,mi}(<\kappa>\beta_{ij}))\rightarrow\chi_i)$
SDL	$\exists k(k\wedge(k\equiv(\text{DL k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i))$ where: $(\text{DL } \kappa\ \Gamma\ \alpha_i:\beta_{ij}/\chi_i)=\text{df } \exists p(p\wedge([p]\Gamma)\wedge\forall_i((([p]\alpha_i)\wedge\wedge_{j=1,mi}<\kappa>\beta_{ij})\rightarrow[p]\chi_i))$
SDR	$\kappa\equiv(\text{DR } \kappa\ \Gamma\ \alpha_i:\beta_{ij}/\chi_i)$ where DR is defined as: $(\text{DR } \kappa\ \Gamma\ \alpha_i:\beta_{ij}/\chi_i)=\text{df }\forall t(\text{R t k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i)$ $(\text{R } 0\ \text{k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i)=\text{df } \Gamma$ $(\text{R t+1 k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i)=\text{df (R t k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i)\wedge\forall_i((([(\text{R t k } \Gamma\ \alpha_i:\beta_{ij}/\chi_i)]\alpha_i)\wedge\wedge_{j=1,mi}(<\kappa>\beta_{ij}))\rightarrow\chi_i)$
SAEL	$\exists k(k\wedge(k\equiv(\text{AEL k } \Gamma))$ where: $(\text{AEL } \kappa\ \Gamma)=\text{df } \Gamma\wedge\forall i((\text{L } '\chi_i)\leftrightarrow([\kappa]\chi_i))$

Figure 9: Modal Representations of Skeptical Logics

The skeptical logics defined in set theory are now related to the corresponding skeptical logics defined in Modal Logic. The following theorems show that the two representation coincide

SFOL1: $(\text{ms}(\cap\{k: k=(\text{fol } '\Gamma)\}))\equiv(\text{ss}(\exists k(k\wedge(k\equiv(\text{FOL}(\text{ms } '\Gamma))))))$

proof: Letting nml be fol and NML be the identity operator makes the three hypotheses to theorem SL3 true:

(1) $('\kappa=(\text{fol } '\Gamma))\rightarrow('\kappa=(\text{fol } '\kappa))$

(2) $((\text{fol } '\kappa)=(\text{fol}(\text{fol } '\Gamma)))\leftrightarrow((\text{ms } '\kappa)\equiv(\text{FOL}(\text{ms } '\Gamma)))$

(3) $(\kappa\equiv(\text{FOL}(\text{ms } '\Gamma)))\rightarrow\exists s(\kappa\equiv(\text{ms s}))$

FOL is the identity functor. (1) is a tautology. Letting s be 'Γ makes (3) a tautology. The left hand of (2) is equivalent to: ((fol 'κ)=(fol 'Γ)). By C4 proven in [Brown 2005] this holds. The conclusion of SL1 with the variable nml instantiated to fol and the variable NML instantiated to FOL must therefore be true. QED. Since k does not occur in (FOL(ms 'Γ)), (k≡(FOL(ms 'Γ)) has only one solution and theorem MSK3 allows SFOL1 to be rewritten as: (ms(fol 'Γ))≡(FOL(ms 'Γ)).

SCWA1: (ms(∩{k: k=(cwa 'Γ 'χi)}))≡(ss(∃k(k∧(k≡(CWA(ms 'Γ)χi)))))

proof: Letting nml be cwa [Reiter 1978] and NML be CWA makes the three hypotheses of theorem SL3 true. These three hypotheses were proven as theorems FOL4, CWA2, and MC6 in [Brown 2005]:

(1) FOL4: ('κ=(cwa 'Γ 'χi))→('κ=(fol 'κ))

(2) CWA2: ((fol 'κ)=(cwa(fol 'Γ)'χi)) ↔((ms 'κ)≡(CWA(ms 'Γ)χi))

(3) MC6: (κ≡(CWA(ms 'Γ)(mg 'χi)))→∃s(κ≡(ms s))

The conclusion of SL1 with the variable nml instantiated to cwa and the variable NML instantiated to CWA must therefore be true. QED. Since k does not occur in (CWA(ms 'Γ)χi)), (k≡(CWA(ms 'Γ)χi)) has only one solution and theorem MSK3 allows SCWA1 to be rewritten as: (ms(cwa 'Γ 'χi))≡(CWA(ms 'Γχi)).

SRL1: (ms(∩{'k: 'k=(rl 'k 'Γ 'αi:'βij/'χi)}))≡(ss(∃k(k∧k≡(RL k(ms 'Γ)αi:βij/'χi))))

proof: Letting nml be rl [Brown 1989] and NML be RL makes the three hypotheses of theorem SL3 true. These three hypotheses were proven as theorems FOL4, RL2, and MR6 in [Brown 2003a]:

(1) FOL4: ('κ=(rl 'κ 'Γ 'αi:'βij/'χi))→('κ=(fol 'κ))

(2) RL2: (ms(rl(fol 'κ)'Γ 'αi:'βij/'χi))≡(RL(ms 'κ)(ms 'Γ)αi:βij/χi))

(3) MR6: (κ≡(RL κ(ms 'Γ)αi:βij/(mg 'χi)))→∃s(κ≡(ms s))

The conclusion of SL1 with nml instantiated to rl and NML instantiated to RL must therefore be true. QED.

SDL1: (ms(∩{'k: 'k=(dl 'k 'Γ 'αi:'βij/'χi)}))≡(ss(∃k(k∧k≡(DL k(ms 'Γ)αi:βij/χi))))

proof: Letting nml be dl and NML be the DL [Reiter 1980] makes the three hypotheses of theorem SL3 true. These three hypotheses were proven as theorems FOL6, DL2, and MD7 in [Brown 2003b]:

(1) FOL6: ('k=(dl 'κ 'Γ 'αi:'βij/'χi))→('κ=(fol 'κ))

(2) DL2: ((fol 'κ)=(dl(fol 'κ)'Γ 'αi:'βij/'χi)) ↔((ms 'κ)≡(DL(ms 'κ)(ms 'Γ)αi:βij/χi)))

(3) MD7: (κ≡(DL κ(ms 'Γ)αi:βij/(mg 'χi)))→∃s(κ≡(ms s))

The conclusion of SL1 with nml instantiated to dl and NML instantiated to DL must therefore be true. QED.

SDR1:(ms(∩{'k: 'k=(dl 'k 'Γ 'αi:'βij/'χi)}))≡(ss(∃k(k∧k≡(DL k(ms 'Γ)αi:βij/χi))))

proof: Letting nml be dl [Reiter 1980] and NML be the DL makes the three hypotheses of theorem SL3 true. These three hypotheses were proven as theorems FOL6, DL2, and MD7 in [Brown 2004]:

(1) FOL6: ('k=(dr 'κ 'Γ 'αi:'βij/'χi))→('κ=(fol 'κ))

(2) DR2: ((fol 'κ)=(dr(fol 'κ)'Γ 'αi:'βij/'χi)) ↔((ms 'κ)≡(DL(ms 'κ)(ms 'Γ)αi:βij/χi)))

(3) MD8: (κ≡(DR κ(ms 'Γ)αi:βij/(mg 'χi)))→∃s(κ≡(ms s))

The conclusion of SL1 with nml instantiated to dr and NML instantiated to DR must therefore be true. QED.

SAEL1: (ms(∩{k: k=(ael k 'Γ)}))≡(ss(∃k(k∧(k≡(AEL k(ms 'Γ))))))

proof: Letting nml be ael [Moore 1986] and NML be AEL makes the three hypotheses of theorem SL true. These three hypotheses were proven as theorems FOL4, AEL2, and MA6 in [Brown 2003c]:

(1) FOL4: ('κ=(ael 'κ'Γ))→('κ=(fol 'κ))

(2) AEL2: ((fol 'κ)=(ael(fol 'κ)'Γ)) ↔((ms 'κ)≡(AEL(ms 'κ)(ms 'Γ)))

(3) MA6: (κ≡(AEL κ(ms 'Γ)))→∃s(κ≡(ms s))

The conclusion of SL1 with nml instantiated to ael and NML instantiated to AEL must therefore be true. QED.

8. Conclusion

Theorem SL1 proves and section 7 exemplifies that many skeptical nonmonotonic logics can all be represented in a single modal quantificational logic which is a slight extension of S5.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers: MIP-9526532, EIA-9818341, and EIA-9972843.

Bibliography

[Bressan 1972] Bressan, Aldo 1972. *A General Interpreted Modal Calculus,* Yale University Press.

[Brown 1977] Brown, F. M., "The Theory of Meaning", Dept. of Artificial Intelligence Research Report 35, Univ. of Edinburgh, June 1977, *Edinburgh Artificial Intelligence Research Papers 1973-1986*, Scientific Datalink microfische, 1989.

[Brown 1986] Brown, Frank M. 1986. "Reasoning in a Hierarchy of Deontic Defaults", *Proceedings of the Canadian Artificial Intelligence Conference CSCI 86*, Montreal, Canada, Morgan-Kaufmann, Los Altos.

[Brown 1987] Brown, Frank M. 1987. "The Modal Logic Z", In *The Frame Problem in AI"; Proc. of the 1987 AAAI Workshop*, Morgan Kaufmann, Los Altos, CA .

[Brown 1989] Brown, Frank M. 1989. "The Modal Quantificational Logic Z Applied to the Frame Problem", advanced paper *First International Workshop on Human & Machine Cognition*, May 1989 Pensacola, Florida. Abreviated version published in *International Journal of Expert Systems Research and Applications, Special Issue: The Frame Problem. Part A.* eds. Kenneth Ford and Patrick Hayes, vol. 3 number 3, pp169-206 JAI Press 1990. Reprinted in *Reasoning Agents in a Dynamic World: The Frame problem*, editors: Kenneth M. Ford, Patrick J. Hayes, JAI Press 1991.

[Brown 2003a] Frank M. Brown "Representing Reflective Logic in Modal Logic", *Information Theories and Applications*, Vol. 10, no. 4, pages: 431-438, December 2003.

[Brown 2003b] Frank M. Brown, "Representing Default Logic in Modal Logic", *Information Theories and Applications*, Vol. 10, no. 4, pages 439-446, December, 2003.

[Brown 2003c]Frank M. Brown, "Representing Autoepistemic Logic in Modal Logic", *Information Theories and Applications*, Vol. 10, no. 4, pages 455-462, December 2003.

[Brown 2004] Brown, Frank M. 2004, "Representing "Recursive" Default Logic in Modal Logic", *Information Theories and Applications*, Vol. 11, no. 4, pages: 431-438, 2004.

[Brown 2005] Brown, Frank M. 2004, "Representing the Closed World Assumption in Modal Logic", University of Kansas Artificial Intelligence Lab Report 2005-1. Submitted to this Conference.

[Carnap 1946] Carnap, Rudolf 1946. "Modalities and Quantification" *Journal of Symbolic Logic*, vol. 11, number 2.

[Carnap 1956] Carnap, Rudolf 1956. *Meaning and Necessity: A Study in the Semantics of Modal Logic*, The University of Chicago Press.

[Fine 1970] Fine, K. 1970. "Propositional Quantifiers in Modal Logic" *Theoria* 36, p336--346.

[Hughes & Cresswell 1968] Hughes, G. E. and Creswell, M. J., 1968, *An Introduction to Modal Logic*, Methuen & Co. Ltd., London.

[Mendelson 1964] Mendelson, E. 1964. *Introduction to Mathematical Logic*, Van Norstrand, Reinhold Co., New York.

[Moore 1985] Moore, R. C. 1985. "Semantical Considerations on Nonmonotonic Logic" *Artificial Intelligence*, 25.

[Parks 1976] Parks, Z. 1976. "An Investigation into Quantified Modal Logic", Studia Logica 35, p109-125.

[Quine 1969] Quine, W.V.O., Set Theory and Its Logic, revised edition, Oxford University Press, London, 1969.

[Reiter 1978] Reiter, R. 1978. "On Closed World Databases", in, *Logic and Databases*, eds. Gallaire and Minker, J., Plenum Press, Ney York.

[Reiter 1980] Reiter, R. 1980. "A Logic for Default Reasoning", *Artificial Intelligence*, 13.

Author's Information

Frank M. Brown – University of Kansas, Lawrence, Kansas, 66045, e-mail: brown@ku.edu.

AUTOMATIC FIXED-POINT DEDUCTION SYSTEMS FOR FIVE DIFFERENT PROPOSITIONAL NONMONOTONIC LOGICS

Frank M. Brown

***Abstract**: The commonality and differences among five different nonmonotonic logics is described by implementing an automatic fixed-point equation solver for their propositional logic versions with finite stream based algorithms involving maps, filters, and accumulators. The result of organizing nonmonotonic computations in this fashion is to make apparent in an elementary way, that different nonmonotonic systems embody many of the same basic ideas and in fact differ by often only a few filters or accumulators. The nonmonotonic systems investigated are the Closed World Assumption, the kernel of Autoepistemic Logic, Frame Logic, Default Logic, and Parallel Circumscription. Scheme code which defines all the fixed-points for all these systems for all propositional problems is given.*

***Keywords**: Automatic Deduction Systems, Fixed-point, Nonmonotonic Logic.*

1. Introduction

Growing out of the need for logics and automatic deduction systems for common sense reasoning phenomena such as default reasoning, reasoning about an agent's knowledge or lack thereof, and reasoning about the consequences of robotic actions, researchers have developed a number of logical systems generally called nonmonotonic logics. The various systems at first appeared to be very different partly because of the different formalisms involved and partly because the interrelationships were not at first understood or were misunderstood. However, there are now a number of known important relationships between the different nonmonotonic systems. These results are often metatheorems which state that certain sets of nonlogical axioms and inference rules in one nonmonotonic system have fixed-point solutions that are related to the fixed-point solutions produced by certain other sets of nonlogical axioms and inference rules in some other nonmonotonic system. Herein we take a different approach to comparing nonmonotonic systems. Our approach is to begin by studying the propositional structure of nonmonotonic systems while ignoring the quantificational structure. Later, we hope to extend this research by restoring widening ranges of quantificational structure to the propositional nonmonotonic systems studied herein. Because herein we limit ourselves to propositional nonmonotonic systems everything is finite and we will be able to present the different nonmonotonic systems as Lambda Calculus functions. These functions, define the propositional versions of these nonmonotonic systems by specifying all the fixed-points for any input sets of nonlogical axioms and inference rules. By writing these functions in the fashion of (finite) stream functions involving generators, maps, filters, and accumulators, we are then able to extract out the differences among different nonmonotonic systems as some minor changes in a few of these Lambda Calculus functions thereby providing an interesting elementary way to compare them. Alternatively a programmer (as opposed to a theoretical logician) may view these functions as being Scheme programs.

Section 2 presents the call to a Propositional Logic Decision Procedure. Section 3 presents the generic functions for testing whether defaults hold. These functions implement the basic ideas of testing defaults and generating potential fixed-points which are common to all the nonmonotonic systems discussed herein. Since we are limiting this study to propositional logic the tableaux algorithm in Section 2 is decidable returning true or false. For this reason we don't need to represent the nonmonotonic default structure itself in a formal logic such as in the modal logic style of [Brown 1986] and [Brown 1989], nor in the tableaux style of [Bonatti et.al 2002] and [Olivetti 1992]. These results in a significant simplification of the automatic deduction systems involved, allowing each to be described in as a few small functions thereby making these systems available to the many AI scientists who understand LISP like languages but are not familiar with modal logic nor settheoretic fixed-point description of nonmonotonic logics. The remaining sections give Propositional Logic versions of the Closed World Assumption (Section 4), the kernel of Autoepistemic Logic (Section 5), Frame Logic (Section 6), Default Logic (Section 7), and Parallel Circumscription (Section 8). Some conclusions are made in Section 9.

2. An Automatic Theorem Prover for Propositional Logic

We assume the existence of a decision procedure for Propositional Logic represented as a Scheme function called `prove`. This function is called with two arguments as `(prove h x)` where `h` is a list of hypotheses and `c` is a theorem to be proven. Each hypothesis and the theorem are sentences of the propositional logic obtained by replacing the variables p, q, p1,...,pn in one of the forms: #t, #f, (and p1,...,pn), (or p1,...,pn), (not p), (if p q), (iff p q) by other sentences. The elementary sentences are distinct symbols such as P or list structures not beginning with a logical symbol such as (Loves John Mary).

3. Nonmonotonic Default Inference Rules

Nonmonotonic inferences in Propositional Logic may be specified by nonlogical inference rules called "defaults" of the form:

$$\frac{\alpha_1,...,\alpha_m: \beta_1,...,\beta_n}{\chi}$$

where $\alpha_1,...,\alpha_m$, $\beta_1,...,\beta_n$, and χ are sentences. Such a nonlogical inference rule is interpreted to mean that if each α_j holds in a given theory A_j and if each β_i is consistent with respect to a given theory B_i then χ may be inferred. If we suppose that all the A_j theories are identical, then the inference rule may be rewritten as:

$$\frac{\alpha : \beta_1,...,\beta_n}{\chi}$$

where α is (and $\alpha_1...\alpha_m$) since α holds in A_i if and only if each α_i holds in that theory. The case where there are no α_j sentences may then be represented by letting α be #t. However, if we suppose that the B_i theories are identical, the $\beta_1,...,\beta_n$ sentences cannot likewise be replaced by one large β since each β_i sentence may be consistent with a theory without (and $\beta_1,...,\beta_n$) itself being consistent with the theory, as for example, in the theory `(not(and p q))` where β_1 is `p` and β_2 is `q`. Thus, a nonlogical inference rule or default will be represented as a three element list of the form: $(\alpha(\beta_1...\beta_n)\chi)$

Assuming that all A_j are identical and that all the B_i are identical, a Scheme function to determine whether a nonlogical inference rule is applicable is given in Figure 2. `test` takes as arguments: A_j, B_i, and a default.

```
(define(test a b d)
       (define(pv x) (prove a x))
       (define(pos x) (not(prove b(list 'not x))))
       (and(pv(car d)) (all pos(cadr d))))
```

Fig. 1. Procedure to determine whether a default is applicable.

`test` assumes that the theories `a` and `b` are known before one tests the default `d`. However, if either of the theories on which the α and $\beta_1,...,\beta_n$ sentences of each default are tested itself includes the χ sentences of those defaults which fired, then we need to already know which χ sentences fired in order to test the α and $\beta_1,...,\beta_n$ sentences. This circularity may be avoided by picking a subset of the set of defaults to be the fired subset and then testing that each default in the chosen set is applicable and that each default not in the chosen subset is not applicable. By successively choosing as the set of fired defaults each subset of the initial set of defaults we get all potential default subsets for generating the fixed-points.

A Scheme function called `testall?` which determines whether a subset `s` of the set of defaults `Ds` is a fired subset of defaults is given in Figure 3. This function calls `test+` which checks to see that every default which should fire (i.e. those in `s`) does and `test-` which checks to see that every default which is not supposed to fire (i.e. those not in `s`), does not. It also calls the function `setdifference` which computes the defaults which are not fired. Also given in Figure 3 are the `fp` and `subsets` functions. The `fp` function produces a potential fixed-point which is a list of all the χ sentences of the fired defaults appended to the set of axioms `As`. The `subsets` function generates the list of all subsets of the set of defaults.

```
(define(testall? a b s Ds)
       (and(test+ a b s)(test- a b(set-difference Ds s))))
(define(test+ a b s)(all(lambda(d)(test a b d))s))
(define(test- a b s)(all(lambda(d)(not(test a b d)))s))
(define(set-difference Ds s) (filter(lambda(x)(not(memq x s)))Ds))
(define(fp As s)(append(map caddr s)As))
(define(subsets L)(accumulate
       (lambda(x m)(append(map(lambda(s)(cons x s))m)m))'(())L))
```

Fig. 2. Procedures to test whether a subset of defaults is a firing set of defaults.

4. The Closed World Assumption

The Closed World Assumption [Reiter 1978] (which is related to Completion systems [Clark 1978]) is a rule which allows the inference of a sentence of the form $\neg(\pi\ \delta_1...\delta_m)$ where π in an m-ary predicate and $\delta_1...\delta_m$ are variable free terms, whenever that sentence is consistent with a given theory:

$$\frac{: \neg(\pi\ \delta_1...\delta_m)}{\neg(\pi\ \delta_1...\delta_m)}$$

Since no variables occur in $(\pi\ \delta_1...\delta_m)$ we may think of it as being a propositional constant χ.

$$\frac{: \chi}{\chi}$$

We interpret this structure to mean that there is a set of axioms As such that if χ is consistent with As then χ. We generalize the closed world assumption so that there may be an α sentence, any number of β_i sentences which may be different from χ:

$$\frac{\alpha: \beta_1,...,\beta_n}{\chi}$$

and interpret this structure to mean that there is a set of axioms As such that if α follows from As and each β_i is consistent with As then χ. A Scheme function to compute the result of applying such defaults to an initial set of axioms is given in Figure 4. This function works by generating all the subsets of the set of defaults, filtering them by removing those subsets which are not fired subsets, and then by constructing the resulting fixed-points.

```
(define(cwa-fixedpoints As Ds)
       (map(lambda(s)(fp As s))
           (filter(lambda(s)(testall? As As s Ds))
                  (subsets Ds))))
```

Fig. 3. Fixed-point Deducer for The Closed World Assumption

Since the χ sentences of the defaults are not part of As which is being used to test the defaults, there is only one set of firing defaults. The resulting theory is the result of appending the χ sentences of the firing defaults to As. Thus, there is a single cwa fixed-point which may be computed by the more efficient function given in Figure 5 which tests each default to see if it fires and then constructs the fixed-point from As and all the firing defaults.

```
(define(cwa-fixedpoint As Ds)
       (fp As(filter(lambda(d)(test As As d))Ds)))
```

Fig. 4. Efficient Fixed-point Deducer for the Closed World Assumption

Example 1: This example has one axiom: A and one default:

$$\frac{:B}{B}$$

The set of axioms is represented as a list of axioms `'(A)` and the set of defaults is represented as a list of defaults `'((#t(B)B))`. Since there is no α part of this default, it is represented as `#t`. To compute the fixed-point we simply apply `cwa-fixedpoint` to the set of axioms and the set of defaults:

```
(cwa-fixedpoint '(A) '((#t(B)B))  => (A B)
```

Example 2: Here is an example illustrating the problem `cwa` has in dealing with two contradictory defaults:

$$\frac{:A}{A} \qquad \frac{:\neg A}{\neg A}$$

```
(cwa-fixedpoint '()'((#t(A)A)(#t((not A))(not A))=>(A(not A))
```

which is a fixed-point from which `#f` may be derived by the prove function.

5. The Kernel of Autoepistemic Logic

Autoepistemic Logic [Moore 1985] for a Propositional Logic includes the syntax of Propositional Logic supplemented with a unary operation L which has the properties of an S4.5 modal logic. These modal properties allow each sentence of this Propositional Autoepistemic Logic to be rewritten [Konolige 1987] as a finite set of sentences of the form: $((L\alpha)\wedge(\neg L\neg\beta_1)\wedge...\wedge(\neg L\neg\beta_n))\rightarrow\chi$ where the $(L\alpha)$ expression is optional in any sentence and n may be 0. Such a sentence can then be thought of as being the inference rule:

$$\frac{\alpha: \beta_1,...,\beta_n}{\chi}$$

where α follows from the given theory k and each β_i is consistent with that theory k. The given theory k will always be a fixed-point. The fixed-points can be obtained by enumerating all the subsets of the set of defaults, filtering out those subsets which are not fired subsets and conjoining all the χ sentences of the firing defaults to the initial axioms. A Scheme function to do this, called `aek-fixedpoints`, is given in Figure 6. [Bouix 1998] gives an analogous function written in Logistica [Brown 2003b]. This procedure is an encoding of the splitting technique on the modal logic representation of nonmonotonic reasoning described in [Brown 1986].

```
(define(aek-fixedpoints As Ds)
       (map(lambda(s)(fp As s))
           (filter(lambda(s)(testall?(fp As s)(fp As s)s Ds))
                  (subsets Ds))))
```

Fig. 6. Fixed-point Deducer for the Kernel of Autoepistemic Logic

Since the χ sentences of the defaults are part of the set being used to test the defaults there is not necessarily a single fixed-point as in the case of cwa. For example, the empty set of axioms and the default A:/¬A has no fixed-points. Likewise the empty set of axioms and the two defaults: :A/A and :¬A/¬A has two fixed-points, namely one consisting of A and the other consisting of ¬A. In general, if there are n defaults then there will be 2^n potential fixed-points produced by `subsets`. Since each call to `testall` involves a finite number of calls to the propositional logic automatic theorem prover which itself is essentially an $O(2^n)$ process, the resulting complexity is essentially: $O(2^{2n})$. This contrasts favourably with [Antoniou 1997] which describes a procedure that generates 2^m subsets where m is the total number of α and β sentences in all the defaults. Thus [Antoniou 1997] is an $O(2^{n+m})$ process where m is greater than or equal to n and is usually much larger. ([Antoniou 1997] does not however claim to be presenting an efficient algorithm). Another approach is presented in [Eiter, Klotz, Tompits, & Woltran 2002] which is based on a quadratic encoding of an Autoepistemic Theory into a propositional logic with propositional quantifiers. [Brown 2003a] compares that approach with the approach used herein. Since quantified propositional logic is at least as difficult as propositional logic, the quadratic representation leads to a proof procedure which is essentially: $O(2^{n*n})$. Both the systems of [Antoniou 1997] and [Eiter, Klotz, Tompits, &

Woltran 2002] produce additional sentences of the form: $L'\chi_i$ whenever χ_i is a theorem of the fixed-point and $\neg L'\chi_i$ whenever χ_i is not a theorem of (all) the fixed-points. This may be obtained in our system by adding the schema: $L'\chi_i \leftrightarrow [k]\chi_i$ to each kernel fixed-point k. However since L does not occur in the kernel fixed-point this addition constitutes a conservative extension of the kernel fixed-point and is therefore irrelevant. The null set of axioms with the n defaults: $\{A_i:/A_i\}_{i=1,n}$ has 2^n fixed-points – one corresponding to each subset of the set of defaults. Since, subsets in Figure 6 produces 2^n subsets, and each subset needs to be checked with `prove`, it is difficult to imagine a more asymptotically efficient algorithm that produces all the fixed-points than the one given hereabove, although producing fixed points by splitting on the modal subformulas and symbolically simplifying at each step, as was used in the nonmonotonic automatic deduction systems described in: [Brown & Araya 1991] and [Leasure 1993] would often allow solutions to be determined without enumerating all the fixed-points.

Aek is a powerful logic capable of representing, as defaults rules, action logics involving both precondition/result pairs of actions and the necessary frame laws:

$$\frac{\text{preconditions-action(t):}}{\text{results-action (t+1)}} \qquad \frac{\alpha(t):\ \chi(t+1)}{\chi(t+1)}$$

t is a number representing the time (represented as an additional argument to each predicate) when the action is applied. The first default states that the results of an action applied at time t holds at time t+1 if the preconditions held at the previous moment of time t. The second law is the frame law which states that a property α which holds at time t causes a property χ to hold at time t+1 if it is consistent for χ to do so. In the simpler cases (usually discussed in the literature) α is identical to χ.

6. Frame Logic

Frame Logic [Brown 1987] represents in a direct manner action logics involving both the precondition/result pairs of actions and the necessary frame laws. A Propositional Frame Logic involves a set of axioms and default laws representing precondition/action pairs and the necessary frame laws of the following forms:

$$\frac{\text{preconditions-action:}}{\text{results-action}} \qquad \frac{\alpha:\ \chi}{\chi}$$

Unlike Autoepistemic Logic (aek) no numeric subscripts representing time are needed because the sentences before the colon in any default rule follow from a given theory representing what holds at the preceding moment in time. Let OLD be the theory which specifies what holds in the previous moment of time. An action law then says that if the preconditions held in OLD then the results hold in the fixed-point (which represents what now holds). Likewise, the frame laws say that a property α which holds in OLD causes a property χ to hold in the fixed-point if χ is consistent for it to do so. Again in the simpler cases (usually discussed in the literature) α is identical to χ. [Brown 1989] discusses more sophisticated cases dealing with Newtonian Mechanics.

We generalize Frame Logic defaults to:

$$\frac{\alpha:\ \beta_1,\ldots,\beta_n}{\chi}$$

where α holds in the old theory and each βi is consistent with a resulting fixed-point which includes the χ sentences of the firing defaults. The fixed-points can be obtained by enumerating all the subsets of the set of defaults, filtering out those subsets which are not fired subsets and conjoining all the χ sentences of the firing defaults to the initial axioms. A Scheme function to do this, called `frame-fixedpoints`, is in Figure 7.

```
(define(frame-fixedpoints old As Ds)
       (map(lambda(s)(fp As s))
           (filter(lambda(s)(testall? old(fp As s)s Ds))
                  (subsets Ds))))
```

Fig. 7. Fixed-point Deducer for Frame Logic

7. Default Logic

Default Logic [Reiter 1980] for a Propositional Logic is essentially a set of axioms As and a finite set of default rules of the form:

$$\frac{\alpha : \beta_1,\ldots,\beta_n}{\chi}$$

where α follows from a theory k and each β_i is consistent with that theory k. The theory k must be constructible by adding to the axioms the χ sentences of those defaults whose α sentences are already deducible from the axioms and previously deduced χ sentences. This requirement does not allow the α sentence of a default to be proven by using its own χ sentence. This supported nature of Default Logic, perhaps, more closely represents defaults such as those used in taxonomies and other cases than do logics such as Autoepistemic logic. Other authors prefer other Default Logics such as Justified Default Logic or Constrained Default Logic. [Antoniou 1997] discusses the merits of different alternatives.

The fixed-points of a system of defaults of Default Logic is obtained by enumerating all the subsets of the set of defaults, filtering out those subsets which are not fired subsets, filtering out those subsets which are not supported, and then conjoining all the χ sentences of each set of firing defaults to the initial axioms. A Scheme function to derive all the fixed-points, called `dl-fixedpoints`, is in Figure 8.

```
(define(dl-fixedpoints As Ds)
       (map(lambda(s)(fp As s))
           (filter(lambda(s)(supported? As s))
                  (filter(lambda(s)(testall?(fp As s)(fp As s)s Ds))
                         (subsets Ds)))))
(define(supported? rn s)
      (define fs(filter(lambda(d)(prove rn(car d)))s))
      (if(null? fs)(null? s)
         (supported?(append(map caddr fs)rn)(set-difference s fs))))
```

Fig. 8. Fixed-point Deducer for Default Logic

Since Default logic as defined herein differs from Autoepistemic kernels by a single filter (i.e. `supported?`) it is obvious that the fixed-points of Default Logic are a subset of the fixed-points of Autoepistemic Kernels as Konolige suggested and eventually proved. (see [Konolige 1987]). This algorithm has many variations. For example, `testall?` can be replaced by:

```
(and(test+ '(#f)(fp As s)s Ds))
    (test-(fp As s)(fp As s)(set-difference Ds s)))
```

making the entailment check in test+ trivially true. This check is not needed since it is implied by the `supported?` filter. The Default Logic algorithms given in both [Schwind 1990] and [Antoniou 1997] use this fact. Another variation is that the order of filters may be swapped. Whereas [Schwind 1990] apply the above filter first followed by the `supported?` filter, [Antoniou 1997] does the reverse. [Antoniou 1997] gives an algorithm (where no default may have more than one β expression) taking this approach but which also combines the `supported?` test into the subset generation procedure. Instead of using the subset generator in Figure 8, which produces 2^n candidates, [Antoniou 1997] generates n! permutations (but suggests that these redundancies may be eliminated). His algorithm has the advantage of not producing candidate fixed-points for defaults whose α expressions do not hold in a fixed-point, and likewise for β expressions which are not consistent with any fixed-point with the expense of redundant tests on the βs.

8. Circumscription

A nonmonotonic system such as the Closed World Assumption produces precisely one fixed-point. In such a case one may determine that a conjecture follows from that fixed point by simply applying the prove function to the fixed-point and the conjecture. However, most nonmonotonic systems do not always produce precisely one fixed-point. In such a case the question arises as to what fixed-point should be used to derive further

consequences. We could choose arbitrarily but a more conservative approach is to require that the theorem hold in all fixed-points. Since (fp_i=>t) & ... & (fp_n=>t) is equivalent to (fp_1 or ... or fp_n) => t we need to prove t from the disjunction of all the fixed-points. From this perspective the problem is to compute this disjunction. It turns out that there is a nonmonotonic system, namely Circumscription [McCarthy 1980, McCarthy 1986, and Lifschitz 1985] that produces just this disjunction [Konolige 1989, Brown 1989] when all the defaults are of the form:

$$\frac{:\chi}{\chi}$$

Specifically, Circumscription for a Propositional Logic may be thought of as being a rule which infers sentences where the form $\neg(\pi\ \delta_1...\delta_m)$ where π in an m-ary predicate and $\delta_1...\delta_m$ are variable free terms, whenever that sentence is consistent with a given theory:

$$\frac{:\neg(\pi\ \delta_1...\delta_m)}{\neg(\pi\ \delta_1...\delta_m)}$$

Such a predicate is said to be Circumscribed. Since no variables occur in $(\pi\ \delta_1...\delta_m)$ we think of it as being a propositional constant. In addition to Circumscribed Predicates, there may be Variable Predicates and Fixed Predicates. Variable Predicates involve no rules but Fixed Predicates involve two contradictory rules of the form:

$$\frac{:\neg(\pi\ \delta_1...\delta_m)}{\neg(\pi\ \delta_1...\delta_m)} \qquad \frac{:(\pi\ \delta_1...\delta_m)}{(\pi\ \delta_1...\delta_m)}$$

These default structures may be interpreted as an Autoepistemic Kernel structure, a Frame Logic structure, or as a Default Logic structure since all three interpretations are the same for the defaults used in Circumscription.

We generalize the defaults of circumscription so that there may be an α sentence, any number of β_i sentences, and so that last occurrences of χ may be any sentence:

$$\frac{\alpha: \beta_1,...,\beta_n}{\chi}$$

A Scheme function for Circumscription by interpreting its defaults as Autoepistemic Kernel defaults is given in Figure 9. This function accumulates the fixed-points together by simply returning the disjunction of the conjunction of all the sentences in each Autoepistemic Kernel fixed-point (see [Brown 1989, Konolige 1989]).

```
(define(circumscription As Ds)
      (define(or-fps s) (cons 'or(map(lambda(k)(cons 'and k))s)))
      (or-fps(aek-fixedpoints As Ds)))
```

Fig. 12. Theory Constructor for Parallel Circumscription with circumscribed, fixed, and variable predicates.

9. Conclusion

The nonmonotonic systems discussed herein are summarized in Table 2 in terms of the filters and accumulators that are used. The main choice is whether testall or `test+` is used and in either case what theories are used to test the α sentences and β sentences of the defaults. There are also the issues of whether the `supported?` filter is used and whether the `or-fps` accumulator is used. Table 2 suggests the existence other nonmonotonic systems with different combinations of filters and accumulators.

Table 1. The differences among the different Nonmonotonic Systems. As is the initial set of axioms, k is the fixed-point, and kβ is the conjunction of the axioms in k and each of the β sentences of all firing defaults.

nonmonotonic system	filters	accumulator
Closed World Assumption	`λs(testall? As As s)`	
Autoepistemic Kernel	`λs(testall? k k s)`	
Frame Logic	`λs(testall? old k s)`	
Default Logic	`λs(testall? k k s),` `λs(supported? As s)`	
Circumscription	`λs(testall? k k s)`	`or-fps`

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers: MIP-9526532, EIA-9818341, and EIA-9972843.

Bibliography

[Antoniou 1997] Antoniou, Grigoris 1997. *NonMonotonic Reasoning*, MIT Press.

[Bonatti, et.al 2002], Nicola, Bonatti, Piero Andrea & Olivetti, Nicola, "Sequent Calculi for Propositional Nonmonotonic Logics", ACM Transactions on Computational Logic, Vol. 3, No. 2, pages 226-278, April 2002.

[Bouix 1998] Bouix, Sylvain, "Solving Autoepistemic Equivalences for Propositional logic", May 7, 1998. personal notes.

[Brown 1986] Brown, F. M., "A Commonsense Theory of Nonmonotonic Reasoning", *Proceedings of the 8th International Conference on Automated Deduction, CADE8,* Lecture Notes in Computer Science, vol. 230, Oxford, England, July 1986, Springer Verlag, New York.

[Brown 1987] Brown, Frank M. 1987. "The Modal Logic Z", In *The Frame Problem in AI"; Proc. of the 1987 AAAI Workshop*, Morgan Kaufmann, Los Altos, CA.

[Brown 1989] Brown, Frank M. 1989. "The Modal Quantificational Logic Z Applied to the Frame Problem", advanced paper *First International Workshop on Human & Machine Cognition*, May 1989 Pensacola, Florida. Abbreviated version published in *International Journal of Expert Systems Research and Applications, Special Issue: The Frame Problem. Part A.* eds. Kenneth Ford and Patrick Hayes, vol. 3 number 3, pp169-206 JAI Press 1990. Reprinted in *Reasoning Agents in a Dynamic World: The Frame problem*, editors: Kenneth M. Ford, Patrick J. Hayes, JAI Press 1991.

[Brown & Araya 1991] Brown, Frank M.& Carlos Araya, "A Deductive System for Theories of Nonmonotonic Reasoning," *Transactions of the Eighth Army Conference on Applied Mathematics and Computing, 1991*.

[Brown 2003a] Frank M. Brown "Decision Procedures for the Propositional Cases of 2nd Order Logic and Z Modal Logic Representations of a 1st Order L-Predicate Nonmonotonic Logic", *Automated Reasoning with Analytic Tableaux and Related Method: Tableaux 2003*, September 2003, Rome, *Lecture Notes in Artificial Intelligence 2796*, ISBN 3-540-40787-1, Springer-Verlag, Berlin, 2003.

[Brown 2003b] Frank M. Brown "Logistica 2.0: A Technology for Implementing Automatic Deduction Systems", *Automated Reasoning with Analytic Tableaux and Related Method: Tableaux 2003*, September 2003, Rome, *Lecture Notes in Artificial Intelligence 2796*, ISBN 3-540-40787-1, Springer-Verlag, Berlin, 2003.

[Clark 1978] Clark, Keith 1978. "Negation as Failure", in, *Logic and Databases*, eds. Gallaire and Minker, J., Plenum Press, New York, pages 293-322.

[Eiter, Klotz, Tompits, & Woltran 2002] Eiter, Thomas & Klotz, Volker & Tompits, Hans and Woltran, Stefan, 2002, "Modal Nonmonotonic Logics Revisited: Efficient Encodings for the Basic Reasoning Tasks", *Automated Reasoning with Analytic Tableaux and Related Method: Tableaux 2002, LNAI 2381, Springer Verlag*.

[Konolige 1987] Konolige, Kurt 1987, "On the Relation between Default Theories and Autoepistemic Logic", IJCAI87.

[Konolige 1989] Konolige, Kurt 1989. "On the Relation between Autoepistemic Logic and Circumscription Preliminary Report", IJCAI89.

[Leasure 1993] Leasure David, "A Logistica Deduction System for Solving NonMonotonic Reasoning Problems Using the Modal Logic Z" Ph.D dissertation, University of Kansas, 1993.

[Lifschitz, 1985] Lifschitz, V. 1985. "Computing Circumscription" *IJCAI 9*,pages 121-127.

[McCarthy 1980] McCarthy, J. "Circumscription -- A Form of Nonmonotonic Reasoning", *Artificial Intelligence*, vol. 13. 1980.

[McCarthy 1986] McCarthy, J. 1986. "Applications of Circumscription to Formalizing Common-Sense Reasoning", *Artificial Intelligence*, vol. 28.

[Moore 1985] Moore, R. C. 1985. "Semantical Considerations on Nonmonotonic Logic" *Artificial Intelligence*, 25.

[Olivetti 1992] Olivetti, Nicola 1992, "Tableaux and Sequent Calculus for Minimal Entailment", Journal of Automated Reasoning 9: 99-139.

[Reiter 1978] Reiter, R. 1978. "On Closed World Databases", in, *Logic and Databases*, eds. Gallaire and Minker, J., Plenum Press, New York.

[Reiter 1980] Reiter, R. 1980. "A Logic for Default Reasoning" *Artificial Intelligence*, 13.

[Schwind 1990] Schwind, Camilla 1990. "A Tableaux-Based Theorem Prover for a Decidable Subset of Default Logic", 10th International Conference on Automated Deduction, Kaiserslautern. Springer Verlag. Lecture Notes in AI vol. 449.

Author's Information

Frank M. Brown – University of Kansas, Lawrence, Kansas, 66045, e-mail: brown@ku.edu.

NONMONOTONIC SYSTEMS BASED ON SMALLEST AND MINIMAL WORLDS REPRESENTED IN WORLD LOGIC, MODAL LOGIC, AND SECOND ORDER LOGIC

Frank M. Brown

***Abstract**: A monotonic representation of a nonmonotonic logic makes it possible for an automatic theorem prover for that monotonic logic to be used to automatically deduce consequences for the nonmonotonic logic. Multiple monotonic representations, allow different automatic deduction approaches to be developed. Herein, we discuss two different nonmonotonic concepts, namely simplest worlds and minimal worlds, and show how each may be represented in three different monotonic logics. These monotonic logics are World Logic, Modal Logic, and Second Order Logic. These representations are more general than those previously described and are related back to less general previous work. In all these representations quantifiers obey the normal laws of both classical logic and S5 Modal Logic and may quantify variables across the scope of the nonmonotonic structures.*

***Keywords**: Smallest Worlds, Minimal Worlds, Nonmonotonic Logic.*

1. Introduction

From the standpoint of the Z Priorian Modal Second Order Logic, we might ask when a sentence is entailed by worlds with certain properties. Two such types of worlds of interest to nonmonotonic reasoning are the nonmonotonic concepts of the Smallest World and the Minimal Worlds. Using the laws of the Z Priorian Modal Second Order Logic, whose notation and axiomatization is given in [Brown 2005], we prove that both Simplest Worlds and Minimal Worlds can be represented in two other ways, namely in a Z MODAL Logic and in a Second Order Logic. This is important because the three equivalent representations of these two nonmonotonic concepts, allow different automatic theorem proving methods to be developed. In the succeeding sections we define Simplest and Minimal Worlds and show that they are representable in all three sub-languages. Smallest Worlds are discussed in Section 2 and Minimal Worlds are discussed in Section 3. Finally, in Section 4 we draw some conclusions.

2. Smallest Worlds

Smallestworlds is the "infinite disjunction" of all the Smallest worlds. A world I is Smallest iff every other world entailing Γ entails more instances of α. These definitions are given below:

(Smallestworlds Γ α_i) =d $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i))

(Smallest I Γ α_i) =df $\forall$J(((world J)$\wedge$([J]Γ))$\rightarrow$(I$\leq_\alpha$ J))

(I $\leq_\alpha$ J) =df $\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow$([J]$\alpha_i$))

The above definitions may be taken as the Z World Logic representation. We now prove the existence of two other equivalent representations of Smallestworlds. We prove that Smallestworlds are equivalent to the Quantified Closed World Assumption which is represented in Z Modal Logic and that it is also equivalent to Completion, which is representable in SOL. The Quantified Closed World Assumption (i.e. QCWA) is written in Z Modal SOL as:[1]

(QCWA Γ α_i) =df $\exists$k(k$\wedge$(k$\equiv\Gamma$)$\wedge\wedge_{i=1,n}\forall\xi_i$((<k>$\neg\alpha_i$)$\rightarrow\neg\alpha_i$))

where ξ_i is the sequence of free variables in α_i. Intuitively, QCWA asserts the negation of each α_i which is not entailed by Γ. A common subcase of QCWA is where each α_i has the form (π_i ξ_i).

[1] QCWA is equivalent to: $\Gamma\wedge\wedge_{i=1,n}\forall\xi_i$((<$\Gamma$>$\neg\alpha_i$)$\rightarrow\neg\alpha_i$). However, this is not necessarily a linear representation because Γ occurs n times. If there are no free variables in any α_i this sentence may be expressed in FOL metatheory as:
(cwa 'Γ 'α_i)=df ('$\Gamma\cup${'$\neg\alpha_i$: ('$\alpha_i\notin$(fol-theorems 'Γ)})) where fol-theorems produces the set of First Order Logic consequences. Thus, if the α_i constitute the set of all sentences beginning with a predicate and followed by a sequence of variable free terms we get the Closed World Assumption (i.e. cwa) as described in [Reiter 1978].

Example: (QCWA ((N 0)$\wedge\forall$x((N x)$\rightarrow$(N(+1 x)))) (N x))
is equivalent to: $\forall$x((N x)$\leftrightarrow$([π 0)$\wedge\forall$x((π x)$\rightarrow$(π(+1 x)))](π x)))

We first reformulate Smallest as an entailment:
QCWA1:(Smallest I Γ α_i)$\leftrightarrow$([Γ]$\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$))
proof: (Smallest I Γ α_i). Unfolding the definitions gives: $\forall$J(((world J)$\wedge$([J]Γ))$\rightarrow$(I $\leq_\alpha$ J))
and then: $\forall$J(((world J)$\wedge$([J]Γ))$\rightarrow\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow$([J]$\alpha_i$)))
Since J is a world, we pull out the J entailments giving: $\forall$J((world J)$\rightarrow$[J]($\Gamma\rightarrow\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$)))
By N1 this is equivalent to: []($\Gamma\rightarrow\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$)) which is equivalent to: [Γ]$\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$) QED.

We now show that the (Smallest Γ α_i) is logically equivalent to the Quantified Closed World Assumption.
QCWA2: (Smallestworlds Γ α_i) $\equiv$ (QCWA Γ α_i)
proof: (Smallestworlds Γ α_i). Unfolding the definitions gives: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i))
Using theorem QCWA1 gives: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$[Γ]$\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$))
which is equivalent to: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge\ \wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow$([$\Gamma$]$\alpha_i$)))
Since J is a world [I] may be pulled out and joined giving: $\exists$I(I$\wedge$(world I)$\wedge$([I]($\Gamma\wedge\ \wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$([$\Gamma$]$\alpha_i$))))).
From theorem D1 we get: $\Gamma\wedge\ \wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$([$\Gamma$]$\alpha_i$)) which may be rewritten as: $\Gamma\wedge\wedge_{i=1,n}\forall\xi_i$((<$\Gamma$>$\neg\alpha_i$)$\rightarrow$($\neg\alpha_i$)).
Pulling out all occurrences of Γ this may be rewritten as: $\exists$k(k$\wedge$(k$\equiv\Gamma$)$\wedge\wedge_{i=1,n}\forall\xi_i$((<k>$\neg\alpha_i$)$\rightarrow\neg\alpha_i$)) which is just (QCWA Γ α_i). QED.

Completion is a representation of Smallest Worlds in Second Order Logic. Completion in SOL is defined as:
(Completion Γ α_i) =df ($\Gamma\wedge\forall P_1\ldots P_m$(($\Gamma${$\pi_j$/$P_j$})$\rightarrow\wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$($\alpha_i${$\pi_j$/$P_j$}))))
where $\pi_i\ldots\pi_m$ are <u>all</u> the unmodalized predicate symbols in Γ and α_i.
Example: (Completion ((N 0)$\wedge\forall$x((N x)$\rightarrow$(N(+1 x)))) (N x)) is equivalent to:
$\forall$x((N x)$\leftrightarrow$($\forall$P(((P 0)$\wedge\forall$x((P x)$\rightarrow$(P(+1 x))))$\rightarrow$(P x)))

CL1: (Smallestworlds Γ α_i)$\equiv$(Completion Γ α_i)
proof: (Smallestworlds Γ α_i). Unfolding Smallestworlds gives: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i))
By QCWA1 this is: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$([Γ]$\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$)))
Since I is a world this is equivalent to: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$[I]([Γ]$\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$)))
By T1 this is equivalent to: $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$[I]$\forall P_1\ldots P_m$($\Gamma\rightarrow\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow\alpha_i$)){$\pi_j$/$P_j$})
where $\pi_i\ldots\pi_m$ are all the unmodalized predicates in Γ or any α_i. Pushing the substitutions in (to the unmodalized subformulas) gives the equivalent sentence:
$\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$[I]$\forall P_1\ldots P_m$((Γ{π_j/P_j})$\rightarrow\wedge_{i=1,n}\forall\xi_i$(([I]$\alpha_i$)$\rightarrow$($\alpha_i${$\pi_j$/$P_j$}))))
Since I is a world, [I] may be pushed in with nestings [I][I] absorbed to [I] and the then pulled out giving:
$\exists$I(I$\wedge$(world I)$\wedge$([I]($\Gamma\wedge\ \forall P_1\ldots P_m$(($\Gamma${$\pi_j$/$P_j$})$\rightarrow\wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$($\alpha_i${$\pi_j$/$P_j$})))))))
By D1 this is equivalent to: $\Gamma\wedge\forall P_1\ldots P_m$(($\Gamma${$\pi_j$/$P_j$})$\rightarrow\wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$($\alpha_i${$\pi_j$/$P_j$}))) QED

(SOL) Parallel Predicate Completion is the instance of Completion where each α_i is a sentence beginning with a distinct predicate: (π_i ξ_i). In this case n$\leq$m. It may be written as:
$\Gamma\wedge\ \forall P_1\ldots P_M$(($\Gamma${$\pi_i$/$P_i$})$\rightarrow\wedge_{i=1,n}\forall\xi_i$(($\pi_i$ ξ_i)$\rightarrow$(P_i ξ_i))) where $\pi_i\ldots\pi_m$ are <u>all</u> the unmodalized predicate symbols in Γ. The Completed Predicates are $\pi_1\ldots\pi_n$ and the Varying Predicates are $\pi_{n+1}\ldots\pi_m$. Rewriting the Varying Predicates with different variable and metavariable symbols and indices gives:
$\Gamma\wedge\ \forall P_1\ldots P_n\forall Z_1\ldots Z_m$(($\Gamma${$\pi_i$/$P_i$}{$\rho_j$/$Z_j$})$\rightarrow\wedge_{i=1,n}\forall\xi_i$(($\pi_i$ ξ_i) $\rightarrow$(P_i ξ_i)))
where $\pi_1\ldots\pi_n$ and $\rho_1\ldots\rho_m$ are <u>all</u> of the unmodalized predicate symbols in Γ.[1] 2.

[1]A procedure for completing predicates [Clark 1978] for a set of Horn Clauses Γ involving separable predicates is as follows: Γ is put into the form ($\chi\wedge\wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$($\pi_i$ ξ_i))) where $\pi_1\ldots\pi_n$ do not occur in χ nor in any α_i. The completion is to replace each $\rightarrow$ by $\leftrightarrow$. In this form the definition of predicate completion in the text becomes:
($\chi\wedge\wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$($\pi_i$ ξ_i))) $\wedge\ \forall P_1\ldots P_m$(($\chi\wedge\wedge_{i=1,n}\forall\xi_i$($\alpha_i\rightarrow$($P_i$ ξ_i))) $\rightarrow\wedge_{i=1,n}\forall\xi_i$(($\pi_i$ ξ_i) $\rightarrow$(P_i ξ_i)))

3 Minimal Worlds

Minworlds is the "infinite disjunction" of all the minimal worlds which entail Γ. A world I is minimal iff every world which entails Γ and is less than or equal to I for all α_i, is such that I is less than or equal to it for all α_i. A world is less than or equal to another world for α_i iff α_i is entailed by the second world whenever it is entailed by the first. These definitions[1] are given below:

(Minworlds Γ α_i) =df $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Minimal I Γ α_i))

(Minimal I Γ α_i) =df $\forall$J(((world J)$\wedge$([J]Γ)$\wedge$(J $\leq\alpha$ I))$\rightarrow$(I $\leq\alpha$ J))

(J $\leq\alpha$ I) =df $\wedge_{i=1,n}\forall\xi_i$(([J]$\alpha_i$)$\rightarrow$([I]$\alpha_i$))

Every Smallest World is a Minimal World. Even though Minimal Worlds are not necessarily Smallest worlds, if the "disjunction" of the smallest worlds is possible then there is only one and it will be the Minimal World. These facts are proven below[2]:

SM1: Smallest worlds entail minimal worlds: [(Smallestworlds Γ α_i)](Minworlds Γ α_i)

proof: Unfolding the definitions of Smallestworlds and Minworlds gives:

[($\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i)))] ($\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Minimal I Γ α_i)))

which by letting I:=I simplifies to just: [(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i))] (Minimal I Γ α_i)

Generalizing gives: (Smallest k Γ α_i)$\rightarrow$(Minimal k Γ α_i). Unfolding Smallest and Minimal gives:

($\forall$J(((world J)$\wedge$([J]Γ))$\rightarrow$(I $\leq\alpha$ J)))$\rightarrow\forall$J(((world J)$\wedge$([J]Γ)$\wedge$(J $\leq\alpha$ I))$\rightarrow$(I $\leq\alpha$ J))

which holds by classical logic. QED

SM2: A Reduction case of Minworlds to Smallestworlds.

(<>(Smallestworlds Γ α_i))$\rightarrow$(Minworlds Γ α_i)$\equiv$(Smallestworlds Γ α_i)

proof: By SM1 it suffices to prove: (<>(Smallestworlds Γ α_i))$\rightarrow$ [(Minworlds Γ α_i)](Smallestworlds Γ α_i)

Unfolding Smallestworlds and Minworlds gives:

(<>$\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i)))$\rightarrow$

[$\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Minimal I Γ α_i)) $\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i))

The hypothesis simplifies as follows: first: <>$\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i)), then

$\exists$I <> (I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i)), then $\exists$I ((<>I)$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i)), and then:

$\exists$W (world W)$\wedge$([W]Γ)$\wedge$(Smallest W Γ α_i)) and the conclusion simplifies as follows:

[$\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Minimal I Γ α_i))]$\exists$I(I$\wedge$(world I)$\wedge$([I]Γ)$\wedge$(Smallest I Γ α_i))

$\forall$U[(U$\wedge$(world U)$\wedge$([U]Γ)$\wedge$(Minimal U Γ α_i))]$\exists$V(V$\wedge$(world V)$\wedge$([V]Γ)$\wedge$(Smallest V Γ α_i)). In Z, this is equivalent to: $\forall$U((world U)$\wedge$([U]Γ)$\wedge$(Minimal U Γ α_i))$\rightarrow\exists$V([U]V]$\wedge$(world V)$\wedge$([V]Γ)$\wedge$(Smallest V Γ α_i))

Since a world entails a world only if they are synonymous we get:

$\forall$U((world U)$\wedge$([U]Γ)$\wedge$(Minimal U Γ α_i))$\rightarrow$ $\exists$V(U$\equiv$V)$\wedge$(world V)$\wedge$([V]Γ)$\wedge$(Smallest V Γ α_i))

Eliminating the $\exists$V gives: $\forall$U ((world U)$\wedge$([U]Γ)$\wedge$(Minimal U Γ α_i))$\rightarrow$((world U)$\wedge$([U]Γ)$\wedge$(Smallest U Γ α_i))

which simplifies to: $\forall$U(((world U)$\wedge$([U]Γ)$\wedge$(Minimal U Γ α_i))$\rightarrow$(Smallest U Γ α_i))

Putting the simplified hypothesis and conclusion together gives:

Letting P_i be $\lambda\xi_i\alpha_i$ allows us to infer: ($\chi\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow(\pi_i\ \xi_i)))\wedge\wedge_{i=1,n}\forall\xi_i((\pi_i\ \xi_i)\rightarrow\alpha_i)$
which is equivalent to: $\chi\wedge\wedge_{i=1,n}\forall\xi_i((\pi_i\ \xi_i)\leftrightarrow\alpha_i)$.

[1] A long-winded definition of Minimal is: (Minimal I Γ α_i β_j) =df ($\forall$J((world J)$\wedge$([J]Γ)$\wedge$(J $\leq\alpha$ I)$\wedge$(J =β I))$\rightarrow$(I $\leq\alpha$ J)), (J $\leq\alpha$ I) =df ($\wedge_{i=1,n}\forall\xi_i$(([J]$\alpha_i$)$\rightarrow$([I]$\alpha_i$))), (J =$\beta$ I) =df (J $\leq\beta$ I)$\wedge$(I $\leq\beta$ J). Since (I $\leq\beta$ J)) is equivalent to (J $\leq\neg\beta$ I) this definition of minimal is equivalent to the one used in the text provided that the α_i used therein is (α_i,β_j,$\neg\beta_j$) for the α_i and β_j used here.

[2] The instance of SM2 where each α_i is of the form (π_i ξ_i), where π_1,...,π_n are all the predicates in Γ, and where the domain is finite is analogous to the result in [Lifschitz 1985] relating Simplest Models (actually cwa) to Minimal Models.

$(\exists W((\text{world } W)\wedge([W]\Gamma)\wedge(\text{Smallest } W\ \Gamma\ \alpha_i)))\rightarrow \forall U(((\text{world } U)\wedge([U]\Gamma)\wedge(\text{Minimal } U\ \Gamma\ \alpha_i))\rightarrow(\text{Smallest } U\ \Gamma\ \alpha_i))$

which follows from:

$((\text{world } W)\wedge([W]\Gamma)\wedge(\text{Smallest } W\ \Gamma\ \alpha_i)\wedge(\text{world } U)\wedge([U]\Gamma)\wedge(\text{Minimal } U\ \Gamma\ \alpha_i))\rightarrow(\text{Smallest } U\ \Gamma\ \alpha_i)))$

Unfolding Smallest and Minimal gives:

$(\text{world } W)\wedge([W]\Gamma)\ \wedge\forall J(((\text{world } J)\wedge([J]\Gamma))\rightarrow(W\leq\alpha\ J)))\wedge$

$(\text{world } U)\wedge([U]\Gamma)\ \wedge(\forall J((\text{world } J)\wedge([J]\Gamma)\wedge(J\leq\alpha\ U))\rightarrow(U\leq\alpha\ J)))\rightarrow\forall J(((\text{world } J)\wedge([J]\Gamma))\rightarrow(U\leq\alpha\ J))))$

Noting that all variables are worlds, which entail Γ let us pretend these are sorted quantifiers of that ilk. Writing the above as sorted quantifiers gives: $(\forall J(W\leq\alpha\ J)\ \wedge\forall J((J\leq\alpha\ U)\rightarrow(U\leq\alpha\ J)))\rightarrow\forall J(U\leq\alpha\ J)$

which is implied by: $(\forall J(W\leq\alpha\ J)\ \wedge\forall J((J\leq\alpha\ U)\rightarrow(U\leq\alpha\ J)))\rightarrow(U\leq\alpha\ J)$

Letting the first J quantifier have the instances: U and J, and letting the second J quantifier have the instance W gives: $(W\leq\alpha\ U)\wedge(W\leq\alpha\ J)\ \wedge((W\leq\alpha\ U)\rightarrow(U\leq\alpha\ W))\rightarrow(U\leq\alpha\ J)$

which simplifies to: $(W\leq\alpha\ U)\wedge(W\leq\alpha\ J)\ \wedge(U\leq\alpha\ W))\ \rightarrow(U\leq\alpha\ J)$

Since $\leq\alpha$ is transitive the second and third hypotheses imply the conclusion. QED

T2 suggests the importance of Minimal Worlds as opposed to Smallest Worlds lies in the case where the Smallestworlds is not possible. A simple example of this case is in dealing with disjunctive information when the negation of the components of the disjunction is defaults:

$(\text{Smallestworlds } ((P\ a)\vee(P\ b))\ \ (P\ x))\equiv \#f$

$(\text{Minworlds } ((P\ a)\vee(P\ b))\ \ (P\ x))\equiv\ (\forall x((P\ x)\leftrightarrow(x=a)))\vee(\forall x((P\ x)\leftrightarrow(x=b)))$

The definition of Minimal Worlds constitutes the World Logic Representation. We now prove the existence of two other equivalent representations of Minimal Worlds. We prove that Minimal Worlds are equivalent to the "infinite disjunction" of a necessary equivalence which is represented in Z Modal Logic and that it is also equivalent to Circumscription which is representable in SOL. Quantified Possibility (i.e., $(Qpos\ \kappa\ \Gamma\ \alpha_i)$) with a sentence κ representing the resulting knowledgebase, a sentence Γ representing the conjunction of the initial nonlogical axioms, and a conjunction of defaults which imply α_i if α_i is possible with κ, is defined as follows:

$(Qpos\ \kappa\ \Gamma\ \alpha_i)\ =df\ \Gamma\wedge\ \wedge_{i=1,n}\forall\xi_i(<\kappa>\neg\alpha_i\rightarrow\neg\alpha_i)$

From this definition we form the necessary equivalence representing the solutions as follow[1],[2]:

$$\kappa\equiv(Qpos\ \kappa\ \Gamma\ \alpha_i)$$

A necessary equivalence may have zero or more solutions including an infinite number of solutions. When a necessary equivalence has more than one solution there the question as to which solution is to be used arises . One way of avoiding this question is to adopt the skeptical approach of only accepting what is common to all the solutions. This is achieved with the "$\exists k(k\wedge$" idiom as in:

$$\exists k(k\wedge\ (k\equiv(Qpos\ k\ \Gamma\ \alpha_i)))$$

This may be read as the (possibly infinite) disjunction of all the solutions k to the necessary equivalence. When there are a finite number of solutions $\beta_1,\ldots,\beta_n$ to such a necessary equivalence represented as: $(k\equiv\beta_1)\vee\ldots\vee(k\equiv\beta_n)$, the sentence $\exists k(k\wedge(k\equiv(Qpos\ k\ \Gamma\ \alpha_i)))$ is equivalent to $\exists k(k\wedge((k\equiv\beta_1)\vee\ldots\vee(k\equiv\beta_n)))$ which by the distribution properties of $\wedge$ and $\vee$ and by the fact that $\exists$ associates through $\vee$ gives the equivalent expression:

$(\exists k(k\wedge(k\equiv\beta_1)))\vee\ldots\vee(\exists k(k\wedge(k\equiv\beta_n))).$

[1] Necessary Equivalences in Modal Logics were first discussed in [Brown 1986].

[2] This particular type of necessary equivalence is a generalization of fixed-point representations where quantifiers are allowed to cross the scope of defaults of both Default Logic [Reiter 1980] consisting only of defaults of the form: $\#t{:}\neg\alpha/\neg\alpha$ and of the Kernel of Autoepistemic Logic where the modal sentences are of the form: $(\neg L\alpha)\rightarrow\neg\alpha$. (see [Konolige 1987, Konolige 1987b, Brown 1987, Brown 1989, Antoniou 1997]. Unlike a number of other generalizations of these logics to the case where where free variables may occur in α [Konolige 1989], this representation in World Logic obeys all the normal laws of FOL, SOL, and S5 Modal Logic.

Since $\exists k(k\wedge(k\equiv\beta_i))$ is logically equivalent to just β_i, the result will then be equivalent to the disjunction of solutions: $\beta_1\vee\ldots\vee\beta_n$. However, if there are an infinite number of solutions the existential quantifier is not eliminatable in this manner.

SK1: $(\text{Minimal I }\Gamma\ \alpha_i)\leftrightarrow([\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))]\ \wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i))$

proof: $(\text{Minimal I }\Gamma\ \alpha_i)$. Unfolding the definitions gives: $\forall J(((\text{world J})\wedge([J]\Gamma)\wedge(J\leq_\alpha I))\rightarrow(I\leq_\alpha J))$

and then: $\forall J(((\text{world J})\wedge([J]\Gamma)\wedge(\wedge_{i=1,n}\forall\xi_i(([J]\alpha_i)\rightarrow([I]\alpha_i))))\rightarrow\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow([J]\alpha_i)))$

Since J is a world we can pull out the J entailments giving:

$\forall J((\text{world J})\rightarrow[J]((\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))\rightarrow\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)))$

By Prior's Law this is equivalent to: $[](((\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))\rightarrow\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i))$

which is equivalent to: $[\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))]\ \wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)$ QED

We now show that the $(\text{Minworlds }\Gamma\ \alpha_i)$ is logically equivalent to the "infinite disjunction" of all solutions to the fixed point equivalence:

SK2: $(\text{Minworlds }\Gamma\ \alpha_i)\equiv\exists k(k\wedge(k\equiv(\text{Qpos k }\Gamma\ \alpha_i)))$

proof: $(\text{Minworlds }\Gamma\ \alpha_i)$. Unfolding the definition of Minworlds gives: $\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge(\text{Minimal I }\Gamma\ \alpha_i))$

which by SK1 is just: $\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge([\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))]\ \wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)))$

This is equivalent to: $\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge([\Gamma\wedge\wedge_{i=1,n}\forall\xi_i((\langle I\rangle\neg\alpha_i)\rightarrow\neg\alpha_i))]\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)))$

Since I is a world it entails: $\wedge_{i=1,n}\forall\xi_i((\langle I\rangle\neg\alpha_i)\rightarrow\neg\alpha_i))$and the above may be rewritten as:

$\exists I(I\wedge(\text{world I})\wedge([I](\Gamma\wedge\wedge_{i=1,n}\forall\xi_i((\langle I\rangle\neg\alpha_i)\rightarrow\neg\alpha_i)))\wedge([\Gamma\wedge\wedge_{i=1,n}\forall\xi_i((\langle I\rangle\neg\alpha_i)\rightarrow\neg\alpha_i)]\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)))$

Using the definition of Qpos this becomes: $\exists I(I\wedge(\text{world I})\wedge([I](\text{Qpos I }\Gamma\ \alpha_i))\wedge([(\text{Qpos I }\Gamma\ \alpha_i)]\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)))$

which we rewrite so as to combine the two Qpos subexpressions:

$(\exists I(I\wedge(\text{world I})\wedge\exists k((k\equiv(\text{Qpos I }\Gamma\ \alpha_i))\wedge([I]k)\ \wedge([k]\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i)))$

and then by S5 modal Logic as: $\exists k\exists I(I\wedge(\text{world I})\wedge([I]k)\wedge(k\equiv(\text{Qpos I }\Gamma\ \alpha_i))\wedge(\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow([k]\alpha_i)))$

Since $([I]k)$ implies that $\wedge_{i=1,n}\forall\xi_i(([k]\alpha_i)\rightarrow([I]\alpha_i))$ the above is equivalent to:

$\exists k\exists I(I\wedge(\text{world I})\wedge([I]k)\wedge(k\equiv(\text{Qpos I }\Gamma\ \alpha_i))\ \wedge(\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\leftrightarrow([k]\alpha_i)))$

Since $\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\leftrightarrow([k]\alpha_i)))$ is equivalent to $\wedge_{i=1,n}\forall\xi_i((\langle I\rangle\neg\alpha_i)\leftrightarrow(\langle k\rangle\neg\alpha_i))$, it allows I in Qpos to be replaced by k giving: $\exists k\exists I(I\wedge(\text{world I})\wedge([I]k)\wedge(k\equiv(\text{Qpos k }\Gamma\ \alpha_i))\ \wedge(\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\leftrightarrow([k]\alpha_i)))$

Again, since $([I]k)$ implies $\wedge_{i=1,n}\forall\xi_i(([k]\alpha_i)\rightarrow([I]\alpha_i)))$ this is equivalent to:

$\exists k\exists I(I\wedge(\text{world I})\wedge([I]k)\wedge(k\equiv(\text{Qpos I }\Gamma\ \alpha_i))\wedge(\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow([k]\alpha_i)))$

which is equivalent to: $\exists k\exists I(I\wedge(\text{world I})\wedge([I]k)\wedge(k\equiv(\text{Qpos k }\Gamma\ \alpha_i))\wedge(\wedge_{i=1,n}\forall\xi_i((\langle k\rangle\neg\alpha_i)\rightarrow\neg([I]\alpha_i)))$

Since $(k\equiv(\text{QPos k }\Gamma\ \alpha_i))$ implies $\wedge_{i=1,n}\forall\xi_i((\langle k\rangle\neg\alpha_i)\rightarrow([k]\neg\alpha_i))$, $[I]k$ implies $\wedge_{i=1,n}\forall\xi_i(([k]\neg\alpha_i)\rightarrow([I]\neg\alpha_i))$, and (world I) implies $\wedge_{i=1,n}\forall\xi_i([I]\neg\alpha_i)\rightarrow\neg([I]\alpha_i))$, from these three hypotheses we infer: $\wedge_{i=1,n}\forall\xi_i((\langle k\rangle\neg\alpha_i)\rightarrow\neg([I]\alpha_i))$.

Thus the formula above is equivalent to: $\exists k\exists I(I\wedge(\text{world I})\wedge([I]k)\wedge(k\equiv(\text{Qpos k }\Gamma\ \alpha_i)))$

By D1 this simplifies to just: $\exists k(k\wedge(k\equiv(\text{Qpos k }\Gamma\ \alpha_i)))$ QED

Circumscription in Second Order Logic is defined as:

$(\text{Circ }\Gamma\ \alpha_i)=\text{df }(\Gamma\wedge\forall P_1\ldots P_m(((\Gamma\{\pi_u/P_j\})\wedge\wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\})\rightarrow\alpha_i)))\rightarrow\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow(\alpha_i\{\pi_j/P_j\})))$

where $\pi_i\ldots\pi_m$ are <u>all</u> the unmodalized predicate symbols in Γ and α_i.

CIRC1: $(\text{Minworlds }\Gamma\ \alpha_i)\equiv(\text{Circ }\Gamma\ \alpha_i)$

proof: $(\text{Minworlds }\Gamma\ \alpha_i)$. Unfolding Minworlds gives: $\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge(\text{Minimal I }\Gamma\ \alpha_i))$

By SK1 this is: $\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge[\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))]\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i))$

Since I is a world this is equivalent to: $\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge[I][\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i))]\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i))$

By T1 this is equivalent to:

$\exists I(I\wedge(\text{world I})\wedge([I]\Gamma)\wedge[I]\forall P_1\ldots P_m(((\Gamma\wedge\wedge_{i=1,n}\forall\xi_i(\alpha_i\rightarrow([I]\alpha_i)))\rightarrow\wedge_{i=1,n}\forall\xi_i(([I]\alpha_i)\rightarrow\alpha_i))\{\pi_j/P_j\}))$

where $\pi_i \ldots \pi_m$ are all the unmodalized predicates in Γ or any α_i. Pushing the substitutions into the unmodalized subformulas gives the equivalent sentence:

$\exists I(I \wedge (\text{world } I) \wedge ([I]\Gamma) \wedge [I]\forall P_1 \ldots P_m\ (((\Gamma\{\pi_j/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow ([I]\alpha_i))) \rightarrow \wedge_{i=1,n}\forall\xi(([I]\alpha_i) \rightarrow (\alpha_i\{\pi_j/P_j\}))))$

Since I is a world, [I] may be pushed in with nestings [I][I] absorbed to [I] and the then pulled out giving:

$\exists I(I \wedge (\text{world } I) \wedge ([I](\Gamma \wedge \forall P_1 \ldots P_m\ (((\Gamma\{\pi_j/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow \alpha_i)) \rightarrow \wedge_{i=1,n}\forall\xi_i(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\})))))$

By D1 this is equivalent to: $\Gamma \wedge \forall P_1 \ldots P_m(((\Gamma\{\pi_j/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow \alpha_i)) \rightarrow \wedge_{i=1,n}\forall\xi_i(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\})))$ QED

Circumscription with explicit fixed predicates:

Our definition of Circumscription requires that $\pi_i \ldots \pi_m$ are all of the unmodalized predicate symbols in Γ and any α_i. This requirement is not necessary since any Circumscription not obeying that requirement (which we will call Circ*) is equivalent to a Circumscription which does:

$(\text{Circ}^*\ \Gamma\ \alpha_i) =_{df} \Gamma \wedge \forall P_1 \ldots P_m(((\Gamma\{\pi_u/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow \alpha_i)) \rightarrow \wedge_{i=1,n}\forall\xi_i(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\})))$

where $\pi_i \ldots \pi_m$ are some the unmodalized predicate symbols in Γ or any α_i.

CIRC2: $(\text{Circ}\ \Gamma\ (\alpha_i, (\rho_j\ \xi_j), \neg(\rho_j\ \xi_j))) \equiv (\text{Circ}^*\ \Gamma\ \alpha_i)$ where the ρ_j predicates are the ones missing from the π_i list.

proof: $(\text{Circ}\ \Gamma\ (\alpha_i, (\rho_j\ \xi_j), \neg(\rho_j\ \xi_j)))$. Unfolding Circ letting Q be the variables for the ρ predicates from gives:

$\forall P_1 \ldots P_m \forall Q_1 \ldots Q_l(((\Gamma\{\pi_j/P_j\}\{\rho_j/Q_j\})$

$\wedge\wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}\{\rho_j/Q_j\}) \rightarrow \alpha_i) \wedge \wedge_{j=1,l}\forall\zeta_i(((\rho_j\ \zeta_i)\{\rho_j/Q_j\}) \rightarrow (\rho_j\ \zeta_j)) \wedge \wedge_{j=1,l}\forall\zeta_i(((\neg\rho_j\ \zeta_j)\{\rho_j/Q_j\}) \rightarrow (\neg\rho_j\ \zeta_j)))$

$\rightarrow (\wedge_{i=1,n}\forall\xi_j(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\}\{\rho_j/Q_j\} \wedge \wedge_{j=1,l}\forall\zeta_j((\rho_j\ \zeta_j) \rightarrow ((\rho_j\ \zeta_j)\{\rho_j/Q_j\})) \wedge \wedge_{j=1,l}\forall\zeta_j((\neg\rho_j\ \zeta_j) \rightarrow ((\neg\rho_j\ \zeta_j)\{\rho_j/Q_j\}))))$

which is equivalent to:

$\forall P_1 \ldots P_m \forall Q_1 \ldots Q_l(((\Gamma\{\pi_j/P_j\}\{\rho_j/Q_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}\{\rho_j/Q_j\}) \rightarrow \alpha_i) \wedge \wedge_{j=1,l}\forall\zeta_i((Q_j\ \zeta_j) \leftrightarrow (\rho_j\ \zeta_j)))$
$\rightarrow (\wedge_{i=1,n}\forall\xi_j(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\}\{\rho_j/Q_j\})) \wedge \wedge_{j=1,l}\forall\zeta_j((\rho_j\ \zeta_j) \leftrightarrow (Q_j\ \zeta_j)))$

The $\wedge_{i=1,l}\forall\zeta_i((Q_j\ \zeta_j) \leftrightarrow (\rho_j\ \zeta_j))$ hypothesis allows each Q_j elsewhere to be replaced by ρ_j giving:

$\forall P_1 \ldots P_m \forall Q_1 \ldots Q_l(((\Gamma\{\pi_j/P_j\}\{\rho_j/\rho_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}\{\rho_j/\rho_j\}) \rightarrow \alpha_i) \wedge \wedge_{j=1,l}\forall\zeta_i((Q_j\ \zeta_j) \leftrightarrow (\rho_j\ \zeta_j)))$
$\rightarrow (\wedge_{i=1,n}\forall\xi_j(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\}\{\rho_j/\rho_j\})) \wedge \wedge_{j=1,l}\forall\zeta_j((\rho_j\ \zeta_j) \leftrightarrow (\rho_j\ \zeta_j)))$

This simplifies to:

$\forall P_1 \ldots P_m \forall Q_1 \ldots Q_l(((\Gamma\{\pi_j/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow \alpha_i) \wedge \wedge_{j=1,l}\forall\zeta_i((Q_j\ \zeta_j) \leftrightarrow (\rho_j\ \zeta_j))) \rightarrow \wedge_{i=1,n}\forall\xi_j(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\}))$

Pushing the Q quantifiers to lowest scope gives:

$\forall P_1 \ldots P_m(((\Gamma\{\pi_j/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow \alpha_i) \wedge \wedge_{j=1,l}\exists Q_j\forall\zeta_i((Q_j\ \zeta_j) \leftrightarrow (\rho_j\ \zeta_j))) \rightarrow \wedge_{i=1,n}\forall\xi_j(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\}))$

Letting each Q_j be ρ_j then gives: $\forall P_1 \ldots P_m((\Gamma\{\pi_j/P_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((\alpha_i\{\pi_j/P_j\}) \rightarrow \alpha_i)) \rightarrow \wedge_{i=1,n}\forall\xi_j(\alpha_i \rightarrow (\alpha_i\{\pi_j/P_j\}))$

Folding Circ* then gives: $(\text{Circ}^*\ \Gamma\ \alpha_i)$. QED

(SOL Single) Formulae Circumscription [McCarthy 1986] is the instance of Circ* where n=1. Dropping the superfluous i subscript, it may be expressed as:

$\Gamma \wedge \forall P_1 \ldots P_m(((\Gamma\{\pi_j/P_j\}) \wedge \forall\xi((\alpha\{\pi_j/P_j\}) \rightarrow \alpha)) \rightarrow \forall\xi(\alpha \rightarrow (\alpha\{\pi_j/P_j\})))$

where $\pi_1 \ldots \pi_m$ are some of the unmodalized predicate symbols in Γ or α.

(SOL) Parallel (Predicate) Circumscription [Lifschitz 1985] is the instance of Circ* where each α_i is a sentence beginning with a distinct predicate: $(\pi_i\ \xi_j)$ In this case n≤m. This may be written as:

$\Gamma \wedge \forall P_1 \ldots P_m(((\Gamma\{\pi_i/P_i\}) \wedge \wedge_{i=1,n}\forall\xi_i((P_i\ \xi_i) \rightarrow (\pi_i\ \xi_i))) \rightarrow \wedge_{i=1,n}\forall\xi_i((\pi_i\ \xi_i) \rightarrow (P_i\ \xi_i)))$

where $\pi_1 \ldots \pi_m$ are some of the unmodalized predicate symbols in Γ. The Circumscribed Predicates are $\pi_1 \ldots \pi_n$ and the Varying Predicates are $\pi_{n+1} \ldots \pi_m$. Rewriting the Varying Predicate with different variable and metavariable symbols and indices gives the usual formulation:

$\Gamma \wedge \forall P_1 \ldots P_n \forall Z_1 \ldots Z_m\ (((\Gamma\{\pi_i/P_i\}\{\rho_j/Z_j\}) \wedge \wedge_{i=1,n}\forall\xi_i((P_i\ \xi_i) \rightarrow (\pi_i\ \xi_i))) \rightarrow \wedge_{i=1,n}\forall\xi_i((\pi_i\ \xi_i) \rightarrow (P_i\ \xi_i)))$

where $\pi_1 \ldots \pi_n$ and $\rho_1 \ldots \rho_m$ are some of the unmodalized predicate symbols in Γ.

(FOL) Parallel (Predicate) Circumscription without any varying predicates: Ruling out any varying predicates and replacing the second order variables in the above sentence with schematic metavariables, we may infer the FOL axiom scheme suggested in [McCarthy 1980]: $\Gamma \wedge (((\Gamma\{\pi_i/\lambda\xi_i\beta_i\}) \wedge \wedge_{i=1,n}\forall\xi_i(\beta_i \rightarrow (\pi_i\ \xi_i))) \rightarrow \wedge_{i=1,n}\forall\xi_i((\pi_i\ \xi_i) \rightarrow \beta_i))$

where $\pi_1 \ldots \pi_m$ are some of the unmodalized predicate symbols in Γ.

4. Conclusion

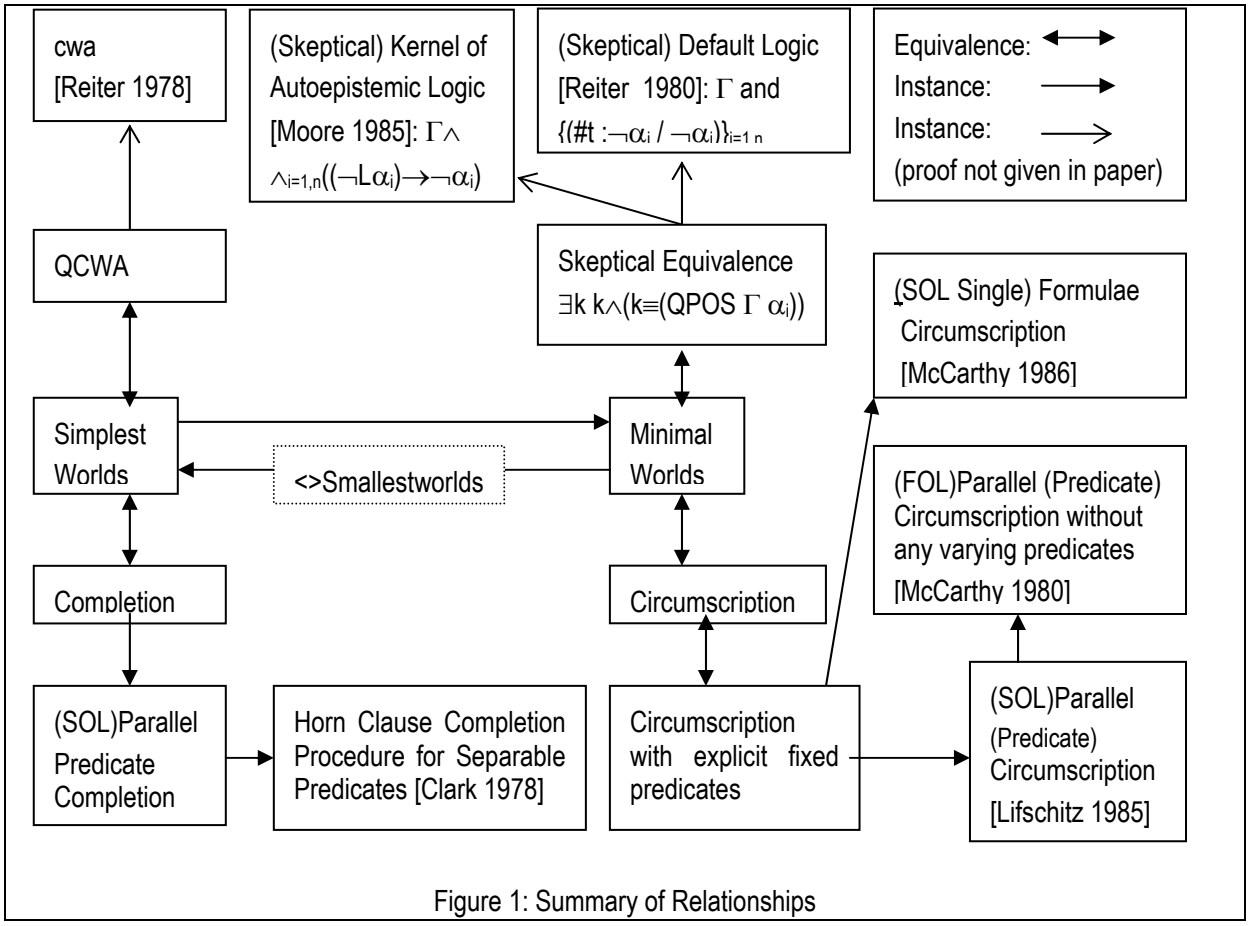


Figure 1: Summary of Relationships

We have proven in the Z Priorian Modal Second Order Logic that Smallestworlds are equivalent to both QCWA (theorem QCWA2) and Completion (theorem CL1). Likewise, we have proven that Minworlds is equivalent to both Skeptical Necessary Equivalences (theorem SK2) and Circumscription (theorem CIRC1). Thus, for each of these two types of worlds we have shown that they are represented in three different ways: namely in Z World Logic, Z Modal Logic, and Second Order Logic. The relationships among these systems are summarized in Figure 1 using the bi-arrows.

In addition, the two horizontal arrows show that Simplest Worlds entail Minimal Worlds (theorem SM1) and that Minimal Worlds entail Simplest worlds if Simplest Worlds are logically possible (theorem SM2). The other arrows show how a number of previously developed nonmonotonic systems are related to the representations discussed herein. These include three versions of Circumscription which are instances of the more general Circumscription with explicit fixed predicates given herein is in turn equivalent to Circumscription as defined herein. Clark's Completion Procedure for Horn Clauses whose consequences are implied by Parallel Predicate Completion given herein which in turn is an instance of Completion as herein defined. The relationships between QCWA and the Closed World Assumption cwa were discussed in [Brown 2005a]. The relationships between Skeptical Necessary Equivalences and the Skeptical Autoepistemic Kernel and Skeptical Default Logic were discussed in [Brown 2005b].

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers: MIP-9526532, EIA-9818341, and EIA-9972843.

Bibliography

Antoniou, Grigoris 1997, *NonMonotonic Reasoning*, MIT. Press.

Brown, Frank M., 1986, "A Commonsense Theory of Nonmonotonic Reasoning", Proceedings of the 8th International Conference on Automated Deduction, Oxford England, July 1986, Lecture Notes in Computer Science 230, Springer-Verlag.

Brown, Frank .M. 1987, "The Modal Logic Z", In *The Frame Problem in AI"; Proceedings of the 1987 AAAI Workshop*, Morgan kaufmann, Los Altos, CA .

Brown, Frank M. 1989, "The Modal Quantificational Logic Z Applied to the Frame Problem", advanced paper *First International Workshop on Human & Machine Cognition*, May 1989 Pensacola, Florida. Abbreviated version published in *International Journal of Expert Systems Research and Applications, Special Issue: The Frame Problem. Part A.* eds. Kenneth Ford and Patrick Hayes, vol. 3 number 3, pp169-206, JAI Press 1990. Reprinted in *Reasoning Agents in a Dynamic World: The Frame problem*, editors: Kenneth M. Ford, Patrick J. Hayes, JAI Press 1991.

Brown, Frank M., 2005a, "Representing the Closed World Assumption in Modal Logic", Submitted to KDS 2005.

Brown, Frank M., 2005b, "Representing Skeptical Logics in Modal Logic", Submitted to KDS 2005.

Brown, Frank M., 2005c, "Z Priorian Modal Second Order Logic", Submitted to KDS 2005.

Clark, K [1978], Negation as failure. In: Logic and Databases, Gallaire, H. and Minker, J. (eds.), Plenum Press, New York, pp 293-322.

Konolige, Kurt, 1987, "On the Relation Between Default Theories and Autoepistemic Logic", *IJCAI87 Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 1987.

Konolige, Kurt, 1987b, "On the Relation Between Default Theories and Autoepistemic Logic", personally circulated new version of IJCAI87 paper correcting it to account for a counterexample found by Gelfond and Przymusinska.

Konolige, Kurt 1989, "On the Relation between Autoepistemic Logic and Circumscription Preliminary Report", IJCAI89.

Lifschitz, V. 1985, "Computing Circumscription" *IJCAI 9*,pages 121-127.

Lifschitz, V. 1985b, "Closed World Databases and Circumscription", *Artificial Intelligence*, 27.

McCarthy, J. 1980, "Circumscription -- A Form of Nonmonotonic Reasoning", *Artificial Intelligence*, volume 13.

McCarthy, J. 1986, "Applications of Circumscription to Formalizing Common-Sense Reasoning", *Artificial Intelligence*, volume 28.

Moore, R. C. 1985, "Semantical Considerations on Nonmonotonic Logic" *Artificial Intelligence*, 25.

Reiter, R 1978, On closed world databases. In: Logic and Databases, Gallaire, H. and Minker, J. (eds.), Plenum Press, New York, pp 55-76.

Reiter, R. 1980, "A Logic for Default Reasoning" *Artificial Intelligence*, 13.

Author's Information

Frank M. Brown – University of Kansas, Lawrence, Kansas, 66045, e-mail: brown@ku.edu.

Z PRIORIAN MODAL SECOND ORDER LOGIC

Frank M. Brown

***Abstract**: Multiple representations of a concept allow different automatic deduction approaches to be developed for theorems involving. Herein, we discuss three different monotonic logics. These monotonic logics are World Logic, Modal Logic, and Second Order Logic. In all these representations quantifiers obey the normal laws of both classical logc.*

***Keywords**: Z Priorian Modal Second Order Logic, Modal Logic, World Logic, Second Order Logic*

1. Introduction

Leibniz [Leibniz 1686] said that the truths of reason are true in all possible worlds. In Modal Logic this idea is explicated by saying that the logically necessary propositions are those which hold in all worlds. In a Priorian Modal Logic, worlds are represented in the logic itself as those propositions which are possible and which for

every other proposition, entail that other proposition or its negation [Prior & Fine 1977]. Since possibility and entailment are defined in terms of necessity, the properties of a world depend on the properties of the necessity symbol. Unfortunartely, even a modal logic as strong as S5 is not sufficient for axiomatizing worlds since its says nothing about what combinations of predicates are possible. For example In S5 we cannot even prove that a sentence consisting of a single predicate symbol is possible. Herein, we give axiom schemes sufficient to address this problem. We call the resulting Logic the Z Priorian Modal Second Order Logic.

Section 2 presents the syntax and axioms of the Z Priorian Modal Second Order Logic and three distinct sublanguages, namely Z World Logic, Z Modal Logic, and Second Order Logic, used for representing nonmonotonic logics in different ways. Some basic properties of the worlds are presented in Section 3. Finally, in Section 4 we draw some conclusions.

2. Z Priorian Modal Second Order Logic

The syntax and laws of the Z Priorian Modal Second Order Logic is given in sections 2.1 and 2.2 respectively. Its subllogis are defined in section 2.3.

2.1 Syntax of Z Priorian Modal Second Order Logic

The syntax of Z Priorian Modal Second Order Logic is an amalgamation of 3 parts:

The first part is a First Order Logic (i.e. FOL) represented as the six tuple: ($\rightarrow$, #f, $\forall$, *vars*, *predicates*, *functions*) where $\rightarrow$, #f, $\forall$, are logical symbols, *vars* is a set of object variable symbols, *predicates* is a set of predicate symbols each of which has an implicit arity specifying the number of terms associated with it, and *functions* is a set of function symbols each of which has an implicit arity specifying the number of terms associated with it. Roman letters x, y, and z possibly indexed with digits are also variables.

The second part is the extension to Second Order Logic (i.e. SOL) which consists of a set of predicate variables *predvars* and quantification over predicate variables (using $\forall$). Roman letters (other than x, y and z) possibly indexed with digits are used as predicate variables.

The third part is Modal Logic [Lewis 1918] which adds the necessity symbol: [].

Greek letters are used as syntactic metavariables. $\pi,\pi_1...\pi_n,\rho,\rho_1...\rho_n$ range over the predicate symbols, ϕ, $\phi_1...\phi_n$ range over function symbols, $\delta,\delta_1...\delta_n$ range over terms, $\gamma,\gamma 1,...\gamma_n$, range over the object variables ξ, $\xi_1...\xi_n,\zeta,\zeta_1...\zeta_n$ range over a sequence of object variables of an appropriate arity, $f,f_1...f_n$ range over predicate variables, and $\alpha,\alpha_1...\alpha_n,\beta,\beta_1...\beta_n,\chi,\chi_1...\chi_n,\Gamma$, κ range over sentences. Thus, the terms are of the forms: γ and (ϕ $\delta_1...\delta_n$), and the sentences are of the forms: ($\alpha \rightarrow \beta$), #f, ($\forall\gamma$ α), (π $\delta_1...\delta_n$), (f $\delta_1...\delta_n$), ($\forall f$ α), and ([]α). A zero arity predicate π, predicate variable f, or function ϕ is written as a sentence or term without parentheses, i.e., π instead of (π), f instead of (f), and ϕ instead of (ϕ). $\alpha\{\pi/\lambda\xi\beta\}$ is the sentence obtained from α by replacing all unmodalized occurences of π by $\lambda\xi\beta$ followed by lambda coversion. $\alpha\{\pi_i/\lambda\xi\beta_i\}_{i=1,n}$, abreviated as, $\alpha\{\pi_i/\lambda\xi\beta_i\}$ represents simultaneous substitutions. $\wedge_{i=1,n}\beta_i$ represents ($\beta_1\wedge\ldots\wedge\beta_n$). The primitive symbols are listed in Figure 1. The defined symbols are listed in Figure 2 with their intuitive interpretations.

Symbol	Meaning
$\alpha\rightarrow\beta$	if α then β.
#f	falsity
$\forall\gamma\ \alpha$	for all γ, α.
$\forall f\ \alpha$	for all f, α.
[] α	α is logically necessary

Figure 1: Primitive Symbols

Symbol	Definition	**Meaning**
$\neg\alpha$	$\alpha \rightarrow$ #f	not α
#t	$\neg$ #f	truth
$\alpha\vee\beta$	$(\neg\ \alpha)\rightarrow\ \beta$	α or β
$\alpha\wedge\beta$	$\neg(\alpha \rightarrow \neg\beta)$	α and β
$\alpha\leftrightarrow\ \beta$	$(\alpha\rightarrow\ \beta)\ \wedge\ (\beta\rightarrow\ \alpha)$	α if and only if β
$\exists\gamma\ \alpha$	$\neg\forall\gamma\ \neg\alpha$	some γ is α
$\exists f\alpha$	$\neg\forall f\ \neg\alpha$	some f is α
<> α	$\neg[]\neg\alpha$	α is logically possible
[β] α	$[](\beta\rightarrow\alpha)$	β entails α
<β> α	$<>(\beta\wedge\alpha)$	α is possible with β
$\alpha\equiv\ \beta$	$[](\alpha\leftrightarrow\beta)$	α and β are synonymous
$\delta_1=\delta_2$	$(\pi\ \delta_1)\equiv(\pi\ \delta_2)$	δ_1 necessarily equals δ_2
$\delta_1\neq\delta_2$	$\neg(\delta_1=\delta_2)$	δ_1 does not necessarily equal δ_2
(det α)	$\forall f(([\alpha]f)\vee([\alpha]\neg f))$ where f is not in α	α is deterministic
(world α)	$(<>\alpha)\wedge$(det α)	α is a world

Figure 2: Defined Symbols of Z

2.2 Axiomatization of Z Priorian Modal Second Order Logic

The law of the Z Priorian Modal Second Order Logic is an amalgamation of five parts:

(a) The first part, which is given in Figure 3, consists of the laws of a FOL. The laws are: FOLR1, FOLR2, FOLA3, FOLA4, FOLA5, FOLA6, and FOLA7 [Mendelson 1964].

(b) The second part, which is given in Figure 4, consists of the additional laws needed for second order quantification. The three following laws SOLR2, SOLA4 and SOLA5 are the analogues of FOLR2, FOLA4 and FOLA5 for predicate variables.

(c) The Third part, which is given in Figure 5, consists of the laws MR0, MA1, MA2 and MA3 which constitute an S5 modal logic [Hughes &Cresswell 1968] [Carnap 1956]. When added to parts 1 and 2, they form Second Order Modal Quantificational logic similar to a Second Order version of [Bressan 1972].

(d) The fourth part, which is given in Figure 6, consists of the Priorian World extension of S5 Modal Logic. The PRIOR law [Prior and Fine 1977] states that a proposition is logically true if it is entailed by every world.

(e) The fifth part, which is given in Figure 7, consists of some laws axiomatizing what is logically possible. The law MA1a allows one to derive theorems such as $<>\forall x((\pi\ x)\leftrightarrow(x=\phi))$ and therefore extends the work on propositional possibilities in [Hendry and Pokriefka 1985] to possibilities in First Order Logic. The Laws MA1b and MA4 allow one to derive $(\phi_1\ \xi_1)\neq(\phi_2\ \xi_2)$ when ϕ_1 and ϕ_2 are distinct function symbols. MA4 states that at least two things are not necessarily equal just as there are at least two propositions: #t and #f.

FOLR1: from α and $(\alpha\rightarrow\ \beta)$ infer β
FOLR2: from α infer $(\forall\gamma\ \alpha)$
FOLA3: $\alpha \rightarrow (\beta\rightarrow\ \alpha)$
FOLA4: $(\alpha\rightarrow\ (\ \beta\rightarrow\ \rho)) \rightarrow ((\alpha\rightarrow\ \beta)\rightarrow\ (\alpha\rightarrow\ \rho))$
FOLA5: $((\neg\ \alpha)\rightarrow\ (\neg\ \beta))\rightarrow\ (((\neg\ \alpha)\rightarrow\ \beta)\rightarrow\alpha)$
FOLA6: $(\forall\gamma\ \alpha)\rightarrow\ \beta$ where β is the result of substituting an expression (which is free for the free positions of γ in α) for all the free occurrences of γ in α.
FOLA7: $(\forall\gamma\ (\alpha \rightarrow \beta)) \rightarrow\ (\alpha \rightarrow(\forall\gamma\ \beta))$ where γ does not occur in α.

Figure 3: The Laws of FOL

SOLR2: from α infer $(\forall f\ \alpha)$
SOLA6: $(\forall f\ \alpha)\rightarrow \beta$ where β is the result of substituting an expression (which is free for the free positions of f in α) for all the free occurrences of f in α.
SOLA7: $(\forall f(\alpha \rightarrow \beta)) \rightarrow (\alpha \rightarrow (\forall f\ \beta))$ where f does not occur in α.

Figure 4: Additional Laws of SOL

MR0: from α infer $([]\ \alpha)$ provided α was derived only from logical laws
MA1: $([]\alpha) \rightarrow \alpha$
MA2: $([\alpha]\beta) \rightarrow (([]\alpha)\rightarrow ([]\beta))$
MA3: $([]\alpha) \vee ([]\neg[]\alpha)$

Figure 5: Additional Laws of S5 Modal logic

PRIOR: $(\forall w((\text{world } w) \rightarrow ([w]\alpha)))\rightarrow []\alpha$ where w does not occur free in α.

Figure 6: Additional Law of a Piorian World Logic

MA1a: $([]\alpha) \rightarrow \alpha\{\pi/\lambda\zeta\beta\}$ where α may not contain an unmodalized occurence of a higher order variable.
MA1b: $([]\alpha) \rightarrow \alpha\{\phi/\lambda\zeta\delta\}$ where α may not contain an unmodalized occurence of a higher order variable. nor an unmodalized free object variable.
MA4: $\exists x\exists y(x\neq y)$

Figure 7: Additional Laws of a Z Modal Logic

2.3 Axiomatization of three sublanguages of the Z Priorian Modal Second Order Logic

The three sublanguages are described as follows:
(1) Z World Logic – The syntax consists of the symbols of FOL, SOL as it pertains to zero arity predicate variables (i.e. propositional variables) which are worlds, and []. The laws consist of the laws in Figure 3, 5, 6, and 7, and in addition the laws in Figure 4 only for the case of zero arity predicate variables which are worlds.[1]
(2) Z Modal Logic - The syntax consists of the symbols of FOL, SOL as it pertains to zero arity predicate variables (i.e. propositional variables), and []. The laws consist of the laws in Figures 3, 5, and 7, and in addition the laws in Figure 4 only for the case of zero arity predicate variables.
(3) Second Order Logic (i.e. SOL) – which consistets of the symbols of FOL and SOL and the laws in Figures 3 and 4.

3. Worlds and their Possibility

In Section 3.1 we show that worlds function as models do. These results follow in S5 Modal Logic. In section 3.2 we illustrate how worlds are related to Necessity using the PRIOR axiom in Figure 7. In Section 3.3 we generalize MA1a and MA1b to involve multiple substitutions. Finally, in section 3.4 we relate necessity to classical Second Order Logic using axioms MA1a and PRIOR.

3.1 World Entailments

Some basic theorems of S5 modal quantificational logic are given in Figure 8. Since many of the relations are only implications, necessity cannot be pushed through all the logical connectives. This makes automatic deduction somewhat difficult [Brown 1978]. However, if necessity is replaced by entailment by a world proposition then in S5 Modal logic that entailment may be pushed through all logical connectives. Thus world propositions play the role that models play in the metatheoretic semantics. Entailment by a world proposition is thus the analogue of satisfibility by a model. This provides a "semantics" divorced from any specific syntax, and thus allows general semantical laws to be expressed more easily without being cluttered up with syntax. Theorems W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, and W11 in Figure 9 show how such entailments are reduced through the symbols of classical logic. Likewise, theorems W12 and W13 show that such entailments

[1] [Areces&Heguiabehere 2002] describes a World Logic with special symbols which may be defined as follows: $@w\alpha$ =df $[w]\alpha$ and $\downarrow w\alpha$ =df $\exists w(w\wedge(\text{world } w)\wedge([w]\alpha))$

disappear in front of modal symbols as one would expect in the framework of a modal logic at least as strong as S5. These entailments were used in [Brown 1979].

B1: $([]($\alpha\wedge$q)) $\leftrightarrow$ (([]α)$\wedge$([]q))
B2 $([]($\alpha\vee$q)) $\leftarrow$ (([]α)$\vee$([]q))
B3: $([]($\alpha\rightarrow$q)) $\rightarrow$ (([]α)$\rightarrow$([]q))
B4: $([]($\alpha\leftrightarrow$q)) $\rightarrow$ (([]α)$\leftrightarrow$([]q))
B5: ([]#t) $\leftrightarrow$ #t
B6: ([]#f) $\leftrightarrow$ #f
B7: $([]\neg\alpha) \rightarrow (\neg([]\alpha))$
B8: $([](\forall\gamma\alpha))\leftrightarrow (\forall\gamma([]\alpha))$
B9: $([](\exists\gamma\alpha))\leftarrow (\exists\gamma([]\alpha))$
B10: $([](\forall f\alpha))\leftrightarrow (\forall f([]\alpha))$
B11: $([](\exists f\alpha))\leftarrow (\exists f([]\alpha))$
B12: $([]([]\alpha)) \leftrightarrow ([]\alpha)$
B13: $([](<>\alpha)) \leftrightarrow (<>\alpha)$

Figure 8: Basic Laws of S5

W1: $([w](\alpha\wedge\beta))\leftrightarrow (([w]\alpha)\wedge([w]\beta))$
W2: (det w) $\rightarrow$ $(([w](\alpha\vee\beta)) \leftrightarrow (([w]\alpha)\vee([w]\beta)))$
W3: (det w) $\rightarrow$ $(([w](\alpha\rightarrow\beta)) \leftrightarrow (([w]\alpha)\rightarrow([w]\beta)))$
W4: (det w) $\rightarrow$ $(([w](\alpha\leftrightarrow\beta)) \leftrightarrow (([w]\alpha)\leftrightarrow([w]\beta)))$
W5: $(([w]\#t) \leftrightarrow \#t)$
W6: (<> w) $\rightarrow$ $(([w]\#f) \leftrightarrow \#f)$
W7: (world w)$\rightarrow$ $(([w](\neg\alpha)) \leftrightarrow (\neg([w]\alpha)))$
W8: $(([w](\forall\gamma\alpha))\leftrightarrow (\forall\gamma([w]\alpha)))$
W9: (det w) $\rightarrow$ $(([w](\exists\gamma\alpha))\leftrightarrow (\exists\gamma([w]\alpha)))$
W10: $(([w](\forall f\alpha))\leftrightarrow (\forall f([w]\alpha)))$
W11: (det w) $\rightarrow$ $(([w](\exists f\alpha))\leftrightarrow (\exists f([w]\alpha)))$
W12: (<> w) $\rightarrow$ $(([w]([]\alpha)) \leftrightarrow ([]\alpha))$
W13: (<> w) $\rightarrow$ $(([w](<>\alpha)) \leftrightarrow (<>\alpha))$

Figure 9: Reduction of World Entailments

3.2 Consequences of Prior's axiom

PRIOR says that a proposition is logically necessary if it is entailed by every world proposition. This law was implied in [Leibniz 1686] and has been used by a number of authors including [Prior&Fine 1977, Fine 1970]. It is equivalent to saying that every possible proposition has a world that entails it:

$$(<>\alpha) \rightarrow \exists w((\text{world } w)\rightarrow([w]\alpha))$$

The PRIOR axiom allows any necessity symbol to be replaced by an entailment by all world propositions in accordance with the laws in Figure 10 below. This is important because the entailments can then be reduced to lowest scope using the laws in Figure 9.

N1: $([]\alpha) \leftrightarrow (\forall w((\text{world } w)\rightarrow([w]\alpha)))$
P1: $(<>\alpha) \leftrightarrow (\exists w((\text{world } w)\rightarrow([w]\alpha)))$
D1: $(\exists w(w\wedge(\text{world } w)\wedge[w]\alpha))\leftrightarrow\alpha$
C1: $(\forall w((w\wedge(\text{world } w))\rightarrow[w]\alpha))\leftrightarrow\alpha$
DC:$(\exists w(w\wedge(\text{world } w)\wedge(fw))\leftrightarrow(\forall w((w\wedge(\text{world } w))\rightarrow(fw)))$ (w may not occur free in α in laws: N1, P1, D1, C1)

Figure 10: Necessity and Possibility as Worlds

3.3 Parallel Versions of MA1a and MA1b

First we note that MA!a and MA1b may be modified to include necessary conclusions:

TMA1a[]: $([]\alpha) \rightarrow []([\alpha\{\pi/\lambda\xi\beta\})$

where α does not contain any unmodalized occurence of a higher order variable

proof: From Axiom MA1a: $([]\alpha) \rightarrow \alpha\{\pi/\lambda\xi\beta\}$ the inference rule MR0 gives: $[](([]\alpha) \rightarrow (\alpha\{\pi/\lambda\xi\beta\}))$. By MA2 this implies: $([]([]\alpha))\rightarrow ([](\alpha\{\pi/\lambda\xi\beta\}))$ which by the laws of S4 Modal logic gives: $([]\alpha)\rightarrow ([](\alpha\{\pi/\lambda\xi\beta\}))$. QED

TMA1b[]: $([]\alpha) \rightarrow []([\alpha\{\phi/\lambda\xi\delta\})$

where α does not contain any unmodalized occurence of a higher order variable

proof: From Axiom MA1b: $([]\alpha) \rightarrow \alpha\{\pi/\lambda\xi\delta\}$ the inference rule MR0 gives: $[](([]\alpha) \rightarrow (\alpha\{\pi/\lambda\xi\delta\}))$. By MA2 this implies: $([]([]\alpha))\rightarrow ([](\alpha\{\pi/\lambda\xi\delta\}))$ which by the laws of S4 Modal logic gives: $([]\alpha)\rightarrow ([](\alpha\{\pi/\lambda\xi\delta\}))$. QED

MA1a and MA1b may be generalized to deal with multiple predicates and functions in parallel. Generalizing MA1a is difficult because unmodalized higher order variables may occur in any of the β_i:

TMA1a*: $([]\alpha) \rightarrow \alpha\{\pi_i/\lambda\xi_i\beta_i\}$ where α does not contain any unmodalized occurence of a higher order variable.

proof: By MA1 it suffices to prove: $([]\alpha)\rightarrow([]\alpha\{\pi_i/\lambda\xi_i\beta_i\})$. By N1 this is:

$([]\alpha)\rightarrow\forall w((\text{world } w)\rightarrow ([w]\alpha\{\pi_i/\lambda\xi_i\beta_i\}))$ which is: $(([]\alpha)\wedge(\text{world } w))\rightarrow([w]\alpha\{\pi_i/\lambda\xi_i\beta_i\}))$. Letting the ρ_j be any other unmodalized predicates in α other than the π_i predicates gives: $(([]\alpha)\wedge(\text{world } w))\rightarrow([w]\alpha\{\pi_i/\lambda\xi_i\beta_i\}\{\rho_j/\lambda\zeta_i\rho_j\}))$. Since α does not contain any unmodalized occurence of a higher order variable, the laws in Figure 9 allow the [w] to be pushed down to the predicates giving: $(([]\alpha)\wedge(\text{world } w))\rightarrow\alpha\{\pi_i/\lambda\xi_i([w]\beta_i)\}\{\rho_j/\lambda\zeta_j([w]\rho_j)\}))$. Starting from $([]\alpha)$, we now use TMA1a[] (many times) to replace, in any order, each predicate π_i or ρ_j with $\lambda\xi_i([w]\beta_i)$ or $\lambda\zeta_j([w]\rho_j)$, respectively, giving: $(([]\alpha\{\pi_i/\lambda\xi_i([w]\beta_i)\}\{\rho_j/\lambda\zeta_j([w]\rho_j)\})\wedge(\text{world } w)) \rightarrow\alpha\{\pi_i/\lambda\xi_i([w]\beta_i)\}\{\rho_j/\lambda\zeta_j([w]\rho_j)\})$ which holds by axiom MA1. QED

TMA1b*: $([]\alpha) \rightarrow \alpha\{\phi_i/\lambda\xi_i\delta_i\}$

proof: Since there are no unmodalized higher order variables in any δ_i we use TMA1b[] to replace each function symbol ϕ_i with δ_i giving $[]\alpha\{\phi_i/\lambda\xi_i\delta_i\}$. The conclusion then follows using MA1. QED

3.4 Relation of [] to SOL

The PRIOR and MA1a laws provide a direct method for translating FOL Modal sentences into sentences of (non modal) Second Order Logic[1]:

T1: If Γ does not contain any unmodalized occurences of a higher order variable and if $\pi_1\ldots\pi_n$ are all the unmodalized predicate symbols in Γ then:

$([]\Gamma) \leftrightarrow \forall P_1\ldots\forall P_n (\Gamma\{\pi_i/P_i\})$

proof: The proof divides into two cases:

(a) $([]\Gamma) \rightarrow \forall P_1\ldots\forall P_n \Gamma\{\pi_i/P_i\}$ which gives: $([]\Gamma) \rightarrow \Gamma\{\pi_i/P_i\}$

By TMA1a* letting π_i be P_i gives: $\Gamma\{\pi_i/P_i\} \rightarrow \Gamma\{\pi_i/P_i\}$ which is a tautology.

(b) $(\forall P_1\ldots\forall P_n \Gamma\{\pi_i/P_i\}) \rightarrow []\Gamma$. By PRIOR we get: $(\forall P_1\ldots\forall P_n \Gamma\{\pi_i/P_i\}) \rightarrow \forall w((\text{world } w) \rightarrow ([w]\Gamma))$

which is equivalent to: $(\forall P_1\ldots\forall P_n \Gamma\{\pi_i/P_i\}) \wedge (\text{world } w) \rightarrow ([w]\Gamma)$. Letting P_i be $\lambda\xi_i[w](\pi_i\ \xi_i)$ gives:

---

[1] This theorem may be constrasted with the Logic described in [Montague 1960] where $([]\Gamma)$ would be $\forall P_1\ldots\forall P_n,\forall F_1\ldots\forall F_m(\Gamma\{\pi_i/P_i\}\{\phi_j/F_j\})$ where Γ does not contain any unmodalized occurences of a higher order variable, $\pi_1\ldots\pi_n$ are all the unmodalized predicate symbols in Γ and $\phi_1\ldots\phi_m$ are all the unmodalized function symbols in Γ. Therein, each F_j is a functional variable of the same arity as ϕ_j. Montague points out that his system does not obey the normal laws of logic across modal scopes (but thinks that that is a virtue). For example, unfolding = in $\forall x\exists y(x=y)$ gives, $\forall x\exists y((\pi\ x)\equiv(\pi\ y))$, which in Montagues system is equivalent to $\forall x\exists y\forall P((P\ x)\leftrightarrow(P\ y))$ which is #t. Likewise, unfolding = in $\exists y(\phi=y)$ gives: $\exists y((\pi\ \phi)\equiv(\pi\ y))$, which in Montague's system is $\exists y\forall P\forall f((P\ f)\leftrightarrow(P\ y))$ which is false if there are at least two distinct things. Thus $\forall x\exists y(x=y)$ cannot imply $\exists y(\phi=y)$ in Montague's system thereby contradicting the law of universal instantiation (over a modal scope).

$(\Gamma\{\pi_i/\ \lambda\xi_i[w](\pi_i\ \ \xi_i)\})\ \wedge$ (world w) $\rightarrow ([w]\Gamma)$. Since Γ does not contain any unmodalized higher order variables and every unmodalized predicate is replaced by a [w] entailment, the laws in Figure 9 allow the [w] entailments to be pulled out giving $([w]\Gamma\{\pi_i/\ \lambda\xi_i(\pi_i\ \ \xi_i)\})\ \wedge$ (world w) $\rightarrow ([w]\Gamma)$ or rather: $([w]\Gamma)\ \wedge$ (world w) $\rightarrow ([w]\Gamma$ which is a tautology. QED.

An instance of this theorem is Leibniz's Law about the identity of indescernables: $(x=y)\leftrightarrow(\forall P((P\ x)\leftrightarrow(P\ y)))$.

4. Conclusion

The existence of the three different representations (Z World Logic, Z Modal Logic and SOL) for representing concepts will allow different automatic theorem proving approaches to be developed for those concepts. The World representation serves as semantics analogous to Model Theory and the Modal representation serves as the analogue of proof Theory, both with the advantage that quantified variables are allowed to cross the scope of the analogues to both the model theoretic satisfaction and proof theoretic concept of being derivable. An application of these three logics to nonmonotonic reasoning is given in [Brown 2005].

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers: MIP-9526532, EIA-9818341, and EIA-9972843.

Bibliography

Areces, Carlos and Heguiabehere, Juan HyLoRes 1.0: Direct Resolution for Hybrid logics, CADE18, LNAI 2392, 2002.

Bressan, Aldo 1972. *A General Interpreted Modal Calculus,* Yale University Press,.

Brown, Frank M., 1978. "A Sequent Calculus for Modal Quantificational Logic", *3rd AISB/GI Conference Proceedings,* Hamburg, July 1978.

Brown, Frank M. 1979. "A Theorem Prover for Meta theory", *Proceedings Fourth Workshop on Automated Deduction*, Austin, Texas.

Brown, Frank M., 1986 "A Commonsense Theory of Nonmonotonic Reasoning", Proceedindings of the 8th International Conference on Automated Deduction, Oxford England, July 1986, Lecture Notes in Computer Science 230, Springer-Verlag.

Brown, Frank .M. 1987. "The Modal Logic Z", In *The Frame Problem in AI"; Proceedings of the 1987 AAAI Workshop*, Morgan kaufmann, Los Altos, CA .

Brown, Frank M. 1989. "The Modal Quantificational Logic Z Applied to the Frame Problem", advanced paper *First International Workshop on Human & Machine Cognition*, May 1989 Pensacola, Florida. Abreviated version published in *International Journal of Expert Systems Research and Applicatiions, Special Issue: The Frame Problem. Part A.* eds. keneth Ford and Pattrick Hayes, vol. 3 number 3, pp169-206 JAI Press 1990. Reprinted in *Reasoning Agents in a Dynamic World: The Frame problem*, editors: Kenneth M. Ford, Patrick J. Hayes, JAI Press 1991.

Brown, Frank M. 2005, "Nonmonotonic Systems based on Smallest and Minimal Worlds represented in World Logic, Modal Logic, and Second Order Logic", Summitted to this conference.

Carnap, Rudolf 1956., *Meaning and Necessity: A Study in the Semantics of Modal Logic*, The University of Chicago Press,.

Fine, k. 1970. "Propositional Quantifiers in Modal Logic" *Theoria* 36, p336--346.

Hendry, Herbert E. and Pokriefka, M. L. 1985. "Carnapian Extensions of S5", *Journal of Phil. Logic 14*,.

Hughes, G. E. and Creswell, M. J., 1968. *An Introduction to Modal Logic*, Methuen & Co. Ltd., London.

Leibniz, G. 1686, "Necessary and Contingent Truths" 1686, *Leibniz Philosophical Writings*, Dent&Sons. 1973.

Lewis, C. I. 1936. Strict Implication, *Journal of Symbolic Logic*, vol I..

Mendelson, E. 1964. *Introduction to Mathematical Logic*, Van Norstrand, Reinhold Co., New York.

Montague, Richard 1960, "Logical Necessity, Physical Necessity, Ethics, and Quantifiers", Inquiry 4: p259-269.

Prior, A. N. & Fine, k. 1977. *Worlds, Times, and Selves*, Gerald Duckworth.

Author's Information

Frank M. Brown – University of Kansas, Lawrence, Kansas, 66045, e-mail: brown@ku.edu.

Section 6. Neural and Growing Networks

6.1. Neural Network Applications

PARALLEL MARKOVIAN APPROACH TO THE PROBLEM OF CLOUD MASK EXTRACTION

Natalia Kussul, Andriy Shelestov, Nguyen Thanh Phuong, Michael Korbakov, Alexey Kravchenko

***Abstract**: An application of Markovian approach to cloud mask extraction is presented. Also parallel algorithm of Markovian segmentation is considered.*

***Keywords**: Meteosat, cloud mask, Markov Random Fields, parallel programming, MPI.*

Introduction

One of the most useful satellite data products is cloud mask. It can be used in a standalone way in applications such as air flights and satellite photography planning. Also it can be used as an input data for various satellite data processing algorithms like Normalized Difference Vegetation Index (NDVI), Sea Surface Temperature (SST) and operational wind vectors maps extraction, or even more complex applications such as numerical weather models.

A common approach for cloud mask extracting is using of multi- and hyperspectral satellites providing data in many spectral bands. Basing on information about radiance intensities a conclusion about cloudiness can be made on per pixel basis. For instance, this approach is widely used for processing of multispectral AVHRR and MODIS data. But temporal resolution of satellites with such equipment on-board is usually quite low thus making impossible on-line monitoring of particular region of Earth disk. The obvious solution of this problem is the use of geostationary satellites.

For European region, the only geostationary satellite that can be used for solving cloud mask extraction problem is Meteosat. Meteosat is operated by EUMETSAT international organization and provides data for solving practical meteorological problems. This satellite's onboard equipment makes one image of earth disk in 30 minutes in three spectral bands – visible, infrared, water vapour.

The temporal characteristics of Meteosat data make their use in solving of problem of cloud mask extracting quite actual. But three spectral bands of Meteosat do not provide enough information for multispectral cloud recognition algorithms operating on per pixel basis. This causes the need for algorithms, which involves temporal and spatial dependencies in data processing. One of such algorithms is a Markov Random Field segmentation, which allows determining pixel's class with regard to its neighborhoods. Markovian approach allows taking into account different possible distributions of intensities per class and do not introduce global parameters such as thresholds, which is often used in multispectral data processing.

One of the main disadvantages of Markovian approach is high computational complexity. Therefore processing of large images requires great amount of time. So, for effective solving of Markovian segmentation problem a parallel computing approach is needed.

Data Preprocessing

Meteosat images come with a lot of noise of two sorts. The first one is the so-called “salt and pepper” noise consisting of noisy pixels uniformly distributed over image. The second one is the impulse burst noise, which distorts images with horizontal streaks of few pixel heights filled with white noise.

In the Space Research Institute NASU-NSAU algorithm for detecting and removing such noise was developed [Phuong, 2004]. On the first step of this algorithm noise streaks is detected and removed by cubic spline interpolation method. During the second step the “salt and pepper” noise is detected by modified median filter and removed by using bit planes approach. The algorithm based on this approach separates an image in 256 planes with binary values. After that, each of these planes is processed separately in order to remove noise.

Cloud Mask Extraction

Following Markovian approach the image is represented as a $n \times m$ matrix of sites S. The neighborhood of site s_{ij} is any subset $\partial_{ij} \subset S$, such that $s_{ij} \notin \partial_{ij}$. With each site s_{ij} two varieties are associated – an intensity X_{ij} (as usual it takes integer value in interval [0; 255]) and a hidden label Y_{ij}. The specific values the varieties take are denoted x_{ij} and y_{ij} respectively. So two sets of varieties defined for image S: $X = \{X_{11},...,X_{nm}\}$ and $Y = \{Y_{11},...,Y_{nm}\}$.

Markov Random Fields (MRFs) are widely used for image segmentation [Li, 1995]. With the Hammersley-Clifford theorem the equivalence of MRF and statistical physics Gibbs models was proved [Li, 1995]. This theorem gives us the equation for probability of specific segmentation $P(Y)$

$$P(Y) = \frac{1}{z} e^{\beta V(Y)} = \frac{1}{z} e^{\beta \sum_{i,j} V_{ij}\left(\partial_{ij} \cap \{y_{ij}\}\right)}$$

In this equation, z is normalizing constant necessary for holding the condition $\sum_{Y} P(Y) = 1$. β denotes the image correlation parameter. V is the so-called potential function. Its structure is highly coupled with optimal segmentation of MRF. Defining a particular potential function it is possible to model physics features of segmentation. The right part of equation shows that potential function V can be represented as a sum of potentials defined at each site: $V = \sum_{i,j} V_{ij}$.

For the cloud mask extraction problem the following Markovian model was used: the observed intensity X_{ij} depends only on local label Y_{ij} and a conditional distribution of variety X_{ij} is Gaussian. The Bayes’ theorem about a priory and a posteriori probabilities’ relation yields a complete model of intensities and labels coupling [Shiryaev, 1989]:

$$P(Y \mid X) \propto P(X \mid Y)P(Y) = \frac{1}{z} \exp\left\{\sum_{i,j} \beta_{ij} n_{ij}\left(y_{ij}\right)\right\} \times \prod_{i,j} \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{1}{2\sigma_{ij}^2}\left(x_{ij} - \mu_{ij}\right)^2\right\}$$

Here σ_{ij} and μ_{ij} are a standard deviation and a mean of random variable X_{ij}, n_{ij} is the number of pixels in neighborhood ∂_{ij} with the label equal to Y_{ij}.

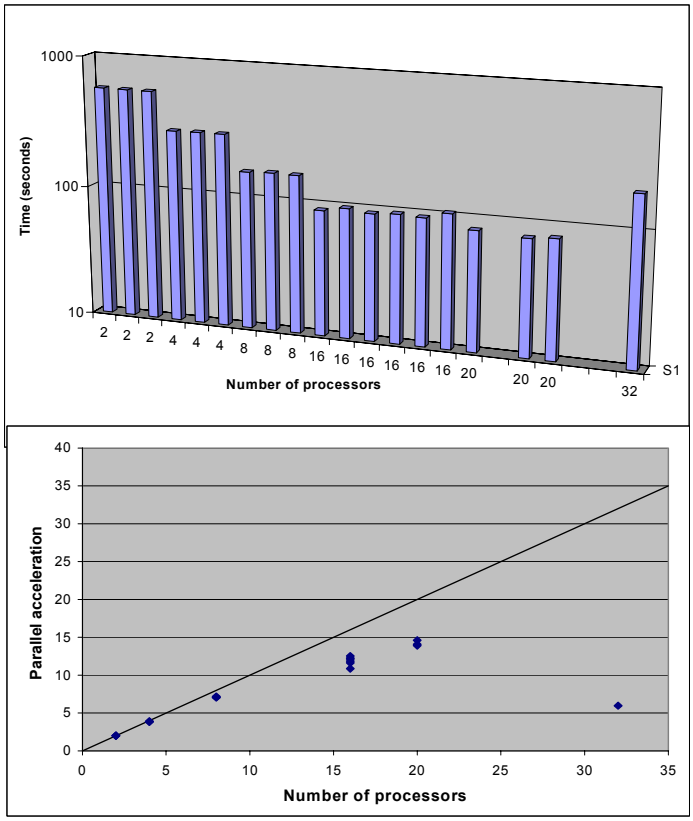
The goal of segmentation is to maximize $P(Y \mid X)$ under particular intensities X. This corresponds to obtaining maximum for a posteriori label’s estimate: $Y : Y^* = \arg\max_y \{P(Y \mid X)\}$

Parallel Execution Results

High computational complexity of Markovian segmentation algorithm together with large sizes of satellite images determines the need for parallel realization of cloud mask extraction process.

Meteosat image filtering and Markovian segmentation algorithms were implemented using MPI parallel programming interface [MPI, 1997]. Due to locality of dependencies in Markovian image model, it is possible to divide image into almost independent rectangular parts. Then each of these parts is processed by different computational node. Synchronization of several global per-class parameters and image part's borders is performed by means of MPI's group communication functions.

The program was run on the cluster of Institute of Cybernetics NASU consisting of 32 Intel Xeon processors. It has demonstrated good level of parallel acceleration giving almost proportional speed boost with increase of number of computational nodes used.



Conclusions and Further Works

Markovian approach has showed its effectiveness in task of cloud mask extraction from Meteosat satellite data. Also parallel Markovian segmentation algorithm performed very well exploiting locality of Markovian image model.

Further works includes implementation of this algorithm in GRID environment, which will connect computational cluster with satellite data archives and method to process them. This GRID design will allow separating algorithms from environment and using them in standalone applications locally. Also, we are planning to port this algorithm implementation to IA-64 processor architecture to utilize upcoming computational cluster with 64 Itanium-2 processors.

Bibliography

[Phuong, 2004] N. T. Phuong. A Simple and Effective Method for Detection and Deletion of Impulsive Noised Streaks on Images. Problems of Controls and Informatics, 2004. (in Russian)

[Li, 1995] S.Z. Li. Markov Random Field Modelling in Computer Vision. Springer-Verlag, 1995.

[Shiryaev, 1989] A. N. Shiryaev. Probability. Nauka, 1989. (in Russian)

[MPI, 1997] MPI-2: Extensions to the Message-Passing Interface. University of Tennessee, 1997

Authors' Information

Natalia Kussul – e-mail: inform@ikd.kiev.ua
Andriy Shelestov – e-mail: shelest@dialektika.com
Nguyen Thanh Phuong – e-mail: ntphuong72@yahoo.com
Michael Korbakov – e-mail: rmihael@ukr.net
Alexey Kravchenko – e-mail: akm3000@mail.ru
Space Research Institute NASU-NSAU, Glushkova St. – 41, Kyiv

ИДЕНТИФИКАЦИЯ НЕЙРОСЕТЕВОЙ МОДЕЛИ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ КОМПЬЮТЕРНЫХ СИСТЕМ[1]

Н. Куссуль, С. Скакун

***Аннотация**: В работе проводилось математическое моделирование поведения пользователей компьютерных систем. Изучалась динамика работы пользователя во время сеанса. Также осуществлялось статистическое моделирование данных, характеризующих его работу за сеанс в целом.*

***Ключевые слова**: модель поведения пользователей, нейронные сети, компьютерные системы.*

1. Введение

Масштабное использование компьютерных технологий практически во всех сферах человеческой деятельности приковывает все большее внимание к самому пользователю. Знание того, какие действия он выполняет (или должен выполнять), может применяться в разных областях, например, в системах безопасности [1], для создания персонализированного окружения для пользователей [2] и так далее.

В работе [3] была предложена комплексная модель пользователя, состоящая из интерактивной и сеансовой частей, которые учитывают, соответственно, динамические и статистические свойства поведения пользователя. В обеих моделях для выявления отклонений от обычного или ожидаемого поведения пользователей используются нейронные сети. Так, интерактивная модель основана на прогнозировании команд пользователя на основе предыдущих (В данной работе под прогнозированием команд будем понимать прогнозирование процессов, порожденных запуском файлов ОС Windows.). Поскольку выбор архитектуры нейронной сети представляет собой нетривиальную задачу, важно знать, на сколько его текущее поведение зависит от предыстории. В случае сеансовой модели возникает проблема с размером выборки, которая используется для обучения нейронной сети. Дело в том, что при небольшом размере обучающего множества нейронная сеть имеет тенденцию к локальному запоминанию образов, что нежелательно. В сеансовой модели на вход нейронной сети подаются данные, собранные за сеанс в целом. Соответственно, размер обучающего множества напрямую определяется количеством сеансов, во время которых проводилось наблюдение за деятельностью пользователя. Однако даже за продолжительный отрезок времени этих данных будет недостаточно для качественного обучения нейронной сети. Поэтому в сеансовой модели для качественного обучения нейронной сети очень важно обеспечить более представительную выборку данных.

Вопросам, связанными с оптимизацией архитектуры нейронной сети, в частности размерностью входного слоя, и моделированием данных, и посвящена данная статья.

2. Комплексная нейросетевая модель пользователя компьютерных систем

Комплексная модель пользователя, предложенная в работе [3], учитывает как динамические (интерактивная часть), так и статистические (сеансовая часть) свойства поведения пользователей. В основу разработанной модели положена нейронная сеть прямого распространения, которая состоит из входного, выходного и одного или нескольких скрытых слоев нейронов [4]. Выход нейрона в слое n определяется следующим отношением:

$$y_j^n = f(s_j^n), \tag{1}$$

где n — номер слоя ($n = \overline{1, p}$); p — количество слоев в нейронной сети; j — индекс нейрона в слое $(j = \overline{1, N_n})$; N_n — число нейронов в слое; f — активационная функция слоя (в нашем случае

[1] Работа выполнена при содействии гранта Президента Украины для поддержки научных исследований молодых ученых № Ф8/323, "Прототип интеллектуальной мультиагентной системы компьютерной безопасности".

для скрытых слоев используется сигмоидная активационная функция $f(x)=\frac{1}{1+e^{-\alpha x}}$, а для выходного слоя — линейная $f(x)=\alpha x$); $y_j{}^n$ — выход j-го нейрона слоя; $s_j{}^n$ — постсинаптический потенциал j-го нейрона слоя, который вычисляется согласно следующей формуле:

$$s_j{}^n=\sum_{k=1}^{N_{n-1}} W_{jk}{}^n y_k{}^{n-1}+b_j{}^n, \quad S^n=W^n\cdot\tilde{y}^{n-1}, \tag{2}$$

где $W_{jk}{}^n$ — весовой коэффициент связи k-го нейрона слоя $n-1$ с j-м нейроном слоя n; $y_k{}^{n-1}$ — выход k-го нейрона слоя $n-1$; $\tilde{y}^{n-1}$ — расширенный вектор с учетом bias-нейрона; $b_j{}^n$ — порог (bias-нейрон) j-го нейрона слоя n.

Интерактивная модель используется для выявления аномальной деятельности во время работы пользователя. Для каждого пользователя компьютерной системы строится и обучается нейронная сеть таким образом, чтобы прогнозировать следующую команду на основе предыдущих. Пусть s_t (где $t\in\{1, 2, ...\}$ номер сеанса) — некоторый сеанс пользователя, для которого имеется следующая последовательность выполненных им команд:

$$\mathbf{c}^{s_t}=\left(c_1^{s_t},\ c_2^{s_t},\ \ldots,\ c_{N_{s_t}}^{s_t}\right),$$

где N_{s_t} — количество команд, введенных за сеанс s_t. Прежде чем подать последовательность из m команд на вход нейронной сети применялось бинарное кодирование. При этом для каждой команды было использовано q бит. Тогда результат работы нейронной сети при выполнении i-1 команды определяется зависимостью:

$$\tilde{c}_i^{s_t}=F(\mathbf{x}_i),\ \mathbf{x}_i=\left(\tilde{\mathbf{c}}_{i-1}^{s_t},\tilde{\mathbf{c}}_{i-2}^{s_t},\ldots,\tilde{\mathbf{c}}_{i-m}^{s_t}\right), \tag{3}$$

где F — нелинейное преобразование, осуществляемое нейронной сетью согласно формулам (1) и (2); $\tilde{\mathbf{c}}_{i-k}^{s_t}\left(k=\overline{1,m}\right)$ — бинарный вектор для команды $c_{i-k}^{s_t}$; $\mathbf{x}_i$, $\tilde{c}_i^{s_t}$ — вход и выход сети, соответственно (в данном случае размерность вектора $\mathbf{x}_i$ составляет q*m, а $\tilde{c}_i^{s_t}$ — 1); $c_i^{s_t}$ — i-тая команда сеанса s_t; m — количество команд, на основе которых происходит прогнозирование следующей (окно прогнозирования).

На основе количества команд, которые были правильно спрогнозированы нейронной сетью, делается вывод о том, соответствует ли текущее поведение пользователя ранее построенной модели.

Пусть

$$\Theta(i)=\frac{1}{(i-m)}\sum_{j=m+1}^{i}\chi(\tilde{c}_j^{s_t},c_j^{s_t}),\quad \chi(\tilde{c}_j^{s_t},c_j^{s_t})=\begin{cases}1, & \textit{если } \tilde{c}_j^{s_t}=c_j^{s_t}\\ 0, & \textit{в противном случае}\end{cases}. \tag{4}$$

Величина $\Theta(i)$ определяет относительное число верно спрогнозированных команд. Если $\Theta(i)<\Theta'$, т.е. значение $\Theta(i)$ меньше некоторого порога, то поведение пользователя следует считать аномальным, в противном случае — нормальным.

При этом необходимо учитывать, что пользователям свойственно изменять поведение с течением времени, поэтому с целью обеспечения адаптации к их поведению нейронную сеть следует периодически дообучать. (Критерий изменения поведения пользователей будет рассмотрен далее.)

Сеансовая модель предназначена для выявления нехарактерной деятельности пользователя за сеанс в целом и для этого использует статистический набор данных. Эта информация, в свою очередь, используется для построения и обучения нейронной сети, которая определяет, на сколько активность пользователя соответствует ранее построенной модели. При этом ожидаемый выход нейронной сети может принимать два значения: 1 — для нормального поведения пользователя и 0 — для аномального, т.е. нейронная сеть работает в качестве классификатора.

Выход же нейронной сети по завершении сеанса s_t определяется следующим соотношением:

$$\Delta_{s_t}=F(x_{s_t}),\quad \mathbf{x}_{s_t}=(n_{s_t},o_{s_t},h_{s_t},d_{s_t},s_{s_t}), \tag{5}$$

где F — нелинейное преобразование, осуществляемое нейронной сетью согласно формулам (1) и (2); $\mathbf{x}_{s_t}$, Δ_{s_t} — вход и выход сети, соответственно (в данном случае размерность вектора $\mathbf{x}_{s_t}$ составляет 5, а Δ_{s_t} — 1); n_{s_t} — количество команд за сеанс, $o_{s_t} = \Theta(N_{s_t})$ — результаты интерактивной модели (процентное соотношение правильно спрогнозированных команд за весь сеанс), h_{s_t} — номер компьютера, d_{s_t} — продолжительность сеанса, s_{s_t} — время начала сеанса.

Следует заметить, что выход нейронной сети Δ_{s_t} не обязательно будет равным 0 или 1, а будет принадлежать отрезку [0; 1] и определять вероятность нормального (соответствующего модели) поведения пользователя.

3. Описание данных

Для моделирования поведения пользователей компьютерных систем использовались реальные данные, которые были собраны в локальной сети Института космических исследований НАНУ-НКАУ в период за два-три месяца. В данной сети рабочие станции функционировали под управлением операционных систем (ОС) Windows 98, XP, 2000. Поскольку эти ОС необходимой информацией об активности пользователя обеспечивали не в полной мере, было разработано специальное программное приложение. Для каждого сеанса пользователя создавался отдельный аудит-файл (название которого однозначно определяло имя учетной записи пользователя и дату его работы), в который сохранялась информация в следующем формате:

время запуска команды|идентификатор команды|название команды|флаг начала или завершения

Необходимо отметить, что идентификатор команды присваивается ОС и является уникальным (для команд с одним и тем же именем он различен, причем от сеанса к сеансу он также меняется). Поэтому при кодировании команд важно обеспечить, чтобы одинаковым командам соответствовали одни и те же значения. С этой целью для интерактивной части комплексной модели на основе собранной информации для каждого пользователя был построен алфавит команд *A* (т.е. набор команд, которые вводились пользователем на протяжении указанного периода времени). Далее каждой команде был присвоен соответствующий десятичный номер, который впоследствии использовался при преобразовании аудит-файлов в последовательности команд ($A = \{1, 2, \ldots, N\}$). В результате для каждого пользователя был представлен следующий набор данных:

$$\left\{c_i^{s_t}\right\}_{t=1 i=1}^{T \quad N_{s_t}}, \tag{6}$$

где $c_i^{s_t} \in A$ — десятичный номер *i*-ой введенной команды сеанса s_t; T — количество сеансов; N_{s_t} — общее количество выполненных команд за сеанс s_t. На рис. 1 приведен пример последовательности команд, вводимых пользователем за один сеанс.

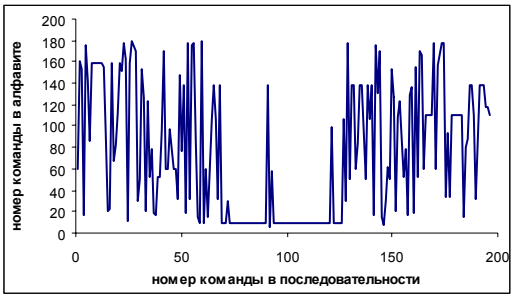


Рис. 1. Пример последовательности команд, вводимых пользователем за один сеанс

В свою очередь, для сеансовой части комплексной модели был получен следующий набор данных, используя информацию из аудит-файлов: $\left\{n_{s_t}, h_{s_t}, d_{s_t}, s_{s_t}\right\}_{t=1}^{T}$, где t — условный номер сеанса; T — количество сеансов; параметры $n_{s_t}, h_{s_t}, d_{s_t}, s_{s_t}$ определены в соотношении (5).

4. Изучение динамики поведения пользователя во время сеанса

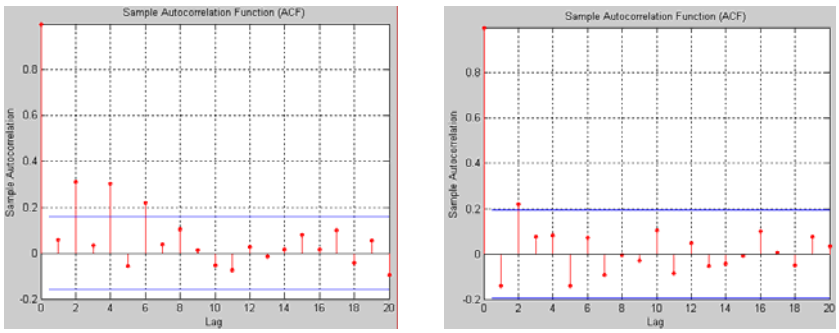
При исследовании динамики поведения пользователя по последовательностям вводимых им команд, нас, в первую очередь, будет интересовать решение следующих задач: определение количества команд, которые следует использовать для прогноза следующей, а также построение критерия изменения поведения пользователя с целью уменьшения ложных тревог и переобучения нейронных сетей.

4.1. Построение автокорреляционных функций

Поскольку интерактивная модель основана на прогнозировании нейронной сетью команд пользователя, важно знать, на сколько его поведение в данный момент времени зависит от предыдущего. Для этого для каждого сеанса пользователя s_t были построены автокорреляционные кривые, определяемые соотношениями следующего вида:

$$\rho_{s_t}(n) = \frac{1}{N_{s_t} - n} \sum_{i=1}^{N_{s_t}-n} \left(c_i^{s_t} - \mu_{s_t}\right)\left(c_{i+n}^{s_t} - \mu_{s_t}\right),\ \mu_{s_t} = \frac{1}{N_{s_t}} \sum_{i=1}^{N_{s_t}} c_i^{s_t}\ .$$

где $\rho_{s_t}(n)$ — коэффициент корреляции для последовательности команд, введенных за сеанс s_t; n — значение лага (временное смещение между элементами последовательности). На рис. 2 приведены примеры автокорреляционных кривых для разных пользователей.



пользователь № 1 пользователь № 2

Рис. 2. Примеры автокорреляционных функций для разных пользователей

Анализ построенных кривых показывает, что с ростом числа лагов автокорреляционные функции убывают. При этом экстремумы наблюдаются при значениях лагов от 1 до 5. Таким образом, при прогнозе команд пользователя следует использовать именно это количество команд. При этом необходимо учитывать следующее: использование слишком большого количества команд приведет к тому, что на протяжении этого периода времени будет невозможно осуществлять прогноз команд, что снизит возможности по выявления аномальной деятельности пользователей.

4.2. Построение критерия изменения поведения пользователя

Рассмотрим теперь вопрос, связанный с построением критерия для определения изменения поведения пользователя компьютерных систем. Это позволит уменьшить количество ложных тревог, связанных с прогнозированием действий пользователя, и определить момент, когда следует переобучать нейронные сети.

При построении интерактивной модели пользователя необходимо учитывать, что с течением времени поведение пользователя меняется. Это может происходить по разным причинам. Например, пользователь установил новое программное обеспечение и начал часто его запускать, а другие программные продукты он стал использовать намного реже. Тогда в этом случае точность прогноза, осуществляемого нейронной сетью, снизится и такое поведение, соответственно, будет интерпретироваться как аномальное. Поэтому при построении интерактивной модели важно знать, что поведение пользователя изменилось и нейронную сеть для него следует переобучить.

Однако с этим необходимо заметить, что аномальное поведение также можно интерпретировать как изменение поведения. Поэтому при построении критерия будем исходить из следующих предположений:

- Будем считать, что аномальное поведение пользователя характеризуется резкими изменениями и скоротечностью;
- В свою очередь, естественное изменение поведения происходит на протяжении нескольких сеансов и не характеризуется резкими перепадами.

Перейдем теперь к формулировке критерия изменения поведения пользователя.

Пусть A — алфавит команд некоторого пользователя (т.е. набор всех команд, которые выполнялись на протяжении некоторого времени T). Предполагается, что за этот промежуток времени поведение пользователя не изменялось. Обозначим посредством $|A| = N$ — количество команд в алфавите. При этом будем считать, что каждой команде присвоен номер от 1 до N, причем номер N зарезервирован за командами, которые отсутствуют в алфавите. Пусть s_t ($t>T$) — текущий сеанс пользователя, для которого имеется следующая последовательность выполненных им команд:

$$c^{s_t} = \left(c_1^{s_t}, c_2^{s_t}, \ldots, c_{N_{s_t}}^{s_t}\right).$$

Для того чтобы определить, изменилось ли поведение пользователя, по окончании сеанса s_t строится вектор $g(s_t)$, компоненты которого определяются таким образом:

$$g_j(s_t) = \begin{cases} 1, & \text{если существует такое } k = 1,\ N_{s_t}, \text{ что } c_k^{s_t} = j,\ j \in A \\ 0, & \text{в противном случае} \end{cases}. \qquad (7)$$

Т.е. если определенная команда была выполнена во время сеанса s_t, то значение соответствующей компоненты вектора $g(s_t)$ становится равным 1, в противном случае — 0. Затем этот вектор попарно сравнивается с аналогичными векторами, построенными для предыдущих сеансов s_{t-1}, s_{t-2}, …, s_{t-l} и т.д. (в данной работе l = 5). В качестве меры сравнения использовалось расстояние Хэмминга:

$$\aleph(g(s_{t'}), g(s_{t''})) = \sum_{j=1}^{N} \chi(g_j(s_{t'}), g_j(s_{t''})),\ \chi(g_j(s_{t'}), g_j(s_{t''})) = \begin{cases} 1, \text{если } g_j(s_{t'}) \neq g_j(s_{t''}) \\ 0, \text{в противном случае} \end{cases}. \qquad (8)$$

Т.е. величина $\aleph$ определяет число компонент двух векторов, значения которых отличны. Получив в результате сравнения l значений, вычисляем среднее и делим его на общее количество команд в алфавите:

$$H_t = \frac{1}{N}\left(\frac{1}{l}\sum_{k=1}^{l} \aleph(g(s_t), g(s_{t-k}))\right). \qquad (9)$$

Если поведение пользователя не изменилось и не аномально, тогда вектор $g(s_t)$ будет отличаться от предшествующих незначительно. Соответственно, значение параметра H_t будет небольшим (меньше некоторого порога H_*). И наоборот, если наблюдается аномальное поведение пользователя, вектор $g(s_t)$ будет значительно отличаться от всех остальных и значение параметра H_t будет большим (больше некоторого порога H^*). Если же значение H_t принадлежит интервалу (H_*; H^*), можно говорить о естественном изменении поведения пользователя.

Таким образом, критерий изменения поведения пользователя можно сформулировать в следующем виде: *для данного (текущего) сеанса пользователя s_t на основе формул (7)-(9) вычисляется значение параметра H_t. Если $H_t < H_*$, считается, что поведение не изменилось, если $H_* < H_t < H^*$ — изменилось, если же $H_t > H^*$ — тогда поведение аномально.*

Для того чтобы проверить адекватность предложенного критерия, для разных пользователей были проведены эксперименты. Для этого использовались реальные данные, описание которых приводилось в разделе 3. Сначала значение H_t вычислялось для сеансов, во время которых поведение пользователя было нормальным (характерным). (На рис. 3 им соответствуют номера сеансов от 1 до 24.) Как видно из рис. 3, значения этого параметра для разных сеансов отличаются незначительно и лежат в пределах (0; 0,15). Для моделирования аномальной работы пользователя осуществлялась подмена данных. Т.е. вектор $g(s_t)$ строился для сеанса другого пользователя и сравнивался с аналогичными векторами, построенными для предшествующих сеансов исходного пользователя. Полученное в этом случае значение параметра H_t резко увеличивалось (на рис. 3 до 0,32).

Для того чтобы смоделировать естественное изменение поведения пользователей, было сделано следующее: значения элементов вектора $g(s_t)$ случайным образом (с вероятностью 0,25) изменялись на противоположные (на рис. 3 им соответствуют сеансы с номерами 35-50). Т.е. моделировалась ситуация, когда пользователь переставал выполнять одни команды и начинал использовать другие. В этом случае

значение параметра H_t увеличивалось до 0,25 (сеанс с номером 35), а затем снова, как и в случае с нормальным поведением, выходило на обычный уровень.

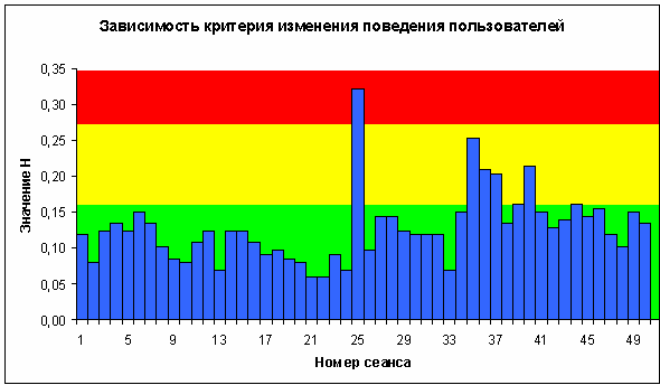


Рис. 3. Изменение значения H_t

Таким образом, результаты проведенных экспериментов показывают, что использование параметра H_t позволяет определить момент времени, когда поведение пользователя изменилось.

5. Моделирование данных о работе пользователя

В общем случае функционирование нейронной сети значительно зависит от качества обучающей выборки. Дело в том, что при небольшом размере обучающего множества нейронная сеть имеет тенденцию к жесткому запоминанию образов, что приводит к уменьшению ее способности к обобщению. Так, при построении интерактивной модели пользователя в нашем случае проблема с представительной выборкой данных не возникала, поскольку даже за непродолжительный период времени работы пользователя может быть собрано достаточное количество образов (представительных) для обучения нейронной сети. (Например, с учетом того, что пользователь в среднем вводит от 80 до 150 команд за сеанс, то за десять сеансов обучающая выборка может насчитывать до 1000 образов.)

В случае сеансовой модели размер обучающего множества напрямую определяется количеством сеансов, во время которых проводилось наблюдение за работой пользователя. Так, за три месяца таковых сеансов может достигать 100, что в нашем случае было недостаточно для качественного обучения нейронной сети и оптимизации ее архитектуры. Для решения этой проблемы можно использовать два подхода. Первый из них состоит в значительном увеличении отрезка времени, в рамках которого происходит наблюдение за поведением пользователя (скажем до 8-10 месяцев). Однако в этом случае велика вероятность того, что за этот промежуток времени оно изменится и, таким образом, обучающееся множество будет содержать противоречивые образы. Второй подход заключается в статистическом моделировании данных на основе имеющейся выборки. Он и будет использован в данной работе.

Поскольку в сеансовой модели для обучения нейронной сети используется информация о количестве вводимых команд за сеанс, номере компьютера, продолжительности и времени начала сеанса, для каждого пользователя проводилось моделирование именно этого набора данных. Для этого проверялось соответствие эмпирического распределения набору теоретических (нормальному, логарифмическому нормальному, равномерному и т.д.). В качестве критерия согласия использовался так называемый критерий χ^2 Пирсона [5, 6].

Критерий согласия χ^2. Пусть для каждого сеанса s_t имеется следующий набор данных о работе пользователя: $\{n_{s_t}, h_{s_t}, d_{s_t}, s_{s_t}\}_{t=1}^{T}$. Каждую из этих величин можно рассматривать как случайную, принимающую T определенных значений. К тому же, будем считать эти величины независимыми. Обозначим посредством X одну из этих случайных величин. Разобьем ее значения по k интервалам и представим в виде следующего статистического ряда:

l_i	$x_1; x_2$	$x_2; x_3$	…	$x_k; x_{k+1}$
p_i^*	p_1^*	p_2^*	…	p_k^*

где l_i — i-ый интервал значений, p_i^* — частота попадания в интервал l_i.

Требуется проверить, согласуются ли эти данные с гипотезой о том, что случайная величина *X* имеет закон распределения, заданный функцией распределения *F(x)* или плотностью *f(x)* (назовем его теоретическим). Зная теоретический закон распределения, можно найти теоретические вероятности попадания случайной величины в тот или иной интервал: $p_1, p_2, \ldots, p_k$.

Проверяя согласованность теоретического и статистического (эмпирического) распределений, возникает вопрос о том, каким же способом следует выбирать меру расхождений. В качестве такой меры в математической статистике обычно выбирают величину:

$$U = \sum_{i=1}^{k} n \frac{\left(p_i^* - p_i\right)^2}{p_i}. \qquad (10)$$

Так, Пирсон показал, что при достаточно больших значениях *n* закон распределения величины *U* практически не будет зависеть от функции распределения *F(x)* и числа *n*, а будет зависеть только от количества интервалов *k* и с увеличением *n* приближаться к распределению χ^2 с плотностью:

$$f(u) = \frac{1}{2^{\frac{r}{2}}\Gamma\left(\frac{r}{2}\right)} u^{\frac{r}{2}-1} \mathrm{e}^{-\frac{u}{2}} \quad (u > 0),$$

где $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}dt$ — гамма-функция; *r* — количество степеней свободы. Поскольку распределение χ^2 зависит от параметра *r* (число степеней свободы), то для его вычисления используют следующее правило:

$$r = k\text{-}s,$$

где *s* — количество независимых условий, наложенных на вероятности p_i^*. К примерам таких условий можно отнести требование, что бы сумма всех частот p_i^* была равна единице, или требование равенства теоретического среднего и дисперсии статическому и т.д.

Для распределения χ^2 составлены специальные таблицы. Пользуясь ими, можно для каждого значения χ^2 и числа степеней свободы *r* найти вероятность *p* того, что величина, распределенная по закону χ^2, превзойдет это значение.

Таким образом, схема применения критерия χ^2 к оценке согласованности статистического и теоретического распределений сводится к следующему:

1. Определяется мера расхождения *U* в соответствии с формулой (10).
2. Определяется число степеней свободы *r* как разность между числом интервалов *k* и количеством наложенных связей *s*.
3. По полученным значениям *U* и *r* (с помощью специальных таблиц для χ^2) определяется вероятность *p*. Если эта вероятность весьма мала, гипотеза о соответствии статистического распределения теоретическому отбрасывается. Если же эта вероятность относительно велика, данную гипотезу можно признать не противоречащей опытным данным.

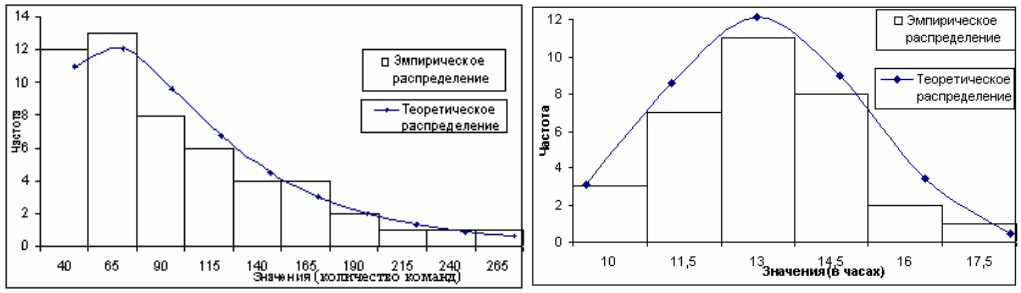
Рассмотрим теперь зависимости, которые были получены при моделировании.

Количество вводимых команд. Был проведен анализ различных распределений. Наилучшее значение критерия χ^2, равное 1,98, было получено для логарифмического нормального распределения. Поскольку в данном случае количество степеней свободы *r* равнялось 7 ($k = 10$, $s = 3$), это значение обеспечивало 97%-ое соответствие гипотезы реальным данным. На рис. 4,а приведено эмпирическое и построенное теоретическое распределение для этого параметра.

Номер компьютера. При анализе значений этого параметра необходимо учитывать следующее: пользователь, как правило, имеет свое основное место работы за компьютером и очень редко работает за остальными. Поэтому всегда существует значение, вероятность которого наибольшая и составляет 0,8…0,95. События же, связанные с работой пользователя за другими рабочими станциями, можно считать равновероятными.

Продолжительность сеанса. Анализ значений этого параметра показывает, что они распределены равновероятно (рис. 6). Значение критерия χ^2, равное 2,19, (при значениях параметров $r = 7$, $k = 10$, $s = 3$) обеспечивает 94%-ое соответствие гипотезы реальным данным.

Время начала сеанса. Наилучшее значение критерия χ^2, равное 0,87, для этого параметра было получено для нормального распределения. При $r = 3$, $k = 6$, $s = 3$ это значение обеспечивало 85%-ое соответствие гипотезы реальным данным. На рис. 4,б приведено эмпирическое и построенное теоретическое распределение.



(а) (б)

Рис. 4. Эмпирическое и теоретическое распределения для количества вводимых команд (а) и времени начала сеанса (б)

На основе полученных теоретических зависимостей была разработана программа, моделирующая работу пользователя за сеанс. Сформированные с ее помощью данные позволили оптимизировать архитектуру нейронной сети и улучшить качество ее функционирования.

6. Заключение

В данной работе проводилось математическое моделирование поведения пользователей компьютерных систем. Для этого использовалась комплексная модель, предложенная в работе [3]. Изучалась динамика работы пользователя во время сеанса. Поскольку определение оптимальной архитектуры нейронной сети является нетривиальной задачей, в интерактивной модели важно знать, сколько команд следует использовать при обучении для прогноза следующей. На основе построенных автокорреляционных кривых было выявлено, что для этого следует учитывать до восьми команд.

Также было проведено статистическое моделирование данных, используемых для обучения нейронной сети в сеансовой модели. Эмпирические распределения были аппроксимированы теоретическими. На их основе были сгенерированы необходимый набор данных, что дало возможность оптимизировать архитектуру нейронной сети и улучшить ее функционирование.

Литература

1. Куссуль Н., Соколов А. Адаптивное обнаружение аномалий в поведении пользователей компьютерных систем с помощью марковских цепей переменного порядка. Часть 2. Методы обнаружения аномалий и результаты экспериментов // Проблемы управления и информатики. — 2003. — № 4. — С. 83–88.
2. Manavoglu E., Pavlov D., Lee Giles C. Probabilistic User Behavior Models // Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003). — Melbourne, Florida (USA). — 2003. — P. 203–210.
3. Скакун С.В., Куссуль Н.Н. Нейросетевая модель пользователей компьютерных систем // Кибернетика и вычислительная техника. — 2004. — Выпуск 143. — С. 55–68.
4. Haykin S. Neural Networks: a comprehensive foundation. — Upper Saddle River, New Jersey: Prentice Hall, 1999. — 842 p.
5. Ширяев А.Н. Вероятность. — М.: Наука, 1980. — 575 с.
6. Вентцель Е.С. Теория вероятностей: Учебник для студ. вузов. — М.: Издательский центр «Академия», 2003. — 576 с.

Информация об авторах

Куссуль Н.Н. – зав. отделом, д.т.н., Институт космических исследований НАНУ-НКАУ; г. Киев, пр. Глушкова 40, тел. 8(044)266-25-53; e-mail: inform@ikd.kiev.ua

Скакун С.В. – м.н.с., Институт космических исследований НАНУ-НКАУ; г. Киев, пр. Глушкова 40, тел. 8(044)266-25-53; e-mail: inform@ikd.kiev.ua

JAMMING CANCELLATION BASED ON A STABLE LSP SOLUTION

Elena Revunova, Dmitri Rachkovskij

***Abstract**: Two jamming cancellation algorithms are developed based on a stable solution of least squares problem (LSP) provided by regularization. They are based on filtered singular value decomposition (SVD) and modifications of the Greville formula. Both algorithms allow an efficient hardware implementation. Testing results on artificial data modelling difficult real-world situations are also provided*

***Keywords**: jamming cancellation, approximation, least squares problem, stable solution, recurrent solution, neural networks, incremental training, filtered SVD, Greville formula*

Introduction

Jamming cancellation problem appears in many application areas such as radio communication, navigation, radar, etc. [Shirman, 1998], [Ma, 1997]. Though a number of approaches to its solution were proposed [McWhirter, 1989], [Ma et al., 1997], no universal solution exists for all kinds of jamming and types of antenna systems, stimulating further active research to advance existing methods, algorithms and implementations.

Consider an antenna system with a single primary channel and n auxiliary channels. Signal in each channel is, generally, a mixture of three components: a valid signal, jamming, and channel's inherent noise. The problem consists in maximal jamming cancellation at the output while maximally preserving valid signal.

Within the framework of weighting approach [Shirman, 1998], the output is obtained by subtraction of the weighted sum of signals provided by the auxiliary channels from the primary channel signal. The possibility of determining a weight vector $\boldsymbol{w}^*$ that minimizes noise at the output while preserving the valid signal to a maximum degree is, in general case, provided by the following. The same jamming components are present both in primary and auxiliary channels, however, with different mixing factors. Valid signal has small duration and amplitude and is almost absent in auxiliary channels. Characteristics of jamming, channel's inherent noise, and their mixing parameters are stable within the sliding "working window".

These considerations allow us to formulate the problem of obtaining the proper $\boldsymbol{w}^*$ as a linear approximation of a real-valued function $y=f(x)$:

$$F(x) = w_1 h_1(x) + w_2 h_2(x) + \ldots + w_n h_n(x) = \sum_{i=1,\,n} w_i\, h_i\,(x), \qquad (1)$$

where $h_1(x),\ldots,h_n(x)$ is a system of real-valued basis functions; $w_1,\ldots,w_n$ are the real-valued weighting parameters, $F(x)$ is a function approximating f(x).

In our case, $h_1(x),\ldots,h_n(x)$ are signals provided by the n auxiliary channels. Information about $y=f(x)$ at the output of the primary channel is given at discrete set of (time) points $k=1,\ldots,m$ (m is the width of the working window) by the set of pairs $(\boldsymbol{h}^k, y^k)$. It is necessary to find $\boldsymbol{w}^*$ approximating $f(x)$ by $F(x)$ using linear least squares solution:

$$\boldsymbol{w}^* = \mathrm{argmin}_{\,w}\, ||\boldsymbol{H}\,\boldsymbol{w} - \boldsymbol{y}||_2, \qquad (2)$$

where $\boldsymbol{H}$ is the so-called $m \times n$ "design matrix" containing the values provided by the n auxiliary channels for all $k=1,\ldots,$m; and $\boldsymbol{y} = [y_1, \ldots, y_m]^T$ is the vector of corresponding y values provided by the primary channel.

After estimating $\boldsymbol{w}^*$, the algorithm's output $\boldsymbol{s}$ is the residual discrepancy:

$$\boldsymbol{s} = \boldsymbol{H}\,\boldsymbol{w}^* - \boldsymbol{y}. \qquad (3)$$

Such a system may be represented as a linear neural network with a single layer of modifiable connections, $n+1$ input and single output linear neurons connected by a weight vector $\boldsymbol{w}$ (Fig. 1). In the case of successful training $\boldsymbol{w}^*$ provides an increased signal-jamming ratio at the output s compared to the input of the primary channel y, at least, for the training set $\boldsymbol{H}$.

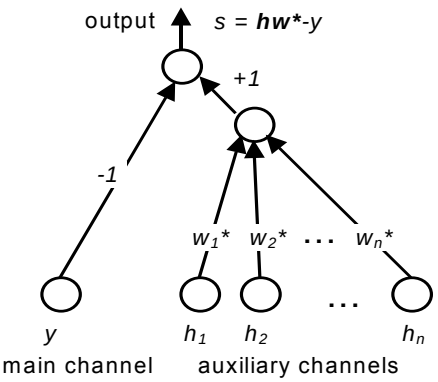


Fig 1. A neural network representation of a jamming canceller

A peculiarity of jamming cancellation problem in such a formulation consists in contamination of both ***y*** and ***H*** by the inherent noise of channels. Existing algorithms for jamming cancellation in the framework of weighting processing (2)-(3) [Shirman, 1998] do not take into account inherent noise contamination of ***y*** and ***H***. This results in instability of ***w*** estimation, leading, in turn, to a deterioration of cancellation characteristics, and often even to amplification of noise instead of its suppression. Therefore, methods for obtaining $\boldsymbol{w}^*$ should be stable to inherent noise contamination of ***y*** and ***H***. Other necessary requirements are real-time operation and simplicity of hardware implementation.

Least Squares Solution and Regularization

Generally, the solution of the least squares problem (LSP) (2) is given by

$$\boldsymbol{w}^* = \boldsymbol{H}^+ \boldsymbol{y}; \tag{4}$$

where $\boldsymbol{H}^+$ is pseudo-inverse matrix. If ***H*** is non-perturbed (noise is absent), then:

$$\text{for } m = n,\ \text{rank}(\boldsymbol{H}) = n = m \Rightarrow \det \boldsymbol{H} \neq 0.\ \boldsymbol{H}^+ = \boldsymbol{H}^{-1}; \tag{5}$$

$$\text{for } m > n,\ \text{rank}(\boldsymbol{H}) = n, \Rightarrow \det \boldsymbol{H} \neq 0:\ \boldsymbol{H}^+ = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T; \tag{6}$$

$$\text{for } m = n,\ m > n,\ \text{rank}(H) < n \Rightarrow \det H = 0,\ \boldsymbol{H}^+ = \lim_{v \to 0} (\boldsymbol{H}^T\boldsymbol{H} + v^2\boldsymbol{I})^{-1}\boldsymbol{H}^T. \tag{7}$$

$\boldsymbol{H}^+$ for (7) can be obtained numerically using SVD [Demmel, 1997] or the Greville formula [Greville, 1960].

The case when ***y*** and elements of matrix ***H*** are known precisely is very rare in practice. Let us consider a case that is more typical for jamming cancellation, i.e. when ***y*** and ***H*** are measured approximately: $\boldsymbol{y} = \boldsymbol{y} + \boldsymbol{\varsigma}$; $\boldsymbol{H} = \boldsymbol{H} + \boldsymbol{\Xi}$; where $\boldsymbol{\varsigma}$ is noise vector, $\boldsymbol{\Xi}$ is noise matrix. In such a case, solutions (5)-(7) may be unstable, i.e. small changes of ***y*** and ***H*** cause large changes of $\boldsymbol{w}^*$ resulting in instable operation of application systems based on (5)-(7).

To obtain a stable LSP solution, it is fruitful to use approaches for solution of "discrete ill-posed problems" [Hansen, 1998], [Jacobsen et al., 2003], [Reginska, 2002], [Wu, 2003], [Kilmer, 2003]. Such a class of LSPs is characterized by ***H*** with singular values gradually decaying to zero and large ratio between the largest and the smallest nonzero singular values. This corresponds to approximately known and near rank-deficient ***H***.

Reducing of an ill-posed LSP to a well-posed one by introduction of the appropriate constraints to the LSP formulation is known as regularization [Hansen, 1998]. Let us consider a problem known as standard form of the Tikhonov regularization [Tikhonov, 1977]:

$$\text{argmin}_{\,w} \{||\boldsymbol{y} - \boldsymbol{H}\,\boldsymbol{w}||_2 + \lambda\, ||\boldsymbol{w}||_2\}. \tag{8}$$

Its solution $\boldsymbol{w}_\lambda$ may be obtained in terms of SVD of ***H*** [Hansen, 1998]:

$$\boldsymbol{w}_\lambda = \sum_{i=1,\,n} f_i\ \boldsymbol{u}_i^T\boldsymbol{y}\,/\sigma_i\ \boldsymbol{v}_i; \tag{9}$$

$$f_i = \sigma^2_i/(\sigma^2_i + \lambda^2), \tag{10}$$

where σ_i are singular values, $\boldsymbol{u}_1 \ldots \boldsymbol{u}_n$, $\boldsymbol{v}_1 \ldots \boldsymbol{v}_n$ are left and right singular vectors of ***H***, f_i are filter factors.

Note that solution of (8) using truncated SVD method [Demmel, 1997] is a special case of (9), (10) with $f_i \in \{0,1\}$.

Algorithmic Implementation of Solutions of Discrete ill-posed LSPs

Requirements of an efficient hardware implementation of jamming cancellation pose severe restrictions on the allowable spectrum of methods and algorithms. In particular, methods of $\boldsymbol{w}^*$ estimation are required to allow for parallelization or recursion. Taking this into account, let us consider some algorithmic implementations of (8).

SVD-based solution. The implementation of SVD as a systolic architecture with paralleled calculations is considered in [Brent, 1985]. We have developed a systolic architecture that uses effective calculation of $\boldsymbol{u}_i$, σ_i, and $\boldsymbol{v}_i$ for obtaining the regularized solution $\boldsymbol{w}_\lambda^*$ (9)-(10). The architecture implements two regularization techniques: truncated and filtered SVD [Hansen, 1998].

Advantages of SVD-based solution are accuracy and parallelism. Drawbacks are connected with the hardware expenses for calculation of trigonometric functions for diagonalization of sub-matrices and implementation of systolic architecture itself.

Solution based on the Greville formula and its modifications. Let us consider another stable method for $\boldsymbol{w}^*$ estimation based on the Greville formula, which can be readily implemented in hardware because of its recursive character. A recursive procedure for the LSP solution [Plackett, 1950] for a full-rank $\boldsymbol{H}$ is as follows:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \boldsymbol{b}_{k+1}(y_{k+1} - \boldsymbol{h}^T_{k+1}\boldsymbol{w}_k); \quad k = 0, 1, \ldots, \tag{11}$$

$$\boldsymbol{b}_{k+1} = \boldsymbol{P}_k \boldsymbol{h}_{k+1}/(1 + \boldsymbol{h}^T_{k+1}\boldsymbol{P}_k \boldsymbol{h}_{k+1}); \tag{12}$$

$$\boldsymbol{P}_{k+1} = (\boldsymbol{H}^T_{k+1}\boldsymbol{H}_{k+1})^{-1} = (\boldsymbol{I} - \boldsymbol{b}_{k+1}\boldsymbol{h}^T_{k+1})\boldsymbol{P}_k; \tag{13}$$

where $\boldsymbol{h}_k$ is the kth row (sample) of $\boldsymbol{H}$; $\boldsymbol{P}_0 = 0$; $\boldsymbol{w}_0 = 0$. Note that this provides an iterative version of training algorithm for a neural network interpretation of Fig. 1.

The Greville formula [Greville, 1960] allows $\boldsymbol{b}_{k+1}$ calculation for (11) without $(\boldsymbol{H}^T_{k+1}\boldsymbol{H}_{k+1})^{-1}$ calculation, thus overcoming the problem of rank-deficiency of $\boldsymbol{H}$:

$$\boldsymbol{b}_{k+1} = (\boldsymbol{I} - \boldsymbol{H}^+_k\boldsymbol{H}_k)\boldsymbol{h}_{k+1} / \boldsymbol{h}^T_{k+1}(\boldsymbol{I} - \boldsymbol{H}^+_k\boldsymbol{H}_k)\boldsymbol{h}_{k+1}; \ \text{if} \ \boldsymbol{h}^T_{k+1}(\boldsymbol{I} - \boldsymbol{H}^+_k\boldsymbol{H}_k)\boldsymbol{h}_{k+1} \neq 0; \tag{14}$$

$$\boldsymbol{b}_{k+1} = \boldsymbol{H}^+_k(\boldsymbol{H}^+_k)^T\boldsymbol{h}_{k+1} / (1 + \boldsymbol{h}^T_{k+1}\boldsymbol{H}^+_k(\boldsymbol{H}^+_k)^T\boldsymbol{h}_{k+1}); \ \text{if} \ \boldsymbol{h}^T_{k+1}(\boldsymbol{I} - \boldsymbol{H}^+_k\boldsymbol{H}_k)\boldsymbol{h}_{k+1} = 0; \tag{15}$$

$$\boldsymbol{H}^+_{k+1} = (\boldsymbol{H}^+_k - (\boldsymbol{b}_{k+1}\boldsymbol{h}^T_{k+1}\boldsymbol{H}^+_k \mid \boldsymbol{b}_{k+1})). \tag{16}$$

$\boldsymbol{w}^*$ obtained by (11)-(13) using (14)-(16) is equivalent to $\boldsymbol{w}^*$ obtained by (7) for precisely specified $\boldsymbol{H}$. Presence of $\boldsymbol{H}^+_k$ and $\boldsymbol{H}_k$ in (14)-(16) makes recursion more resource- and computation-expensive than (12)-(13). As a new sample arrives, it is necessary to calculate $\boldsymbol{H}^+_k\boldsymbol{H}_k$ or $\boldsymbol{H}^+_k(\boldsymbol{H}^+_k)^T$ that requires calculation of $\boldsymbol{H}^+_k$ and storage of $\boldsymbol{H}^+_k$ and $\boldsymbol{H}_k$. These drawbacks are overcome by an improvement of the Greville formula proposed recently in [Zhou, 2002].

For $\boldsymbol{h}^T_{k+1}\boldsymbol{Q}_k = 0$ calculations of $\boldsymbol{b}_{k+1}$ and $\boldsymbol{P}_{k+1}$ are made by (12)-(13). If $\boldsymbol{h}^T_{k+1}\boldsymbol{Q}_k \neq 0$

$$\boldsymbol{b}_{k+1} = \boldsymbol{Q}_k\boldsymbol{h}_{k+1}/(\boldsymbol{h}^T_{k+1}\boldsymbol{Q}_k\boldsymbol{h}_{k+1}); \tag{17}$$

$$\boldsymbol{P}_{k+1} = (\boldsymbol{I} - \boldsymbol{b}_{k+1}\boldsymbol{h}^T_{k+1})\boldsymbol{P}_k(\boldsymbol{I} - \boldsymbol{b}_{k+1}\boldsymbol{h}^T_{k+1})^T + \boldsymbol{b}_{k+1}\boldsymbol{b}^T_{k+1}; \tag{18}$$

$$\boldsymbol{Q}_{k+1} = (\boldsymbol{I} - \boldsymbol{b}_{k+1}\boldsymbol{h}^T_{k+1})\boldsymbol{Q}_k. \tag{19}$$

Here $\boldsymbol{P}_k = \boldsymbol{H}^+_k(\boldsymbol{H}^+_k)^T$ is Hermitian $n \times n$ matrix; $\boldsymbol{P}_0 = 0$, $\boldsymbol{Q}_k = \boldsymbol{I} - \boldsymbol{H}^+_k\boldsymbol{H}_k$; $\boldsymbol{Q}_0 = \boldsymbol{I}$.

We further modified the Greville formula so that $\boldsymbol{w}_{k+1}$ is equivalent to the regularized solution $\boldsymbol{w}_\lambda^*$ (9). This is achieved by comparison of vector norm $\boldsymbol{h}^T_{k+1}\boldsymbol{Q}_k$ not with 0, but with some threshold value 0_{eff} calculated from noise matrix $\boldsymbol{\Xi}$. We name such an algorithm "pseudo-regularized modification of the Greville formula" (PRMGF).

The algorithm (11)-(13), (17)-(19) calculates $\boldsymbol{w}^*$ using all previous samples. However, for a non-stationary case it is necessary to process only a part of the incoming samples inside a sliding working window. Full recalculation of $\boldsymbol{H}^+_{k+1}$ for estimation of $\boldsymbol{w}_{k+1}$ as each new sample arrives can be avoided by using inverse recurrent representation of [Kirichenko, 1997]. For the purpose of removing the row $\boldsymbol{h}^T_1$ from $\boldsymbol{H}_{k+1}$, $\boldsymbol{H}^+_k$ is represented through $\boldsymbol{H}^+_{k+1} = (\boldsymbol{b}_1 | \boldsymbol{B}_{k+1})$ as follows.

For a linear independent row:

$$(\boldsymbol{H} + \boldsymbol{h}\boldsymbol{e}^T)^+ = \boldsymbol{H}^+ - \boldsymbol{H}^+\boldsymbol{h}\boldsymbol{h}^T\boldsymbol{Q}/(\boldsymbol{h}^T\boldsymbol{Q}\boldsymbol{h}) - \boldsymbol{Q}\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{H}^+/\boldsymbol{e}^T\boldsymbol{Q}\boldsymbol{e} + \boldsymbol{Q}\,\boldsymbol{e}\boldsymbol{h}^T\boldsymbol{Q}(\boldsymbol{H}^T)\,(1 + \boldsymbol{e}^T\boldsymbol{H}^+\boldsymbol{h})/\,\boldsymbol{h}^T\boldsymbol{Q}(\boldsymbol{H}^T)\boldsymbol{h}\,\boldsymbol{e}^T\boldsymbol{Q}\boldsymbol{e}; \tag{20}$$

and for a linear dependent row:

$$(\boldsymbol{H} + \boldsymbol{h}\boldsymbol{e}^T)^+ = (\boldsymbol{I} - \boldsymbol{z}\boldsymbol{z}^T / \|\boldsymbol{z}\|^2)\, \boldsymbol{H}^+;\ \boldsymbol{z} = \boldsymbol{H}^+\boldsymbol{h} - \boldsymbol{e} / \|\boldsymbol{e}\|^2. \quad (21)$$

Thus, we propose to use PRMGF with a sliding window for the case, when it is required to obtain $\boldsymbol{w}^*$ not for the whole training set, but for its subset of a fixed size. For initial $k < K$ samples $\boldsymbol{h}_k$ (where K is the size of working window) $\boldsymbol{w}^*_k$ is recursively calculated by PRMGF. For $k > K$, (20)-(21) are used for updating $\boldsymbol{H}^+$ by removing the sample that has left a working window, and the incoming sample s is taken into account using PRMGF as earlier.

Advantages of PRMGF with a sliding window include:

- natural embedding into recursive algorithm for $\boldsymbol{w}^*$;
- increase of calculation speed due to using $\boldsymbol{h}^T_{k+1}$ instead of $\boldsymbol{H}_k$, also resulting in reduction of required memory;
- additional memory reduction since $\boldsymbol{P}_k$, $\boldsymbol{K}_k$ and $\boldsymbol{Q}_k$ have fixed $n \times n$ dimension for any k;
- further increase of calculation speed when sliding window is used due to the Greville formula inversion;
- considerably smaller hardware expenses in comparison with SVD;
- $\boldsymbol{w}^*$ close to the Tikhonov regularized solution for noisy, near rank-deficient $\boldsymbol{H}$ (at least, for small matrices);
- natural interpretation as an incrementally trained neural network.

Example of Modelling a Jamming Cancellation System

Let's compare the following jamming cancellation algorithms: ordinary LS-based (6); non-truncated SVD-based [Demmel, 1997]; truncated SVD-based (9) with $f_i = \{0, 1\}$; PRMGF-based (section 3.2). We use near rank-deficient $\boldsymbol{H}$, which is critical for traditional jamming cancellation algorithms – e.g., for ordinary LS-based ones.

Testing scheme and cancellation quality characteristics. In a real situation, all antenna channels receive jamming signals weighted by the gain factor that is determined by the antenna directivity diagram in the direction of particular jamming. We simulated signals in antenna channels as follows:

$$\boldsymbol{X} = \boldsymbol{S}\,\boldsymbol{M} + \boldsymbol{\Xi}; \quad (22)$$

where $\boldsymbol{X}$ is $L \times (n+1)$ matrix of signals in antenna channels ($\boldsymbol{H}$ is sub-matrix of $\boldsymbol{X}$); L is the number of samples; n is the number of auxiliary channels; $\boldsymbol{S}$ is jamming signals' matrix; $\boldsymbol{\Xi}$ is channel inherent noise matrix; $\boldsymbol{M}$ is mixing matrix.

Jamming signals and channels' inherent noise are modeled by normalized centered random variables with normal and uniform distribution correspondingly. $\boldsymbol{M}$ is constructed manually, values of its elements are about units, rank deficiency was achieved by entering identical or linearly dependent rows. For ideal channels without inherent noise, rank deficiency of $\boldsymbol{M}$ gives rise to strict rank deficiency of $\boldsymbol{H}$. Inherent noise results in near rank-deficient $\boldsymbol{H}$. Tests were carried out for n=8 auxiliary channels.

The main characteristics of jamming cancellers are: jamming cancellation ratio (K^c) and jamming cancellation ratio vs inherent noise level in auxiliary channels K^{naux}: $K^c = f(K^{naux})$ [Bondarenko, 1987] at fixed inherent noise at the primary channel K^{n0}.

$$K^c = P^{in}/P^{out}, \quad (23)$$

where P^{in} and P^{out} is power of jamming in the primary channel and in the output of jamming canceller, respectively. In all tests, the valid signal with amplitude not exceeding amplitude of primary channel jamming was present at the input for 5 nearby samples. L=1000; m=16; K^{n0} = {0.1, 0.2, 0.3}, $K^{n0} >> K^{naux}$ to complicate the algorithm's operation.

Testing results. A family of jamming cancellation characteristics $K^c = f(K^{naux})$ for rank-deficient $\boldsymbol{M}$ and near rank-deficient $\boldsymbol{H}$ is shown in Fig.2. K^{naux} varied from $1.6\ 10^{-9}$ up to $6.4\ 10^{-6}$. K^c for the ordinary LS did not exceed 1 at $K^{naux} < 2.5\ 10^{-8}$. For truncated SVD and PRMGF $K^c \approx 10$ ($K^{n0} = 0.1$) are nearly constant over the whole range of K^{naux} and close to each other. Note that for a full-rank matrix $\boldsymbol{H}$, K^n for all algorithms was approximately the same and large in the considered range of K^{naux}.

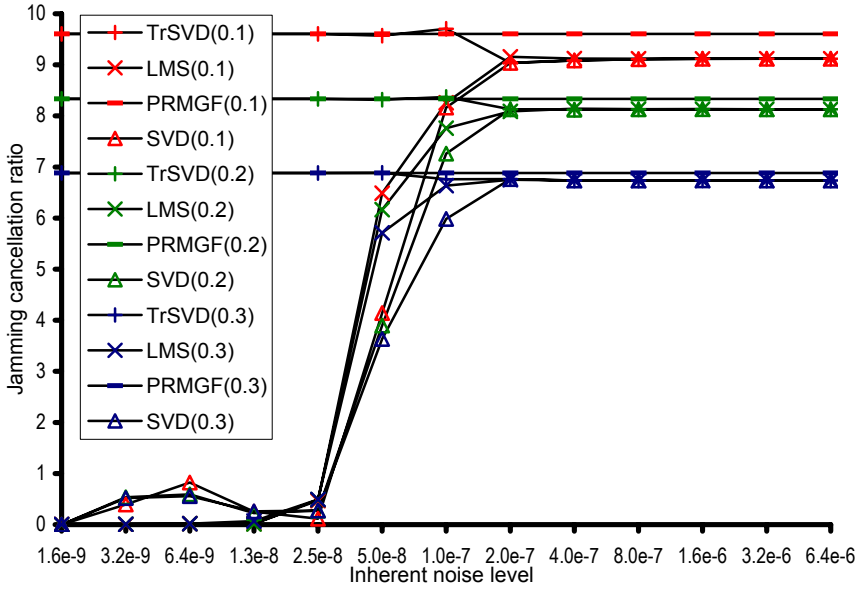


Fig. 2. Kc= f(Knaux) for near rank-deficient H

It may seem from the analysis of the shown results that one may use the ordinary LS algorithm at increased level of K^{naux}. However, roll-of of cancellation characteristic is also observed when jamming intensity in auxiliary channels is much more than the inherent noise level. To show that, let us consider $K^c(P^{in})$ for near rank-deficient $\boldsymbol{H}$ (Fig.3). Jamming power P^{in} changed from $9 \cdot 10^3$ to $1.2 \cdot 10^7$ by step $2 \cdot 10^2$, $K^{naux} = 0.1$. For $P^{in} > 10^4$, K^c for ordinary LS and non-truncated SVD decreases. For truncated SVD and PRMGF, $K^c \approx 9$ ($K^{n0} = 0.1$) are constant and close to each other. In this case, we cannot artificially increase inherent noise level because it will completely mask the valid signal.

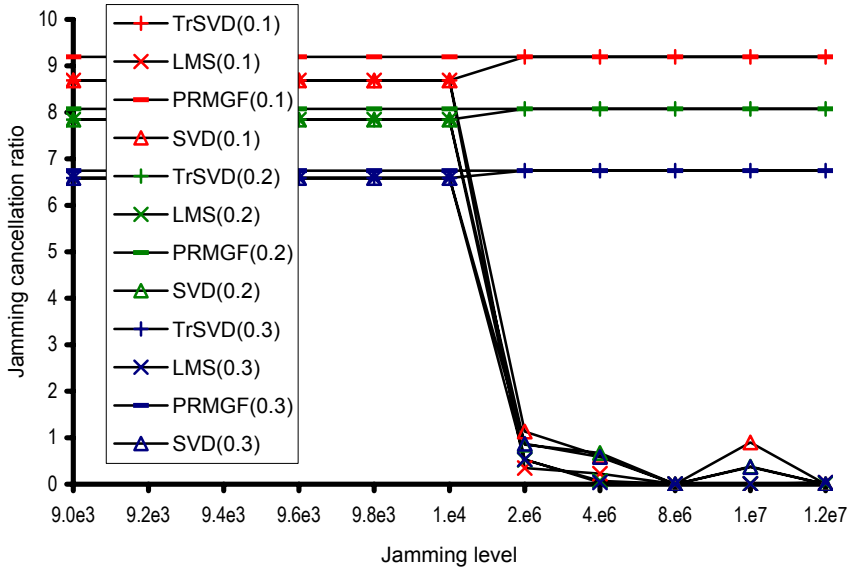


Fig. 3. Kc(Pin) for near rank-deficient H

Conclusions

In the framework of this work, two new jamming cancellation algorithms have been developed based on the so-called weighting approach. Special requirements to the problem have resulted in its classification as a discrete ill-posed problem. That has allowed us to apply an arsenal of the regularization-based methods for its stable solution - estimation of weight vector $\boldsymbol{w}^*$.

The standard form of Tikhonov regularization based on SVD has been transformed to efficient hardware systolic architecture. Besides, pseudo-regularized modification of the Greville formula allowed us to get weight vector estimations very close to estimations for a truncated SVD based regularization - at least for ***H*** of about tens of columns. Testing on near rank-deficient ***H*** has shown that distinctions in ***w**** obtained by both algorithms are of the order 10^{-5}. A combined processing technique based on a regularized modification of the Greville formula and inverse recurrent representation of Kirichenko permits a more efficient processing of data for a sliding working window.

Testing on artificial data that model real-world jamming cancellation problem has shown an efficient cancellation for near rank-deficient ***H***. For the developed PRMGF-based algorithm the jamming cancellation ratio is near constant and considerably higher than 1 in the whole range of variation of auxiliary channels' inherent noise and jamming amplitude. On the contrary, for the non-regularized LS method the ratio roll-offs to less than 1, meaning jamming amplification.

A straightforward neural network interpretation of such a system is provided. The developed algorithms and computer architectures for their implementation can be applied to solution of other discrete ill-posed LS problems and systems of linear algebraic equations.

Bibliography

[Bondarenko, 1987] B.F. Bondarenko, Bases of radar systems construction, Kiev, 1987 (in Russian).

[Brent, 1983] R.P. Brent, H. T. Kung and F. T. Luk, Some linear-time algorithms for systolic arrays, in Information Processing 83, 865–876, North-Holland, Amsterdam, 1983.

[Brent, 1985] R.P. Brent and F. T. Luk, The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays, SIAM J. Scientific and Statistical Computing 6, 69–84, 1985.

[Demmel, 1997] J.W. Demmel, Applied Numerical Linear Algebra. SIAM, Philadelphia, 1997.

[Greville, 1960] T. N. E. Greville, Some applications of the pseudoinverse of a matrix, SIAM Rev. 2, pp. 15-22, 1960.

[Hansen, 1998] P.C. Hansen, Rank-deficient and discrete ill-posed problems. Numerical Aspects of Linear Inversion, SIAM, Philadelphia, 1998.

[Jacobsen et al., 2003] M. Jacobsen, P.C. Hansen and M.A. Saunders, Subspace preconditioned LSQR for discrete ill-posed problems, BIT Numerical Mathematics 43: 975–989, 2003.

[Kilmer, 2003] M. Kilmer, Numerical methods for ill-posed problems, Tufts University, Medford, MAPIMS Workshop, 2003.

[Kirichenko, 1997] N.F. Kirichenko, Analytical representation of pseudo-inverse matrix perturbation, Cybernetics and system analysis, 2, pp. 98-107, 1997 (in Russian).

[Ma et al., 1997] J. Ma, E.F. Deprettere and K.K. Parhi, Pipelined cordic based QRD-RLS adaptive filtering using matrix look ahead. In Proc. IEEE Workshop on Signal Processing Systems, pp.131-140 Lecester, UK, November, 1997.

[Ma, 1997] J. Ma, Pipelined generalized sidelobe canceller, Technique Report, Circuits and Systems, Delft University of Technology, 1997.

[McWhirter, 1989] J.G. McWhirter and T.J. Shepherd, Systolic array processor for MVDR beamforming, In IEEE Proceedings vol.136, pp. 75-80, April 1989.

[Plackett, 1950] R. L. Plackett, Some theorems in least squares, Biometrika, 37, pp. 149-157, 1950.

[Reginska, 2002] T. Reginska, Regularization of discrete ill-posed problems, IM PAN Preprints, 2002, http://www.impan.gov.pl/Preprints.

[Shirman, 1998] J.D. Shirman, Radio-electronic systems: bases of construction and the theory, Moscow, 1998 (in Russian).

[Tikhonov, 1977] A.N. Tikhonov, V.Y. Arsenin, Solution of ill-posed problems. V.H. Winston, Washington, DC, 1977.

[Wu, 2003] L. Wu, A parameter choice method for Tikhonov regularization, Electronic Transactions on Numerical Analysis, Volume 16, pp. 107-128, 2003.

[Zhou, 2002] J. Zhou, Y. Zhu, X. R. Li, Z. You, Variants of the Greville formula with applications to exact recursive least squares. SIAM J. Matrix Anal. Appl. Vol.24, No.1, pp.150-164, 2002.

Authors' Information

Elena G. Revunova, Dmitri A. Rachkovskij - International Research and Training Center of Information Technologies and Systems, Pr. Acad. Glushkova 40, Kiev 03680, Ukraine. email: helab@i.com.ua, dar@infrm.kiev.ua

GRAPH REPRESENTATION OF MODULAR NEURAL NETWORKS

Michael Kussul, Alla Galinskaya

***Abstract**: Modular neural networks are widely used for applied tasks solving due to its flexibility and big potential abilities. As a result, development of modelling aids for modular neural networks become very important. Networks that contain cycles are of particular interest. However, for the networks with cycles there is necessity to have tools for formal analysis, which allow defining sequence of run of modules in the networks. We propose representation of modular neural networks with help of directed graphs. This representation is intended for general analysis of modular architectures and, first of all, for analysis with automatic systems. On the basis of proposed representation we give definitions of cycles, build its classification and examine properties of cycles in modular neural networks.*

***Keywords**: neural networks, modular neural networks, graph of neural network, cycle.*

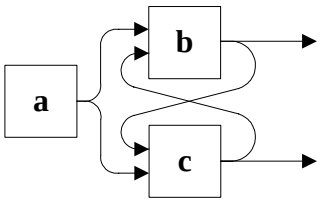
Introduction

Modular approach is actively developed in the field of application of artificial neural networks during last decade [1]. Modular networks, in particular, allow dividing a complex task into simple subtasks that are solved by individual modules and construct whole task decision as a combination of solved subtasks. Because of growing interest to modular architectures, development of modelling aids for modular neural networks becomes very important. By amodular network, we imply arbitrary set of algorithms for data processing, including artificial neural networks that combined for some task solving.

We should solve several problems both general and particular nature for modular neural networks application. There are questions of choosing the architecture of modular network for concrete task solving among particular problems. General problems are methods of combination of heterogeneous algorithms to one modular network, run sequence determination and methods of training of modular network as a whole.

The most of combination problems are solved in software neural computer MNN CAD [2]. Training of modular networks is a separate task, which solving is often architecture-dependent. Different approaches of modular networks training are given in articles [3] and [4].

Modular network can contain both algorithms that process data in parallel and modules that process data sequentially. Also, modular architectures can contain cycles. By the run sequence of modular network, we'll imply a sequence, in which modules process input data. Run sequence is obvious when there are no cycles in modular network. However, if we use cycles and run sequence is not defined by outer conditions it is possible that some contradictions occur. An example of such contradiction is a trigger scheme of modules' combination shown at the picture 1. The result of data processing will differ depending on what algorithm b or c was applied first or they work in parallel.



Pic. 1. Trigger scheme of modules' combination

Determination and resolution of contradictions in modular architectures is important task of modular networks' application. This task is of particular significance for development of interactive aids for modular network modelling such as MNN CAD. CAD system should automatically determine and interdict creation of contradiction architectures or demand contradiction resolution from user.

The main goal of this work is an attempt to propose model of modular neural networks, which allow creating consistent architectures and automatic determination of run sequence of modules.

We propose representation of modular networks with help of directed graphs. On the basis of proposed representation we give definitions of cycles, build its classification and examine properties of cycles in modular neural networks.

Digraph Representation of Modular Neural Networks

Modular network can be described by digraph $G(V,E)$, where vertices $v \in V$ correspond to network modules and edges $e \in E$ are connections between modules. We'll consider that digraph edge $e(v_1,v_2)$ correspond to all connections between modules v_1 and v_2. That is if module v has n inputs, some k of which connected to the outputs of module v_1 and remaining $n-k$ to the outputs of module v_2, then v will have two incoming edges $e_1(v_1,v)$ and $e_2(v_2,v)$. Similar to the definition in graph theory, we'll denote a walk as an alternating sequence of vertices and edges, with each edge being incident to the vertices immediately preceding and succeeding it in the sequence. A path (elementary path) is a walk with no repeated vertices. Walk and path from vertex v_1 to vertex v_2 we'll denote $w(v_1,v_2)$.

Let's introduce two special "virtual" modules representing inputs and outputs of whole modular network. All inputs of modular net we'll consider as input module and outputs – as an output module. Input and output modules are virtual in the sense that they do not realize data processing algorithms.

By the analogy with graph theory, we'll consider vertex I, which corresponds to virtual input module of network, to be **input** or source of graph G. Input of graph is determined by the condition $s_{in}(I)=0$, where $s_{in}(I)$ - the number of incoming edges.

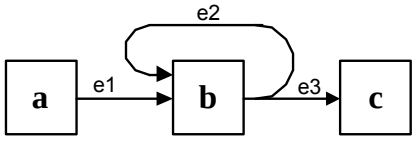
Output or sink of graph we'll denote vertex O that corresponds to the virtual output module. The next assertion is true for output: $s_{out}(O)=0$, where $s_{out}(O)$ - the number of outgoing edges. It should be noted that the number of the module outputs (in fact outputs of network) can be arbitrary.

In order that digraph G can be regarded as some modular network, it must satisfy the next conditions:

1) Digraph G is weakly connected, i.e. undirected graph corresponded to G is connected;
2) Graph G has not multiple edges;
3) There is single input vertex $I \in V$ in G and every vertex is accessible from input: $\forall v \in V : v \neq I \Rightarrow \exists\, w(I,v)$ and therefore $\forall v \in V : v \neq I \Rightarrow s_{in}(v) \geq 1$
4) There is single output vertex $O \in V$ in G and output is accessible from every vertex: $\forall v \in V : v \neq O \Rightarrow \exists\, w(v,O)$ and therefore $\forall v \in V : v \neq O \Rightarrow s_{out}(v) \geq 1$

The last two conditions are obvious if we take into consideration that modular network are built for data processing. If network don't meet the condition 3 then there are modules that have not input data. Condition 4 guarantees that result of every module processing will reach output of network directly or after another modules' processing.

In contrast to the definition accepted in the graph theory we'll consider walk w from vertex a to vertex b as a set of vertices and edges not including final vertex. According to such definition we can pick out six walks on picture 2: $w_1(a,b)=\{a;e_1\}$; $w_2(a,c)=\{a;e_1;b;e_3\}$; $w_3(a,c)=\{a;e_1;b;e_2;b;e_3\}$; $w_4(b,b)=\{b;e_2\}$; $w_5(b,c)=\{b;e_3\}$ and $w_6(b,c)=\{b;e_2;b;e_3\}$. There is no walk $w(a,b)=\{a;e_1;b;e_2\}$ because vertex b cannot belong to the walk by our definition. Similar, path not includes final vertex. Correspondingly, there are four paths on picture 2: w_1, w_2, w_4 and w_5.



Pic. 2 Example of cycle

Specification of the walk definition pursues the next aim: if we consider subgraphs of modular network as a set of vertices and edges then exclusion of final vertex in the walk definition let us use difference operation without explicit indication of membership of vertices in the resulting set. That is, if we have walk represented by the set of vertices and edges $w(a,c)=\{a;e_1;b;e_2\}$, then this walk can be represented by two non-overlapping subsets $w_1(a,b)=\{a;e_1\}$ and $w_2(b,c)=\{b;e_2\}$, each of which is also a walk due to our definition.

In the subsequent text, we'll require two definitions: projection P of one module onto another and degree of module uncertainty U.

Definition 1: **Projection** $P(b,a)$ of vertex (module) $a \in V$ onto vertex (module) $b \in V$ is a directed subgraph of digraph $G(V,E)$ inclusive all possible **paths** $w(b,a)$ from b to a:

$$P(b,a) = \bigcup w(b,a), \quad \forall w(b,a) = \{v_1, v_2, \ldots, v_n\} : v_1 = b, v_{n+1} = a, \quad v_i \neq v_j \quad \forall i, j = \overline{1,n}.$$

That is this subgraph contains all vertices and corresponding edges of all paths form b to a, except vertex a according to our definition.

One of the most important projections is a projection of some module a onto network input I: $P(I,a)$.

Definition 2: If we associate every vertex $v \in V$ with a parameter $t(v)$: $t(v) = 1$ if outputs of corresponding network module are not computed at given run step and $t(v) = 0$ if module is already calculated, then **uncertainty** of vertex a will be defined as

$$U(a) = \sum_{v \in P(I,a)} t(v).$$

Uncertainty of input module $U(I)$ is always equal to zero. As input module is virtual we consider its outputs to be always calculated.

Necessity of such definition directly follows from goals putted by. Checking modular architectures on consistency and construction of sequence of network modules' run requires introducing of some comparison criterion, which of modules should be calculated first. Defined uncertainty can be such criterion.

Axiom 1: Module with less uncertainty should be calculated first if there are no special outer conditions.

Let's illustrate requirement of this axiom by the example of sequential modules' connection (picture 3). Outputs of input module I are determined by our definition and outputs of module a are not calculated at the beginning of modular network's run. Hence, the uncertainty of module a is equal to zero and uncertainty of module b is equal to one according to Definition 2. Therefore, according to Axiom 1, module a should be calculated before module b as it obvious for sequential modules' connection.

Pic 3. Sequential connection of modules

Necessity of Axiom 1 becomes even more evident for modular neural network training. If untrained module (neural network) a is found on a walk from input to module b, then training of module b is meaningless while network a is untrained.

Cycles in Modular Networks

Cycles are of special interest with relation to application tasks. If input data are organized as time series we mean "recurrent" cycles. If modules, contained in a cycle, realize associate fields or other algorithms that require several run iterations with one input vector then we mean "iterative" cycles. Hence, cycle can be recurrent or iterative.

It's necessary to introduce clock of modular network run for recurrent cycles description but for iterative cycles it's enough to define the number of cycle's round, where cycle's round is a single run of all modules contained in a cycle. In this article we discuss only general definitions related both to iterative and recurrent cycles.

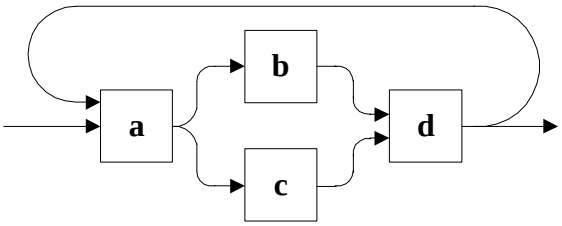
Primary types of cycles and basic definitions

First, we need to introduce general definition of cycle in modular network.

Definition 3: **Cycle** with first vertex a is a directed subgraph $C_a(V_a, E_a)$ of a modular network digraph $G(V, E)$ that contains all possible walks $w(a,a)$ from vertex a into itself:

$$C_a(V_a, E_a) = \bigcup w(a,a), \quad \forall w(a,a) = \{v_1, v_2, \ldots, v_n\} : v_1 = a, v_{n+1} = a$$

It should be noticed that walks but not paths are appearing in a cycle definition in contrast to definition of projection. Definition 3 generalizes cycle definition that is common in graph theory [5] in the sense that it includes all possible walks $w(a,a)$. It is not difficult to show that cycles according to Definition 3 correspond to strictly connected components in a digraph and have obvious property: $C_a \bigcap C_b = \varnothing$.



Pic 4. Simple cycle scheme

Architecture on picture 4 illustrates the necessity of definition of a cycle as a set of all walks. It's obvious that while modules' run in the cycle $C_a : V_a = \{a, b, c, d\}$ it's necessary to calculate both modules b and c before calculating module d. Note that Axiom 1 requires the same run sequence

It's directly follows from a cycle definition that virtual modules (input and output) can never belong to a cycle because input module has not incoming edges and output module has not outgoing edges. Moreover, according to the conditions imposed on a modular network graph, any cycle must have both incoming and outgoing external edges.

Definition 4: **First** vertex f of a cycle $C(V, E)$ we'll denote such vertex that $d_f = \min_{v \in V} \min_{w = w(I,v)} d(w)$, where $d(w)$ is a length of walk *w.* **Last** vertex l of a cycle is such vertex that $d(w(l, f)) = 1$.

Given definitions do not guarantee the uniqueness of the vertices but nevertheless such definition allows classifying cycles.

Lets introduce definitions for two types of cycles that fundamentally important from a point of modular networks' run.

Definition 5: Ordinary cycle with first vertex a is a cycle $Co_a(V_a, E_a)$ where there are no pair of vertices that belong to each other projections onto input, i.e. $\forall v_1, v_2 \in V_a : v_1 \in P(I, v_2) \quad \Rightarrow \quad v_2 \notin P(I, v_1)$. **Crossed** cycle with first vertex b is a cycle $Cc_b(V_b, E_b)$ that contains at least one pair of vertices belonging to each other projections onto input, i.e. $\exists v_1, v_2 \in V_b : v_1 \in P(I, v_2) \quad \& \quad v_2 \in P(I, v_1)$.

Introducing of term "ordinary cycle" is necessary for order determination of run sequences for modular network, and first of all for separation of all possible trigger schemes (picture 1). It can be easily shown that cycle on the picture 1 is crossed and has two first vertices.

Properties of cycles

The development of algorithms for automatic analysis is substantively simplified if such algorithms are built on the basis of inherent cycles' properties.

Proposition 1: For each vertex v contained in a cycle C_f one can find at least one walk $w(v,v)$ that contains first vertex f of a cycle.

This proposition is trivial but not always evident because cycle can contain arbitrary number of walks buy our definition.

Proof: Let v belongs to C_f, then

$\exists\, w(f,f) = w(f,v) \bigcup w(v,f) \quad \Rightarrow \quad \exists\, w(v,v) = w(v,f) \bigcup w(f,v)$ ■

Theorem 1: For every two vertices a and b of a cycle $C_f(V_f, E_f)$, projection $P(a,b)$ always exists and this projection is entirely belongs to a cycle:

$\forall a,b \in V_f : \exists\, P(a,b) \subseteq C_f$

Proof: Lets show that as long as vertices a and b belong to the cycle then at least one walk $\tilde{w}(a,b)$ exists. From Proposition 1, it is follows that

$a \in C_f \quad \Rightarrow \quad \exists\, w(a,a) = w(a,f) \bigcup w(f,a) \quad \Rightarrow \exists\, w(a,f)$

Then if $b \in w(a,f) \Rightarrow w(a,f) = \tilde{w}(a,b) \bigcup \tilde{w}(b,f)$. That is walk from a to b exists. If vertex b does not belong to the walk from a to f then as long as $b \in C_f \quad \Rightarrow \quad \exists\, w(b,b) = w(b,f) \bigcup w(f,b)$, therefore walk $\tilde{w}(a,b) = w(a,f) \bigcup w(f,b)$ exists.

Let walk $\tilde{w}(a,b)$ is not an elementary path. That is vertex $\tilde{v}$ is included in this walk at least two times:

$\tilde{w}(a,b) = \{a, v_1 ... v_i, \tilde{v}, ..., \tilde{v}, v_{i+k}, ..., v_n\}, v_{n+1} = b$. Then it is possible to construct a path from a to b:

$w(a,b) = \{a, v_1 ... v_i, \tilde{v}, v_{i+k}, ..., v_n\}, v_{n+1} = b$. That is at least one path exists $w(a,b)$ and $P(a,b) \neq \varnothing$.

Let's show that all paths from a to b belong to the cycle. Since both vertices belong to the cycle then $a \in C_f \Rightarrow \exists\, w_1(f,a)$ and $b \in C_f \Rightarrow \exists\, w_2(b,f)$. Then walk $\exists\, w(f,f) = w_1(f,a) \bigcup w(a,b) \bigcup w_2(b,f)$ exists. Hence, every path $w(a,b)$ belongs to the cycle as it contains all possible walks from first vertex to itself by the definition. ■

Theorem 2: (Sign of cycle existence) Graph of a modular network will contain cycle (cycles) if this graph contains at least one vertex having at least one edge not belonging to the projection of this vertex onto input.

Proof: Let network digraph has vertex a having incoming edge $e(b,a)$ not belonging to the projection of this vertex onto input. According to projection definition, this means that for the vertex b, from which this edge comes out, there are no walks from network input not containing vertex a. As by the conditions imposed on a graph, every vertex is accessible from the input, there is at least one walk $w(I,b)$ for the vertex b. Hence, if there is edge $e(b,a)$, then vertex a belongs to every walk

$$w(I,b) : \exists\, e(b,a) : e \notin P(I,a) \quad \Leftrightarrow \quad a \in w(I,b) \quad \forall\, w(I,b).$$

Then at least one walk from vertex a to itself exists: $w(a,a) = w(a,b) \bigcup e(b,a)$ and, hence, cycle exists. ■

It should be noticed that sign of cycle existence proposed in the Theorem 2 gives only sufficient condition for cycle existence. It allows finding all ordinary cycles in a modular network but only some crossed cycles.

Let's consider some important properties of ordinary cycles.

Theorem 3: (Theorem on uniqueness of ordinary cycle). If ordinary cycle exists, then the first vertex of this cycle can be selected uniquely.

Proof: Lets carry out the proof ex adverso. Let $C_{f1}(V_{f1}, E_{f1})$ is an ordinary cycle. Let there is two first vertices in this cycle, i.e. $\exists\, f2 \in V_{f1}, f2 \neq f1 : \quad d_{f1} = d_{f2} = \min_{w(I,v)}(d_v) \quad \forall v \in V_{f1}$ (from the definition of first vertex).

As minimal distances from the network input to the first vertices are equal then for the vertex $f1$ there is a walk from the network input not containing vertex $f2$, i.e. $\exists\, w_1(I, f1) : f2 \notin w_1$. In the same way, there is a walk $\exists\, w_2(I, f2) : f1 \notin w_2$, because $f2$ is also a first vertex. From the other side, as long as vertices $f1$ and $f2$ both belong to the cycle then according to the Proposition 1 there are walks $w_3(f1, f2)$ and $w_4(f2, f1)$. And according to a walk definition $f2 \notin w_3$ and $f1 \notin w_4$. Then next walks exist:

$(f1 \notin w_2 \;\&\; f1 \notin w_4) \Rightarrow \exists\, w(I,f1) = w_2(I,f2) \bigcup w_4(f2,f1)$ and

$(f2 \notin w_1 \;\&\; f2 \notin w_3) \Rightarrow \exists\, w(I,f2) = w_1(I,f1) \bigcup w_3(f1,f2)$.

The next follows from the definition of projection of a vertex onto input: $f1 \in P(I,f2) \quad \& \quad f2 \in P(I,f1)$. That is cycle $C_{f1}(V_{f1},E_{f1})$ is crossed by the definition and therefore there is only one first vertex in an ordinary cycle. ■

Consequence 3.1: If ordinary cycle Co_f exists: $V_f \neq 0$, then for a given set of vertices it can be determined uniquely.

Indeed, as long as all possible walks from first vertex of a cycle to itself are contained in the cycle by the definition, then every cycle is determined accurate to a first vertex. Since first vertex in an ordinary cycle is unique then cycle is determined unambiguously.

Theorem 4: There is the only vertex in an ordinary cycle connected by incoming edges with vertices outside the cycle and this vertex is the first vertex.

$v \in Co_f: \quad \exists\, e(p,v),\, p \notin Co_f \Rightarrow v = f$

Proof: Lets carry out the proof ex adverso. Let $C_f(V_f,E_f)$ is an ordinary cycle and there is a vertex v (except the first vertex) having outer incoming edge $e(p,v)$, i.e. $\exists\, v \in V_f, v \neq f: \quad \exists\, e(p,v),\, p \notin V_f$. Since every vertex is accessible from the input then walk $w(I,p)$ exists. Lets show that if vertex p does not belong to the cycle then for every walk $\forall\, w(I,p),\, p \notin V_f \Rightarrow f \notin w$.

Let walk exists that contains first vertex of the cycle: $\exists\, w(I,p): f \in w \quad \Rightarrow \quad w(I,p) = w(I,f) \bigcup w(f,p)$. Correspondingly, walk $w(f,p)$ exists. According to the initial assumption $v \in V_f \Rightarrow \exists\, w(v,f)$ and edge $e(p,v)$ exists. Hence, walk from the first vertex to itself through vertex p exists: $w(f,f) = w(f,p) \bigcup e(p,v) \bigcup w(v,f)$. Therefore, $p \in V_f$ that conflicts with initial condition.

As first vertex of the cycle does not belong to a walk from the input to the vertex p then there is a walk from the network input to the vertex v not containing first vertex f: $w_2(I,v) = w(I,p) \bigcup e(p,v)$. Also there is a walk $\exists\, w_1(I,f): v \notin w_1$, because f is the first vertex. On the other hand, both vertices f and v belong to the cycle and according Proposition 1 there are walks $w_3(f,v)$ and $w_4(v,f)$. According to a walk definition $v \notin w_3$ and $f \notin w_4$. Then next walks exist:

$(f \notin w_2 \;\&\; f \notin w_4) \Rightarrow \exists\, w(I,f) = w_2(I,v) \bigcup w_4(v,f)$ and

$(v \notin w_1 \;\&\; v \notin w_3) \Rightarrow \exists\, w(I,v) = w_1(I,f) \bigcup w_3(f,v)$.

From the existence of such walks and definition of a projection of a vertex onto input the next follows: $f \in P(I,v) \quad \& \quad v \in P(I,f)$. Because of this, the cycle $C_f(V_f,E_f)$ is crossed that conflicts with condition. Hence, there is the only vertex connected by incoming edges with vertices outside the ordinary cycle. ■

Consequence 4.1: For an ordinary cycle projection of first vertex into network input does not contain any vertex of the cycle.

Proof: Let there is a vertex in the ordinary cycle $C_f(V_f,E_f)$ contained in the project of its first vertex onto input: $\exists v \in V_f: \; v \in P(I,f)$. From the definition of a projection of a vertex onto input we'll find that walk $w(I,f): v \in w \;\&\; f \notin w$ exists and correspondingly this walk can be broken down into two: $w(I,f) = w(I,v) \bigcup w(v,f) \quad \Leftrightarrow \quad \exists w(I,v): f \notin w$. Also, there is the shortest walk $\exists\, w(I,f): v \notin w$, because f is the first vertex. By analogy with the proof of the Theorem 4, we'll find that the cycle is crossed. ■

Consequence 4.1: For an ordinary cycle C_f, all paths from the input contain the first vertex for every vertex $v \neq f$, i.e.:

$$P(I,v) = P(I,f) \bigcup P(f,v) \quad \forall v \in C_f$$

Proof: As every vertex is accessible from the input and according to the Theorem 4 none of the cycle's vertices have outside edges (except the first); therefore all paths (walks) from the input to vertices of the ordinary cycle contain the first vertex. ■

Conclusion and Further Work

Given definitions are related mainly to the architectures of modular networks containing cycles. It's obvious that it's necessary to have some assumptions about rules and sequence of cycle's traversal for defining the run sequence of such architectures. Properties of considered types of cycles could be used as a basis of such rules.

Besides the classification based on architectures' types, cycles vary in types of processed data and used algorithms of neural networks. Cycles' types regarded above allow to take into consideration the specificity of processed data and neural algorithms in the form of corresponding modular architectures.

General cycles' properties, considered in this article, allow formal analyzing of modular neural networks. Use of cycle's properties allows constructing algorithms for analysis of architectures with automatic system. Moreover, automatic search of all modules contained in a cycle allows formal attaching of some properties to every cycle. This turn out to be very helpful in CAD systems where user can, first, specify run sequence and, second, define each cycle as recurrent or iterative.

Data processing with a help of modular neural network means that run sequence is determined. The definitions given above are assigned, first of all, for derivation of rules for automatic construction of run sequence and for the controlling of architectures created by user. Cycles' properties allow determining, which architectures of modular networks are valid without introducing additional conditions.

Proposed theory was successfully used in the subsystem of automatic analysis of modular architectures in the MNN CAD [2]. Formal definitions of valid and forbidden architectures of modular networks and ways of settlement of contradictions in forbidden architectures, algorithms of modular networks analysis, based on the properties of proposed graphs, will be subject of future works.

Bibliography

A. Galinskaya. Modular neural networks: the review of modern state of the developments. In: Mathematical Machines and Systems.-2003. -№ 3,4. –C.87-102.

1. M. Kussul, A. Riznyk, E. Sadovaya, A. Sitchov, Tie-Qi Chen. “A visual solution to modular neural network system development”, Proceedings of the 2002 International Joint Conference on Neural Networks IJCNN'02, Honolulu, HI, USA, 12-17 May 2002, vol.1
2. M. Kussul, A. Galinskaya. Comparative study of modular classifiers and its training // Proc. of the X-th International Conference “Knowledge – Dialog - Solution” KDS 2003, Varna, Bulgaria, 16-26 June 2003. –P.168-174.

A. Galinskaya. Architectures and training of modular classifiers for applied tasks In: Mathematical Machines and Systems. -2003. -№ 2. –C.77-86.

3. R. Diestel. Graph Theory. 2nd edition, Springer-Verlag, New York, 2000, P. 312

Authors' Information

Michael Kussul – PhD., Institute of Mathematical Machines and Systems, NASU, Glushkova ave. 42, 03680, Kiev, Ukraine. kussul@nnteam.org.ua

Alla Galinskaya – PhD. student, Institute of Mathematical Machines and Systems, NASU, Glushkova ave. 42, 03680, Kiev, Ukraine. alla@nnteam.org.ua

ГЕТЕРОГЕННЫЕ ПОЛИНОМИАЛЬНЫЕ НЕЙРОННЫЕ СЕТИ ДЛЯ РАСПОЗНАВАНИЯ ОБРАЗОВ И ДИАГНОСТИКИ СОСТОЯНИЙ[1]

Адиль В.Тимофеев

Аннотация: *Рассмотрены различные параллельные архитектуры и методы самоорганизации и минимизации сложности гетерогенных полиномиальных нейронных сетей (ПНС) в задачах распознавания образов и диагностики состояний. Получены конструктивные оценки степени гетерогенности и параллелизма в процессе автономного принятия классифицирующих решений с помощью ПНС различных типов. Показано, что параллелизм, самоорганизация и робастность гетерогенных ПНС могут значительно возрасти при коллективном (мульти-агентном) решении сложных задач распознавания образов, анализа изображений, развернутой (векторной) диагностики состояний и адаптивной маршрутизации информационных потоков*

Введение

Одним из наиболее эффективных средств массового распараллеливания и ускорения процессов обработки и передачи потоков данных в задачах распознавания образов, классификации данных и диагностики состояний являются искусственные нейронные сети (НС) и нейросетевые технологии. Естественным прототипом искусственных НС является биологический мозг и центральная нервная система человека и животных как сложная гетерогенная нейронная сеть, обеспечивающая высокую степень параллелизма, самоорганизации и робастности при решении различных интеллектуальных задач (распознавание образов, классификация данных, поиск закономерностей, анализ изображений, диагностика состояний, прогнозирование явлений и т.п.). Возможности искусственных и биологических НС могут значительно расшириться при коллективном (мульти-агентном) решении сложных интеллектуальных задач.

1. Основные идеи и принципы построения гетерогенных полиномиальных нейронных сетей

Высокая сложность и размерность многих задач распознавания образов, классификации данных, анализа изображений и диагностики состояний, а также часто возникающая необходимость их решения в реальном времени требуют массового параллелизма и самоорганизации распределённых вычислений на базе НС. С этой точки зрения особый интерес представляют гетерогенные полиномиальные нейронные сети (ПНС) с самоорганизующейся архитектурой и их разновидности, предложенные в работах [1–9].

Основные идеи, математические модели, методы обучения и принципы самоорганизации гетерогенных ПНС были описаны в [1–3] и развиты в [4–9]. К ним прежде всего относится следующее:

- архитектура ПНС гетерогенна и многослойна;
- наличие слоя полиномиальных нейронных элементов (П-нейронов);
- возможность и целесообразность самоорганизации архитектуры ПНС различных типов;
- детерминированные и вероятностные методы обучения и самоорганизации гетерогенных ПНС;
- принципы минимальной сложности и высокой экстраполяции гетерогенных ПНС;
- алгебраическое требование диофантовости (целочисленности синаптических весов) гетерогенных ПНС.

В процессе дальнейшего развития теории гетерогенных ПНС были предложены модели многозначных нейронных элементов (М-нейронов) и связанных с ними конъюнктивных, полиномиальных, дизъюнктивных и суммирующих нейронных элементов (МК-, МП-, МД- и МΣ-нейронов), а также новые разновидности гетрогенных ПНС (генно-нейронные сети, квантовые, нейронные сети, мульти-агентные ПНС и т.п.), описанные в [4–14].

В настоящей статье анализируются гетерогенность архитектуры, возможности самоорганизации распределённых вычислений и степень параллелизма ПНС различных типов, предназначенных для распознавания образов, диагностики состояний и решения других интеллектуальных задач (data mining, knowledge discovery и т.п.). Полученные теоретические (априорные) оценки степени параллелизма и самоорганизации гетерогенных ПНС подтверждаются экспериментальными результатами решения прикладных задач. Значительный интерес представляют также новые проблемы организации коллективных (мульти-агентных) решений на базе гетерогенных ПНС.

2. Задачи классификации и идентификации образов и их обобщение

ПНС с гетерогенной архитектурой предназначены для решения сложных интеллектуальных задач. Примерами таких задач могут служить задачи распознавания образов, классификации данных, диагностики состояний, идентификации объектов и т.п. Обычно эти задачи формулируются следующим образом.

Пусть задано конечное множество объектов $\{\omega\} = \Omega$ и существует (но неизвестно) его разбиение на К непустых подмножеств (классов) вида

$$\Omega = \bigcup_{k=1}^{K} \Omega_k, \Omega_k \neq \varnothing, \Omega_k \cap \Omega_j = \varnothing \text{ при } k \neq j. \quad (1)$$

Это означает, что существует неизвестная классифицирующая функция R(ω), ставящая в соответствие каждому объекту $\omega \in \Omega$ номер класса k, к которому он принадлежит, т.е.

$$R(\omega) = k, \omega \in \Omega_k, k = 1,2,...,K. \quad (2)$$

Кроме того, существует множество неизвестных идентифицирующих функций вида

$$R_k(\omega) = \begin{cases} 1, \text{если} \omega \in \Omega_k, \\ 0 - \text{в противном случае}, \end{cases} \quad (3)$$

являющихся характеристическими функциями классов Ω_k, k = 1,2,…,K, которые отделяют все объекты $\omega \in \Omega_k$ от остальных.

Предположим, что имеется некоторая измерительная система (датчики информации, сенсоры, измерительные приборы и т.п.), которая в результате измерения свойств или определения характеристик любого объекта $\omega \in \Omega$ однозначно ставит в соответствие объекту ω его информационное описание в виде вектора признаков $\mathrm{x}(\omega) = \left| x_i(\omega) \right|_{i=1}^{n}$. Тогда векторная функция x: $\Omega \to R^n$ определяет отображение множества объектов Ω в n-мерное пространство признаков R^n. Это отображение порождает разбиение описаний объектов в пространстве признаков R^n на классы (образы) вида

$$X_k = x(\Omega_k), k = 1,2,...,K.$$

Отображение x: $\Omega \to R^n$ будем называть корректным или информативным, если выполняются следующие условия:

1) если $x(\omega_i) = x(\omega_j)$, то $R(\omega_i) = R(\omega_j)$, (4)

т.е. объекты ω_i и ω_j с одинаковым описанием принадлежат одному классу,

2) если $R(\omega_i) \neq R(\omega_j)$, то $x(\omega_i) \neq x(\omega_j)$, (5)

т.е. объектам ω_i и ω_j из разных классов соответствуют различные описания.

Таким образом, корректное отображение объектов в пространство признаков не приводит к потере информации о классах. Это информативное отображение определяет следующее разбиение описаний объектов на классы (образы)

$$\mathrm{X} = \bigcup_{k=1}^{K} \mathrm{X}_k, \ \mathrm{X}_k = \mathrm{x}(\Omega_k) \neq \varnothing. \quad (6)$$

Из соотношений (1) – (6) следует, что если $\omega \in \Omega_k$, то $x(\omega) \in X_k$, и наоборот. В том случае, когда $|\Omega| > |X|$, корректное (информативное) отображение $x(\omega)$ является сжимающим.

Будем считать, что отображение $x: \Omega \to R^n$ является информативным и, возможно, сжимающим. Тогда задачи классификации и идентификации образов сводятся к восстановлению (определению) неизвестных решающих функций вида (2) и (3). При этом единственной доступной информацией о классах (1) являются табличные базы данных вида

$$\{x(\omega_h), R(\omega_h)\}_{h=1}^{m} \text{ или } \{x(\omega_h), R_1(\omega_h), R_2(\omega_h), ..., R_K(\omega_h)\}_{h=1}^{m}, \quad (7)$$

называемые обучающей выборкой.

Мощность этой выборки $m \geq K$ должна быть достаточно большой, т.е. множество (7) должно быть репрезентативным. Поэтому будем предполагать, что множество (7) содержит достаточную информацию об априорном разбиении объектов (1) и их описании (6) на классы (образы). В частности, важно, чтобы обучающее множество (7) было полным, т.е. содержало хотя бы по одному объекту из каждого класса, и непротиворечивым, т.е. не содержало одинаковых описаний, относящихся к объектам из разных классов.

Наряду с классическими задачами классификации и идентификации образов с помощью скалярных решающих функций вида (2) и (3), значительный интерес для практики представляет их обобщение на случай векторного распознавания образов и диагностики состояний. Примером таких задач могут служить сложные задачи медицинской диагностики, когда требуется для каждого больного $\omega \in \Omega$ с вектором симптомов признаков $x(\omega)$ определить не только диагноз, т.е. класс заболеваний Ω_k, к которому он относится, но и дать его "расшифровку", т.е. найти ряд уточняющих и детализирующих характеристик в виде вектора

$$z(\omega) = |z_j(\omega)|_{j=1}^{q}, \quad (8)$$

где $z_j(\omega)$ - многозначные предикаты, принимающие целочисленные значения в интервале $[0, p_j]$.

Формально развернутый (детализирующий) диагноз можно представить в виде (q+1) – мерного вектора

$$L(\omega) = |R(\omega), z_1(\omega), z_2(\omega), ..., z_q(\omega)|, \quad (9)$$

Компонентами этого вектора является классифицирующий предикат $R(\omega)$ с целочисленными значениями (кодами) основного диагноза из интервала [1,2,...,K] и детализирующие этот диагноз дополнительные многозначные предикаты $z_j(\omega)$, j= 1,2,...,q, каждому из которых можно поставить в соответствие локальную "расшифровку" основного диагноза на естественном языке.

При обобщенной векторной классификации образов главный предикат $R(\omega)$ описывает априорное разбиение множества Ω на классы (1), а дополнительные предикаты $z_j(\omega)$ определяют разбиение каждого класса Ω_k на подклассы вида:

$$\Omega_k = \bigcup_{j=1}^{q} \Omega_{k,j} . \quad (10)$$

Предикаты $R(\omega)$ и $z_j(\omega)$ заранее неизвестны. Однако известны из значения на обучающей выборке вида

$$\{x(\omega_h), R(\omega_h), z_1(\omega_h), z_2(\omega_h), ..., z_q(\omega_h)\}_{h=1}^{m}. \quad (11)$$

По этим данным требуется восстановить (определить) неизвестную вектор-функцию (9) и её компоненты (2) и (8). Некоторые методы решения этой задачи векторной (расширенной) классификации образов на базе обучения и самоорганизации гетерогенных ПНС предложено в работах [11 – 13].

3. Нейросетевые архитектуры и алгоритмы

В настоящее время известно много математических и эвристических подходов к решению сформулированных задач классификации и идентификации образов. Среди них важную роль играют нейросетевые подходы, основанные на синтезе различных моделей (архитектур) и алгоритмов обучения НС для распознавания образов и диагностики состояний.

Однако во многих случаях эти НС имеют гомогенную архитектуру. При этом заранее не известно, какое число слоёв и нейроэлементов необходимо для решения задачи. Алгоритмы обучения таких гомогенных НС не всегда сходятся к решению задачи за конечное число шагов.

Примерами гомогенных НС могут служить однослойные НС Хопфилда или Хемминга или многослойные перцептроны, использующие пороговые или сигмоидальные нейроэлементы. Популярные сегодня градиентные алгоритмы обучения гомогенных НС типа Backpropagation и его модификации (Quick propagation, Rpro и т.п.) медленно сходятся или вообще не приводят к решению за конечное время.

Описываемые ниже гетерогенные ПНС различных типов являются эффективным средством восстановления (определения) неизвестных классифицирующих и идентифицирующих функций (2) и (3) по обучающим базам данных (7) и их реализации на базе ПНС с самоорганизующейся архитектурой минимальной сложности. В основе теории этих гетерогенных ПНС лежат идеи и принципы, сформулированные выше, а также в работах [1 – 9].

Предлагаемые гетерогенные ПНС позволяют также решать обобщенные задачи векторной классификации и описания образов, т.е. восстанавливать (определять) неизвестные векторные функции (9) по расширенным обучающим выборкам вида (11). Кроме того, они могут успешно использоваться в качестве нейросетевых агентов при коллективном (мульти-агентном) решении сложных (глобальных) задач распознавания образов, расширенной (векторной) диагностики и адаптивной маршрутизации информационных потоков [7 –14].

4. Гетерогенность, параллелизм и самоорганизация порогово-полиномиальных нейронных сетей

Архитектура ПНС гетерогенна и представляет собой последовательность нескольких однородных слоёв (непересекающихся подмножеств) параллельно работающих нейроэлементов (НЭ) различных типов. В различных слоях гетерогенной ППС могут использоваться разные НЭ, но каждый слой (подмножество НЭ) является однородным (гомогенным). При этом обработка информации в каждом слое НЭ осуществляется параллельно.

Каналы связи между предыдущим и последующим слоями гетерогенной ПНС являются однонаправленными (односторонними) и имеют регулируемые веса (синаптические параметры). Эти веса каналов связи настраиваются в процессе обучения и самоорганизации архитектуры ПНС по имеющимся экспериментальным данным или прецедентам вида (7), называемым обучающей базой данных (ОБД).

Опишем формально гетерогенную архитектуру трёхслойной порогово-полиномиальной нейронной сети (ППНС) [1-4] и рассмотрим принципы её самоорганизации.

Первый слой ППНС состоит из n пороговых нейроэлементов (НЭ), на вход которых параллельно поступают сигналы $y_1(\omega)$, …, $y_p(\omega)$, характеризующие различные свойства объекта ω, а на выходе формируется вектор двоичных сигналов (бинарный код) $x(\omega)=\left|x_i\right|_{i=1}^{n}$, где $x_i=x_i(y_1,\ldots y_p)$. Этот бинарный код является выходом НЭ первого слоя ППНС.

На входы каждого “ассоциативного” НЭ второго слоя поступает вектор $x(\omega)$ бинарных сигналов с “рецепторных” (кодирующих) пороговых НЭ первого слоя. Эти “ассоциативные” НЭ реализуют полиномиальные преобразования $a_j(x)$ входных сигналов и называются П-нейронами или мультипликативными НЭ. Выходом второго слоя ППНС является вектор двоичных сигналов $a(x)=\left|a_j(x)\right|_{j=1}^{N}$.

Третий слой ППНС состоит из “решающих” пороговых НЭ, на вход которых поступает вектор выходных сигналов a_j(x) полиномиальных НЭ второго слоя. Затем эти сигналы умножаются на синаптические веса u_j, суммируются и преобразуются в выходные сигналы НЭ третьего слоя в виде суперпозиций функций

$$R_k(\omega) = sign\ (\sum_{j=1}^{N} u_{k,j} a_j(\mathrm{x}(y(\omega)))),\ k = 1, \ldots \mathrm{K}, \tag{12}$$

где $u_{k,j}$ – настраиваемые синаптические параметры, а K – число непересекающихся классов (образов).

Каждая решающая (идентифицирующая) функция $R_k(\omega)$ вида (8) является характеристической функцией k-го класса. Поэтому она отделяет объекты k-го класса от остальных. Множество таких идентифицирующих функций (8) решает задачу классификации образов.

Принцип самоорганизации ППНС с гетерогенной архитектурой заключается в построении “ассоциативных” полиномов a_j(x) непосредственно по обучающей базе данных (ОБД) (7) в виде одночленов [1–3]

$$a_j(\mathrm{x}(y)) = \prod_{i=1}^{n} x_i^{\,x_i(y_j)}(y),\ j = 1,\ldots,m \tag{13}$$

Здесь m – число элементов (мощность) ОБД, определяющее оценку сверху на число N необходимых полиномиальных НЭ второго слоя ППНС, т.е. $m \geq \mathrm{N} \geq \mathrm{K}$.

Чтобы минимизировать сложность ППНС, нужно найти в идентифицирующих функциях вида (8) векторы синаптических параметров $u_k = \left|u_{k,j}\right|_{j=1}^{N}$ с максимальным числом нулевых компонент. Это приводит к автоматической ликвидации неинформативных (избыточных) синаптических связей в гетерогенной ППНС без потери точности принимаемых решений на ОБД (7).

Быстрые рекуррентные алгоритмы обучения и минимизации сложности гетерогенной ППНС были предложены в [1–4]. Они осуществляют отображение ОБД вида (7) на множество синаптических параметров пороговых НЭ третьего слоя ППНС, причём число шагов алгоритма обучения $r \leq m$.

Степень параллелизма при распознавании и идентификации образов определяется тем, что принятие решений осуществляется ППНС за 3 такта одновременных вычислений в каждом слое НЭ независимо от размерности решаемой задачи $D=n\times m\times K$.

Самоорганизация ППНС обеспечивается тем, что НЭ второго слоя формируются согласно (9) непосредственно по ОБД (7).

Гетерогенность архитектуры ППНС определяется тем, что первый и третий слой состоят из пороговых НЭ, а второй слой – только из полиномиальных НЭ.

5. Гетерогенность, параллелизм и самоорганизация диофантовых и сплайновых нейронных сетей

Гетерогенные ПНС с целочисленными синаптическими параметрами называются диофантовыми НС [1,5]. Примерами диофантовых НС с самоорганизующейся архитектурой, могут служить гетерогенные ПНС, предложенные в [1–3].

Рассмотрим другой класс диофантовых НС. Гетерогенность их архитектуры определяется использованием слоёв, состоящих из полиномиальных или сплайновых НЭ.

Первый слой диофантовой НС состоит из пороговых НЭ. Выходом первого слоя НЭ является двоичный образ (описание) x(ω) объекта ω.

НЭ второго слоя кодируют двоичные образы x(ω) натуральными числами вида

$$d(\omega) = \sum_{i=1}^{n} x_i(\omega) \cdot 2^{n-i}. \tag{14}$$

Для упрощения алгоритмов обучения и самоорганизации гетерогенной архитектуры диофантовой НС построим НЭ третьего слоя в виде ортогональных полиномов (одночленов) вида

$$a_j(d(\omega))=\prod_{\substack{h=1\\ j\neq h}}^{m_k}(d(\omega)-d(\omega_h))(d(\omega_j)-d(\omega_h))^{-1}. \quad (15)$$

Другие способы построения НЭ третьего слоя предложены в [5].

Четвёртый слой гетерогенной диофантовой НС состоит из “решающих” пороговых НЭ с синаптическими параметрами, настраиваемыми по ОБД (7) согласно быстрым алгоритмам обучения, предложенным в [5].

Описанные диофантовые НС целесообразно применять на сравнительно коротких ОБД вида (7). Чтобы снять это ограничение, рассмотрим гетерогенные сплайновые НС.

Сплайновые НС основаны на кусочно-полиномиальной или сплайновой аппроксимации решающих (идентифицирующих) функций вида (8). Идея состоит в синтезе НЭ по ОБД (7) в виде независимых друг от друга полиномов или сплайнов на каждой паре натуральных чисел $\{d(\omega_i),d(\omega_{i+1})\}$ и их коммутации по определённым правилам в четвертом слое НС, состоящем из пороговых НЭ с настраиваемыми по ОБД (7) синаптическими параметрами [5].

Синтезированные диофантовые и сплайновые НС имеют четырёхслойную гетерогенную архитектуру, описываемую суперпозицией параллельных преобразований в каждом слое НЭ.

Самоорганизация и минимизация сложности гетерогенной архитектуры таких НС обеспечиваются самонастройкой полиномиальных и сплайновых НЭ второго и третьего слоёв (например, вида (10) и (11)) и быстрыми алгоритмами обучения пороговых НЭ четвертого слоя, требующими только однократного использования ОБД вида (7), т.е. $r \le m$.

При распознавании образов и диагностике состояний принятие решений с помощью описанных диофантовой или сплайновой ПНС осуществляется за 4 такта параллельных вычислений в каждом слое НЭ независимо от размерности задачи D = n × m × K.

Гетерогенность архитектуры диофантовых и сплайновых НС определяется тем, что первый и четвертый слой состоят из пороговых НЭ, а второй и третий слой включают полиномиальные или сплайновые НЭ.

6. Гетерогенность, параллелизм и самоорганизация классифицирующих полиномиальных нейронных сетей

Рассмотрим четырёхслойную ПНС, предназначенную для классификации образов.

Первый слой этой ПНС состоит из функциональных НЭ (F-элементов) $F_i(y(\omega),\gamma_i)$ с синаптическими параметрами γ_i.

Второй слой этой гетерогенной ПНС состоит из полиномиальных НЭ (П–элементов) вида

$$a_j(\omega)=\prod_{i\in I_j}F_i(y(\omega),\gamma_i), \quad (16)$$

где $F(y,\gamma)=0$ при $y<\gamma$ и $F(\gamma,\gamma)\neq 0$ [6,9].

Третий слой ПНС состоит из одного суммирующего НЭ (Σ –нейрона).

Четвёртый слой ПНС состоит из одного многозначного НЭ (М-нейрона), описываемого К-значным предикатом М, принимающем значения 1,2,…,К и определяющим принадлежность объекта ω к одному из классов Ω_k.

Четырёхслойная архитектура ПНС реализует следующее последовательно-параллельное преобразование вектора входных сигналов $y(\omega)$ в выходной целочисленный сигнал $R(\omega,u,\gamma)$ вида

$$R(\omega,u,\gamma)=M(u_0+\sum_{j=1}^{N}u_j\prod_{i\in I_j}F_i(y(\omega),\gamma_i)), \quad (17)$$

определяющий номер класса, к которому будет отнесён распознанный объект ω.

Задачи обучения, минимизации сложности и самоорганизации гетерогенных классифицирующих ПНС с аналитическим описанием вида (16),(17) заключаются в определении скалярных функций y,F_i и векторов

синаптических параметров $\gamma = \left|\gamma_i\right|_{i=1}^{I_j}$ и $u = \left|u_i\right|_{j=1}^{N}$ по ОБД (7) таким образом, чтобы обеспечивалась не только безошибочная классификация объектов из ОБД (7), но и других распознаваемых (контрольных) объектов. Для решения этой задачи нужно конструктивно задать функциональные НЭ $F_i(y, \gamma_j)$ и векторы синаптических параметров γ и u с возможно меньшим числом ненулевых компонент [5-7].

Степень параллелизма в классифицирующих ПНС, описываемых многозначным предикатом (17) определяется тем, что распознавание образов, классификация данных или диагностика состояний осуществляются за 4 такта параллельных вычислений в НЭ разных слоёв независимо от сложности решаемой задачи $D = n \times m \times K$.

Гетерогенность архитектуры этих ПНС характеризуется тем, что первый слой состоит из функциональных НЭ, второй слой включает полиномиальные НЭ, а третий и четвертый слои содержат по одному суммирующему и многозначному НЭ.

7. Гетерогенность и параллелизм в генно-нейронных сетях с самоорганизующейся архитектурой

Пусть имеется популяция $\{\omega\} = \Omega$ особей ω, каждая из которых может принадлежать одному из К образов (классов) согласно разбиению (1). Особь ω характеризуется набором признаков x (ω), которые принимают значения, соответствующие одному из дискретных состояний генов $x_i(\omega)$. При этом состояния каждого гена описываются многозначным предикатом.

Вектор $x(\omega) = \left|x_i(\omega)\right|_{i=1}^{n}$ состояний генов назовём хромосомой (локальным описанием) особи ω. Будем говорить, что две особи имеют одинаковый генотип, если у них совпадают состояния всех генов, т.е.

$$x_i(\omega_j) = x_i(\omega_h),\ \forall i = 1,2,...,n, j \neq h. \tag{18}$$

Множество хромосом образует класс генотипов, соответствующих некоторому генетическому образу с характеристической функцией $R_k(\omega)$ вида (3).

Задачи генетического описания и распознавания образов сводятся к аппроксимации неизвестных функций $R(\omega)$, $R_k(\omega)$ по генетическим (целочисленным) ОБД вида (7). Для решения этих задач можно использовать трёхслойную гетерогенную генно-нейронную сеть (ГНС) полиномиального типа, описанную в [7].

Обученные ГНС обеспечивают массовый параллелизм при обработке генетических данных, а именно: принятие оптимальных решений при распознавании образов или диагностике состояний осуществляется в ГНС за 3 такта параллельных вычислений в НЭ каждого слоя независимо от размерности задачи $D = n \times m \times K$.

Гетерогенность архитектуры ГНС проявляется в том, что первый и третий слои состоят из пороговых НЭ, а второй слой содержит полиномиальные НЭ.

Другой метод обучения и самоорганизации многослойных ГНС основан на логико-вероятностных генетических алгоритмах селекции информативных генов и конъюнктивных НЭ (К-нейронов) вида

$$a_j^n(\omega) = \mathop{\&}_{i=1}^{n} \xi_i^{\xi_i(\omega_j)},\ j = 1,\ldots, m, \tag{19}$$

а также логико-вероятностных решающих (идентифицирующих) правил импликативного типа

$$\{a_j^r(\omega) \xrightarrow{P_k} R_k(\omega)\}_{i=1}^{N}, 1 \le r_1 \le \ldots \le r_N < n. \tag{20}$$

Здесь P_k - максимальная апостериорная вероятность принадлежности особи ω к k-ому классу Ω_k, оцененная по ОБД вида (7).

Синтезированные генетические решающие (идентифицирующие) правила в случае, когда i-ый ген может иметь только два состояния, можно представить в виде многослойной ГНС минимальной сложности, представляющий собой бинарное “дерево решений”. Ветвям этого дерева соответствует К-нейроны вида (19), а листьям - номера классов генотипов [3,7].

В работах [7 –9] представлено обобщение описанных ГНС на случай, когда каждый ген имеет не два, а произвольное число дискретных состояний, т.е. ген описывается многозначным предикатом вида

$$x_i(\omega) \in \{0,1,...l_i\}, l_i \geq 2. \quad (21)$$

Для увеличения параллелизма в процессе принятия оптимальных (в смысле критерия Байеса) решений описанные многослойные ГНС с древовидной архитектурой можно преобразовать в трёхслойные диофантовые ГНС с целочисленными синаптическими весами с помощью методов, описанных в [3,7]. В этом случае распознавание образов или диагностика состояний произойдут не за $r \leq r_N$ шагов согласно (19) и (20), а за 3 такта параллельных целочисленных вычислений в каждом слое НЭ независимо от размерности задачи D = $n \times m \times$ K.

Гетерогенность архитектуры синтезированной диофантовой ГНС характеризуется тем, что первый и третий слои состоят из пороговых НЭ, а второй слой включает конъюнктивные НЭ вида (19).

8. Параллелизм и самоорганизация в мульти-агентных нейронных сетях с гомогенной или гетерогенной архитектурой

Традиционно гомогенные или гетерогенные НС используются для автономного принятия решений в задачах распознавания образов, диагностики состояний, классификации данных и т.п. По существу эти НС являются обучаемыми интеллектуальными агентами, которые настраиваются на индивидуальное (одно-агентное) решение конкретных задач по ОБД вида (7).

В то же время существует большой класс интеллектуальных задач, требующий не только индивидуальных (одно-агентных), но и коллективных (мульти-агентных) решений. Классическим примером этого могут служить особенно сложные и ответственные задачи медицинской диагностики, когда врачи вынуждены прибегать к помощи своих коллег для совместной постановки окончательного диагноза. При этом формируется “консилиум”, т.е. профессиональная группа врачей, интегрирующая знания и опыт входящих в неё членов для коллективного принятия наиболее правильных и сбалансированных диагностических решений.

Другим примером сложных задач, требующих коллективных решений, являются глобальные задачи, допускающие естественную (например, иерархическую или мультифрактальную) декомпозицию на множество локальных задач. В этом случае решение сложной (глобальной) задачи может быть распределено между интеллектуальными НС-агентами, специализирующимися на решении M частных (локальных) задач. Параллельная работа M таких НС-агентов может значительно ускорить обработку информации и повысить надежность решения общей (глобальной) задачи.

В роли интеллектуальных агентов могут выступать гомогенные или гетерогенные НС различных типов. Однако они должны быть взаимосвязаны с помощью каналов обмена информацией в процессе принятия коллективных решений. В этом случае можно создать мульти-агентную (глобальную) систему обработки и передачи информации, интегрирующую в себе возможности входящих в неё локальных НС как агентов.

Архитектура таких мульти-агентных НС может быть гомогенной или гетерогенной. В гомогенной архитектуре в качестве агентов используется НС одного типа. Например, это могут быть гомогенные НС-агенты типа “перцептрон” или гетерогенные диофантовые ПНС. В гетерогенной архитектуре используются НС-агенты различных (смешанных) типов. Например, они могут содержать разные типы гетерогенных ПНС или могут иметь специальных агентов-координаторов, организующих целенаправленную работу локальных НС-агентов.

Агенты-координаторы могут принимать коллективные (мульти-агентные) решения на основе локальных (одно-агентных) решений остальных НС как автономных агентов с помощью мажоритарных принципов или процедур голосования (например, по “большинству голосов”) [11 – 14]. При этом все локальные решения принимаются параллельно, что ускоряет принятие глобального (коллективного) решения в M раз.

В ряде случаев глобальная самоорганизация НС-агентов обеспечивается иерархической, фрактальной или мультифрактальной декомпозицией общей задачи на M подзадач. При этом степень внешнего (глобального) параллелизма в мульти-агентной нейросетевой системе определяется параметром M, характеризующем одновременную работу M локальных НС-агентов, каждый из которых обладает

внутренним (локальным) параллелизмом при решении интеллектуальных задач, характеризующимся числом слоёв НЭ.

Мульти-агентное распознавание сложных 2D-изображений или 3D-сцен в ряде случаев основывается на их декомпозиции на самоподобные (фрактальные) компоненты и на обучении и самонастройке локальных гетерогенных ПНС на распознавание фрагментов по локальным ОБД, характеризующим эти фрагменты. В результате внутренней и внешней самоорганизации ПНС как НС-агентов достигается высокая степень параллелизма в процессе распознавания и анализа сложных изображений и сцен.

Необходимость в использовании НС - агентов и мульти-агентных технологий возникает в глобальных телекоммуникационных и компьютерных системах [10]. В этом случае НС-агенты обучаются и самоорганизуются по локальным ОБД вида (7) и передают по каналам связи накопленные “нейрознания” и “опыт” другим НС-агентам. Для эффективного (в частности, оптимального) управления потоками данных между удалёнными сетевыми пользователями как внешними агентами (клиентами) и локальными НС как внутренними агентами целесообразно использовать нейросетевые маршрутизаторы потоков данных[10], позволяющие адаптироваться к непредсказуемым изменениям структуры и праметров глобальных телекоммуникационных и компьютерных систем в процессе их функционирования в реальном времени.

Заключение

Описанные гетерогенные архитектуры и быстрые алгоритмы обучения ПНС разных типов обеспечивают высокий параллелизм и самоорганизацию нейровычислений в процессе решения интеллектуальных задач. Они успешно применялись для решения ряда прикладных задач распознавания образов (распознавание деталей на конвейере, класификация дорожных ситуаций и т.д.), медицинской диагностики (диагностика и оценка эффективности лечения артритов, векторная диагностика и расшифровка гастритов и т.д.), прогнозирования явлений (прогнозирование исхода черепно-мозговых травм и т.д.) и нейросетевого представления генетического кода [1–9, 11, 12]. Модифицированные НС Хопфилда успешно использовались для решения задач мульти-агентной маршрутизации параллельных потоков данных в глобальных телекоммуникационных и компьютерных сетях с переменной структурой [10]. Мульти-агентные нейросетевые технологии на базе гетерогенных диофантовых НС и процедур голосования позволили значительно повысить точность и робастность векторной (расширенной) диагностики гастритов [12-14].

Важное значение для эффективного распознавания образов и диагностики состояний в реальном времени представляет тот факт, что аккумулируемые в гетерогенных ПНС с самоорганизующейся архитектурой “нейрообразы” и решающие (классифицирующие и идентифицирующие) правила обеспечивают массовый параллелизм, хорошую экстраполяцию и высокое быстродействие при принятии оптимальных или субоптимальных решений.

Коллективное (мульти-агентное) использование гетерогенных ПНС в качестве нейросетевых агентов позволяет дополнительно распараллелить и распределить между локальными НС-агентами процессы решения сложных (глобальных) задач распознавания образов, анализа изображений и сцен, расширенной (векторной) диагностики состояний и адаптивной маршрутизации информационных потоков.

Список литературы

1. Тимофеев А.В. Об одном классе полиномиальных разделяющих функций в задачах опознавания и диагностики. - Методы вычислений. – Л.: Изд-во ЛГУ, 1971, вып.7, с. 106-121.
2. Пшибихов В.Х., Тимофеев А.В. Алгоритмы обучения и минимизации сложности полиномиальной опознающей системы. - Изд. АН СССР. Техническая кибернетика, 1974, № 5, с. 214-217.
3. Timofeev A.V. Intelligent Control Applied to Non-Linear Systems and Neural Networks with Adaptive Architecture. – International Journal on Intelligent Control, Neurocomputing and Fuzzy Logic, 1997, Vol.1, N1, pp. 1-18.
4. Каляев А.В., Тимофеев А.В. Методы обучения и минимизации сложности когнитивных нейро-модулей нейрокомпьютера с программируемой архитектурой. - Доклады АН, 1994, т. 237, с. 180-183.
5. Тимофеев А.В. Методы синтеза диофантовых нейросетей минимальной сложности. - Доклады АН, 1995, т.301, № 3, с.1106-1109.
6. Тимофеев А.В., Шибзухов З.М. Методы синтеза и минимизации сложности диофантовых нейронных сетей над конечным полем. – Автоматика и телемеханика, 1997, № 4, с. 204-212.

7. Тимофеев А.В. Оптимальный синтез и минимизация сложности генно-нейронных сетей по генетическим базам данных. - Нейрокомпьютеры: разработка и применение, № 5-6, 2002, с. 34-39.
8. Тимофеев А.В., Шибзухов З.М., Шеожев А.М. Синтез нейросетевых архитектур по многозначному дереву решений. - Нейрокомпьютеры: разработка и применение, № 5-6, 2002, с. 44-49.
9. Timofeev A.V. Polynomial Neural Network with Self-Organizing Architecture. – International Journal on Optical Memory and Neural Networks, 2004, N 2.
10. Timofeev A.V. Multi-Agent Information Processing and Adaptive Control in Global Telecommunication and Computer Networks. – Journal on Information Theories and Applications, 2003, vol. 10, N 1, pp.54–60.
11. Тимофеев А.В., Шеожев А.М. Методы построения обучающих выборок для развернутой медицинской диагностики на базе нейросетевых технологий. – Доклады АМАН, 2000, т. 5, № 1, с. 69 – 71.
12. Тимофеев А.В., Шибзухов З.М., Шеожев А.М. Применение диофантовых нейронных систем для генетического анализа и диагностики - Труды 6-ого Санкт-Петербургского симпозиума по теории адаптивных систем “SPAS-99”, СПб: Омега, 1999, т.2, с. 169 –171.
13. Тимофеев А.В., Шибзухов З.М., Шеожев А.М. Проектирование и обучение мульти-агентных диагностических систем. – Труды I-ой международной конференции по мехатронике и робототехнике “M&R – 2000”, СПб, 2000, т. 2, с. 342 – 345.
14. Timofeev A.V. Parallelism and Self-Organization in Polynomial Neural Networks for Image Recognition. – Pattern Recognition and Inage Analysis, 2005, vol. 15, No.1, pp. 97 – 100.

Сведения об авторе

Тимофеев Адиль Васильевич – доктор технических наук, профессор, Заслуженный деятель науки Российской Федерации. e-mail: tav@iias.spb.su

NEURONAL NETWORKS FOR MODELLING OF LARGE SOCIAL SYSTEMS. APPROACHES FOR MENTALITY, ANTICIPATING AND MULTIVALUEDNESS ACCOUNTING.

Alexander Makarenko

Abstract. *It is consider the new global models for society of neuronet type. The hierarchical structure of society and mentality of individual are considered. The way for incorporating in model anticipatory (prognostic) ability of individual is considered. Some implementations of approach for real task and further research problems are described. Multivaluedness of models and solutions is discussed. Sensory-motor systems analogy also is discussed. New problems for theory and applications of neural networks are described.*

1. Introduction

There is one principal feature of the present state of contemporary World: their evolutionary nature. That is the rate of changes that is accelerating rapidly now and the problems of evolution of global systems became more and more complicated. So, the applicability of existing theories and models of society are under question. One of the main tools for the investigation of evolution is the approach from the physical theories - that is synergetic.

There also exists the great variety of the mathematical models. It is known that the above models present mostly three types of global blocks (biospherical, climate and anthropological). The block of human (anthropogenic) factors actually seems to be the less developed one. The artificial intelligence theory can give the answers on some questions, but there is the lack of practical operational models with artificial intellect.

We may say that in spite of many successes of system analysis and mathematical modelling there is the necessity to have socio-economics models. So, main basic items for the theories and models of the World exist: the society as the whole object, the evolutionary nature of the society, the mentality problems and some propositions on the laws of their behaviour. In proposed report, we briefly consider the principles of new models construction, some applications and further scientific problems. The main goals of this report in to describe the ways of mentality accounting and especially the anticipatory property accounting consequences.

2. Short Description of Models

Let us take that society consisting of N >> 1 individuals and each individual characterising by vector of state $S_i = \{s_1^i,..,s_{k_i}^i,s_{k_i+1}^i,..,s_{M_i}^i\}, s_l^i \in M_i^l, l = 1,...,M_i$ where M_i^l is a set of possible values s_i . There are many possibilities to compose the elements in blocks and levels in such models. In sufficiently developed society individuals have many complex connections. Let us formalise this. We assume that there are connections between i and j individuals. Let J_{ij}^{pq} is the connection between p components of element i and q component of element j Thus the set Q = ($\{s_i\}$,$\{J_{ij}^{pq}\}$, i,j = 1,...,N) characterises state of society. Analysis of recent models for media from sets of elements and bonds shows the resemblance of such society models to neural network models.

2.1 Possible Structures of Models

Now we follow the description of hierarchical systems similarly the one in papers by Mesarovich and Takahara. We suppose initially that there are M hierarchical levels in the socio-economical system with N_j elements on j-th level. Each I-th element on j-th level have description by vector of parameters Q_i^j i=1,2,...,N_j ; j=1,2,...,M. Some elements on chosen levels can be in associations, marked by set of possible indexes in associations $L_i^j \subset \{1,2,..., N_j\}$. Many elements in developed society have a vast number of interconnections on there and on upper and lower levels. We may denote connections (bonds) between i1 elements on j1 level with i2 element on j2 level by J(i1,j1;i2,j2). Remark that other fields of interest (political, social, educational and so on) have similar network representation and society, as a whole is a union of such networks.

The bonds from the connection sets may be very different on the nature. The values of bonds may represent the normalisation of economical, informational, control channels, nationality, family bonds, and participation in professional associations and so on. The general model of system as in general system theory can be introduced with the help of input X1, X2, …, XM and output Y1, Y2, …, YM spaces for every level with input variables xi ∈ Xi and output variables yk ∈ Yk.

In reality society is evolutionary system with dynamical changes on time. Further we for simplicity will consider only discrete time models with moments of time: 0,1,2,….,n,… . Following evolutionary nature of systems considered it is natural to consider as input of system in moment n the values of parameters from X in n-th time moment and as output the values at next (n+1) time moment (for n=0, 1, 2, …). Remark that in developing society the content of elements set may changes. For example in economics the list of firms and corporations changes gradually by bankruptcy and by creating of coalitions. Social, political, governmental networks are often in transformations. This lead in general to changing the number of elements N_j (n) and number of hierarchical levels M(n) for different moments of time. Next if we wish to take into account the past states of society explicitly we should introduce to equation (1) or (2) the values X(0), X(1), X(2), … , X((n-1)). Than the system description takes the form

$$Y(n)=f(X(0),X(1),X(2),...,X((N-1)), X(n), P, E).$$

2.2 Dynamics in Model

The equation above is rather general but for further investigations and practical applications we should have more developed models. Because we should consider evolutionary problems the main difficulties consist in searching the principles for modelling dynamics.

The author's models consider the Society as large complex object constructed from many elements with interconnections. The considerations of Society properties allow picking out some interesting properties and then to propose the models, which can imitate society behaviour. Surprisingly the models are familiar with models of brain activity - the neuronets [1]. Such models are under investigation by author since 1992 and yet had some interesting applications. In the processes of model consideration author continuously tried to take into account recent state of above mentioned sciences.

Now let us briefly describe the models. The first step of model building consists in the choice of model element and their description. Because it is need to take into account mentality of peoples in simplest models as the elements was took the individual with their description by series of mental and other (economical, demographic, and other parameters). These parameters may be evaluated in some scales psychology, sociology and other humanity sciences.

Next there are a lot of interconnections between elements in society - informational, business, relationship, and infrastructure. The elements are connected by bounds. The bounds correspond to influence by individual, the money flows and others. Such connections are created historically. The set of element states and bounds give the description of society in some period of time. Remark that such description is familiar with verbal description in humanity sciences. For example the pictures in L.White's works remember the pictures for global socio-economical models. But if we wish to describe the dynamics of society and should to evaluate the influence of control than we must to know or dynamical laws or tendency in dynamics. The proposed models have such dynamical principles that they can imitate the behaviour of global culture in time. This is because the models have the property of associative memory. That is it can learns from historical processes the bounds and tends to very stable constructions- to so called attractor in pattern recognition in informatics and neuroscience. It is important that many social sub-processes in society also have the properties above allow considering the separate sub-models.

In earlier papers author introduced new class of society models as modification of neuronet models such as Hopfield, Potts, Ising. It is well known that Hopfield model is derived from the functional called 'energy'. In case of hierarchical systems and symmetrical bonds between different elements and different levels there also exist functional – counterpart of 'energy'. Remark that there also may be formulated generalisation of Hebbian learning rules.

3. Mentality Accounting

The mentality accounting requires considerations internal structures and incorporating them in global hierarchical models. There are many approaches for mentality accounting (see review of some aspects in [2,3]). The most natural way for implementing this task is to consider as model for internal structure also neuronet models. Remember that originally neuronet models were introduced in the investigation of brain. Firstly we can change the basic laws. On phenomenological level it may be implemented by introducing subdivision of elements parameters on external and internal variables and establishing separate laws for two blocks of parameters. But one of the most prospective ways for mentality account lies in searching equation also in neuronet class. Here proposed to introduce the intrinsic mental models of World in elements, which represent the individuals or decision-making organisations with human participation. The simplest way consists in representing image of World in the individual's brain or in model as collection of elements and bonds between elements. In such World pattern there exist place for representing individual himself with personal beliefs, skills, knowledge, preferences. The mental structures on other individuals are also represented. Then the laws for element evolution should depend on such representation.

4. Anticipatory Property

The next step of developing models consists in accounting anticipatory aspects of individuals. It is evident that individuals in decision-making processes have prognoses on future. In such case the states of elements in model should depend on the images of the future described in internal representation. As in usual reflexive system there may exist some stages of iteration in anticipating future. We call such case as hyperincursion.

The verbal description of internal structure was described in previous section. Now we give the possible structure of models and some corollary. First we describe the model structure with one element with internal structure. If there were no internal structure it was the system in section above for dynamical law. Let the individual with internal structure has the index i=1. Their dynamic is determined by two components. First component determines by external mean field as above. Second part of dynamic is connected with internal dynamics of first individual. Remark that this dynamic partially account the willing of individual. There exist many models for such part of dynamics but it is useful to put the neuronet models for our purposes.

Let us named the pattern of society $Q^{(1)}(t)$ in section above as 'image of real world ' in discrete moment of time t. We also introduce the $Q_{wish}(t)$ - ' desirable image of world in moment t by first individual' as the set of element states and bonds wishes by first individual in moment t. $Q^{(1)}{}_{wish}(t)=(\{s_l{}^{wish}(t)\},J_{ij}{}^{wish}(t)\})$. Then we assume that the

change of first individual state depend on difference between real and desirable image of the world. The resulting system takes the form:

$$S_l(t+1)=G_l(\{s_l(t)\},\{s_l(t+1)\},\ldots,\{s_l(t+g(l))\}, R),$$

where R the set of remaining parameters. It is very prospect that the structure of system above coincides with anticipatory systems with incursion [4]. This follows possible similarity in properties.

5. Multivaluedness in Neural Networks

So far the neuronet approach had followed after the original problem formalisation. But with spreading neural network methodology some new mathematical problems had aroused which may have long- term influence on the development of neurocomputing and not only it. Such problems follow from models above. First topic concerned the neuronet models with hierarchical structure. The second and very interesting connected with possible multivaluedness in neuronet.

The main source of multivaluedness lies in neural elements with internal structure with anticipatory property when the dynamical behaviour of element depends from desired pattern of future [3]. Some preliminary results were received with R.Pushin on modelling unique multivalued neuron. Also the principles for dynamics were considered.

In parallel (and forwards) some possible range of applications may be proposed. Some such issues are brain processes and conscious, quantum mechanical analogies, many worlds concept in physics, logic and philosophy, complexity, multivalued solutions of differential equations.

6. Some Relations to Sensor-Motor Robotics

The models described in previous sections already were applied to some practical problems. Further development will follow by exploiting concepts from another research fields. Surprisingly such enrichment leads to considering some fundamental problems of cybernetics. The main tool is inter- disciplinarily methodology.

One of the sources of new ideas is the psychological investigation of visual perception. Remember that the perception process not only include the reception of signal by visual sensor system but also include internal comparison with patterns. Such patterns are internal constructs of visual perception system [5,6]. Further development of proposed models will lead to complication of internal models and to modelling of process of norms learning. It is directly connected the investigations on norms ruled behaviour. Society norms, morals, religion and so on determine the rules. Remark that till now there was a little investigation on such topics mostly of model character [7].

From another side further development of proposed models needs further re- considering of basic principles of artificial intelligence in application to such problems. From such point of view especially interesting are investigations in formally different research field in automation and control – that is from behaviour theory of mobile robot. The short list of investigations (see [8,9]) includes analysis of sensor- motor robotics; comparison of formal language's and behavioural approaches; internal representation of external environment; role of sensor information channels. Some of such concepts may be transfer to the neural type models of large socio-economical systems. One of the basic concepts in the mobile robot theory is physical landscape. In social systems case the evolution takes place in many- dimensional space constructed from physical and mental space. The points in this space are representation of system space in different time moment. Remark that the description of environment as network from [10] may be useful building internal description of world.

Conversely, the author's model may be interesting for considering mobile robot problems. The neuronet description of external environment is first example for such application. But more prospects may be the investigation on anticipatory agents. As already had formulated above, anticipating property account leads to multivaluedness of behaviour scenarios in systems with self- reference. Concerning mobile robot it may lead to more intelligent behaviour (in definition of intelligence from [11]).

The next perspective approach follows from the considering neuronal models with many agents. As background for modelling large systems it allows to solve the optimal control and game- theoretical problems. The robot soccer may be one of such issues. The second is the control problem for many vehicles with internal structure in 2D and 3D space cases.

7. Applications and Discussion

Now we should discuss some issues connected to above problems. It should be stressed some relations to another topics in artificial intellects. One of such item is so called artificial agent's theory. Now there are many investigations on artificial agents. In this approach some non-classical logic are accepted. Moreover our neuronet type models follows to consideration of some non-classical logic. Another interesting aspect is the structure of neuronet models itself. Our investigations lead to the necessity of considering set valued neural networks.

The possible applications of models with mentality account to election processes, negotiations, public relations, education are discussed. Also pure mathematical problems on multi-valued maps and on conflictly-controlled systems are posed. As application it were considered the modelling future geopolitical relations and collective security system structure in World after the destruction of the USSR, sustainable development, epidemiology, conflict theory, stock market and others [12,13]. It was created as mathematical model as computer program implementation. Remark that recently we had received interesting result on models with internal structure application to the stock market process. This is the example of mental agent application. Recently new possible fields of applications are outlined. Moreover the connected to multi–agent modelling, cellular automata, decision-making in social systems became visible. Also new analogies of quantum mechanics and social system behaviour are found. Besides the theory of distributed reflexive systems can receive the strict models for consideration. Ontology of knowledge and systems description may easy take into account mentality aspects. All this follows to new problems in neural network design and in neural network theory, which the author supposes, discuss in the report. One of the main conclusions is that the new proposed fields of neural network applications can lead to reconsidering some backgrounds of the network considerations.

The paper had been partially supported by Ukrainian Grants 0205U000622, 0105U000490

References

1. Makarenko A. About the models of Global Socio- Economic process. Proceed. of Ukrainian Acad. of Sci. (1994). N.12.p.85-87. (in Russian).
2. Makarenko A. New Neuronet Models of Global Socio- Economical Processes. In "Gaming /Simulation for Policy Development and Organisational Change" (J.Geurts, C.Joldersma, E.Roelofs eds.), Tilburg Univ. Press, (1998). pp.133-138.
3. Makarenko, A. Models with anticipatory property for large socio-economical systems. Proc. 16 th World Congress of IMACS, Lausanne, 21-25 August, (2000). Switzerland, Paper n. 422-1
4. Dubois D. Introduction to Computing Anticipatory Systems. Chaos Newsletter (Liege), January, (1998). n. 2. pp.5-6.
5. Zinchenko T. Perseption and coding. Leningrad, Leningr. Univ.Publish, (1981). 183 p. (in Russian).
6. Perception. Mechanisms and Models. W.H.Freeman and Company, San Francisco, (1972).
7. Epstein J.M. Learning to be thoughtless: social norms and individual computation. Santa Fe Institute, CA, USA, Working Paper n.6. January, (2000). 31 pp.
8. Tani Jun Model- based learning for mobile robot navigation from the dynamical system perspective. IEEE Trans. Syst.Man.Cyb., Part.B. (1996). V.26, n. 3. Pp.421-436.
9. Lect.Notes. in Artificial Intelligence, Springer- Verlag, (1998). N.1515. 260 pp.
10. Makarenko A. Complexity of individual object without including probability concept. Fourth Int.Conf. on Computing Anticipatory Syst., Abstract Book. Liege, Belgium, August, (2000). 3pp.
11. Albus J.S. Outline for a theory of intelligence. IEEE Trans. On Syst. Man. Cyb. , (1991), v.21. n.3. pp. 473-504.
12. Levkov S., Makarenko A. Geopolitical relations in post USSR Europe as a subject of mathematical modelling and control. Proceed. 7 IFAC/IFORS/IMACS Symposium: Large scale Systems. L.UK: Vol.2. (1998). p.983-987.
13. Levkov S., Makarenko A., Zelinsky V. Neuronet type models for stock market trading patterns. Proc. 5th Ukrainian Conf. AUTOMATICA'98. Kiev, May, 1998. Pp.162-166.

Author's Information

Alexander Makarenko – National Technical University of Ukraine "KPI", Institute for Applied System Analysis Department of Mathematical Methods of System Analysis
37, Pobedy Avenue, 03056, Kiev-56, Ukraine, makalex@i.com.ua, makalex@mmsa.ntu-kpi.kiev.ua

6.2. Neural Network Models

ПРЕДСТАВЛЕНИЕ НЕЙРОННЫХ СЕТЕЙ ДИНАМИЧЕСКИМИ СИСТЕМАМИ

Владимир С. Донченко, Денис П. Сербаев

***Abstract**: Рассматривается представление нейронных сетей в виде динамических систем. Предложен метод обучения нейронных сетей с помощью теории оптимального управления.*

***Keywords**: нейронные сети, динамические системы, обучение.*

Введение

В настоящее время нейронные сети получили самое широкое распространение и успешно применяются для решения различных сложных задач таких как, например, управление и идентификация нелинейными системами, анализ финансового рынка, моделирование сигналов и т.д. Качество работы нейронной сети во многом зависит от эффективности выбранного алгоритма определения весов сети для достижения требуемой точности на обучающей и тестовой выборках. Ниже предложен метод нахождения весов нейронной сети на основе теории оптимального управления и представления нейронной сети в виде динамической системы.

Представление сети системой рекуррентных соотношений

Как известно [4], нейронная сеть – это совокупность однотипных элементов - нейронов, – разбитых на части-слои, определенным образом последовательно связанные между собой. Каждый из нейронов, из которых складывается каждый из слоев, собственно, отвечает скалярной функции векторного аргумента $y=F(w^Tx)$, что является суперпозицией линейной формы с вектором линейной формы w, который называют вектором весом, – и скалярной функции F. Последнюю называют функцией активации нейрона. Аргумент x – векторный – вход нейрона, скалярное значение y – выход. Входы нейрона, связанные с их принадлежностью тому или иному слою, на которые разбита сеть. Эти слои упорядочены последовательно так, что выходы всех нейронов предшествующего слоя подаются на входы любого из нейронов следующего слоя. Входом первого слоя является сигнал, который является входным для всей сети. Для стандартизации обозначений, будем считать, что вход образовывает слой с номером 0. Этот слой, в отличие от других, не содержит нейронов и, собственно, задает входной сигнал $x(0)=x_0$, который состоит из l_0 компонент. Слой с номером N есть выходным. Каждый из слоев с соответствующими номерами k: k=0,…,N имеет количество нейронов l_k. Скалярные выходы нейронов одного слоя объединяются в один вектор x(k), k=0,…,N, который будем считать выходом соответствующего слоя. Размерность такого вектора совпадает с количеством l_k, k=0,…,N нейронов в соответствующем слое.

Будем считать, что все нейроны одного и того же слоя имеют одинаковые веса. Общий для всех нейронов одного слоя вес будем обозначать соответственно номера слоя w(k): $w(k)=(w(k)_1,\dots,w(k)_{l_{k-1}})^T$, k=1,…,N... Размерность вектора весов, естественно, совпадает с количеством l_{k-1} нейронов слоя-предшественника.

Что же касается функций активации любого из нейронов соответствующего слоя, будем считать, что они являются разными для каждого нейрона и будут обозначаться $F_i^{(k)}(z)$, и=1,…,l_k, k=1,…,N... Напомним, что функции активации являются скалярными функциями скалярных аргументов.

Таким образом, преобразование входного сигнала $x(0)=x_0$ последовательными слоями нейронной сети описывается системой рекуррентных соотношений:

$$x(k+1)=\begin{pmatrix} F_1^{(k+1)}(w(k+1)^T x(k)) \\ \dots \\ F_{l_{k+1}}^{(k+1)}(w(k+1)^T x(k)) \end{pmatrix}, \quad k=0,\dots,N-1... \tag{1}$$

Обозначив через g(z,k+1) векторную функцию $(F_1^{(k+1)}(z),...,F_{l_k}^{(k+1)}(z))$, перепишем (1) в более компактном виде:

$$x(k+1)=g(w(k+1)^T x(k), k+1), k=0,\ldots,N-1. \quad (2)$$

С учетом того, что каждый слой сети осуществляет отображение из одного линейного пространства в другое в соответствии с (1) или (2), схема нейронной сети может быть представлена рисунком 1.

На рисунке 1 слои представлены прямоугольниками, которые осуществляют преобразование в соответствии с (1) или (2). В каждом таком

прямоугольнике выделенные части, которые имеют естественную интерпретацию: w(k) отвечает входным синапсам, части с функциями активации $F_1^k,...,F_{l_k}^k$ отвечают нейронам, часть с x(k) отвечает за концентрацию выходов нейронов слоя в единый выход всего слоя в целом.

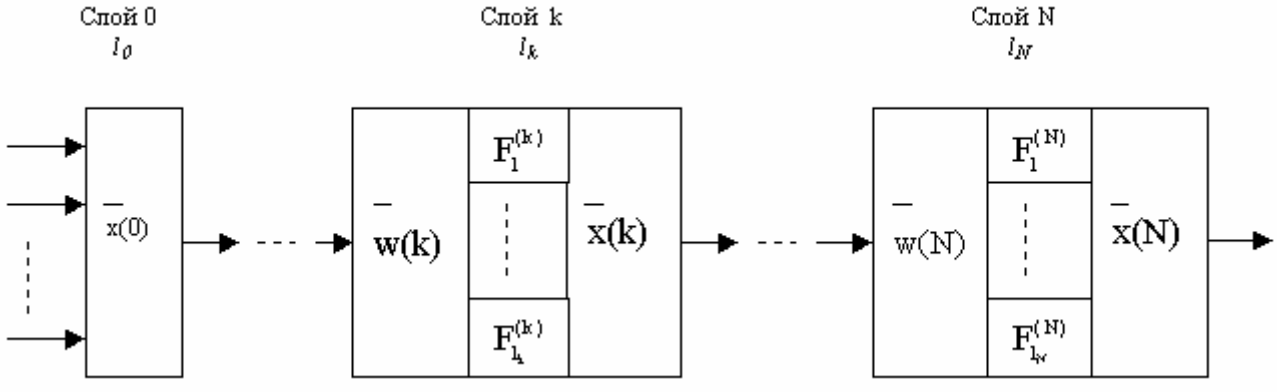


Рисунок 1.

Задача обучения нейронной сети

Задача обучения нейронной сети состоит в том, чтобы подобрать веса w(k), k=1,...,N слоев сети так, чтобы на заданной учебной выборке: последовательности пар $(x_0^{(1)},y^{(1)}),...,(x_0^{(M)},y^{(M)}), x_0^{(i)}\in R^{l_0}, y^{(i)}\in R^{l_N}$, $i=\overline{1,N}$, в которых первая компонента интерпретируется как один из вариантов входа сети и имеет размерность входного слоя l_0, а вторая – как желательный выход сети и имеет размерность выходного слоя l, – достигалось наименьшее отклонение выходных сигналов сети от желательных. Таким образом обучение нейронной сети состоит в том, чтобы минимизировать функционал J(w(1),...,w(N)),который определяется соотношением:

$$J(w(1),\ldots,w(N))=\sum_{i=1}^{M}||y^{(i)}-x^{(i)}(N)||^2, \quad (3)$$

где $x^{(и)}$ (N), $i=\overline{1,N}$ – выход сети на i-том для i-того элемента учебной выборки: объединенный выход нейронов последнего слоя нейронной сети, когда на вход подается соответствующее входное значение элемента учебной выборки.

Отметим, что если учебная выборка состоит из одного элемента, то есть если на вход подается сигнал, а сеть должна обучиться на выходной сигнал, функционал качества обучения приобретает вид:

$$J(w(1),\ldots, w(N))=||y-x(N)||^2. \quad (4)$$

Стандартный подход к решению задачи обучения нейронной сети

Стандартным подходом к обучению нейронной сети есть такой, согласно которому по фиксированному входу из возможных вариантов входа учебной выборки и первоначально фиксированных на каком-то уровне весов слоев происходит последовательное изменение этих весов каждого из слоев в

направлении, противоположном градиенту функционала J(w(1),…,w(N)) по весу соответствующего слоя w(k), k=1,…,N... Коэффициенты, которые определяют длину шага в соответствующем направлении, берутся сравнительно небольшими.

Решение задачи обучения нейронной сети применением результатов для обобщенной системы управления

Принципиальными для решения задачи обучения есть утверждение теорем 1, 2 ниже о представлении задачи обучения сети обобщенной системой управления для соответственно: одной траектории и пучка таких траекторий.

Теорема 1. Задача обучения нейронной сети на один входной сигнал представляет собою оптимизационную задачу для обобщенной системы управления, в которой соответственно:

- фазовые переменные x(k), k=0,...,N является выходами слоев с соответствующими номерами;
- управление u(k) с соответствующим номером k=0,...,N-1 совпадает с весом k+1 слоя и определяется соотношением: u(k) = w(k+1), k=0,…,N–1;
- функции f, которые описывают рекуррентную связь между значениями фазовой переменной, определяются функциями g выходов слоев по соотношениям:

$$f(x(k), u(k), k)=g(w(k+1)^{T}x(k),k+1),\ k=0,\ldots,N-1 \quad (5)$$

- функционал $I(u_0,\ldots,u_{N-1})$ совпадает с J(w(1),…,w(N)) с (2).

Доказательство теоремы приведено в [2].

Следствием теоремы 1 является возможность использовать теорему о виде градиентов для обобщенной системы управления для подсчета градиентов функционала качества обучения в задачи обучения нейронной сети. [2]

Теорема 2. Задача обучения нейронной сети на учебную выборку произвольного объема M представляет собою оптимизационную задачу для пучка траекторий обобщенной системы управления, в которой соответственно:

- фазовые переменные X(k), k=0,…,N являются матричными и состоят из столбцов $x^{(и)}(k),i=\overline{1,M}$, каждый из которых есть выход слоя с соответствующими номером, если на вход подается соответствующий входной элемент $x_0^{(и)},i=\overline{1,M}$ из учебной выборки;
- управление u(k) с соответствующим номером k=0,…,N–1 совпадает с весом k+1 слоя и определяется соотношением: u(k) = w(k+1), k=0,…,N–1;
- функции F= F(X(k), u(k), k), которые описывают рекуррентную связь между значениями фазовой переменной, являются матричными и состоят из столбцов $(f(X(k), u(k), k))_{и,}$ $i=\overline{1,M}$, которые определяются функциями g выходов соответствующих слоев нейронной сети в соответствии с соотношениями:

$$(f(X(k), u(k), k))_{и}=g(w(k+1)^{T}x^{(и)}(k),k+1),\ k=0,\ldots,N-1, \quad (6)$$

  где $x^{(и)}(k),i=\overline{1,M}$ – выход слоя с номером k: $k=\overline{1,M}$, как реакция на i-тый элемент учебной выборки;
- функционал $I(u_0,\ldots,u_{N-1})$ совпадает с J(w(1),…,w(N)).

Доказательство. Доказательство проводится так же, как и для предшествующего результата и приведено [2].

Следствием теоремы 2 является возможность использовать теорему о виде градиентов для обобщенной системы управления с пучком траекторий, для подсчета градиентов функционала качества обучения в задаче обучения нейронной сети. [2] Это, собственно, результат следующей теоремы.

Теорема 3. Градиенты функционала качества обучения нейронной сети определяются соотношениями:

$$grad_{w(k)} J(w(1),...,w(N)) = -grad_{w(k)} \sum_{i=1}^{M} H^{(i)}(x^{(i)}(k), w(k+1), p^{(i)}(k+1), k) \quad k=1,...,N. \qquad (7)$$

Теорема 3. служит основой алгоритма Error Back Propagation для обучения нейросетей.

Выводы

В статье описан метод нахождения весов нейронной сети на основе теории оптимального управления и представления нейронной сети в виде динамической системы. Использование представления нейронной сети в виде динамической системы позволяет эффективно определять веса нейронной сети и таким образом решать задачу обучения.

Литература

1. Кириченко Н. Ф., Крак Ю. В., Полищук А.А. Псевдообратные и проекционные матрицы в задачах синтеза функциональных преобразователей.// Кибернетика и системный анализ –2004.–№3.
2. Кириченко Н.Ф., Донченко В.С., Сербаев Д.П. Нелинейные рекурсивные регрессионные преобразователи: динамические системы и оптимизация. // Кибернетика и системный анализ – 2005.–№1.
3. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems, 2, 303–314.
4. Haykin, S. (1999). Neural networks: A comprehensive foundation (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
5. Kohonen, T. (1995). Self-organizing map. Heidelberg, Germany: Springer-Verlag.
6. Бублик Б.Н., Кириченко Н.Ф. Основы теории управления. – К.: Высшая школа, 1975.–328 с.

Информация об авторах

Владимир С. Донченко – профессор, Киевский национальный университет имени Тараса Шевченко, кафедра Системного анализа и теории оптимальных решений, e_mail: vsdon@unicyb.kiev.ua

Денис П. Сербаев – аспирант, Киевский национальный университет имени Тараса Шевченко, кафедра Системного анализа и теории оптимальных решений, e_mail: sdp@unicyb.kiev.ua

GENERALIZATION BY COMPUTATION THROUGH MEMORY

Petro Gopych

***Abstract**: Usually, generalization is considered as a function of learning from a set of examples. In present work on the basis of recent neural network assembly memory model (NNAMM) a biologically plausible 'grandmother' model for vision has been proposed within which each separate memory unit itself can generalize. For such a generalization by computation through memory analytical formulae and numerical procedure are found to calculate exactly the perfectly learned memory unit's generalization ability. The model's memory has complex hierarchical structure and can be learned by one-step process from one example. A simple binary neural network for bell-shaped tuning is described.*

***Keywords**: generalization, 'grandmother' model for vision, neural network assembly memory model, one-step learning, learning from one example, neuron receptive field, bell-shaped tuning.*

1. Introduction

We know from our everyday experience that even under difficult observation conditions the recognition of complex visual objects occurs in practice immediately, in an on-line regime. The ability to recognize visual objects regardless of the side of view, their illumination, occlusion or particular distortion is called generalization ability; up to present its brain mechanisms remain unclear.

In real life any two successive images cannot coincide literally, point-by-point, although they correspond to the same object. To overcome this difficult computational problem it is supposed that it is enough to remember labels of only some typical images (examples) and to learn the common memory/generalization system to predict to a huge amount of unknown, do not storing in memory, images. Such a statement of the problem implies that particular image can continuously be transformed, possibly not too sharp, into any other image of the same object through an infinite continuous series of intermediate images.

The learning theory gives a definition of generalization and rules to ensure it. So, generalization provides the best possible functional relationship between an input image, x, and its label, y, by learning from a set of examples, x_i, y_i. This problem is similar to the problem of fitting a continuous smooth function of some arguments to measurement data x_i, y_i or, in other words, the ability of estimating correctly the values of this function in points where data are not available. For this purpose standard interpolating methods are usually used [1].

Above approach is not the only possible. It is naturally to assume that the real world is actually represented in human visual system as a series of 'frames,' discrete and only perceived continuously, as in a movie. If so then the amount of information needed to be maintained reduces crucially and for this reason memory system, subserving vision and dealing with a finite set of discrete images, may become simpler. This work follows such an alternative approach.

2. Generalization by Interpolating from Examples

Within the classic learning theory generalization by interpolating among examples supports a popular neural network (NN) architecture which combines the activity of some hidden broadly tuned 'units' (local NN circuits) learned to respond to one of presented training examples and to a variety of other images but at sub-maximal firing rates. This idea is consistent with the fact that bell-shaped tuning is common among neurons in visual cortex and that in infero-temporal cortex, IT, there exist neurons tuned to different complex objects or their parts.

Mathematically, using the method of regularization, the learning from examples may be formulated as measurement data approximation by a smooth function $f(x) = \sum w_i k(x,x_i)$ which minimizes the error of training; it is a weighted sum (weights w_i) of basis functions $k(x,x_i)$ depending on a new (unknown) image x. Function $k(x,x_i)$ may be, for example, a radial Gaussian centered on x_i, representing the ith neuron's receptive field and responding optimally to (memorizing) x_i. The width of $k(x,x_i)$ defines also the unit's selectivity as a memory device: for broadly tuned k its selectivity is poor but a linear combination of such functions provides good generalization ability, for sharply tuned k (e.g., a delta function or very narrow Gaussian) its selectivity is perfect but such a k cannot be used for generalization. In $f(x)$ functions $k(x,x_i)$ may be learned from their inputs, x_i, in a passive (without the feedback) regime while weights, w_i, depend also on outputs, y_i, and demand more complicate iterative learning from examples, x_i, y_i. That is learning splits into two parts: leaning the basis functions (memory units and simultaneously neuron receptive fields) and learning the weights of the whole network (learning to generalize using already learned memory units). The algorithm described can implement a feed forward NN with one hidden layer containing as many units as training examples; parameters w_i are interpreted as synaptic weights between corresponding units and the output, $f(x)$; for further references see [1].

3. A 'Grandmother' Model for Vision

In the classic 'grandmother' theory for vision, an image recognition happens when the combination of all its features precisely coincides with such a combination associated to particular grandmother neuron, i.e. in this case between the input image and different memory records a direct literally comparison is needed. The lack of generalization is the basic problem of such a model. To solve it the model was essentially extended: it is supposed that 'generalization emerges from linear combinations of neurons tuned to an optimal stimulus' [1] (see also Section 2). We propose another extension solving the generalization problem under assumption that each memory unit itself can generalize.

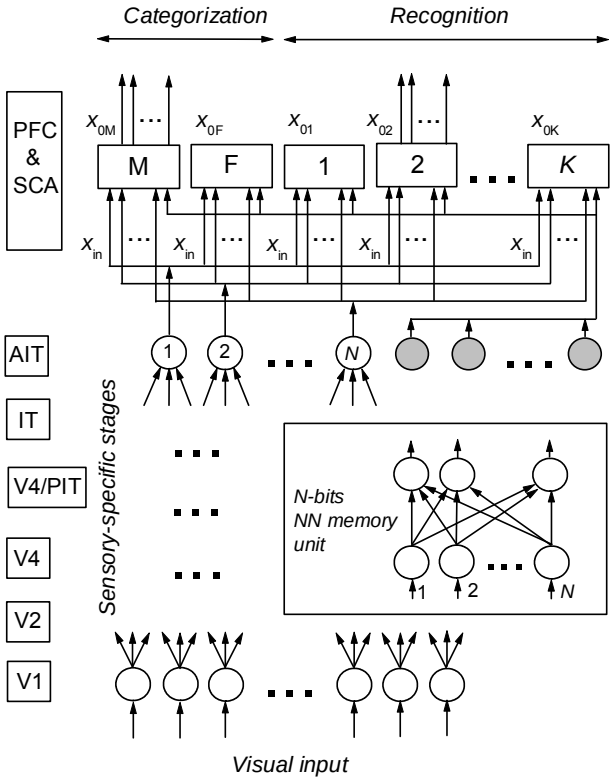


Figure 1. An oversimplified scheme of a 'grandmother' model for vision based on the NNAMM. At the bottom, in V1, cells have small receptive fields and respond preferably to oriented bars, along the ventral visual stream they increase gradually their receptive fields and complexity of their preferable images and at the top, in AIT, neurons respond optimally already to rather complex objects. AIT neurons 1,…,*N* (open circles) code the image of current interest, e.g. a face, as a binary (±1) vector x_{in}; other similar neurons (filled circles) can code (respond optimally to) other complex objects. Boxes M and F correspond to assembly memory units, AMUs (Figure 2), storing reference codes of the 'ideal' male, x_{0M}, and female, x_{0F}, faces; boxes 1,…,*K* denote AMUs storing the codes $x_{01},\ldots,x_{0K}$ which represent known (previously encountered) faces 1,…,*K*, regardless of their categorization. The case is presented where a current face code x_{in} extracted from current visual input is recognized as the face number 2 and categorized as a male face (x_{in} initiates the correct retrieval of memory traces x_{0M} and x_{02} signified as output arrows from boxes M and 2, respectively). In the insertion a feed forward NN (box 2 in Figure 2) related to particular AMU_i and storing the code x_{0i} is shown (AIT neurons 1,…,*N* may be exit-layer neurons of such an NN). If x_{in} is absent among the codes $x_{01},\ldots,x_{0K}$ but recognized as x_{0M} or x_{0F} then it can be remembered in the (*K* + 1)th empty AMU, AMU_{K+1}, which is not shown. V1, primary visual cortex; V2 and V4, extra striate visual areas; IT, infero-temporal cortex; AIT, anterior IT; PIT, posterior IT; PFC, prefrontal cortex; SCA, sub-cortical areas (e.g., hippocampus, see Section 4.2).

As Figure 1 demonstrates, in our model all sensory-specific stages of input visual data processing completely coincide with those that Poggio & Bizzi [1] discussed and, consequently, in this part both models are biologically equally plausible. In particular, in AIT neurons (open circles) tuned to respond to complex images are used although our local circuits employed for tuning are quite different (Section 5). The main distinction between our Figure 1 and Figure 2 in [1] consists in the structure of their sensory-independent parts: in Figure 1 it is implemented on the basis of the neural network assembly memory model, NNAMM, discussed in Section 4 [2].

Visual memory is constructed as a set of the NNAMM's assembly memory units, AMUs (Figure 2 in Section 4.2), interconnected between each other and storing one memory trace per one AMU. Memory traces are *N*-dimensional binary (±1) vectors represented particular images (e.g., known faces, $x_{01},\ldots,x_{0K}$) or categories of such images (e.g., the 'ideal' male, x_{0M}, and female, x_{0F}, faces). Tuned neurons 1,…,*N* (open circles) convey the code x_{in}, extracted from current visual input at sensory-specific stages of data processing and representing a current face, to all AMUs. Similar codes of other images presented in current visual input are also extracted and other tuned neurons (filled circles) convey them to all AMUs. But by means of a spatio-temporal synchrony mechanism, the AMUs shown select only the code of their interest, x_{in}; other similar codes may be the codes of interest for other AMUs which are not shown.

With the probability defined by Equations 5 and/or 6, for example, AMU_2 can perfectly recognize current x_{in} as x_{02} (can interpret x_{in} as a damaged x_{02}) even in the case $x_{in} \neq x_{02}$ and thanks to this fact the ability to generalize occurs (i.e. in contrast to Section 2, an AMU per se provides as perfect memory selectivity as well a generalization by computation through memory). Because particular AMU contains a 'grandmother' neuron (see also Section 4), we can consider our model for vision as a 'grandmother' one.

4. The NNAMM as a Memory Model Used

P.M.Gopych has proposed [3] a ternary/binary 0,±1/±1 data coding and demonstrated [3] that corresponding NN decoding algorithm (inspired by J.J.Hopfield [4]) is simultaneously the retrieval mechanism for an NN memory. As NNs used for data decoding and memory storing/retrieval are the same (see insertion in Figure 1), they have also common data-decoding/memory-retrieval performance (Section 4.3). Later this data coding/decoding approach was developed into the binary signal detection theory (BSDT) [5] and neural network assembly memory model (NNAMM) [2] closely interrelated in their roots and providing the best quality performance. The price paid for the NNAMM optimality is the fact that it places each memory trace in its own AMU (an estimation of human memory capacity, though it is possibly too optimistic — 10^{8432} bits [6], supports this assumption).

4.1 Formal Background

Let us denote a vector with components x^i (i = 1,…,N), whose magnitudes are ±1, as x. It can carry N bits of information and its dimension N is the size of a local receptive field for the NN/convolutional feature discrimination algorithm [7] or the size of an NN memory unit discussed below. If x represents information stored or that should be stored in the NN then we term it reference vector x_0. If the signs of all components of x are randomly chosen with uniform probability, ½, then that is random vector x_r or binary noise. We define also a damaged reference vector $x(d)$ with damage degree of x_0 d and components

$$x_i(d) = \begin{cases} x_0^i, if\, u_i = 0, \\ x_r^i, if\ u_i = 1 \end{cases} \qquad d = \sum u_i / N, \quad i = 1,..., N. \tag{1}$$

Marks u_i take magnitudes 0 or 1 which may randomly be chosen with uniform probability, ½. If the number of marks u_i = 1 is m then the fraction of noise components of $x(d)$ is $d = m/N$; $0 \le d \le 1$, $x(0) = x_0$, and $x(1) = x_r$. The fraction of intact components of x_0 in $x(d)$, $q = 1 - d$, is *intensity of cue* or cue index; $0 \le q \le 1$, $q + d = 1$, d and q are discrete. For a given $d = m/N$ the number of different vectors $x(d)$ is $2^m C^N_m$, $C^N_m = N!/(N - m)!/m!$; for d ranged $0 \le d \le 1$, complete finite set of all vectors $x(d)$ consists of $\sum 2^m C^N_m = 3^N$ elements (m = 0,1,…,N).

For decoding the data coded as described, we use a two-layer NN with N McCalloch-Pitts model neurons in its entrance and exit layers; these neurons are linked 'all-inputs-to-all-outputs' as the insertion in Figure 1 demonstrates. For learned NN, synapse matrix elements w_{ij} are

$$w_{ij} = \xi x_0^i x_0^j \tag{2}$$

where $\xi > 0$ (below ξ = 1), x^i_0 and x^j_0 are the ith and the jth components of x_0, respectively. Hence, vector x_0 and Equation 2 define the matrix w unambiguously. We refer to w as a perfectly learned NN and stress the crucial importance of the fact that it remembers only *one* pattern x_0 (the available possibility of storing other memories in the same NN was intentionally disregarded). It is also assumed that the NN's input vector x_{in} is decoded (reference or state vector x_0 is extracted) successfully if learned NN transforms an x_{in} into the output vector $x_{out} = x_0$ (an additional *'grandmother'* neuron mentioned in Section 3 checks this fact).

The transformation algorithm is the following. For the jth exit-layer neuron, its input signal h_j is

$$h_j = \sum w_{ij} v_i, \quad i = 1,..., N \tag{3}$$

where v_i is an output signal of the ith entrance-layer neuron. The jth exit-layer neuron's output, x^j_{out}, is calculated by a rectangular response function with the neuron's triggering threshold $\theta \ge 0$:

$$x_{out}^j = \begin{cases} +1, & if \quad h_j > \theta \\ -1, & if \quad h_j \le \theta \end{cases} \tag{4}$$

where for $h_j = \theta$ the value $v_j = -1$ was arbitrary assigned.

Since entrance-layer neurons of the NN used play only the role of input fan-outs which convey their inputs to all exit-layer neurons, in Equation 4 $v_i = x^i_{in}$. Of this fact and Equations 3 and 4 for the jth exit layer neuron we have: $h_j = \sum w_{ij} x^i_{in} = x^j_0 \sum x^i_0 x^i_{in} = x^j_0 Q$ where $Q = \sum x^i_0 x^i_{in}$ is a convolution of x_0 and x_{in}. The substitution of $h_j = x^j_0 Q$ into Equation 4 gives that $x_{out} = x_0$ and an input vector x_{in} is decoded (reference vector x_0 is extracted) successfully if $Q > \theta$. Since for each x_{in} exists such a vector $x(d)$ that $x_{in} = x(d)$, inequality $Q > \theta$ can also be written as a function

of $d = m/N$: $Q(d) = \sum x^i_0 x_i(d) > \theta$ ($i = 1,2,\ldots,N$) where θ is the threshold of Q and, simultaneously, the neuron's triggering threshold. Hence, for perfectly learned intact NNs above NN and convolutional decoding algorithms are equivalent.

Since $D = (N - Q)/2$ where D is Hamming distance between x_0 and specific $x(d)$, the inequality $D < (N - \theta)/2$ is also valid and NN, convolutional, and Hamming distance decoding algorithms mentioned are equivalent. As Hamming decoding algorithm is the best (optimal) in the sense of statistical pattern recognition quality (that is no other algorithm can outperform it), NN and convolutional algorithms described are also optimal (the best) in that sense. Moreover, similar decoding algorithms based on locally damaged NNs may also be optimal [2,8] (see also Table 1 in Section 6).

4.2 The AMU Architecture

We saw that a two-layer NN (as in the insertion in Figure 1) can be used for optimal one-trace memory storing/retrieval although randomly chosen $x_{in} = x(d)$ initiates successful retrieval only randomly. Thus, to implement the model's possibilities completely, the retrieval should be initiated by a *series* of different vectors x_{in} and it will be successful if one of the next x_{in} leads suddenly to $x_{out} = x_0$ emergence. Figure 2 exhibits the minimal architecture needed to provide optimal memory trace retrieval from the learned NN (box 2). As retrieval is initiated by $2^m C^N_m$ different vectors $x(d)$ (they are labels of images or 'frames' from their finite set mentioned in Section 1), it gives also optimal generalization by computation through memory. The internal loop 1-2-3-4-1 ensures the generation of different (e.g., random) vectors $x_{in} = x(d)$ with a given value of d while the external loop 1-2-3-4-5-6-1 maintains the internal one.

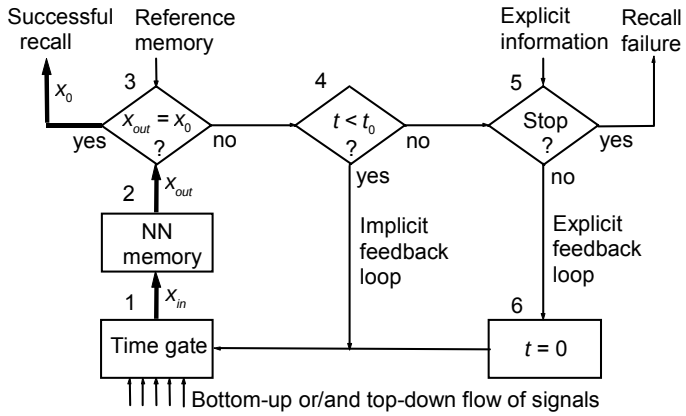


Figure 2. The flow chart (the architecture) of an assembly memory unit, AMU, and its short-distance environment adapted from [2]. The structure of the NN memory unit (box 2) specifies the insertion in Figure 1. Pathways and connections are shown in thick and thin arrows, respectively.

Within the NNAMM, the whole memory is a very large set of interconnected AMUs of rather small capacity (N ~ 100 or less) organized hierarchically. An AMU (Figure 2) consists of boxes 1,2 and 6, diamonds 3,4 and 5, and their internal and external pathways and connections designed for propagation of synchronized groups of signals [vectors $x(d)$] and asynchronous control information, respectively; it implements the BSDT decoding algorithm for solving the problem of optimal memory storing/retrieval directly.

Box 1 (a kind of N-channel time gate) transforms initial ternary (0,±1) sparsely coded very-high-dimensional vectors into binary (±1) densely coded rather low-dimensional ones. Here from the flood of asynchronous input spikes, synchronized pattern of signals in the form of N-dimensional feature vector $x_{in} = x(d)$ is prepared by a dynamical spatiotemporal synchrony mechanism. Box 2 is an NN learned according to Equation 2 (or Equation 7 from Section 4.4) where each input, x_{in}, is transformed into its corresponding output, x_{out}. Diamond 3 performs the comparison of x_{out} just now emerged with reference vector (trace) x_0 from *reference memory* (RM, see below). If $x_{out} = x_0$ then the retrieval is successful and it is finished. In the opposite case, if current time of retrieval t is less than its maximal value t_0 (this fact is checked in diamond 4) then the loop 1-2-3-4-1 is activated, retrieval starts again from box 1, and so forth. If t_0, a parameter of time dependent neurons, was found as insufficient to retrieve x_0 then diamond 5 examines whether an external reason exists to continue retrieval. If it is then the loop 1-2-3-4-5-6-1 is activated, the count of time begins anew (box 6), and internal cycle 1-2-3-4-1 is repeated again with a given frequency f, or time period $1/f$, while $t < t_0$.

The trace x_0 is held simultaneously in a particular NN memory (box 2) and in its auxiliary RM that may be interpreted as a *tag* of corresponding NN memory or as a *card* in a long-term memory catalog and performs two

interconnected functions: verification of current memory retrieval results (diamond 3) and validation of the fact that a particular memory record actually exists in the long-term memory store. Thus, specific RM is a part of memory about memory or *'metamemory'*. In contrast to the NN memory which is a kind of computer register and is conventionally associated with real biological networks, particular RM is a kind of slot devoted to the comparison of a current vector x_{out} with the reference pattern x_0 and may be associated with a coincidence integrate-and-fire *'grandmother'* neuron (cf. Section 3).

All elements of the internal feedback (reentry) loop 1-2-3-4-1 run routinely in an automatic regime and for this reason they may be interpreted as respected to *implicit* (unconscious) memory. That means that under the NNAMM all operations at synaptic and NN memory levels are unconscious. External feedback (reentry) loop 1-2-3-4-5-6-1 is activated in an unpredictable manner because it relies on external (environmental and, consequently, unpredictable) information and in this way provides unlimited diversity of possible memory retrieval modes. For this reason an AMU can be viewed as a particular *explicit* (conscious) memory unit. An external information in diamond 5 used can be thought of as an explicit or conscious one.

Recent evidences demonstrate that learning induces molecular changes in neocortex and hippocampus and this finding, along with based on it physiological theory assuming that long-term memory is stored in parallel in the neocortex and hippocampus [9], supports the NNAMM's idea of storing each memory record simultaneously in an NN (a counterpart to a neocortex network) and in a 'grandmother' neuron (probably, a cell in hippocampal structures). For other arguments in favor of the NNAMM's biological plausibility see ref. [2].

4.3 The AMU Basic Performance

For the best data-decoding/memory-retrieval algorithms considered their quality performance function is $P(d,\theta)$, the probability of correct decoding/retrieval conditioned under the presence or absence of x_0 in the data analyzed against d (or q) and θ (all notations are as in Section 4.1).

The finiteness of the set of vectors $x(d)$ makes possible to find $P(d,\theta)$ by multiple computations [3]:

$$P(d,\theta) = n(d,\theta)/n(d) \tag{5}$$

where $n(d)$ is a given number of different inputs $x_{in} = x(d)$ with a given value of d; $n(d,\theta)$ is a number of $x(d)$ from $n(d)$ leading to the NN's response $x_{out} = x_0$ if the NN decoding/retrieval algorithm with triggering threshold θ is applied. For small N, $P(d,\theta)$ can be calculated exactly because $n(d) = 2^m C^N_m$, complete set of $x(d)$, is small and all possible inputs can be taken into account. For large N, $P(d,\theta)$ can be estimated by multiple computations approximately but, using a sufficiently large set $n(d)$ of randomly chosen inputs $x(d)$, with any given accuracy.

For intact perfectly learned NNs, convolutional (Hamming) version of the BSDT/NNAMM formalism allows to derive analytical expression for $P(d,\theta)$ [8]:

$$P(d,\theta) = \sum_{k=0}^{k\max} C_k^m / 2^m, \quad k\max_0 = \begin{cases} (N-\theta-1)/2, & if \quad N \quad is \quad odd \\ (N-\theta)/2-1, & if \quad N \quad is \quad even. \end{cases} \tag{6}$$

Here if $k\max \le m$ then $k\max = m$ else $k\max = k\max_0$ and C^m_k denotes binomial coefficient.

Since θ, F (false-alarm probability), Q, and D, d and q are related, $P(d,\theta)$ can, for example, be written as ROC curves, $P_q(F)$, or as basic memory performance functions, $P_F(q)$ [2].

4.4 The AMU Learning

Equation 2 defines perfect one-step learning from one example as for the NN considered its input and its output (the label, 'teacher,' or 'supervisor') are exactly known, x_0. But it is often necessary to have unsupervised learning.

Let us use the traditional delta learning rule in the form

$$w_{ij}^{(n+1)} = w_{ij}^{(n)} + \eta\, v_j^{(n)} h_i^{(n)} \tag{7}$$

where n and $\eta > 0$ are an iteration number and a learning parameter, respectively; $v_j = x^i_{in}$; $h_i = \sum w_{ik} v_k$, $k = 1,\ldots,N$; the training set consists of only one sample, $x_{in} = x_0$ (such an iteration process does not feedback the NN's output x_{out} to the NN's input, it estimates w_{ij} using the previous value of w_{ij} and the values of η and components of x_0).

If η is small ($\eta < 1$) then the learning rate achieved is low and asymptotic values of w_{ij} are not reached. This case has no essential practical significance. If η is large ($\eta > 100$) then the iteration process leads to a fast, one-trial, without the 'catastrophic forgetting' learning because already the first iteration gives the result which is close to the asymptote and next iterations do not lead to the essential advance.

Let us consider the NN with $N = 40$, continuous w_{ij}, v_j, h_i, x^i_{in}, x^i_{out} and all initial values of w_{ij} chosen randomly with uniform probability from the range [–1,1]. If the initial learning pattern is $x_{in} = x_0$ then after each next iteration an NN with the next version of its weight matrix w_{ij} provides the emergence of the next version of x_{out} (the next approximation of x_0). For $\eta = 400$ already the first iteration gives the approximation's quality estimation $\sum |x^i_{out} - x^i_0| < 10^{-30}$ ($i = 1,\ldots,N$). The NN's specific RM related to the same AMU should also be learned simultaneously.

5 Neuron RFs and NNs for the Tuning

Figure 3 illustrates visual data processing using the NN of Section 4.1 together with its 'grandmother' neuron checking whether $x_{out} = x_0$. Binarization of y gives x_{in} (e.g., if $y_i > bd_i$ then $x^i_{in} = 1$ else $x^i_{in} = -1$) with no loss of information important for the following feature discrimination [7]. Binarization of components of y or h means spike generation; h may be interpreted as a simplified 1D profile of a 'grandmother' neuron's receptive field (RF), it results in an internal weighted network process (Equation 3). Such RFs can be typical for on-cells (panels *a*, *c*, *d*) or off-cells (panel *b*) and, as *a*) and *b*) demonstrate, noise x_{in} can initiate the reverse of RF polarity (these predictions are consistent with current physiological results [10]). The set of outputs of 'grandmothers' of different NNs (the top raw in Figure 3) reduces the redundancy of initial data and can constitute x_{in} for NNs at the next level of data processing hierarchy and etc; in particular, in AIT x_{in} could already represent a face (Section 3).

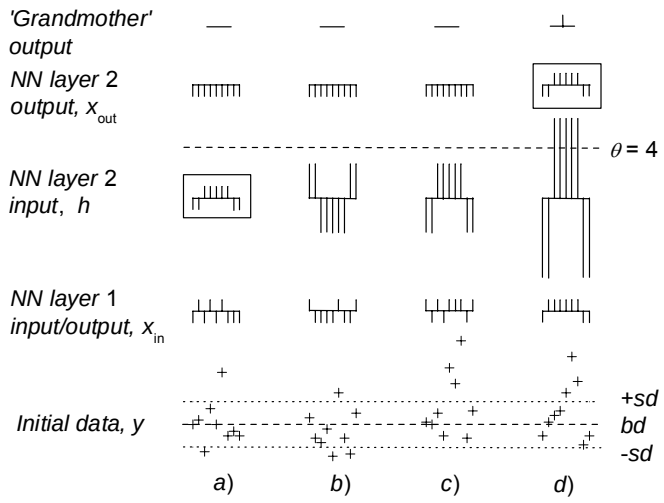


Figure 3. Computer simulated samples of initial visual data (y, an electric output of light-sensitive retina cells) and their processing results (x_{in}, h, x_{out}) in four N-channel windows: *a*) and *b*), y is a background, bd = 100, damaged by Poisson-like noise; *c*) and *d*), y is a Gaussian peak (a = 20, $fwhm$ = 5) on bd damaged by noise (crosses, values of y in each channel). Vectors y, x_0 (boxed), x_{in}, h, and x_{out} are N-dimensional ones (N = 9); positive and negative components of x_0, x_{in}, h, x_{out} correspond to upward and downward bars, respectively; intact NN and its 'grandmother' hold $x_0 = (-1, -1, 1, 1, 1, 1, 1, -1, -1)$, a kernel for the convolutional decoding/retrieval ($Q(d) > \theta$, θ = 4); peak is identified in panel *d*. $sd = bd^{1/2}$, standard deviation of bd; a, amplitude; $fwhm$, full width at half maximum.

The learned NN [Equations 2-4, Figures 1 (the insertion) and 3] provides a bell-shaped tuning to a 'grandmother' neuron's specific set of its input activities, x_0, as it is a tool for computing the convolution, Q, or Hamming distance, D, between x_0 and current set of the neuron's input activities, $x_{in} = x(d)$; i.e. it performs simply a given (Inequality 6) normalization of a current input and threshold the result (see [7] for numerical examples).

6 Generalization by Computation through Memory Performance

Since $P(q,\theta)$ defines (Equation 5) the fraction of vectors $x_{in} \neq x_0$ leading, along with x_0, to successful retrieval of the trace x_0 from the learned NN (the insertion in Figure 1 and box 2 in Figure 2), the probability of memory retrieval, $P(q,\theta)$, and generalization ability by computation through memory, $g(q,\theta)$, are numerically equal, $g(q,\theta) = P(q,\theta)$. In Table 1 generalization abilities for AMUs containing intact and damaged NNs mentioned are compared.

Usually, generalization is considered as a function of the relative size $\alpha = k/N$ of the training set of k examples and the learning strategy. It was found that for very large networks ($N \rightarrow \infty$) and $\alpha >> 1$ the error of generalization decreases as $\sim \alpha^{-1}$ [11] but the problem of generalization by learning from very few examples

remains unsolved in theory [1]. In the approach proposed learning even from one example is easily possible (Section 4.4). Values of $g(q,\theta)$ in Table 1 provide optimal (the best in the sense of pattern recall/recognition quality) generalization abilities by computation through memory; $g(0,6) = g(0,0) \sim 1\%$ was for example chosen as that is typical for professionals [7].

Table 1

Generalization ability, $g(q,\theta) = n(q,\theta)/n(q)$, for an AMU storing the trace $x_0 = (-1, -1, 1, 1, 1, 1, 1, -1, -1)$[1].

q	Intact NN, $g(q,6)$,%[2]	Damaged NN, $g(q,0)$,%[3]	q	Intact NN, $g(q,6)$,%	Damaged NN, $g(q,0)$,%
1	2	3	4	5	6
0/9	10/512 = 1.953	10/512 = 1.953	5/9	5/16 = 31.250	630/2016 = 31.250
1/9	9/256 = 3.516	81/2304 = 3.516	6/9	4/8 = 50.000	336/672 = 50.000
2/9	8/128 = 6.250	288/4608 = 6.250	7/9	3/4 = 75.000	108/144 = 75.000
3/9	7/64 = 10.938	588/5376 = 10.938	8/9	2/2 = 100.000	18/18 = 100.000
4/9	6/32 = 18.750	756/4032 = 18.750	9/9	1/1 = 100.000	1/1 = 100.000

[1] $q = 1 - d = 1 - m/N$ ($0 \le m \le N$, $N = 9$), intensity of cue ($q = 0$, free recall; $0 < q < 1$, cued recall; $q = 1$, recognition); θ, the neuron's triggering threshold; for definitions of $n(q,\theta)$ and $n(q)$ see Section 4.3.

[2] Values of $g(q,6)$ were calculated by Equation 5 or 6, results are equal.

[3] Values of $g(q,0)$ were calculated by Equation 5; 30 disrupted interneuron connections (entrance-layer neuron, exit-layer neuron) are the follows: (2,1), (4,1), (5,1), (6,1), (8,1), (3,2), (5,2), (7,2), (1,3), (4,3), (5,3), (2,4), (4,4), (2,5), (3,5), (7,5), (9,5), (3,6), (7,6), (8,6), (9,6), (1,7), (2,7), (4,7), (8,7), (1,8), (5,8), (3,9), (6,9), (7,9); this set was chosen to illustrate the fact that similar to intact NNs damaged NNs can also provide the best decoding/retrieval/generalization performance (in columns 2(5) and 3(6) generalization abilities coincide completely).

7 Conclusion

A solution of the problem of generalization by computation through memory has been illustrated by a 'grandmother' theory for vision introduced using the recent NNAMM [2]. Exact optimal calculations of such a generalization as a function of the cue index q and neuron's triggering threshold θ are performed; the approach considered provides generalization ability by learning from one example. A binary NN discussed could also be a universal circuit underlying the bell-shaped tuning of neurons in different brain areas.

Acknowledgments

I am grateful to the Health InterNetwork Access to Research Initiative (HINARI) for free on-line access to current full-text research journals and my family and my friends for their help and support.

Bibliography

1. T.Poggio and E.Bizzi. Generalization in vision and motor control. *Nature*, 2004, 431(7010), 768-774.
2. P.M.Gopych. A neural network assembly memory model based on an optimal binary signal detection theory. *Programming Problems*, 2004, no. 2-3, 473-479; http://arXiv.org/abs/cs.AI/0309036.
3. P.M.Gopych. Determination of memory performance. *JINR Rapid Communications*, 1999, 4[96]-99, 61-68 (in Russian).
4. J.J.Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings National Academy of Sciences, USA*, 1982, 79(8), 2554-2558.
5. P.M.Gopych. Sensitivity and bias within the binary signal detection theory, BSDT. *Int. Journal Information Theories & Applications*, 2004, 11(4), 318-328.
6. Yingxu Wang, D.Liu, Ying Wang. Discovering the capacity of human memory. *Brain and Mind*, 2003, 4(2), 189-198.
7. P.M.Gopych. Identification of peaks in line spectra using the algorithm imitating the neural network operation. *Instruments and Experimental Techniques*, 1998, 41(3), 341-346.
8. P.M.Gopych. ROC curves within the framework of neural network assembly memory model: some analytic results. *Int. Journal Information Theories & Applications*, 2003, 10(2), 189-197.
9. P.K.Dash, A.E.Hebert and J.D.Runyan. A unified theory for systems and cellular memory consolidation. *Brain Research Reviews*, 2004, 45(1), 30-37.

10. G.C.DeAngelis, I.Ohzava and R.D.Freeman. Receptive-field dynamics in the central visual pathways, *Trends in Neurosciences*, 1995, 18(10), 451-458.
11. M.Opper. Statistical mechanics of generalization. In *The Handbook of Brain Theory and Neural Networks,* Michael A. Arbib editor, pages 922-925. The MIT Press, Cambridge, Massachusetts, 1995.

Author's Information

Petro Mykhaylovych Gopych – V.N.Karazin Kharkiv National University; Svoboda Sq., 4, Kharkiv, 61077, Ukraine; e-mail: pmg@kharkov.com

NEURAL NETWORK BASED APPROACH FOR DEVELOPING THE ENTERPRISE STRATEGY

Todorka Kovacheva, Daniela Toshkova

Abstract: *Modern enterprises work in highly dynamic environment. Thus, the developing of company strategy is of crucial importance. It determines the surviving of the enterprise and its evolution. Adapting the desired management goal in accordance with the environment changes is a complex problem. In the present paper, an approach for solving this problem is suggested. It is based on predictive control philosophy. The enterprise is modelled as a cybernetic system and the future plant response is predicted by a neural network model. The predictions are passed to an optimization routine, which attempts to minimize the quadratic performance criterion.*

Keywords: *enterprise strategy, model predictive control, neural network, black-box modelling, business trends.*

Introduction

In the present paper, a Generalized Strategy Development (GSD) approach is suggested. Designing of the enterprise strategy is a very complicated process. It depends on many factors, which require a lot of variables to be taken into account. The relationships between them are complex and non-linear.

In the decision making process the managers need to know the environment characteristics in order to adapt the developed strategy. Therefore, the predictions of the environment changes are needed. They enable businesses make better strategic decisions and manage their activity more efficiently. It can also identify new opportunities for increased revenues and entering new markets. The prediction of the environment changes is a very difficult task. Price, advertising, goods seasonality, customers and competitors behaviour, global economic trends etc. are all factors that influence the overall performance of the enterprise.

Traditional forecasting methods such as regression and data reduction models are limited in their effectiveness as they make assumptions about the distribution of the underlying data, and often fail to recognize the inter-relatedness of variables. Now, a new forecasting tool is available – artificial neural networks (ANN). They are a form of artificial intelligence, which provide significant potential in economic applications by increasing the flexibility and effectiveness of the process of economic forecasting [Tal, Nazareth, 1995]. They are successfully used in various economic studies including investment, economic and financial forecast [Hsieh, 1993; Swales and Yoon, 1992; Hutchinson, Lo, and Poggio, 1994; Shaaf, 2000].

The enterprise strategy development requires not only predictions but also have to be optimized and adapted according to the environment changes. A suitable control design algorithm is needed. During the last years a number of methods for automatic control synthesis are applied for managing business processes. Many authors suggest the Model-Based Predictive Control (MBPC) algorithm to be used as a decision-making tool for handling complex integrated production planning problems [Tzafestas, Kapsiotis, Kyriannakis, 1997] and supply chain management [Braun et al., 2003]. MBPC is a very popular controller design method in the system engineering. It is a suitable technique for prediction of future behaviour of a plant.

Both, ANN and MBPC, are tools for solving complex problems under uncertainty by providing the ability to learn from the past experience and use information from various sources to control the enterprise performance.

Generalized Strategy Development approach combines the advantages of artificial neural networks and Model-Based Predictive Control algorithm to increase the effectiveness of the enterprise management in the entire decision making process and development of all functional strategies (incl. production-, marketing-, financial-, sales-, innovation strategy etc.).

Business Trends and Management Theory

The contemporary business is accomplished in highly dynamic environment. The continuous changes in the internal and external environment of the enterprise force it to apply a number of adaptation mechanisms, which contribute to its surviving and competitive power. These adaptation mechanisms are based on the degree of information availability. This makes providing the information a necessary condition for adaptation process and the adaptation itself – the most important characteristic of each system. In this regard the developing and the implementing of tools, which enables the corporate adaptation according to the environment changes becomes a strategic need.

Globalization [Кирев, 2001; Голдщейн 2002, 2003; Ганчев, 2004; Стоилова, 2004; Стоянов, 2003; Краева, 2003], ***virtualization*** [Мейтус, 2004; Баксанский, 2000; Манюшис, Смольянинов, Тарасов, 2003; Вютрих, Филипп, 1999], ***Internet*** and the developing of the ***Information Technologies*** [Христова, 1997; Върбанов, 2000; Илиев, 2003; Седлак, 2001] have a deep impact on the economic and social life of the society. These global trends determine the transition from the traditional industrial society to the information age society. A ***new economic based on knowledge*** [Applegate et al., 1996] appears and as a result the traditional managerial hierarchy seize to exist and a horizontal relationships are formed. Enterprises of a new type appear, which accomplish their activity on the global market from their founding. They overcome the spatial and time boundaries. The common name for such structures is ***"globally born"*** [Андерссон, Виктор, 2004]. These enterprises have their own mechanisms for developing, which substantially differ from those of the traditional industrial enterprises. Thus, the small national companies become multinational very fast.

The adaptation to environment changes requires new knowledge for its elements, the relationships between them, and characteristics of their functioning. Thus the concept of ***"Learning enterprise"*** [Senge, 1990] comes into being. It is based on the continuous acquiring new knowledge regarding the environment, using it for innovation strategies and this process applies to the enterprise as a whole.

The global trends in business development mentioned above cannot be considered partially. There are mutual relationships and dependencies between them. The existing of certain trend is a prerequisite for appearing and developing of another one and vice versa. Therefore, they influence the contemporary enterprise activities by forming an integrated set of strategies.

Now let us consider the modern business trends in a management theory point of view. Many authors [Каменов, 1984; Камионский, 1998; Рубцов, 2001] state that an unified management theory does not exist. There are different managerial concepts. Some of them claim to be universal, other are a tool for solving particular problems, some are not developed enough, other are just catchwords, some contribute and expand each other, and other contradict each other [Айвазян, Балкинд, Баснина, 1998]. This causes difficulties for the development the enterprise strategy, which strongly depends on the environment changes.

The experience shows that there is time delay between the problems, which arise in the practice and the developing of methods for their solving, which constitute the theory. In this regard, the new structures mentioned above – "globally born" and "learning enterprise" are not considered in the general management theory. Therefore, there is a lack of methodologically developed and scientifically based management approaches. These enterprises do not respond to the traditional rules and concepts as they arise and perform in strongly uncertain and highly dynamic environment. They need new management, approaches, which have to correspond to their characteristics and meet their requirements. These enterprises can be presented as complex nonlinear cybernetic systems. Thus, the laws of system and control theory can be applied to their management.

Model-Based Predictive Control

Model-Based Predictive Control has established itself in industry as an important form of advanced control [Townsend, Irwin, 2001]. An overview of industrial applications of advanced control methods in general can be found in Takatsu et al. [Takatsu et al, 1998] and in Qin and Badgwell [1998].

The main advantage of MBPC algorithm is the simplicity of the basic scheme, forming a feedback, which combines with adaptation capabilities. This determines its successful applying in the practice of designing control systems.

MBPC is an efficient methodology to solve complex constrained multivariable control problems in the absence, as well as in the presence of uncertainties [Mayne et al., 2000]. It makes possible the uncertainty of the plant and disturbances to be taken into account and enables the on-line optimization and control synthesis.

In general, it is used to predict the future plant behaviour. According to this prediction in the chosen period (prediction horizon), the MBPC optimizes the manipulated variables to obtain an optimal future plant response. The input of chosen length (also known as control horizon) is sent into the plant and then the entire sequence is repeated again in the next period. An important advantage of MBPC is that it allows the inclusion of constraints on the inputs and outputs.

The prediction plant model is realized with neural network. It provides predictions of the future plant response over a specified time horizon. The predictions are passed to an optimization routine to determine the control signal that minimizes the following performance criterion over the specified time horizon:

$$J = \sum_{j=N_1}^{N_2} \left(y_r(t+j) - y_m(t+j) \right)^2 + \rho \sum_{j=N_1}^{N_u} \left(u'(t+j-1) - u'(t+j-2) \right)^2$$

subject to the constraints, which are imposed on the state and control variables. The constants N_1, N_2, N_u define the horizons over which the tracking error and control increments are evaluated. The u' variable is the tentative control signal, y_r is the desired response and y_m is the network model response. The ρ value is weight coefficient.

Generalized Strategy Development approach will be introduced in Model-Based Predictive Control framework.

Generalized Strategy Development Approach

The purpose of the Generalized Strategy Development Approach is to transform the incomplete information about the environment and the processes inside the enterprise into complete strategy for its adaptation and evolution. From cybernetic point of view, this can be considered as a control system. The functional structure is given in Fig.1

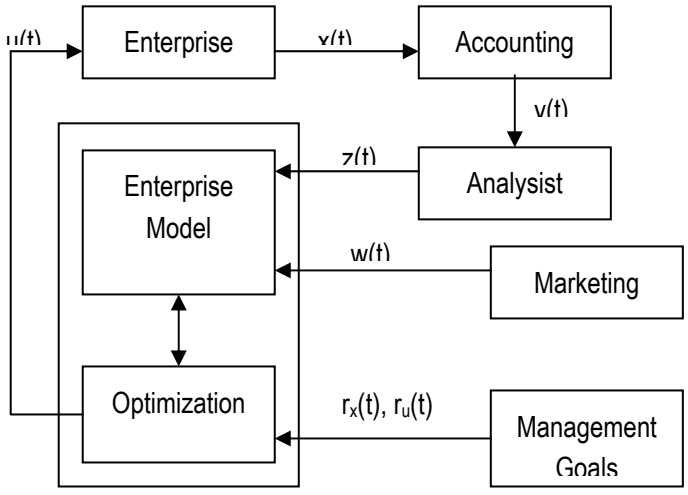


Fig.1 Global Strategy Development functional structure

Enterprise

The enterprise is a dynamic system with a high complexity. In order, the management and control [*u(t)*] to be effective we need to know its physical structure, the relationships between the constituting elements, their dynamic behaviour and the characteristics of the environment. We have to compare the current state of the enterprise to the desired state. In case they coincide entirely the management goals are achieved. In order to

register the difference we need to measure the state of the enterprise. This could be realized by performance measurement system.

Performance management is the prerequisite to performance improvement. For the enterprises to improve their performance, they must be able to measure how they are performing at present, and how they are performing after any changes. So, the companies will have the possibility to monitor if a chosen strategic direction is appropriate.

Traditional performance management systems are frequently based on cost and management accounting. There are five main difficulties with traditional management accounting techniques for performance measurement [Maskell, 1991]:

1. Management accounting reports are not relevant to strategy development;
2. Some of the data which are used for decision-making process can be distorted by cost accounting;
3. Traditional accounting reports are inflexible and are usually received too late to be of value;
4. The information about the pay-back on capital projects comes late;
5. To be of value, management accounting systems must be based on different methods and assumptions than on the financial accounts.

As traditional performance measurement systems are based on management accounting, they are primarily concerned with cost. But in today's manufacturing environment, cost based measures are no longer the only basis for decision making in enterprises. The new performance measurement systems should have some additional characteristics [Maskell, 1991].

Accounting

In Fig.1 the Accounting is the traditional performance measurement system. Therefore, the current state of the enterprise [*y(t)*] is represented by the measured one from the accounting and measurement error. The reports, which are formed by the accounting, need to be interpreted in order to be useful for the management. This task is performed by analyst.

Analyst

The accounting information is now manipulated for giving proper estimate for the current enterprise state [*z(t)*]. The manipulations include: recapitulation, generalization, estimation, recalculation etc. in order to analyze the entire enterprise activity. The results are used by managers to make decisions about the future behaviour of the enterprise.

The analysis is performed on the basis of incomplete information about the environment changes. Another error is formed. The obtained information is passed to the prediction model of the enterprise in order to minimize the tracking error.

Enterprise model

The model is used to determine the direction in which changes in the manipulated variables will improve performance. The plant operating conditions are then changed by a small amount in this direction, and a new, updated model is evaluated. The enterprise is a complex, dynamic and non-linear plant. Also different disturbances affect the its performance. Because of that, there is a lack of knowledge on the function or construction of the system.

The process output can be predicted by using a model of the process to be controlled. Any model that describes the relationship between the input and the output of the process can be used and a disturbance or noise model can be added to the process model [Duwaish, Naeem, 2001]. We can build a model using the observations of the enterprise activities.

Therefore the enterprise can be viewed as a black-box [Sjoberg et al., 1995] which aims to describe the relationships between input/output data. The non-linearities and the disturbances are taken into consideration.

During the past few years, several authors [Narendra and Parthasarathy, 1990; Nerrand et al. 1994] have suggested neural networks for nonlinear dynamical black-box modelling. To date, most of the work in neural black-box modelling has been performed making the assumption that the process to be modeled can be described accurately by neural models, and using the corresponding input-output neural predictors [Rivals,

Personnaz, 1996]. Therefore, artificial neural networks are used as effective black-box function approximators with learning and adaptation capabilities.

Marketing

We could receive information [*w(t)*] about the environment changes from the Marketing Information System (MIS) which is used in the enterprise. It is passed to the enterprise model by taking into account the forming of a new error. This error is due to impossibility of MIS to register all trends in global economy and social life of the society.

Optimization and Management Goals

Using the enterprise model, we predict the future plant response and taking into consideration the management goals [rx(t), ru(t)] we optimize it and develop a new management strategy. This is an iterative process, which provides the continuous enterprise adaptation to the environment changes.

Conclusion

Strategy development is a complex task in the continuously changing environment. The enterprise management must combine internal and external information in order to survive and evaluate. Therefore, the company needs an efficient control and strategy development and evaluation system to work in rapidly changing business conditions.

The Generalized Strategy Development approach suggested here is very suitable for this problem, namely for optimization and adaptation of the strategy development process. Thus, the effectiveness of management is increased.

Bibliography

[Айвазян, Балкинд, Баснина, 1998] Айвазян, С. А., Балкинд, О. Я., Баснина, Т. Д., и др., Стратегии бизнеса: Аналитический справочник. / Под ред. Г. Б. Клейнера. – М.: КОНСЭКО,1998

[Андерссон, Виктор, 2004] Андерссон, С., Виктор, И., Инновационная интернационализация в новых фирмах, Международный журнал “Проблемы теории и практики управления”, 1/2004

[Баксанский, 2000] Баксанский, О.Е., Виртуальная реальность и виртуализация реальности,//Концепция виртуальных миров и научное познание, СПб.: РХГИ, 2000 (http://sociology.extrim.ru)

[Върбанов, 2000] Върбанов, Р., Електронният бизнес като отражение на Интернет революцията в деловата сфера, Бизнес управление 3/2000, стр.83-95, СА “Д.А.Ценов”, Свищов

[Вютрих, Филипп, 1999] Вютрих, Х., Филипп, А., Виртуализация как возможный путь развития управления, Международный журнал “Проблемы теории и практики управления”, 5/1999

[Ганчев, 2004] Ганчев, П., Глобализацията и българските национални интереси, Диалог, СА “Д.А.Ценов” – Свищов, 2/2004, стр.32-47

[Гольдштейн, 2003] Гольдштейн, Г. Я., Основы менеджмента, ТРТУ, Таганрог, 2003

[Гольдштейн, 2002] Гольдштейн, Г. Я., Стратегический иновационный менеджмент: тенденции, технологии, практика, ТРТУ, Таганрог, 2002

[Илиев, 2003] Илиев, П., Виртуална организация на електронния бизнес, Сборник доклади от научна конференция “Иновации и трансформации на организираните пазари в България”, ИУ-Варна, 2003, стр.147-153

[Каменов, 1984] Каменов, С., Ефективност на управлението, Издателство “Г.Бакалов”, Варна, 1984

[Камионский, 1998] Камионский, С. А., Менеджмент в российском банке: опыт системного анализа и управления/ Общая ред. и предисловие Д. М. Гвишиани. М.: Деловая библиотека “Омскпромстройбанка”, 1998

[Кирев, 2001] Кирев, Л., Относно факторите за глобализация на научноизследователската дейност от транснационалните корпорации, Бизнес управление, 4/2001, стр.49-64, СА “Д.А.Ценов” – Свищов

[Краева, 2003] Краева, В., Глобални аспекти на електронната търговия, Сборник доклади от научна конференция “Иновации и трансформации на организираните пазари в България”, ИУ-Варна, 2003, стр.192-197

[Манюшис, Смольянинов, Тарасов, 2003] Манюшис, А., Смольянинов, В., Тарасов, В., Виртуальное предприятие как эффективная форма организации внешнеэкономической деятельности компании, Международный журнал “Проблемы теории и практики управления”, 4/2003

[Мейтус, 2004] Мейтус, В., Виртуализация производства, Международный журнал “Проблемы теории и практики управления”, 1/2004

[Рубцов, 2001] Рубцов, С., Целевое управление корпорациями, Москва, 2001

[Седлак, 2001] Седлак, Я., Мировая экономика: возможность неожиданных потрясений, Международный журнал “Проблемы теории и практики управления”, 5/2001

[Стоилова, 2004] Стоилова, М., Глобализацията – познати плюсове и минуси, Диалог, СА "Д.А.Ценов" – Свищов, 1/2004, стр.63-68
[Стоянов, 2003] Стоянов, В., Пазар, трансформация, глобализация, нов световен ред, Галик, София, 2003
[Христова, 1997] Христова, С., Материали от дискусия, Предизвикателствата на информационните технологии на прага на 21-и век, Национална научна конференция АИ'97, Автоматика и информатика 5/6 – 1997, САИ, София, стр.10-11
[Abdallah, Al-Thamier, 2004] Abdallah, J., Al-Thamier, A., The Economic-environmental Neural Network Model for Electical Power Dispatching, Journal of Applied Sciences 4 (3): 340-343, 2004
[Applegate et al, 1996] Applegate, L. M., McFarlan, F. W., McKenney, J. L., Corporate information systems management: text and cases, 4th ed., IRWIN, USA, 1996
[Braun et al., 2003] Braun, M., Rivera, D., Carlyle, W., Kempf, K., A Model Predictive Control Framework For Robust Management Of Multi-Product, Multi-Echelon Demand Networks, Annual Reviews in Control, 2003, vol.27, no. 2, pp.229-245(17)
[Duwaish, Naeem, 2001] Duwaish, H., Naeem, W., Nonlinear model predictive control of Hammerstein and Wiener Models using genetic algorithms, In Proceedings of the 2001 IEEE International Conference on Control Applications (CCA'01), 5-7 September, Mexico City, Mexico, pp.465-469, IEEE
[Hsieh, 1993] Hsieh, C., Some potential applications of artificial neural system in financial management. Journal of System Management, April, 1993, 12-15
[Hutchinson, Lo, and Poggio, 1994] Hutchinson, J., Lo, A., and Poggio, T., A nonparametric approach to the pricing and hedging of derivative securities via learning networks. Journal of Finance, 49, 851-99, 1994
[Maskell, 1991] Maskell, B. H., Performance Measurement for World Class Manufacturing, Productivity Press, Cambridge, Massachusetts, 1991
[Mayne et al., 2000] Mayne, D. Q., J. B. Rawlings, C. V. Rao and P. O. M. Scokaert, Constrained model predictive control: Stability and optimality, Automatica, 36, 2000, 789-814
[Narendra and Parthasarathy, 1990] Narendra, K. S., Parthasarathy, K., Identification and control of dynamical sysems using neural networks, IEEE Trans. on Neural Networks 1 (1990), pp. 4-27
[Nerrand et al. 1994] Nerrand, O., Roussel-Ragot, P., Urbani, D., Personnaz, L., Dreyfus, G., Training recurrent neural networks: why and how? An Illustration in Process Modelling, IEEE Trans. on Neural Networks 5 (1994), pp. 178-184
[Qin, Badgwell, 1998] Qin, S. Joe and Badgwell, T.A., An Overview of Non-linear Model Predictive Control Applications, Proc. Workshop on NMPC, Ascona, Switzerland, 1998
[Rivals, Personnaz, 1996] Rivals, I., Personnaz, L., Black-box modelling with state-space neural networks. In: "Neural Adaptive Control Technology", R. Zbikowski and K. J. Hunt eds., World Scientific (1996), pp. 237-264
[Senge, 1990] Senge, P. M., The Fifth Discipline. The Art & Practice of The Learning Organization. – Currency Doubleday, 1990
[Sjoberg et al., 1995] Sjoberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.Y., Hjalmarsson, H., Juditsky, A., Nonlinear black-box modelling in system identification: an unified overview. Automatica, 12. 1691-1724, 1995
[Shaaf, 2000] Shaaf, M., Predicting recession using the yield curve: an artificial intelligence and econometric comparison. Eastern Economic Journal, 26(2), Spring, 2000
[Swales, Yoon, 1992] Swales, G. S., Yoon, Y., Applying artificial neural networks to investment analysis. Financial Analyst Journal, September-October, 70-80, 1992
[Takatsu et al., 1998] Takatsu, Haruo, Itoh, Toshiaki and Araki, Mituhiko; Future needs for the control theory in industries – report and topics of the control technology survey in Japanese industry, J.Proc.Cont., Vol.8, no 5-6, 369-374, 1998
[Tal, Nazareth, 1995] Tal, B., Nazareth, L., Artificial Intelligence and Economic Forecasting, Canadian Business Economics, Volume 3, Number 3, Spring 1995
[Townsend, Irwin, 2001] Townsend , S., Irwin, G., Non-linear model based predictive control using multiple local models. In: B. Kouvaritakis, & M. Cannon (Eds.), Non-linear predictive control: Theory and practice, IEE Control Engineering Series, IEE, London, 61(11), 47-56
[Tzafestas, Kapsiotis, Kyriannakis, 1997] Tzafestas, S., G. Kapsiotis and E. Kyriannakis (1997). Model-based predictive control for generalized production planning problems. Computers in Industry 34(2), 201-210

Authors' Information

Todorka Kovacheva – Economical University of Varna, Kniaz Boris Str, e-mail: todorka_kovacheva@yahoo.com
Daniela Toshkova – Technical University of Varna, 1, Studentska Str., Varna, 9010, e-mail: daniela_toshkova@abv.bg

NEURO-FUZZY KOLMOGOROV'S NETWORK WITH A HYBRID LEARNING ALGORITHM

Yevgeniy Bodyanskiy, Yevgen Gorshkov, Vitaliy Kolodyazhniy

Abstract. *In the paper, a novel Neuro-Fuzzy Kolmogorov's Network (NFKN) is considered. The NFKN is based on and is the development of the previously proposed neural and fuzzy systems using the famous Kolmogorov's superposition theorem (KST). The network consists of two layers of neo-fuzzy neurons (NFNs) and is linear in both the hidden and output layer parameters, so it can be trained with very fast and simple procedures: the gradient-descent based learning rule for the hidden layer, and the recursive least squares algorithm for the output layer. The validity of theoretical results and the advantages of the NFKN in comparison with other techniques are confirmed by experiments.*

1. Introduction

According to the Kolmogorov's superposition theorem (KST) [1], *any* continuous function of *d* variables can be *exactly* represented by superposition of continuous functions of one variable and addition:

$$f(x_1,\ldots,x_d)=\sum_{l=1}^{2d+1} g_l\left[\sum_{i=1}^{d}\psi_{l,i}(x_i)\right], \tag{1}$$

where $g_l(\bullet)$ and $\psi_{l,i}(\bullet)$ are some continuous univariate functions, and $\psi_{l,i}(\bullet)$ are independent of *f*. Aside from the exact representation, the KST can be used as the basis for the construction of parsimonious universal approximators, and has thus attracted the attention of many researchers in the field of soft computing.

Hecht-Nielsen was the first to propose a neural network implementation of KST [2], but did not consider how such a network can be constructed. Computational aspects of approximate version of KST were studied by Sprecher [3], [4] and Kůrková [5]. Igelnik and Parikh [6] proposed the use of spline functions for the construction of Kolmogorov's approximation. Yam *et al* [7] proposed the multi-resolution approach to fuzzy control, based on the KST, and proved that the KST representation can be realized by a two-stage rule base, but did not demonstrate how such a rule base could be created from data. Lopez-Gomez and Hirota developed the Fuzzy Functional Link Network (FFLN) [8] based on the fuzzy extension of the Kolmogorov's theorem. The FFLN is trained via fuzzy delta rule, whose convergence can be quite slow. The authors proposed a novel KST-based universal approximator called Fuzzy Kolmogorov's Network (FKN) with simple structure and training procedure with high rate of convergence [9–11]. However, this training algorithm may require a large number of computations in the problems of high dimensionality. In this paper we propose an efficient and computationally simple learning algorithm, whose complexity depends linearly on the dimensionality of the input space.

2. Network Architecture

The NFKN is comprised of two layers of neo-fuzzy neurons (NFNs) [12] and is described by the following equations:

$$\hat{f}(x_1,\ldots,x_d)=\sum_{l=1}^{n} f_l^{[2]}(o^{[1,l]}),\quad o^{[1,l]}=\sum_{i=1}^{d} f_i^{[1,l]}(x_i),\quad l=1,\ldots,n, \tag{2}$$

where *n* is the number of hidden layer neurons, $f_l^{[2]}(o^{[1,l]})$ is the *l*-th nonlinear synapse in the output layer, $o^{[1,l]}$ is the output of the *l*-th NFN in the hidden layer, $f_i^{[1,l]}(x_i)$ and is the *i*-th nonlinear synapse of the *l*-th NFN in the hidden layer.

The equations for the hidden and output layer synapses are

$$f_i^{[1,l]}(x_i)=\sum_{h=1}^{m_1}\mu_{i,h}^{[1]}(x_i)w_{i,h}^{[1,l]},\quad f_l^{[2]}(o^{[1,l]})=\sum_{j=1}^{m_2}\mu_{l,j}^{[2]}(o^{[1,l]})w_{l,j}^{[2]},\ l=1,\ldots,n,\quad i=1,\ldots,d, \tag{3}$$

where m_1 and m_2 is the number of membership functions (MFs) per input in the hidden and output layers respectively, $\mu_{i,h}^{[1]}(x_i)$ and $\mu_{l,j}^{[2]}(o^{[1,l]})$ are the MFs, $w_{i,h}^{[1,l]}$ and $w_{l,j}^{[2]}$ are tunable weights.

Nonlinear synapse is a single input-single output fuzzy inference system with crisp consequents, and is thus a universal approximator [13] of univariate functions. It can provide a piecewise-linear approximation of any functions $g_l(\bullet)$ and $\psi_{l,i}(\bullet)$ in (1). So the NFKN, in turn, can approximate any function $f(x_1,\ldots,x_d)$.

The output of the NFKN is computed as the result of two-stage fuzzy inference:

$$\hat{y} = \sum_{l=1}^{n}\sum_{j=1}^{m_2}\mu_{l,j}^{[2]}\left[\sum_{i=1}^{d}\sum_{h=1}^{m_1}\mu_{i,h}^{[1]}(x_i)w_{i,h}^{[1,l]}\right]w_{l,j}^{[2]}. \tag{4}$$

The description (4) corresponds to the following two-level fuzzy rule base:

$$\text{IF } x_i \text{ IS } X_{i,h} \text{ THEN } o^{[1,1]} = w_{i,h}^{[1,1]}d \text{ AND}\ldots\text{AND } o^{[1,n]} = w_{i,h}^{[1,n]}d,\ i=1,\ldots,d,\quad h=1,\ldots,m_1, \tag{5}$$

$$\text{IF } o^{[1,l]} \text{ IS } O_{l,j} \text{ THEN } \hat{y} = w_{l,j}^{[2]}n,\quad l=1,\ldots,n,\quad j=1,\ldots,m_2, \tag{6}$$

where $X_{i,h}$ and $O_{l,j}$ are the antecedent fuzzy sets in the first and second level rules, respectively. Each first level rule contains *n* consequent terms $w_{i,h}^{[1,1]}d,\ldots,w_{i,h}^{[1,n]}d$, corresponding to *n* hidden layer neurons.

Total number of rules is

$$N_R^{FKN} = d\cdot m_1 + n\cdot m_2, \tag{7}$$

i.e., it depends *linearly* on the number of inputs *d*.

The rule base is complete, as the fuzzy sets $X_{i,h}$ in (5) completely cover the input hyperbox with m_1 membership functions per input variable. Due to the linear dependence (7), this approach is feasible for input spaces with high dimensionality *d* without the need for clustering techniques for the construction of the rule base.

Straightforward grid-partitioning approach with m_1 membership functions per input requires $(m_1)^d$ fuzzy rules, which results in combinatorial explosion and is practically not feasible for $d > 4$.

3. Learning Algorithm

The weights of the NNFKN are determined by means of a batch-training algorithm as described below. A training set containing *N* samples is used. The minimized error function is

$$E(t) = \sum_{k=1}^{N}\left[y(k) - \hat{y}(t,k)\right]^2 = \left[Y - \hat{Y}(t)\right]^T\left[Y - \hat{Y}(t)\right], \tag{8}$$

where $Y = \left[y(1),\ldots,y(N)\right]^T$ is the vector of target values, and $\hat{Y}(t) = \left[\hat{y}(t,1),\ldots,\hat{y}(t,N)\right]^T$ is the vector of network outputs at epoch *t*.

Since the nonlinear synapses (3) are linear in parameters, we can employ recursive least squares (RLS) procedure for the estimation of the output layer weights. Re-write (4) as

$$\begin{aligned}&\hat{y} = W^{[2]T}\varphi^{[2]}(o^{[1]}),\quad W^{[2]} = \left[w_{1,1}^{[2]}, w_{1,2}^{[2]},\ldots,w_{n,m_2}^{[2]}\right]^T,\\ &\varphi^{[2]}(o^{[1]}) = \left[\mu_{1,1}^{[2]}(o^{[1,1]}),\mu_{1,2}^{[2]}(o^{[1,1]}),\ldots,\mu_{n,m_2}^{[2]}(o^{[1,n]})\right]^T.\end{aligned} \tag{9}$$

Then the RLS procedure will be

$$\begin{cases}W^{[2]}(t,k) = W^{[2]}(t,k-1) + P(t,k)\varphi^{[2]}(t,k)(y(k) - \hat{y}(t,k)),\\ P(t,k) = P(t,k-1) - \dfrac{P(t,k-1)\varphi^{[2]}(t,k)\varphi^{[2]^T}(t,k)P(t,k-1)}{1+\varphi^{[2]^T}(t,k)P(t,k-1)\varphi^{[2]}(t,k)},\end{cases} \tag{10}$$

where $k=1,\ldots,N$, $P(t,0) = 10000\cdot I$, and *I* is the identity matrix of corresponding dimension.

Introducing after that the regressor matrix of the hidden layer $\Phi^{[1]} = \left[\varphi^{[1]}(x(1)),\ldots,\varphi^{[1]}(x(N))\right]^T$, we can obtain the expression for the gradient of the error function with respect to the hidden layer weights at the epoch *t*:

$$\nabla_{W^{[1]}}E(t) = -\Phi^{[1]^T}\left[Y - \hat{Y}(t)\right], \tag{11}$$

and then use the well-known gradient-based technique to update these weights:

$$W^{[1]}(t+1) = W^{[1]}(t) - \gamma(t)\frac{\nabla_{W^{[1]}}E(t)}{\left\|\nabla_{W^{[1]}}E(t)\right\|}, \tag{12}$$

where $\gamma(t)$ is the adjustable learning rate, and

$$\begin{gathered} W^{[1]} = \left[w_{1,1}^{[1,1]}, w_{1,2}^{[1,1]}, \ldots, w_{d,m_1}^{[1,1]}, \ldots, w_{d,m_1}^{[1,n]}\right]^T, \\ \varphi^{[1]}(x) = \left[\varphi_{1,1}^{[1,1]}(x_1), \varphi_{1,2}^{[1,1]}(x_1), \ldots, \varphi_{d,m_1}^{[1,1]}(x_d), \ldots, \varphi_{d,m_1}^{[1,n]}(x_d)\right]^T, \\ \varphi_{i,h}^{[1,l]}(x_i) = a_l^{[2]}(o^{[1,l]})\mu_{i,h}^{[1,l]}(x_i), \end{gathered} \tag{13}$$

and $a_l^{[2]}(o^{[1,l]})$ is determined as in [9–11]

$$a_l^{[2]}(o^{[1,l]}) = \frac{w_{l,p+1}^{[2]} - w_{l,p}^{[2]}}{c_{l,p+1}^{[2]} - c_{l,p}^{[2]}}, \tag{14}$$

where $w_{l,p}^{[2]}$ and $c_{l,p}^{[2]}$ are the weight and center of the *p*-th MF in the *l*-th synapse of the output layer, respectively. The MFs in an NFN are chosen such that only two adjacent MFs *p* and *p*+1 fire at a time [12].

Thus, the NFKN is trained via a two-stage optimization procedure without any nonlinear operations, similar to the ANFIS learning rule for the Sugeno-type fuzzy inference systems [14]. In the forward pass, the output layer weights are adjusted. In the backward pass, adjusted are the hidden layer weights. An epoch of training is considered 'successful' when root mean squared error (RMSE) on the training set is reduced in comparison with the previous epoch. Only successful epochs are counted. If RMSE is not reduced, the training cycle (forward and backward passes) is repeated until RMSE is reduced or the maximum number of cycles per epoch is reached. Once it is reached, the algorithm is considered to converge, and the parameters from the last successful epoch are saved as the result of training. Otherwise the algorithm is stopped when the maximum number of epochs is reached.

The number of tuned parameters in the hidden layer is $S_1 = d \cdot m_1 \cdot n$, in the output layer $S_2 = n \cdot m_2$, and total $S = S_1 + S_2 = n \cdot (d \cdot m_1 + m_2)$.

Hidden layer weights are initialized deterministically using the formula [9–11]

$$w_{h,i}^{[1,l]} = \exp\left\{-\frac{i\left[m_1(l-1)+h-1\right]}{d(m_1 n - 1)}\right\}, h = 1, \ldots, m_1, i = 1, \ldots, d, l = 1, \ldots, n, \tag{15}$$

broadly similar to the parameter initialization technique proposed in [6] for the Kolmogorov's spline network based on rationally independent random numbers.

Further improvement of the learning algorithm can be achieved through the adaptive choice of the learning rate $\gamma(t)$ as was proposed for the ANFIS.

4. Experimental Results

To verify the theoretical results and compare the performance of the proposed network to the known approaches, we have carried experiments with prediction and emulation of the Mackey-Glass time series [15].

The Mackey-Glass time series is generated by the following equation:

$$\frac{dy(t)}{dt} = \frac{0.2\, y(t-\tau)}{1 + y^{10}(t-\tau)} - 0.1\, y(t), \tag{16}$$

where τ is time delay. In our experiments, we used $\tau = 17$.

In the first experiment, the values $y(t-18), y(t-12), y(t-6)$, and $y(t)$ were used to predict $y(t+85)$. The NFKN used for prediction had 4 inputs, 9 neurons in the hidden layer with 3 MFs per input, and 1 neuron in the output layer with 9 MFs per synapse (189 adjustable parameters altogether). The training algorithm converged after 43 epochs.

The NFKN demonstrated similar performance as multilayer perceptron (MLP) with 2 hidden layers each containing 10 neurons. The MLP was trained for 50 epochs with the Levenberg-Marquardt algorithm. Because the MLP weights are initialized randomly, the training and prediction were repeated 10 times.

Root mean squared error on the training and checking sets (trnRMSE and chkRMSE) was used to estimate the accuracy of predictions. For the MLP, median values of 10 runs for the training and checking errors were calculated. The results are shown in Table 1 and Fig.1.

Table 1. Results of Mackey-Glass time series prediction

Network	Parameters	Epochs	trnRMSE	chkRMSE
MLP 10-10-1	171	50	0.022	0.0197
NFKN 9-1	189	43	0.018455	0.017263

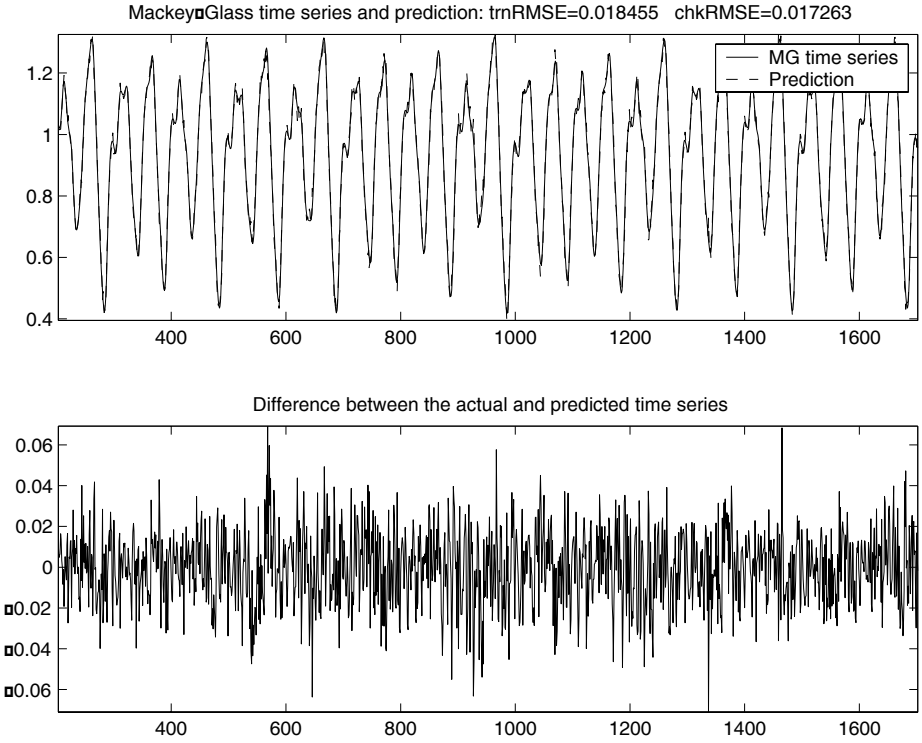


Fig. 1. Mackey-Glass time series prediction

With a performance better than that of the MLP, the NFKN with the hybrid learning algorithm requires much less computations as it does not require matrix inversions for the tuning of the synaptic weights.

The second experiment consisted in the emulation of the Mackey-Glass time series by the FKN. The values of the time series were fed into the NFKN only during the training stage. 3000 values for $t = 118,...,3117$ were used as the training set. The FKN had 17 inputs corresponding to the delays from 1 to 17, 5 neurons in the hidden layer with 5 MFs per input, and 1 neuron in the output layer with 7 MFs per synapse (total 460 adjustable parameters). The NFKN was trained to predict the value of the time series one step ahead.

The training procedure converged after 28 epochs with the final value of $RMSE_{TRN}$=0.0027 and the last 17 values of the time series from the training set were fed to the inputs of the FKN. Then the output of the network was connected to its inputs through the delay lines, and subsequent 1000 values of the NFKN output were computed. As can be seen from Fig.2, the NFKN captured the dynamics of the real time series very well. The difference between the real and emulated time series becomes visible only after about 500 time steps. The emulated chaotic oscillations remain stable, and neither fades out nor diverge. In such a way, the FKN can be used for long-term chaotic time series predictions.

Two-level structure of the rule base helps the FKN avoid the combinatorial explosion in the number of rules even with a large number of inputs (17 in the second experiment).

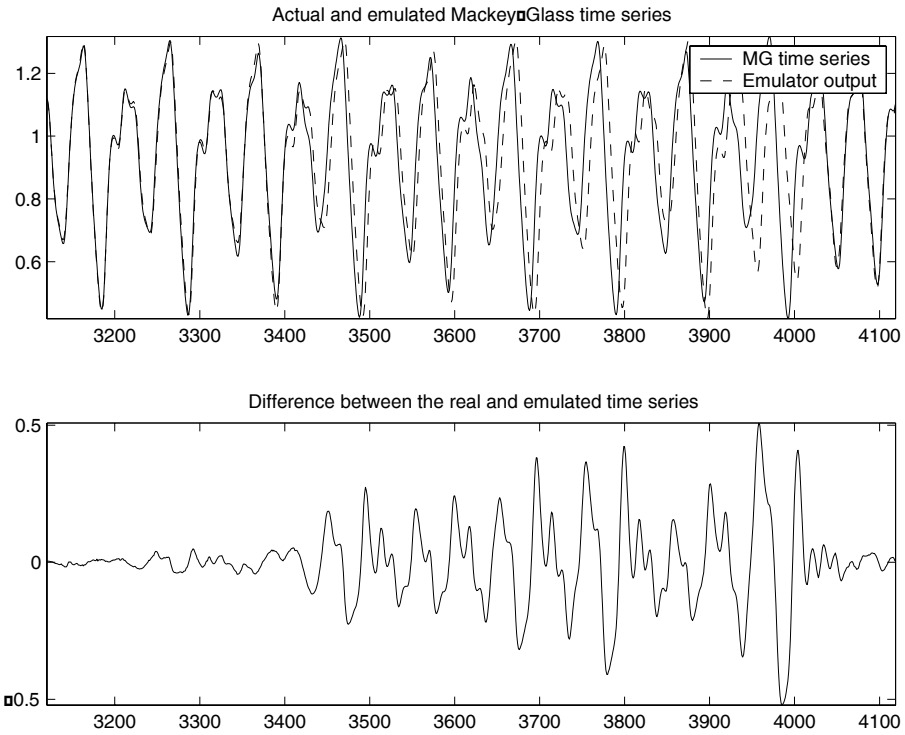


Fig. 2. Mackey-Glass time series emulation

5. Conclusion

In the paper, a new simple and efficient training algorithm for the NFKN was proposed. The NFKN contains the neo-fuzzy neurons in both the hidden and output layer and is not affected by the curse of dimensionality because of its two-level structure. The use of the neo-fuzzy neurons enabled us to develop fast training procedures for all the parameters in the NFKN.

Bibliography

1 Kolmogorov, A.N.: On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. Dokl. Akad. Nauk SSSR 114 (1957) 953-956
2 Hecht-Nielsen, R: Kolmogorov's mapping neural network existence theorem. Proc. IEEE Int. Conf. on Neural Networks, San Diego, CA, Vol. 3 (1987) 11-14
3 Sprecher, D.A.: A numerical implementation of Kolmogorov's superpositions. Neural Networks 9 (1996) 765-772
4 Sprecher, D.A.: A numerical implementation of Kolmogorov's superpositions II. Neural Networks 10 (1997) 447–457
5 Kůrková, V.: Kolmogorov's theorem is relevant. Neural Computation 3 (1991) 617-622
6 Igelnik, B., and Parikh, N.: Kolmogorov's spline network. IEEE Transactions on Neural Networks 14 (2003) 725-733
7 Yam, Y., Nguyen, H. T., and Kreinovich, V.: Multi-resolution techniques in the rules-based intelligent control systems: a universal approximation result. Proc. 14th IEEE Int. Symp. on Intelligent Control/Intelligent Systems and Semiotics ISIC/ISAS'99, Cambridge, Massachusetts, September 15-17 (1999) 213-218
8 Lopez-Gomez, A., Yoshida, S., Hirota, K.: Fuzzy functional link network and its application to the representation of the extended Kolmogorov theorem. International Journal of Fuzzy Systems 4 (2002) 690-695
9 Kolodyazhniy, V., and Bodyanskiy, Ye.: Fuzzy Neural Networks with Kolmogorov's Structure. Proc. 11th East-West Fuzzy Colloquium, Zittau, Germany (2004) 139-146
10 Kolodyazhniy, V., and Bodyanskiy, Ye.: Fuzzy Kolmogorov's Network. Proc. 8th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES 2004), Wellington, New Zealand, September 20-25, Part II (2004) 764-771
11 Kolodyazhniy, V., and Bodyanskiy, Ye.: Universal Approximator Employing Neo-Fuzzy Neurons. Proc. 8th Fuzzy Days, Dortmund, Germany, Sep. 29 – Oct. 1 (2004) CD-ROM
12 Yamakawa, T., Uchino, E., Miki, T., and Kusanagi, H.: A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. Proc. 2nd Int. Conf. on Fuzzy Logic and Neural Networks "IIZUKA-92", Iizuka, Japan (1992) 477-483
13 Kosko, B.: Fuzzy systems as universal approximators. Proc. 1st IEEE Int. Conf. on Fuzzy Systems, San Diego, CA (1992) 1153-1162

14 Jang, J.-S. R.: Neuro-Fuzzy Modelling: Architectures, Analyses and Applications. PhD Thesis. Department of Electrical Engineering and Computer Science, University of California, Berkeley (1992)
15 Mackey, M. C., and Glass, L.: Oscillation and chaos in physiological control systems. Science 197 (1977) 287-289

Authors' Information

Yevgeniy Bodyanskiy – bodya@kture.kharkov.ua
Yevgen Gorshkov – ye.gorshkov@gmail.com
Vitaliy Kolodyazhniy – kolodyazhniy@ukr.net
Control Systems Research Laboratory; Kharkiv National University of Radioelectronics;
14, Lenin Av., Kharkiv, 61166, Ukraine

НЕЙРОСЕТЕВАЯ КЛАССИФИКАЦИЯ ЗЕМНОГО ПОКРОВА НА ОСНОВАНИИ СПЕКТРАЛЬНЫХ ИЗМЕРЕНИЙ

Алла Лавренюк, Лилия Гнибеда, Екатерина Яровая

Аннотация: *разработка метода классификации наземных объектов*

Ключевые слова*: нейронные сети, спектральные кривые, классификация*

Вступление

При дистанционном исследовании объектов спектральные характеристики отраженного света могут оказаться удобным и высокоинформативным источником данных. Например, с их помощью можно оценить состояние растительности для определения степени зараженности и загрязненности [Куссуль, 2003]. Так же они представляют значительный интерес для задач изучения грунтов. Использование спектральных кривых поможет классифицировать земной покров и создать электронную карту поверхности Земли, что является актуальным на сегодняшний день. Эта задача слабоформализуема, поэтому для ее решения предлагается использовать неростетевой подход [Kussul, 2003].

Постановка задачи

Исследования проводились по данным библиотеки спектральных кривых USGS (USGS Digital spectral library). Имеется множество экспериментальных данных о спектральных характеристиках отраженного излучения некоторых наземных объектов, представленных набором кривых, иллюстрирующих зависимость интенсивности излучения от длины волны. Если анализировать общий вид спектральных кривых, то их условно можно разделить на следующие классы: вода, снег, растительность. Кривые каждого класса имеют одинаковый образ, с той лишь разницей, что различия в химическом составе объектов приводят к смещениям кривых относительно осей координат.

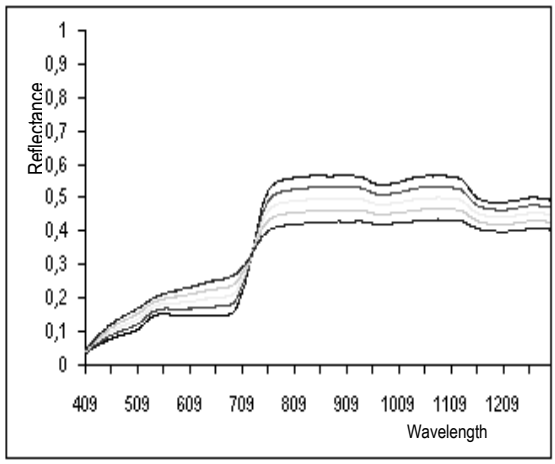


Рисунок 1. Спектральные характеристики травы

Из графиков растительности видно, что они имеют общую черту – характерный для всей зеленой растительности скачок в области видимого красного света. Но при этом их можно разделить еще на три класса: трава, хвойная растительность и лиственная, т.к. они имеют заметные различия, если анализировать образ кривой в целом (рис. 1-3).

Спектральные кривые снега делятся на два типа – одни ниспадают, другие имеют скачок в области 690-750 нм (рис. 4). Это объясняется прозрачностью снега. В тех случаях, когда сквозь снег просвечивает зеленая растительность на кривых появляется скачок в области красной составляющей видимой части

спектра, который характерен только для растений. Поэтому необходимо разделить спектральные кривые снега на два класса – снег и тающий снег с видной растительностью.

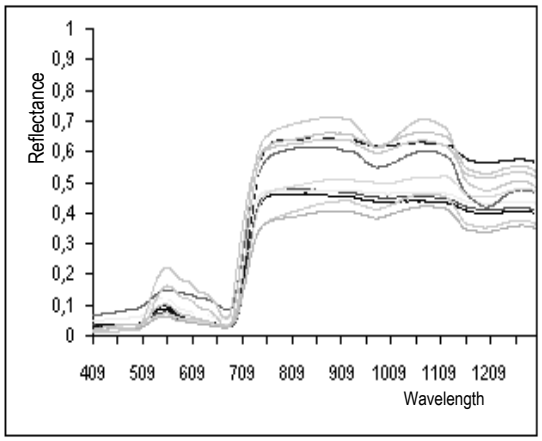


Рисунок 2. Спектральные характеристики лиственных растений

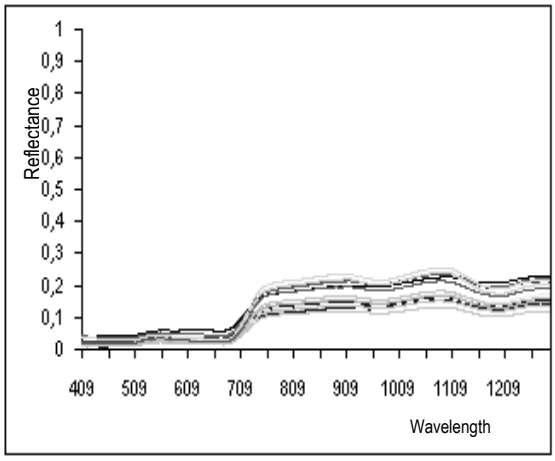


Рисунок 3. Спектральные характеристики хвойных растений

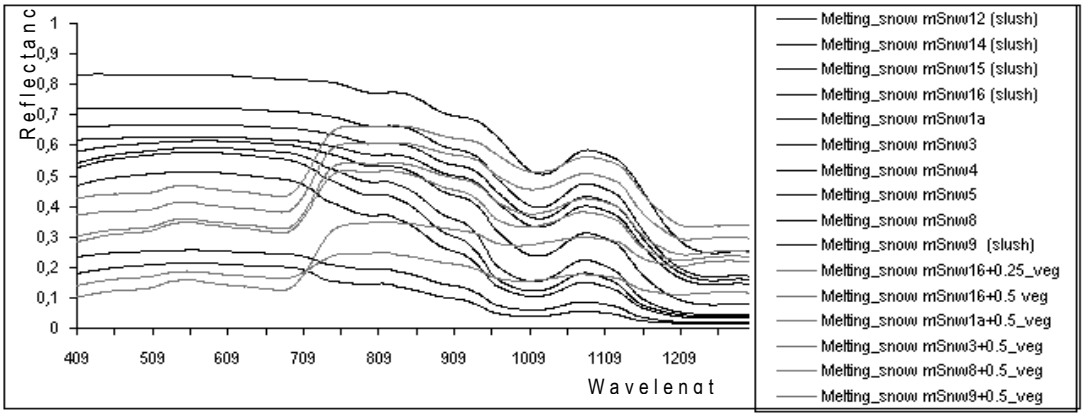


Рисунок 4. Спектральные характеристики снега

В итоге получили шесть классов: вода, снег, тающий снег с видной растительностью, трава, хвойные, лиственные. На каждый класс имеется порядка 10 спектральных кривых.

Задача идентификации наземных объектов сводится к распознаванию образов спектральных кривых и разделению их на классы. Требуется разработать эффективный метод классификации кривых. Поэтому целесообразно решать эту задачу на основе интеллектуальных методов, в частности нейронных сетей.

Решение задачи классификации с помощью нейронной сети

Основу нейронной сети (НС) составляют относительно простые, элементы, имитирующие работу нейронов мозга. Каждый нейрон характеризуется своим текущим состоянием по аналогии с нервными клетками головного мозга, которые могут быть возбуждены или заторможены. Он обладает группой синапсов – однонаправленных входных связей, соединенных с выходами других нейронов, а также имеет аксон – выходную связь данного нейрона, с которой сигнал (возбуждения или торможения) поступает на синапсы следующих нейронов. Каждый синапс характеризуется величиной синаптической связи или ее весом w_i, который по физическому смыслу эквивалентен электрической проводимости [Каллан, 2001].

Текущее состояние нейрона определяется как взвешенная сумма его входов:

$$s = \sum_{i=1}^{n} x_i \cdot w_i$$

Активационная функция

Выход нейрона есть функция его состояния: $y = f(s)$

Нелинейная функция f называется активационной и может иметь различный вид. Одной из наиболее распространенных является нелинейная функция с насыщением, так называемая логистическая функция или сигмоид (т.е. функция S-образного вида):

$$f(x) = \frac{1}{1 + e^{-\alpha x}}$$

При уменьшении α сигмоид становится более пологим, в пределе при α=0 вырождаясь в горизонтальную линию на уровне 0.5, при увеличении α сигмоид приближается по внешнему виду к функции единичного скачка. Из выражения для сигмоида очевидно, что выходное значение нейрона лежит в диапазоне [0,1]. Кроме того, сигмоид обладает свойством усиливать слабые сигналы лучше, чем большие, и предотвращает насыщение от больших сигналов, так как они соответствуют областям аргументов, где сигмоид имеет пологий наклон.

Центральная точка сигмоидной зависимости может сдвигаться вправо или влево по оси X. Это смещение может принимать произвольное значение, которое подбирается на стадии обучения вместе с весовыми коэффициентами. Такое смещение обычно вводится путем добавления к слою нейронов еще одного входа, возбуждающего дополнительный синапс каждого из нейронов, значение которого всегда равняется 1. Присвоим этому входу номер 0. Тогда текущее состояние нейронов будет определяться по следующей формуле:

$$s = \sum_{i=0}^{n} x_i \cdot w_i \quad \text{где } w0 = -T,\ x0 = 1.$$

Параметры обучения

Для обучения сетей прямого распространения применяются итеративные методы, основанные на постепенной коррекции веса межнейронных связей в направлении уменьшения ошибки реакции сети. Каждая итерация обучения, называемая эпохой, состоит в повторении всего запоминаемого материала. Поскольку число эпох может исчисляться тысячами, обучение сетей этого класса занимает много времени. Сети прямого распространения имеют многослойную организацию, причем выход каждого нейрона предыдущего слоя имеет связи с входами всех нейронов последующего слоя.

В процессе обучения может возникнуть ситуация, когда большие положительные или отрицательные значения весовых коэффициентов сместят рабочую точку на сигмоидах многих нейронов в область насыщения. Малые величины производной от логистической функции приведут к остановке обучения, что парализует НС. Эта проблема связана еще с одной, а именно – с выбором величины скорости обучения. Доказательство сходимости обучения в процессе обратного распространения основано на производных, то есть приращения весов и, следовательно, коэффициенты скорости обучения должны быть бесконечно малыми, однако в этом случае обучение будет происходить неприемлемо медленно. С другой стороны, слишком большие коррекции весов могут привести к постоянной неустойчивости процесса обучения. Поэтому в качестве коэффициента скорости обучения обычно выбирается число меньше 1, но не очень маленькое, например, 0.1, и он, вообще говоря, может постепенно уменьшаться в процессе обучения. Кроме того, для исключения случайных попаданий в локальные минимумы иногда, после того как значения весовых коэффициентов застабилизируются, коэффициент скорости обучения кратковременно сильно увеличивают, чтобы начать градиентный спуск из новой точки. Если повторение этой процедуры несколько раз приведет алгоритм в одно и то же состояние НС, можно более или менее уверенно сказать, что найден глобальный максимум, а не какой-то другой.

Фактор момента для каждого из скрытых слоев и выходного слоя определяет степень учета изменений весовых коэффициентов на предыдущем шаге обучения. Чем выше этот коэффициент, тем большее влияние на изменение весовых коэффициентов оказывают изменения на предыдущем шаге.

Организация сети

Теоретически число слоев может быть произвольным, однако фактически оно ограничено ресурсами компьютера или специализированной микросхемы, на которых обычно реализуется НС. Чем больше количество нейронов и слоев, тем шире возможности сети, тем медленнее она обучается и работает. Если нейронов и слоев слишком много, быстродействие будет низким, а памяти потребуется много; сеть будет неспособна к обобщению: в области, где нет или мало точек активационной функции выходной вектор будет случаен и непредсказуем и не будет адекватен решаемой задаче. Поэтому было

использовано небольшое количество слоев – 3 (один скрытый), чтобы сеть обобщала входную информацию, не запоминая каждую спектральную кривую, а создавая образ кривой для каждого класса.
Количество нейронов входного и выходного слоя вычисляется автоматически на основе информации о входных и выходных данных обучающей выборки, а для каждого скрытого слоя число нейронов определяется в процессе экспериментов.
Размерность входного слоя нейронной сети – это количество точек, которые содержит спектральная кривая. Каждая кривая состояла из 850 точек. При столь высокой размерности входных данных для обучения нейронной сети на большой выборке требовалось использовать сложную нейросетевую архитектуру, и процесс обучения был длительным. Поэтому мы сократили количество входных данных в две итерации. Во-первых, отбросили область кривых, где не очень заметна разница между ними (диапазон длины волны от 350 до 410 нм) и, во-вторых, усреднили кривую, заменив каждые четыре точки их средним значением по оси X и У.
В результате проведения серии экспериментов были найдены оптимальные параметры сети прямого распространения с обратным распространением ошибки:

- размерность входного слоя нейронной сети - 222
- размер скрытого слоя – 20 нейронов
- размер выходного слоя – 6 нейронов (в соответствии с количеством классов)
- функция активации – сигмоида
- веса и пороговые уровни инициализируются случайными значениями
- смещение активационной функции относительно оси X – 0,001

Значения коэффициентов:
Коэффициент скорости обучения

- Скрытый слой 0,1;
- Выходной слой 0,05;

Фактор момента

- Скрытый слой 0,15;
- Выходной слой 0,0025;

Чтобы достичь стопроцентного разделения кривых на классы при таких значениях коэффициентов количество эпох обучения составило 1000. Размер каждой эпохи обучения зависит от количества спектральных кривых предназначенных для обучения и составляет 44. В тестовой выборке 6 спектральных кривых.

Выводы

Эксперименты проводились для различных начальных значений весовых коэффициентов и различных методов обучения. По результатам экспериментов можно судить, что применение нейронных сетей для классификации земного покрова по спектральным характеристикам является эффективным. Целесообразно продолжить исследования, увеличив количество экспериментальных данных, что поможет более точно классифицировать земную поверхность.

Список литературы

[Каллан, 2001] Р. Каллан. Основные концепции нейронных сетей. Издательский дом "Вильямс", 2001
[Куссуль, 2003] Куссуль Н.Н., Марковски Дж., Яценко В.А., Саченко А.А., Скакун С.В., Сидоренко А.В.Определение содержания хлорофилла в растениях с помощью модульных нейронных сетей // Межведомственный сборник научных трудов “Кибернетика и вычислительная техника”. — 2003. — Выпуск 141. — С. 49–57.
[Kussul, 2003] Natalia Kussul, Michael Kussul, Anatoly Sachenko, George Markowsky, Sergiy Ganzha. Remote Sensing of Vegetation Using Modular Neural Networks // Proc. of the 3nd International Conference on Neural Networks and Artificial Intelligence (ICNNAI 2003). – Minsk, Belarus, November 12-14, 2003. – P. 232-234.

Информация об авторах

Алла Лавренюк – научный сотрудник, к.т.н., Институт космических исследований НАНУ-НКАУ, пр. Глушкова, 40, Киев-03187, inform@ikd.kiev.ua
Лилия Гнибеда – аспирантка, Национальный авиационный университет, пр. Комарова, 1, Киев-03058
Екатерина Яровая – аспирантка, Национальный авиационный университет, пр. Комарова, 1, Киев-03058

Section 7. Philosophy and Methodology of Informatics

7.1. Knowledge Market

THE STAPLE COMMODITIES OF THE KNOWLEDGE MARKET

Krassimir Markov, Krassimira Ivanova, Ilia Mitov

Abstract: *In this paper, the "Information Market" is introduced as a payable information exchange and based on it information interaction. In addition, special kind of Information Markets - the Knowledge Markets are outlined. The focus of the paper is concentrated on the investigation of the staple commodities of the knowledge markets. They are introduced as kind of information objects, called "knowledge information objects". The main theirs distinctive characteristic is that they contain information models, which concern sets of information models and interconnections between them.*

Keywords*: Information Market, Knowledge Market, Knowledge Information Objects, General Information Theory*

"The speaker doesn't deliver his thought to the listener, but his sounds and performances provoke the thought of the listener. Between them performs a process like lighting the candle, where the flame of the first candle is not transmitted to another flame, but only cause it."

Pencho Slaveikov, Bulgarian poet,
the beginning of the XX-th century

Introduction

The main characteristic of the Information Markets is payable information exchange and based on it information interaction. Special kinds of Information Markets are the Knowledge Markets. The main goal of this paper is to continue the investigation of the Knowledge Markets started in [Ivanova et al, 2001], [Markov et al, 2002]. Now, our attention will be paid to the staple commodities of the Knowledge Markets. The usual talk is that at the Knowledge Market one can buy knowledge. But, from our point of view, this is not so correct.

The investigation presented in this paper is based on the *Theory of Information Interaction*, which is one of the main parts of the *General Information Theory* [Markov, 1984], [Markov, 1988], [Markov et al, 1993], [Markov et al, 2003].

Firstly, let remember some basic concepts of the Theory of Information Interaction.

At the first place, we need to remember the concept "INFOS". Its genesis started from the understanding that the concept *"Information Subject"* is perceived as a single human characteristic. It is clear that in the nature there exist many creatures, which may be classified to this category, especially groups of persons and societies.

To exclude the misunderstandings we decide to introduce new word to denote all possessors of the characteristics of the Information Subject. This word is **"*INFOS*"** [Markov et al, 2003], [Markov et al, 2004].

On given level of complexity of the entities, a new quality becomes - the possibility of self-reflection and internal activity appears. One special kind of activity is the secondary (information) one. The secondary activity need to be resolved by relevant possibilities of the entities from the environment. So, not every entity may be used for resolving the secondary activity. This way, the entity needs a special kind of (information) contacts and (information) interaction for resolving the information activity.

The entity, which has:

- ***(primary) activity*** for external interaction;
- possibility for ***reflection***, i.e. possibility for collecting the information;
- possibility for ***self-reflection***, i.e. possibility for generating secondary (information) activity;
- ***information expectation*** i.e. available (secondary) information activity for internal or external contact for resolving it

is called ***Infos***.

The resolving of the information activity is ***the goal*** of the Infos. This goal may be achieved by the establishment and providing (information) contacts and (information) interaction, which are remembered below.

Information Objects

When the Infos interact with the entities around in the environment, there exist at least two cases of reverberation:

- the contacts and interaction are casual and all reflections in the Infos as well as in the entities have casual origin;
- the contacts and interactions are determined by the information activity of the Infos.

In the both cases, the contacted entity may reflect any information model from Infos. The concept “information model” has been defined in [Markov et al, 2001]. In general, the information model is a set of reflections, which are structured by Infos and, from his point of view, represents any entity.

An entity, in which one or more information models are reflected, is called ***"information object".***

The information objects can have different properties depending on:

- the kind of influence over the entities - by ordering in space and time, by partial or full modifying, etc.,
- the way of influence over the entities - by direct or by indirect influence of the Infos on the object,
- the way of development in time - static or dynamic,

etc.

Information Operations

The information is reflected relationship, i.e. it is a kind of reflection [Markov, 1988]. Therefore, the only way Infos to operate with information is to operate with the entity that contains it. Every influence on the entity may cause any internal changes in it and this way may change the information already reflected. Another type of influence is to change the location of entity or to provoke any contact between given entity and any other.

The influence over the information object is called "***information operation***" if it is determined by any Infos information activity.

The information operations may be of two main types:

- the Infos internal operations with the sub-entities that contain information,
- external operations with the information objects that contain information.

The internal operations with the sub-entities closely depend of the Infos' possibilities for self-reflection and internal interaction of its sub-entities.

The self-reflection (self-change) of the Infos leads to the creating of new relationships (and corresponding entities) in it. These are *subjectively* defined relationships, or shortly - *subjective relationships*. When they are reflected in the memory of the Infos they may initiate any new information model on a higher level. In such case, a relation between reflected relationships appears. The high-level information models may have not real relationships and real entities that correspond to them.

For instance, the possibility for creating the information models of similarity is a basis for realising such very high level operations as "comparing elements or substructures of the information models", "searching given substructure or element pattern in the part or in the whole structure of the information model", etc.

It is clear, the Infos is built by entities some of which may be also Infoses, but on the lowest levels. For instance the society and single human who belongs to it. So, the internal operations are determined by the concrete internal level and from the point of view of these low levels, they may be assumed as external operations. Because of this, we will concentrate out attention on the second type of operations.

The external operations with information objects may be differed in two main subtypes:

– basic information operations;
– service information operations.

There are two basic information operations which are called I-operations:

- **I-reflection** (reflecting the information object by the Infos, i.e. the origination of a relevant information model in the memory of the Infos);
- **I-realisation** (creating the information object by the Infos).

In the process of its activity, the Infos S reflects (perceives) information from the environment (entities $O_{I,}$ i=1,2...) by proper sub-entities (sensitive to video, acoustic, tactile, etc. influences) called "***receptors***" R_i (i=1,2...). Consequently, the Infos may receive some information models.

The Infos subjective reflection is called "***I-reflection***".

When necessary, the Infos can realise in its environment (entities O'_j, i=1,2...) some of the information models, which are in his memory, using some sub-entities called "***effectors***" E_j (j=1,2...). Consequently, new or modified already existing entities may reflect the information, relevant to these information models.

The Infos subjective realisation is called ***"I-realisation"***.

There are several operations, which can be realised with the information objects: transfer in space and time, destroying, copying, composition, decomposition, etc. Because of the activity of the Infoses, these operations are different from other events in reality. In this case, such Infos determined operations with information objects are called "service information operations".

For example, some of the very high-level service operations are based on the external influence on the information object to change any existing reflection:

- Including and removing an element in and from the object's structure;
- Copying or moving object's substructures from one place to another;
- Building new object's structure using as a basis one or several others;
- Composing or decomposing of object's elements or substructures;

Etc.

Information Processes

Let "O" is a set of real information objects i.e. $O = \{O_{ij} \mid i= 1,..,n; j=1,..,m\}$.

Let "I_s" is a set of information models in Infos S, i.e. $I_s = \{i_p \mid p=1,..,q\}$.

If the opposite is not stated, we will consider:

- every set of information objects is an information object,
- every set of information models is an information model.

Every information operation "t" can be treated as a function between two sets of information objects, which may be coincidental, i.e. t: $O_d \rightarrow O_r$.

I-realisation can be considered as a function m: $I_s \rightarrow O$.

I-reflection - in the opposite - as a function r: $O \rightarrow I_s$.

Let t_1, t_2,..,t_n are information operations. The consequence of information operations P created by any composition, i.e.

$$P = t_1 \circ t_2 \circ \ldots \circ t_n$$

is called "***information process***".

It is possible that some of t_i, i=1,..,n may be I-realisation or I-reflection.

In particularly an information process can include only one operation.

Information Contact

If an information model from the Infos is reflected in another entity, there exist possibility, during the "a posterior" interactions of the given entity with another Infos, to transfer this reflection in it. This way an information model may be transferred from the Infos to another.

If the second Infos has already established information expectation, the incoming reflection will be perceptible for him. The information expectation will be resolved in some degree and the incoming information model and information in it will be received by the second Infos.

Let S_1 and S_2 are Infoses and O is an arbitrary entity.

The composition of two real contacts Θ_1 and Θ_2 :

$$S_1 \xrightarrow{\Theta_1} O \xrightarrow{\Theta_2} S_2$$

is called "***information contact***" between Infos S_1 and Infos S_2 iff during the contacts any information model from I_{S1} is reflected in the I_{S2} through the entity O.

The Infos S_1 is called "***information donor***", the Infos S_2 is called "***information recipient***", and, of course, the entity O is called "***information object***".

In this case, when the donor and the recipient are different Infoses the information contacts between them consist of the composition of at least two information operations - I-realisation and I-reflection. For the realisation of any direct information contact between two different Infoses is necessary the execution of the composition of these two "basic" operations. All the rest information operations are necessary for supporting the basic ones i.e. they are auxiliary (service) operations.

This way the elementary communicative action will be provided.

In general, every information process "c", having as a start domain the set I_{Sd} of information models of the Infos S_d and as a final domain the set I_{Sr} of information models of the Infos S_r, (I_{Sd} and I_{Sr} may be coincidental)I,

$$c: I_{Sd} \rightarrow I_{Sr}$$

is called ***"information contact"*** between S_d and S_r:

Note that for the realisation of one information contact at least one information object is necessary.

Information Interaction

The set "R" of all information contacts between two Infoses S_a and S_b

$$R= \{c_i \mid i=1,2..;\ c_i: I_{Sa} \rightarrow I_{Sb}\}$$

is called "***information interaction***".

When S_a and S_b are coincident, we call it Information interaction of the Infos with itself (through the space and time).

The set "B" of all information objects, used in the information interaction between given Infoses is called "***information base***".

Information Society

The **"*Information Group*"** (IG) is a set of Infoses, with common Information base of the information interactions between them.

In the small IG the service information operations may be provided by the every Infos. In the large IG this is impossible or not optimal. In such case, some Infoses became as "*information mediators*" between the others. They start to provide the service information operations. They realise **"*Information Service*"**.

The "Information Society" is an IG with internal Information Service.

Information Market

Now we are ready to continue with introducing the basic ideas of the Information Markets.

Up to this moment, the discussion about essence of the information society has not resulted in uniform definition. Everyone from his point of view defines this stage of development of a society.

It is clear, at the stage of social growth, called “information society”, for existence of the separate individuals or social teams the information and information activity get decisive value. Certainly, at earlier stages of development of mankind, the information had the important value too. But never, in all known history, the other means for the existence have been so dominated by the information means as it is in the information society.

So, the direct conclusion is the understanding that ***the information society differs from the other levels of the human been growth by the domination of the information interests above all others.***

From the origin, the human society has been the "information" one, but the levels of the information service differ in the different periods of the existence of the societies. So, it is possible to allocate the following levels:

- **Primitive information society** (people having knowledge, letters on stones etc.);
- **Paper information society** (books, libraries, post pigeons, usual mail etc.);
- **Technology information society** (telephone, telegraph, radio, TV, audio- and video-libraries etc.);
- **High-Technology information society** (automated systems of information service, local computer information networks etc.);
- **Global information society** (global systems for information service, opportunity for every body to use the information service with help of some global network etc.).

The information society does not assume compulsory usage of the information services by the part or all inhabitants of given territory. One very important feature thus is emphasized: for everyone will be necessary diverse and qualitative (from his point of view) information, but also everyone can not receive all necessary information. The enterprising experts will accumulate certain kinds of the information and will provide the existence through favourable to them information exchange with the members of the society. Thus, in one or other form, they will carry out **payable information service (granting of information services for some income)** [Ivanova et al, 2001]. This is the background of the Information Market.

The payable information exchange and services regulated by the corresponded laws and norms as well as by the government protection of the rights of the participants (members) of this kind of social interactions form the ***Information Market***.

So, at the centre of discussion, we have discovered a simple true: ***in the information society the payable information exchange and services will dominate above all other market activities.*** In other words, **the Information Market** dominates over all other type of markets of the information society.

Knowledge Information Objects

V.P. Gladun correctly remarks that the concept “knowledge” does not have common meaning, especially after beginning of it’s using in the technical lexicon in 70-ies years of the last century. Usually, when we talk about the human knowledge we envisage all information one has in his mind. Another understanding sets the “knowledge” against the “data”. We talk about data when we are solving any problem or are making logical inference. Usually the concrete values of the given quantities are used as a data as well as the descriptions of the objects and interconnections between objects, situations, events, etc. During decision making or logical inference we operate with data involving some other information like descriptions of the solving methods, rules for inference of the corollaries, models of the actions from which the decision plan is formed, strategies for creating decision plans, and general characteristics of the objects, situations, and events. In accordance with this understanding, the “knowledge” is information about processes of decision making, logical inference, regularities, etc., which applying to the data creates any new information. [Gladun, 1994].

The usual understanding of the verb "to know" is: "to have in the mind as the result of experience or of being informed, or because one has learned"; "to have personal experience of smt.” etc. The concept "knowledge" usually is connected to concepts "understanding" and "familiarity gained by experience; range of information" [Hornby et al, 1987] or "organized body of information" [Hawkins, 1982].

In other words, the knowledge is a structured or organised body of information models, i.e. the knowledge is information model, which concerns a set of information models and interconnections between them.

In accordance with this the information objects, which contain such information models are called ***“knowledge information objects”***.

This definition corresponds to everyday understanding of the concept “knowledge”. For instance, during the process of education the presented above operations **I-realization** and **I-reflection** correspond to creating and perceiving the ***“knowledge information objects”***.

Knowledge Market

The growth of the societies shows that the knowledge information objects become important and necessary articles of trade. The open social environment and the market attitudes of the society lead to arising of the knowledge customers and knowledge sellers, which step-by-step form the **"Knowledge Markets"** [Markov et al, 2002].

As the other markets, the **Knowledge Market** is the organised aggregate of participants, who operate following common rules and principles. The knowledge market structure is formed by a combination of mutually-connected elements with simultaneously shared joint resources.

The staple commodities of the knowledge market are the knowledge information objects.

The knowledge information bases and tools for processing the knowledge information objects, such as tools for collecting, storing, distributing, etc., form the **knowledge environment**.

The network information technologies enable to construct uniform ***global knowledge environment***.

It is very important, that it will be friendly for all knowledge market participants and open for all layers of the population without dependence from a nationality, social status, language of dialogue, place of residing. The decision of this task can become the important step of humanization of all world commonwealths.

In the global information society, on the basis of modern electronics, the construction of the global knowledge market**,** adapted to the purposes, tasks and individual needs of the knowledge market participants is quite feasible, but the achievement of this purpose is connected to the decision of a number of scientific, organizational and financial problems.

For more clear explanation let consider an example about the correspondence between concepts “information object” and “knowledge information object”. When an architect develops any constructive plan for future building, he creates a concrete “information object”. Of course, he will sell this plan. This is a transaction in the area of the Information Market. Another question is from where the architect has received the skills to prepare such plans.

It is easy to answer – he has studied hardly for many years and received knowledge is the base for his business. So, we see that the textbooks are not concrete information for building concrete house, but they contain the information needed for creating such plans. The textbooks written by the lecturer in the architectural academy are special kind of "information objects" which contain special generalized information models. They are "knowledge information objects" and these textbooks have been sold to the students. It is clear; here we have a kind of transactions at the "Knowledge Market".

At the end, we need to take into consideration the difference between responsibility of the architect and the lecturer. If the building collapses the first who will be responsible is the architect, but never the lecturer!

Conclusion

In this paper, we introduced the "Information Market" as a payable information exchange and based on it information interaction. In addition, special kind of Information Markets - the Knowledge Markets were outlined.

The identifying of the staple commodities of the knowledge markets was a step of the process of investigation of contemporary situation in the global knowledge environment.

The investigation of the staple commodities of the knowledge markets is very difficult but useful task. In this paper we introduced them as kind of information objects, called "knowledge information objects". The main theirs distinctive characteristic is that they contain information models, which concerns sets of information models and interconnections between them.

This way, we have seen the usual talk that at the Knowledge Market one can buy knowledge is not so correct. But in everyday language in is accepted to say "knowledge" with the meaning of the "knowledge information object". We need specially to say that there exists another meaning of knowledge, which points to the information models into the Infos memory. Usually we do not distinct these two meanings. When we say "receiving of knowledge" we assume the I-reflection operation; when we say "generating the knowledge" we assume the I-realization operation; and at the end, when we say simply "knowledge" it is context depended to understand what is the meaning – knowledge information models into the Infos memory or those which the knowledge information objects contain. However, if we remember that the Infoses are built by entities some of which may be also Infoses, but on the lowest levels, the internal memory of given level of the organisation of the Infos may be considered as external set of information objects on the lower levels.

This has important role for future research of this social and information phenomenon.

Acknowledgements

This paper is partially supported by the project **ITHEA-XXI**.

Bibliography

[Gladun, 1994] Гладун В.П. Процессы формирования знаний. София, Педагог 6, 1994.

[Hawkins, 1982] Hawkins J.M., *The Oxford Paperback Dictionary*. Oxford University Press, 1982, ISBN 0-19-281209-2.

[Hornby et al, 1987] Hornby A.S., A.P. Cowie, A.C. Gimson, *Oxford Advanced Learner's Dictionary*. Oxford University Press. 1987, ISBN 0-19-431106-6.

[Ivanova et al, 2001] N. Ivanova, K. Ivanova, K. Markov, A. Danilov, K. Boikatchev. *The Open Education Environment on the Threshold of the Global Information Society*. IJ ITA, 2001, V.8, No.1 pp.3-12. (Presented at Int. Conf. KDS 2001 Sankt Petersburg, 2001, pp.272-280, in Russian, Presented at Int. Conf. ICT&P 2001, Sofia, pp.193-203)

[Ivanova et al, 2003] Кр.Иванова, Н.Иванова, А.Данилов, И.Митов, Кр.Марков. *Обучение взрослых на рынке профессиональных знаний*. Сборник доклади на Национална научна конференция "Информационни изследвания, приложения и обучение" (i.TECH-2003), Варна, България, 2003. Стр. 35-41.

[Markov, 1984] K.Markov. *A Multi-domain Access Method.* Proceedings of the International Conference on Computer Based Scientific Research. PLOVDIV, 1984, pp.558-563.

[Markov, 1988] Кр.Марков. *От миналото към бъдещето на определението на понятието "информация".* Сборник "Програмиране '88". БАН, Варна, Дружба, 1988, стр.150.

[Markov et al, 1993] K.Markov, K.Ivanova, I.Mitov. *Basic Concepts of* a *General Information Theory.* IJ ITA {International Journal "Information Theories & Applications"}; 1993; v.1, n.10; pp.3-10.

[Markov et al, 2000] Kr. Markov, Kr. Ivanova, I. Mitov, N. Ivanova, K. Bojkachev, A. Danilov. *Co-operative Distance and Long-Live Learning*. ITA-2000, Bulgaria, Varna, 2000.

[Markov et al, 2001] K. Markov, P. Mateev, K. Ivanova, I. Mitov, S. Poryazov. *The Information Model*. IJ ITA, 2001, V.8, No.2 pp.59-69 (Presented at Int. Conf. KDS 2001 Sankt Petersburg, 2001, pp.465-472)

[Markov et al, 2002] K. Markov, K. Ivanova, I. Mitov, N. Ivanova, A. Danilov, K. Boikatchev. *Basic Structure of the Knowledge Marke*t. IJ ITA, 2002, V.9, No.4, pp.123-134 (Presented at Int. Conf. ICT&P, 2002, Приморско)

[Markov et al, 2003] Kr.Markov, Kr.Ivanova, I.Mitov. *General Information Theory. Basic Formulations.* FOI-COMMERCE, Sofia, 2003. ISBN 954-16-0024-1.

[Markov et al, 2004] K. Markov, K. Ivanova, I. Mitov, E. Velikova-Bandova. *Formal Definition of the Concept "INFOS"*. Int. Journal "Information Theories and Applications", 2004, ISSN 1310-0513. V.11, N.1, pp.16-19. (Presented at i.TECH 2004, Varna, Bulgaria. pp. 71-74)

Authors' Information

Krassimir Markov – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Institute of Information Theories and Applications FOI ITHEA, foi@nlcv.net

Krassimira Ivanova – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, foi@nlcv.net

Ilia Mitov – Institute of Information Theories and Applications FOI ITHEA, foi@nlcv.net

BASIC INTERACTIONS BETWEEN MEMBERS OF THE KNOWLEDGE MARKET

Krassimira Ivanova, Natalia Ivanova,
Andrey Danilov, Ilia Mitov, Krassimir Markov

Abstract: *The interconnections and information interactions between main members of the Knowledge Market are presented in the paper.*

Keywords: *Knowledge Market, Components of the Knowledge Market, Information Interaction*

Introduction

The growth of the global information society shows that the information and, especially – knowledge, becomes important and necessary article of trade. The open environment and the market attitudes of the society lead to arising of the knowledge customers and knowledge sellers, which step-by-step form the "Knowledge Markets". As the other markets the Knowledge Market is the organised aggregate of participants, which operates in the environment of common rules and principles [Markov et al, 2002].

The Structure of the Knowledge Market was presented in [Markov et al, 2002]. Let's remember its basic elements.

Usually a person or enterprise, called **Employer (Er)**, hires **Employees (Ee)**, who have exact skills and knowledge and transform them in real products or services during the work processes. This process is served by the Manpower Market. But the Employees, even owning a high education level, need additional knowledge to solve the new tasks of the Employers. In this moment they became **customers of new knowledge**, who arouse the necessity of the Knowledge Market, which should rapidly react to the customers' requests. **In other words, the Manpower's Market causes the appearance of the Knowledge Market.** These two members of KM form one side of the market – the knowledge customers.

The continuous changing of technological and social status of the society leads to appearance of new category – **Consultants (C)** – peoples/organisations, who have two main tasks:

- to promote new technologies to Employers in convenient way to implement them in practice;
- to determine the educational methods for training the staff for using the new technologies.

The educational process is carried out by the **Lecturers (L)**, who transform new scientific knowledge into the pedagogical grounded lessons and exercises.

During the realising the concrete educational process the Lecturer is assisted by **Tutor (T)** who organises the educational process and supports the Employees to receive the new knowledge and to master theirs skills.

At the end of the educational process, a new participant of KM appears – **Examiner (E)** – who tests the result of the process and answers to the question "have the necessary knowledge and skills been received".

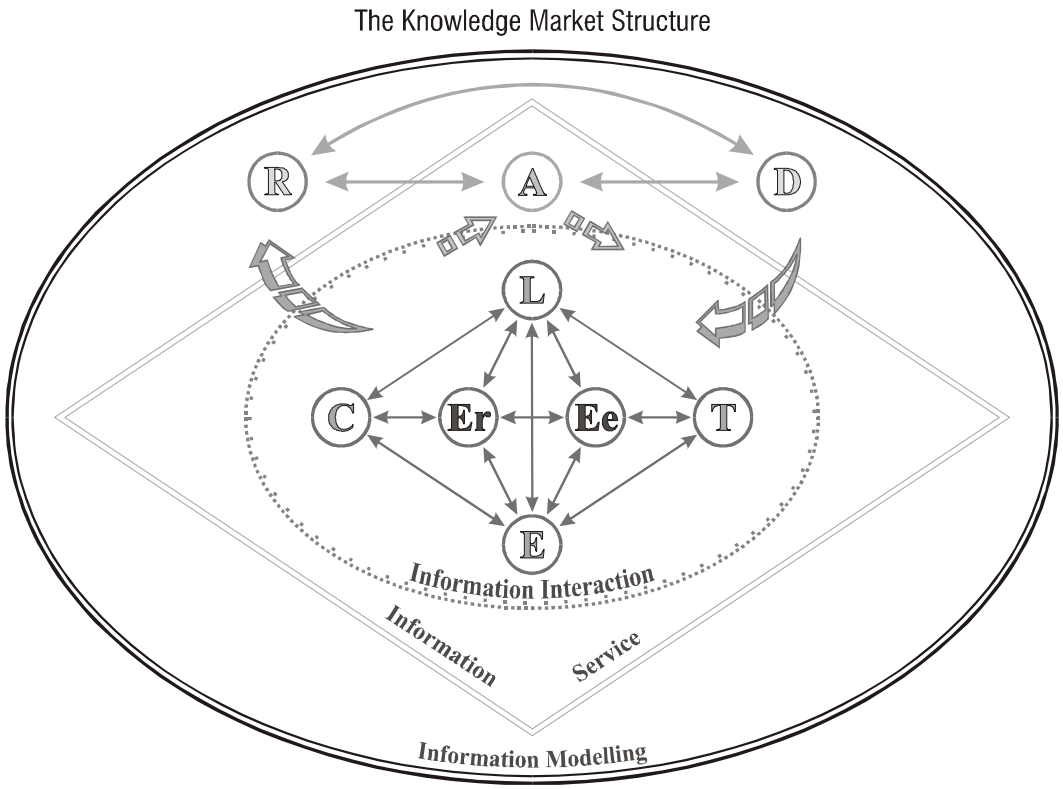
These six Components of the Knowledge Market, which contact each other via global information network, form the first level of the knowledge market, called the level of the **"information interaction"**.

As far as these components are too much and are distributed in the world space, the organisation and co-ordination of theirs information interaction needs adequate **"information service"**. It is provided by a new component called **Administrator (A)**. Usually the Administrators are Internet and/or Intranet providers or organisations.

The rising activity of the knowledge market creates the need of developing modern tools for the information service in the frame of the global information network. This causes the appearance of the high knowledge market level, which allows the observing the processes, as well as developing and implementing new systems for information service. This is the **"information modelling"** level.

It consists of two important components – the **Researchers (R)** and the **Developers (D)**.

On the figure below is presented the scheme of the basic structure of Knowledge Market.



Of course, the Knowledge Market as a kind of Market follows the rules and laws given by the social environment. The interrelation between government and social structure and Knowledge Market need to be studied in separated investigation.

In several papers we have already investigate different problems of the Knowledge Market.

Some of the results given below are received during these works. For five years we have seen that the Knowledge Market is very important scientific area and need to be investigated. The main received results are given in [Markov et al, 2000a], [Markov et al, 2000b], [Ivanova et al, 2001], [Boikatchev et al, 2001a], [Boikatchev et al, 2001b], [Markov et al, 2002], [Ivanova et al, 2003], [Markov et al, 2003].

In global information society the e-commerce becomes as fundamental way for financial support of the Knowledge Market. The advantages of e-commerce are obvious. In the same time there exist many risks for beginners at this kind of market. From this point of view the society need to provide many tasks for training the citizens to use properly opportunities of the new environment [Markov, 1999].

The Interconnections between Members

In this section, the interrelations between members of the Knowledge Market are outlined. For easy reading we will denote the Knowledge Market's members with abbreviated names as they were given in the structure above. The next convention concerns the style of description of the corresponded interconnections. We consider that in every such interconnection the both participants have equal rights and the order of writing is not important. For instance, the denotation "L – T" represents the interconnection between Lecturers and Tutors without any precedence of any of them.

Er – Er

The group of the Employers in given domain causes the growing of the Knowledge Market in this direction.

The products of some Employers develop the new technologies, which are used from other Employers. This causes the development of the Knowledge Market in these new directions.

The role of the Employer from the point of view of the Manpower market is to buy and hold more qualified and skilled workers for less money in competition with other Employers.

Er – Ee

The Employer buys the result of the work of the Employee, which is closely depended to the received knowledge and skills, in one hand, and the possibilities of the Employee to realise them in concrete results in everyday work, on other hand.

The Employee informs the Employer for already received knowledge and skills or his intention for future. He need to prove this knowledge by correspond qualification procedure.

The Employee has to realise the requested knowledge and skills during the working process, because the Employer pays for the real results of the work, but not for the possibilities.

Er – C

The Employer can take advises from the Consultant for optimisation or future expansion of his business.

The Employer may ask the Consultant about information for the other components of the Knowledge Market or for the Knowledge Market as whole.

Special Interest for the Employer may be the data about the degree of the knowledge and skills of the Examiners, because of the importance of theirs role for the employers business.

One of the main tasks of the Consultants is to advise the Employers who can be the appropriate Lecturers or Examiners for the manpower they need.

During the interaction with the Consultant the Employer defines his intentions for the future of his business. This way the Consultant receives information about the future growing of the Knowledge Market. The Consultant investigates this growing by the interaction with the two types of Employers:

- Manufacturers of new products and technologies, which production will income on the real market;
- Users of the existing products and technologies.

The first type of Employers grows up a new niche of the Knowledge Market – the knowledge about using, service, etc. of theirs new products, technologies or theories.

The second type influences over the increasing or decreasing of that niche. From other side, the investigation of this group can focus the attention to the new problem, which solving may generate a new niche in the Knowledge Market.

Er – L

The Employer may consult the Lecturer about the present' and future' needs and changes of the business, he manages. The initiative for such consulting process usually belongs to the Lecturer, or to the Consultant. The Employer becomes as the corrective of the futures Lecturers' work.

During the interaction with the Employers, the Lecturer receives theirs:

- requests for training and additional education of new skilled Employees;
- requests for consulting the specialists already appointed in the enterprises of the Employers;
- requirements for preparing adequate educational programs and courses for theirs Employees;
- In the same time, the Lecturer gives information to the Employers for the level of the skills received by the Employees during the implemented courses.

Er – T

The direct interactions between Employer and the Tutor are not so strong, i.e. the direct interaction between the Tutor and the Employer is not always present.

In one side, the Tutor organises the educational process in correspondence to the requirements of the Employers' business work environment. Usually the Consultant defines these conditions directly or through the Lecturer. The Employer defines the concrete duties, which usually are subset of the Consultant' recommendations.

Meanwhile the Tutor contacts directly with the Employer, if it's a question of consultation, check of the level of the skills and knowledge or re-education of already hired Employees.

The Tutor has to establish appropriate educational environment for training the Employees to satisfy given requirements.

In other side, the Tutor can receive the concrete information for the current Employees, which information may facilitate the learning process.

Er – E

The interaction between Employer and Examiner is aimed to define the specific characteristics, which the Employee has to satisfy. These characteristics are subset of the general requirements and rules for given business activity. Usually the Consultant in collaboration with the Employer generates the definition of the specific rules. It is possible, during the process of qualification of the Employee, the set of rules to be changed in direction to extend or to shrink in correspondence to:

- the real needs and possibilities of the Employer;
- the set of candidates for given workplace.

The Employer can ask the Examiner to examine some of the employees, which already work for the Employer in accordance with the current (or near-future) duties-rules.

The Examiner informs the Employer for the Employees results in the education course.

Er – A

It is clear that the concrete needs of the given Employer are subset of the general Employers' information model. In every concrete case as well as for every concrete event the Employer needs specialised administration support.

The Administrator interacts directly with the Employer for:

- co-ordinating and service the information interaction between him and other Employers as well as with all other Knowledge Market participants;
- advising him how to use the information service tools to establish interconnections;

- placing the tools and lines for distant information exchange at Employers' disposal;
- supporting the creation of the specialised internet and/or intranet sites and other communication possibilities for representing the Employers;
- Software and hardware support of Employers' activities.

Er – R&D

There are several sources for financing the scientific research and its implementation in the Knowledge Market.

The main sources are the Employers. They may order specialised research and developing of correspond tools, which may extend service of the Knowledge Market in the new directions.

Another source is the government or public financial support. In such case the corresponded authorised institutions became as specific Employers but the scientific results could not be implemented immediately. This possibility is very important because of the increase in the production of scientific "garbage".

As corollary we have two types of interaction between Researchers, respectively – Developers, and Employers:

- Interaction based on concrete request and corresponded scientific or practical projects;
- Interaction determined by future prognostic government, social or scientific goals, needs, expectations, etc.

Ee – Ee

The connections between Employees are mainly in two dimensions:

- collaboration with others during the educational process;
- competition with others in the Manpower market as well as during the educational process.

The collaboration is very important because of possibility to receive new knowledge from “colleagues” and the need to discuss in free style the problems arisen from the tasks given by the Lecturer and the Tutor.

Competition has significant role in growing the personal knowledge and skills.

Ee – C

The Employee receives information from the Consultant usually through the other components of the Knowledge Market and in rare case directly. This information concerns the present and future growth of the Manpower and Knowledge Markets and serves the orientation of the Employee in the social environment. This knowledge may help the Employee to make decision for his future professional and social growing.

In addition, the Consultant makes studies on the state of the Employees’ commonwealth and generates conclusions about the advance or reduction of any parts of Knowledge Market.

Ee – L

From point of view of the Employee the Lecturer is integrated set of education sources, from which he can receive general or specialized knowledge and skills.

The Lecturer transfers knowledge to the Employees, using modern methods and devices of delivering. He participates in development of the training program, study materials, study tasks and exercises that the Employees should be carried out during the training before beginning of the educational process.

During preparation of a course the Lecturer should give study materials, which can be written in appropriate formats and media and the references to the additional sources of information.

The Lecturer may be connected with each Employee directly and has access to the discussion forum and the Employees’ diaries where he can estimate the Employees’ knowledge and make the comments.

Ee – T

One very important point of the interaction between the Employee and the Lecturer is to determine the currently received knowledge and skills level of the Employee for choosing the appropriate approach and capacity of the material for education. This process needs to be supported by the Tutor.

The tutor directs and plans the education of the Employee. He also plans and organizes Employee's interaction with the Lecturer and informs the Employee about all the courses that he can get from the Lecturer.

The Employee demands from the Tutor maximal effective organization of the educational process for acceptable expenses.

Let's remark that the concept of acceptable expenses closely depends of financial status of the Employee. These expenses include not only concrete charges for education, but also the charges for existence of the Employee during the educational process.

Ee – E

The Employee must prove to the Examiner that his received knowledge and skills cover chosen qualification level. If the Employee does not cover all the requirements, the Examiner need to explain in appropriate form what the Employee have to learn in addition. These explanations need to be directed to the Tutor (respectively to the Lecturer) for future extension of the education of this Employee, his group or whole educational process.

The Examiner conducts the Employee's attestation on different levels of the education course. Within the interaction with the Employee, he gives him the terms and the conditions of the test or exam Employee's about take. The Examiner sets the tests and gets the Employee's answers.

At the end, the Examiner takes a decision for a final or level test and gives the Employee a certificate based on the tests.

Ee – A

The Employee interact with the Administrator for receiving the appropriate information service for every of given above information interactions.

The concrete needs of the given Employee are subset of the general Employee' information model.

The specific characteristics of the Employee such as physiological characteristic, language specific, professional background, national and religious affiliations, the level of skills for using the information service, etc. strongly influence to the type and degree of the automated information service, provided and conducted by the Administrator.

Ee – R&D

The interaction of the Employee with the Researches and Developers mainly is based on the using of the tools and approaches for information service of the education and self-education.

In this interaction the Employee usually play passive role.

As a rule, the Researchers carry out specialized research or general investigation and developing of corresponding tools for information service. These tools are provided for the specific characteristic of given group of Employees or, respectively, for common (future) needs of the Employees' group.

C – C

The interaction between Consultants may be structured at least on three levels:

- scientific level;
- business level;
- state (government) and social level.

At the scientific level, the Consultants interact to investigate new theoretical or practical domains and to generate new common knowledge. As a rule, this interaction is based on the scientific norms for collaboration and exchange of the new ideas. Some political or socioeconomic processes as well as the business restrictions in the given field may restrict this interaction.

At the business level the main goal of interaction between Consultants is to extend already existing knowledge for decision of concrete practical problems. It is important to remark that often the experts in the same domain do not collaborate because of the concurrency between them.

At the state (government) and social level the Consultants play subordinate role to the Government or Social Institutions, such as the parliament, ministry, syndicates, associations, foundations etc. The knowledge of the consultants plays advising role for the decision making of the governing organization or institution. In such way this knowledge became active in the Knowledge Market.

C – L

The interaction between Consultants and Lecturers has significance place in the Knowledge Market. The Consultants are the sources of the new knowledge and the Lecturers need to be continuously in touch with them.

In many cases the Lecturers collect parts of knowledge from different Consultants and integrate it in educational materials. This way the Consultants may affect to the educational processes in the Knowledge Market.

Beside of this, the Consultants may influence the Knowledge Market by theirs ability to investigate the results of the previous cycle of education and to draw conclusions for eventual corrections in the educational processes.

At the end, the Consultants are the persons or organisations that can and have to test and certify the Lecturers in given knowledge domain.

C – T

At glance, the Consultants and the Tutors not interact at all. The interaction may be realised indirectly by other components of the Knowledge Market – mainly by the Lecturers.

Nevertheless, they exchange via Knowledge Environment much information for practical implementation of the Consultants recommendations for Employees training.

C – E

The main goal of the interaction between Consultants and Examiners is to clear the real needs of the Employers for high qualified and skilled workers. The Examiners need exact information what the Employees really have to know. The qualification procedure includes large scale of tests and other examination steps. For every one, the Consultant may be the source of the knowledge.

In this interaction the Consultants may influence by:

- conclusions for the results of the previous qualification cycles;
- certifying the Examiners in given knowledge domain;
- playing role of the Examiner.

The Examiner receives from the Consultant new test materials and education methods of conducting the attestation activities and informs the Tutor for the results of the use of the new attestation materials and methods.

C – A

The Consultant interact with the Administrator for receiving the appropriate information service for every of given above information interactions.

A specific characteristic of this interaction is the ability of the Consultant to interrupt any of the educational processes in the Knowledge Market via the Administrator's support. This interruption need to be done when from point of view of the Consultant:

- any new events, external for the educational process, appear and the structure of given steps of the educational process need to be changed;
- any deviations in the educational process have registered by the Consultant.

In this role the Consultant can be qualified as a regulator of the Knowledge Market. This is very important that the main way the Consultant can play this role is by the support of the Administrator.

C – R&D

The interaction of the Consultants with the Researchers and Developers usually is provided in two directions:

- the Consultants became as a source for information modelling of theirs work;
- the Consultants assist the creation and verify the implementation of the information models of the other components of the Knowledge Market. In this interaction the Consultant usually play active role.

Let's remark that the second function is very important not only of the work of the Researchers and Developers, but also for the stability and growing of the Knowledge Market.

The interaction of the Researchers and Developers with the Consultants is caused by:

- the needs of the scientific information modelling of the consultants' work as well as the activities of all connected to them Knowledge Market subjects;
- the implementation of the information models – in this case the Consultants need to play active role to assist the creation and to verify the realisations of scientific models.

The integration of the Researchers' and Developers' knowledge with the Consultants' one is very important for the making correct and appropriate solutions.

L – L

The main goal of the interaction between Lecturers is to deliver concrete material and pedagogical methods as well as to exchange theirs educational experience.

In other hand, the competition between Lecturers in one and the same subject can harm the Knowledge Market stability. In the same time, just the competition is the source of stimulus for the Lecturers advance and growth.

L – T

The Lecturer keeps in touch with the Tutor in mean to present him the information and educational materials needed for developing the curriculum during the education course as well as consults him in given application domain.

The Tutor is the Lecturer's main assistant – he has a permanent contact with the Lecturer and plans the interaction between him and the Employees. The Tutor gets the latest information about Lecturer's knowledge and education courses, and helps him to build the advertising strategy for the Knowledge Market. The Tutor also helps the Lecturer in using modern information technologies.

L – E

The Lecturer consults the Examiner about the plans for providing the tests and exams, as well as about the methodical recommendations for the contents of the tests and the educational methods for organizing and providing them.

The Examiner informs the Lecturer for the results of the attestation activities and suggests corrections to the current educational process.

L – A

Owning the basic wares of the knowledge market, namely the knowledge information objects, the Lecturer needs everyday information service for each one of the information interactions.

The Lecturer has to receive the actual knowledge and data and accumulate them, transforming in a resource for selling as a ware at the Knowledge Market.

The Lecturer always has to take into consideration all the specific characteristics of the customer of every ware (such as language specification, professional orientation, nationality and religion, information services' skill level etc.). So he has to own this kind of data as well. This type of information service of the Lecturer should be realised by the Administrator.

The interaction between the Administrator and the Lecturer consists in realization the hardware and software support as well as for all participants of the Knowledge Market, and in particular – service for realization the main function of the Lecturer – receiving new knowledge information objects and their transformation in the new articles of trade.

L – R&D

Special attention needs to be paid to the interaction of the Researchers and Developers with the Lecturers.

For some of the specific types of activity (like rapid and/or network education) the Lecturer can order special scientific research or a development accorded to the devices and the technologies, which can enlarge the service of the Knowledge Market in new direction. Such special scientific investigations and/or developments the Lecturer can order to the Researcher and the Developer, cooperating with the Employer. The Lecturer has to receive the information for the new scientific results and research-based technologies created from that participant in the Knowledge Market, and deliver it to the Employees.

The Lecturers play leading role for Employees' education and training. Because of this, the continual investigation of theirs possibilities and activities is very important.

The main goal of the interaction between Researchers and Developers and Lecturers is to enhance the effectiveness of the Lecturers work.

The achieving of this goal requires to build appropriate models of the Lecturer activities and to develop convenient subsidiary tools for receiving the new information and knowledge and creating educational courses and additional illustrative material.

T – T

The Tutors transfer to each other the educational plans, and compare the information needed for the coordination of the different courses of the whole education process.

In the same time, the competency of the Tutors is not enough to control the educational process. Because of this the collaboration between Tutors need to be provided through the Lecturers and under theirs management. Individual Tutors' initiative may cause embarrassments during the knowledge exchange.

T – E

The Examiner receives from the Tutor the plans for the expected tests and exams conducting as well as the test materials. He also informs the Tutor for the results of the attestation activities.

T – A

The Administrator supplies the Tutor with all needed hardware and software support for realisation of his activities – mainly connection with other participants, especially with the Lecturers, Employees and Examiners, for organising the education cycle.

As far as the Tutor's role is often played by the automated systems, the Administrator may perform a part of the Tutor duties – especially in the cases of the self-education.

T – R&D

The interaction between the Tutors and the Researchers and Developers is usually realised through the Lecturers. The Tutor doesn't take decisions about using new methods or technologies without Lecturer's opinion. He can get the new information from the Researchers and Developers and transfer it to the Lecturer. Only the Lecturer can decide how to use it.

The accent of the investigation of the Tutor's activities lies on the increasing the effectiveness of the organization of educational process as hole. The major element of this organization is the coordination with the other Knowledge Market participants (mainly with the Employee, Lecturer and Examiner) and external structures for optimal planning the activities of the participants in time and in place.

E – E

The Examiners transfer to each other the results of using attestation materials and methods.

At the state (government) and social level there exist similar structures, which are aimed to examine the educational organizations and to give them rights to provide the education in accordance to government and social requests.

E – A

To keep the knowledge level needed for conducting the Employee's exams, as well as all the information interactions, the Examiner needs everyday information service. It should be realised by the Administrator. The Administrator supplies the Examiner with all needed hardware and software support for realisation of his activities – mainly connection with other participants of the Knowledge Market for preparing and processing the process of the educational testing.

E – R&D

Taking into account that the main activity of the Examiner is to estimate the available knowledge and skills of the learner in the beginning, intermediate and final stage of the educational process in order:

- to evaluate the possibilities of the exanimate person (Employee);
- to guide and correct the educational process, provided by the Lecturer and the Tutor;
- to estimate the adequacy of the received possibilities in given learning cycle to the requirements of the Employer, which usually come to Examiner via Consultant,

the major tasks of the Researcher and Developer are:

- to investigate pointed Examiner's activities;
- to build theirs appropriate information models;
- to develop correspond tools for information servicing of these activities.

A – A

As a rule the single Administrator could not support and serve all information activities in the global net. Because of this he needs of participation of other Administrators in his concrete information service activity.

The Administrators of the Knowledge Market are the part of the global administrators' society for the information service and support of the business and social activities.

A – R&D

Every Administrator became as a source of information for building the Knowledge Market Administrator abstract information model by the Researchers and Developers.

The Administrator is the integrating component that is always in connection with all participants in the Knowledge Market. Because of this, he is valued helper of the Researchers' and Developers' work for investigation and modelling the components and communication processes between them.

In other hand, the implementation of the results of the Researchers' and Developers' work closely depends on Administrator's activity.

R&D – R&D

As a rule the single Researcher could not support and serve all investigation activities in the global net. Because of this, he needs of participation of other Researchers in his concrete scientific work. The same is true for Developers. So, the Researchers and Developers of the Knowledge Market need to be part of the global society of Researchers and Developers.

Special interest for us is the interaction between Researchers and Developers.

As it is remarked above, the Developers are obligated to implement new scientific knowledge for advance of the Knowledge Market information service and to abide by the scientific rules and recommendations.

Contrariwise, the Researchers need to interact with the Developers for collecting knowledge for implementation of the scientific results and theirs possibilities as well as for the new and not investigated areas of the Knowledge Market information interaction and service. Such way the Researchers may extend theirs knowledge and generate new scientific results to be implemented in the practice. So, the cycle "practice – science – practice – etc." may be complete.

Conclusion

In this paper the interconnections and information interactions between main members of the Knowledge Market have been presented. The discussion has been based on the direct interconnections between Knowledge Markets' participants. It is clear that there exist many indirect relations and influences which need to be investigated.

The Knowledge Markets are social phenomenon and, because of this, one very important task, not included in this paper, is the investigation of the relations of Knowledge Markets and other social formations. At the first place, it is obligatory for the government organizations to regulate and control the functionality of the Knowledge Markets.

Till now there not exist special lows and other normative documents aimed to regulate the interactions at the Knowledge Markets. Our expectations are turned to future research just in this area.

At the end we need to point one crucial direction for future research work. The new kind of human interconnections at the Knowledge Markets requires new pedagogical approaches. We are at the beginning of new kind of knowledge exchange based on the market principles and regulated by market lows. Many beginners will be embarrassed by the unknown environment and this may cause great social problems. The payable electronic way for exchanging the knowledge information objects will throw aside the destitute groups of people as well as nations of the world.

Acknowledgements

This paper is partially supported by the project **ITHEA-XXI**.

Bibliography

[Markov, 1999] Кр. Марков. *Относно вредите от електронната търговия*. IECC'99: International e-Commerce Conference. ADIS & VIA EXPO, Bulgaria, Sofia, 1999.

[Markov et al, 2000a] Кр. Марков, Кр. Иванова, И. Митов. *Требования к автоматизации обучения на пороге информационного общества*. Новые информационные технологии в электротехническом образовании (НИТЭ-2000): Сборник научных трудов пятой международной научно-методической конференции. – Россия, Астрахань: Изд-во ЦНТЭП, 2000.

[Markov et al, 2000b] Kr. Markov, Kr. Ivanova, I. Mitov, N. Ivanova, K. Bojkachev, A. Danilov. *Co-operative Distance and Long-Live Learning*. ITA-2000, Bulgaria, Varna, 2000.

[Ivanova et al, 2001] N. Ivanova, K. Ivanova, K. Markov, A. Danilov, K. Boikatchev. *The Open Education Environment on the Threshold of the Global Information Society*. IJ ITA, 2001, V.8, No.1 pp.3-12. (Presented at Int. Conf. KDS 2001 Sankt Petersburg, 2001, pp.272-280, in russian, Presented at Int. Conf. ICT&P 2001, Sofia, pp.193-203)

[Boikatchev et al, 2001a] K. Boikatchev, N. Ivanova, A. Danilov, K. Markov, K. Ivanova. *Authoring Tools for Courseware Designing*. IJ ITA, 2001, V.8, No.3 pp.115-121. (Presented at Int. Conf. KDS 2001 Sankt Petersburg, 2001, pp.32-37)

[Boikatchev et al, 2001b]K. Boikatchev, N. Ivanova, A. Danilov, K. Markov, K. Ivanova. *Teacher's and Tutor's Role in Long Life Distance Learnin*g. IJ ITA, 2001, V.8, No.4 pp.171-175. (Presented at Int. Conf. KDS 2001. Sankt Petersburg, 2001, pp.38-41)

[Markov et al, 2002] K. Markov, K. Ivanova, I. Mitov, N. Ivanova, A. Danilov, K. Boikatchev. *Basic Structure of the Knowledge Marke*t. IJ ITA, 2002, V.9, No.4, pp.123-134 (Presented at Int. Conf. ICT&P, 2002, Primorsko)

[Ivanova et al, 2003] Кр.Иванова, Н.Иванова, А.Данилов, И.Митов, Кр.Марков. *Обучение взрослых на рынке профессиональных знаний*. Сборник доклади на Национална научна конференция "Информационни изследвания, приложения и обучение" (i.TECH-2003), Варна, България, 2003. Стр. 35-41.

[Markov et al, 2003] Кр. Марков, А.Данилов, Кр.Иванова, Н.Иванова, И.Митов. *Массовое профессиональное обучение в условиях рынка знаний*. Пленарен доклад. Сборник доклади на: VI-та Международна научно-методическа конференция «Новые информационные технологии в электротехническом образовании», НИТЭ-2003, Астрахан, Русия, октомври 2003. стр. 9-18.

Authors' Information

Krassimira Ivanova – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, foi@nlcv.net

Natalia Ivanova – Petersburg State Transport University, natali@main.ru

Andrey Danilov – St. Petersburg State University of Aerospace Instrumentation, Russia, danilov@main.ru

Ilia Mitov – ITHEA – FOI Institute of Information Theories and Applications, foi@nlcv.net

Krassimir Markov – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Institute of Information Theories and Applications FOI ITHEA, foi@nlcv.net

7.2. Information Theories

ЦЕННОСТЬ ИНФОРМАЦИИ

Андрей Данилов

***Аннотация**: В этой статье обсуждается один из важнейших вопросов теории информации и практики ее применения, а именно, какова ценность информации и как ее можно оценить?*

***Ключевые слова**: ценность информации; количественная оценка информации; критичность информации.*

Введение

Известно, что с определенной точки зрения феномена информации, которая возникает в интеллектуальном и неинтеллектуальном мире, ее можно классифицировать как **контекстно-независимую и контекстно-зависимую.** При этом **неинтеллектуальному** миру принадлежит только контекстно-независимая информация, а **интеллектуальному** миру – контекстно-независимая и контекстно-зависимая информация.

В этой статье будет обсуждаться один из важнейших вопросов теории информации и практики ее применения, а именно, какова ценность информации и как ее можно оценить?

Понятно, что этот вопрос важен не только для экономики предприятия, организации, государства и т.д., но и для решения задачи зашиты информации.

Рассмотрим, а какой фундаментальный смысл скрывается под термином **ценность информации**?

В работе ”Синергетика и информация” [1] Д. С. Чернавский приводит следующее определение, которое объясняет смысл термина ценность информации.

*«Ценность информации зависит **от цели**, которую преследует **рецептор**. Чем в большей мере информация помогает достижению цели, тем более ценной она считается.»*

В этом определении есть два важных ключевых слова **рецептор -** получатель или приемник информации и **цель,** которую хочет реализовать рецептор с помощью полученной информации.

На основе использования объяснения термина ценность информации проведем анализ ценности информации для **неинтеллектуального** получателя информации. Ранее мы обсуждали основные отличия интеллектуального и неинтеллектуального объекта, из которых следует, что неинтеллектуальный объект **не может** сам поставить перед собой задачу и искать ее решение. Если приемник информации не может поставить для себя цели, то и информация, которую он получает, никак не может влиять на постановку или достижение какой либо цели.

Следовательно, для неинтеллектуального мира понятие ценность информации **отсутствует** и это важно для понимания смыслового значения термина «ценность информации»**.**

Покажем это на довольно простом примере. Представим себе, что необходимо определить, на сколько ценна информация, полученная с помощью поисковой системы в информационных ресурсах Интернет, непосредственно **для самой** поисковой системы. Ответ довольно очевиден - ценность информации **отсутствует.**

Процессы, в которых участвует интеллектуальный получатель информации

Ранее мы определили и показали, что интеллектуальному миру принадлежит как контекстно-зависимая, так и контекстно-независимая информация. Чтобы понять, чем определяется ценность информации для **интеллектуального получателя**, например, для Человека, рассмотрим, а какие основные процессы происходят при этом?

К таким процессам, в которых участвует интеллектуальный получатель информации, целесообразно отнести следующие:

1. **Определение или постановка цели**, которую хочет реализовать интеллектуальный получатель (приемник) информации.
2. **Получение информации** интеллектуальным приемником.
3. **Преобразование** интеллектуальным приемником **смысла**, заключенного в полученной информации, в свои **знания** на основе использования собственных интеллектуальных возможностей (процесс получения знаний).
4. **Аналитическая оценка** возможности применения полученных **знаний** для достижения поставленной **цели** на основе использования интеллектуальных возможностей и ранее полученных знаний интеллектуальным приемником**.**
5. **Принятие решения** об использовании или не использовании, полученных знаний для достижения поставленной **цели**:
   - если принято решение об не использовании, полученных знаний для достижения поставленной **цели,** то полученная информация для этого интеллектуального приемника **не имеет ценности** с точки его зрения достижения поставленной **цели;**
   - если принято решение об использовании, полученных знаний для достижения поставленной **цели,** то
6. **Осуществление действий** по достижению поставленной **цели** на основе использования интеллектуальных возможностей и ранее полученных знаний интеллектуальным приемником**.**
7. **Оценка результатов** достижения поставленной **цели** на основе использования определенных критериев, которые удовлетворяют интеллектуального получателя информации:
   - если **оценка** результатов, по достижению поставленной цели, **удовлетворяет** интеллектуального получателя информации, то эта информация обладала определенной **положительной** ценностью для решения этой **задачи.**
   - если оценка результатов, по достижению поставленной цели, **не удовлетворяет** интеллектуального получателя информации, то эта информация также обладала определенной ценностью для решения этой задачи, но эта ценность может оказаться даже **отрицательной.**

Субъективность процесса определения ценности информации

Если посмотреть, как происходит процесс определения ценности информации интеллектуальным приемником для достижения поставленной цели, то становится понятно, что этот процесс всегда **субъективный**.

Субъективность этого процесса определяется:

- **индивидуальной способностью** интеллектуального приемника получить знания из поступившей информации;
- **индивидуальной способностью** интеллектуального приемника провести анализ, полученных знаний, для принятия решения об их использовании или не использовании для достижения поставленной цели;
- **индивидуальной способностью** интеллектуального приемника реализовать поставленную цель на основе использования знаний, полученных из поступившей информации.

Кроме этого, необходимо отметить, что если информация, полученная интеллектуальным приемником из внешнего мира, **всегда** является **контекстно-независимой**, то знания, полученные этим интеллектуальным приемником из поступившей информации и результаты проведения анализа возможности их использования для достижения поставленной цели, **всегда** являются **контекстно-зависимой** информацией.

Таким образом, становится понятно, что **собственная** контекстно-зависимая обработка поступившей информации интеллектуальным приемником является **первым этапом** оценки ее ценности для достижения поставленной цели. Результатом этого этапа является принципиально важное решение, а именно, принятие интеллектуальным приемником решения о **потенциальной целесообразности** использования или не использования полученных знаний из поступившей информации. При этом первый этап **принципиально** включает два процесса:

- получение интеллектуальным приемником **знаний** из поступившей информации;
- принятие интеллектуальным приемником **решения** о потенциальной целесообразности использования или не использования полученных знаний для достижения поставленной цели.

Из анализа **первого этапа** оценки интеллектуальным приемником ценности поступившей информации для достижения поставленной цели следует, что эта информация ценна только тем, что из нее **потенциально** можно получить знания и **не более.** В этом и заключается самое важное свойство контекстно-независимой информации для интеллектуального приемника. Это свойство и является **ее ценностью**, которую можно определить, как **безусловная ценность** информации, и никакой **другой** ценности для интеллектуального приемника в этой информации принципиально **нет**.

Этот вывод нам очень важен, так как на его основе становится понятно, что:

«Только знания и возможность их использования могут иметь ценность для реализации, поставленной цели»

Второй этап оценки ценности поступившей информации непосредственно связан **с реализацией** полученных знаний из этой информации **для достижения** поставленной цели интеллектуальным приемником информации. На этом этапе уже непосредственной ценностью обладают **знания**, полученные из поступившей информации, ранее полученные **знания**, а также **способности** и **возможности** интеллектуального приемника реализовать поставленную цель на их основе.

Таким образом, становится понятно, что знания, полученные из поступившей информации, сами по себе **могут и не иметь ценность**, если интеллектуальный приемник не имеет способностей, знаний и других возможностей для их дальнейшего использования при реализации поставленной цели. Следовательно, ранее полученные знания и интеллектуальные способности являются **фундаментальной базой** для использования знаний, полученных из поступившей информации.

Этот вывод является **весьма важным**, так как он показывает, что при определенных условиях **знания**, полученные интеллектуальным приемником из поступившей информации, **не могут иметь ценность** для реализации, поставленной цели.

Такая ситуация довольно часто встречается в информационно-аналитических центрах предприятий. Например, если аналитик имеет знания, которые могут быть им реализованы для подготовки управляющих решений, но он не имеет соответствующего программного обеспечения для этой реализации, то такие **знания** в этих условиях **не обладают** ценностью.

Следовательно, на ценность знаний влияют **условия их применимости** для достижения поставленной цели. При этом эти условия могут быть как внешними по отношению к интеллектуальному приемнику, так и внутренними. Например, к внутренним условиям применимости знаний для достижения определенной цели можно отнести болезнь человека или его сон и т.д.

Мы рассмотрели только основные процессы, которые возникают при определении ценности информации, поступившей интеллектуальному приемнику. Но одну и туже информацию могут получить несколько интеллектуальных приемников и в одно время. При этом у этих интеллектуальных приемников могут быть как одинаковые, так и разные цели. Но ценность полученной информации всегда будет **индивидуальна** для каждого интеллектуального приемника информации. Это можно объяснить простым примером. В мире каждый человек индивидуален и имеет свою шкалу ценности информации при реализации, поставленной цели. Однако основные процессы, которые возникают при определении ценности информации и, которые мы описали, имеют место для любого интеллектуального приемника и их объединений.

Способы определения количественной оценки информации

Обсудим следующую важную задачу, а можно ли дать количественную оценку **ценности информации**?

В настоящее время известны несколько способов количественного определения ценности информации. Все они основаны на представлении о цели, достижению которой способствует полученная информация. Чем в большей мере информация помогает достижению цели, тем **более ценной** она считается [1].

Если провести анализ процессов, возникающих при определении ценности информации, то можно выделить следующую триаду, которая в наиболее общем виде отражает эти процессы:

"(информация) – (индивидуальные знания и способности, внешние и внутренние условия) – (цель)".

Эта триада нам поможет системно подойти к вопросу количественного определения ценности информации, так как от момента получения интеллектуальным приемником информации до момента ее использования или не использования для достижения поставленной цели необходимо реализовать определенную цепочку интеллектуальных шагов.

1. Затратный способ определения ценности информации (предложен Р.Л.Стратоновичем [2])

Основной принцип этого способа состоит в следующем. Если известно, что цель наверняка может быть достигнута и притом несколькими путями, то возможно определение ценности информации **V** , например, по уменьшению материальных или временных затрат, благодаря **использованию информации.**

$$V = \min [(\$n), (Tm)] \quad n \geq 2,\ m \geq 2 \tag{1}$$

где **V-** количественная мера ценности информации, например, в денежном измерении -**$** или временном измерении **-T**; **n –** возможное число путей решений задачи по минимизации материальных затрат; **m-** возможное число путей решений задачи по минимизации временных затрат.

При этом предполагается, что временные или материальные затраты, которые возникают при получении знаний из поступившей информации **существенно** ниже получаемого эффекта от их использования. Но мы понимаем, что **безусловная** ценность структурированной информации заключается только в том, что она может быть использована для получения **знаний**.

Если учитывать материальные или временные затраты, возникающие при получении знаний из поступившей информации, то выражение (1) примет следующий вид

$$V = \max [(\$n), (Tm)] - \min [(\$n), (Tm)] - [(\$к), (Tк)] \quad n \geq 2,\ m \geq 2 \tag{2}$$

где **$к, Тк –** соответственно материальные и временные затраты, возникшие при получении знаний из поступившей информации.

Такую задачу по минимизации затрат для достижения поставленной цели на основе использования информации практически ежедневно решает каждый взрослый человек при покупке продуктов. Эта же задача является одной из важнейших при управлении предприятием или организацией. Поэтому аналитики и руководители предприятий особое внимание уделяют поиску информации, которая может позволить решить эту задачу с наименьшими затратами и наилучшим способом.

2.Вероятностный способ определения ценности информации

Необходимо заметить, что вероятностные способы количественной оценки ценности информации хороши только в том случае, если известны численные значения вероятности, в противном случае их практическое использование не имеет смысла и нужно применять другие способы, например, экспертные оценки.

Вероятностный способ определения меры ценности информации для достижения цели, предложенный М.М.Бонгартом [3] и А.А.Харкевичем [4] можно сформулировать следующим образом. Если достижение цели вероятно и известно значение этой вероятности до получения информации, а также после получения информации, то меру ценности информации можно определить следующим образом.

$$V = \log2 (P/p), \quad (3)$$

где **V-** мера ценности информации; **p** - вероятность достижения цели до получения информации; **P** - вероятность достижения цели после получения информации.

Представим себе, что до получения информации вероятность достижения цели **p** = 0 и после получения информации вероятность достижения цели **P** = 0. Если эти значения вероятностей подставить в выражение **(**3), то легко заметить, что возникает неопределенность (0/0), которую необходимо разрешать. Однако мы и так понимаем, что ценность полученной информации для достижения цели нулевая.

В.И.Корогодиным [5] было предложено следующее выражение для оценки меры ценности информации, если достижение цели вероятно и известно значение этой вероятности до и после получения информации

$$V = (P-p) / (1-p) \quad (4)$$

Если провести анализ этого выражения, то становится понятно, что оно также имеет определенный недостаток. В этом легко убедится, если представить себе, что **P = p.**

Однако все эти недостатки выражений (4 и 3) не носят принципиального характера. На наш взгляд есть более важные моменты при определении ценности информации, а именно **субъективность** оценки.

Это легко представить, так как для одного человека полученная информация может оказаться более ценной, чем для другого, даже если они хотят достигнуть одну цель. В этом проявляется индивидуальность любого человека, которую мы учли во втором элементе триады **"(информация) – (индивидуальные знания и способности, внешние и внутренние условия) – (цель)".** Ранее мы обсуждали, что эта триада определенным образом отражает процессы, которые существуют при использовании или не использовании информации для достижения цели.

Таким образом, становится понятно, что ценность информации всегда связана с ее **конкретным** получателем и с **конкретной** целью, которую он хочет реализовать. Поэтому ценность одной и той же информации **варьиротивна** по отношению к ее потребителю и этот **принципиально**. В результате нашего анализа мы пришли к выводу, что контекстно-независимая информация обладает следующими свойствами:

- **Безусловной ценностью** по отношению к **знанию**, так как она является источником получения определенного знания интеллектуальным или интеллектуальными приемниками. Поэтому безусловная ценность информации принципиально **не может иметь** количественной оценки.
- **Варьиротивной ценностью** по отношению к интеллектуальному приемнику**,** которая субъективна и может изменятся в любых пределах **количественного измерения** в зависимости от индивидуальных знаний, интеллектуальных и других способностей, внешних и внутренних условий интеллектуального приемника, а также во времени и в пространстве.

Важность этих результатов анализа заключается в том, что если даже сделать техническое устройство для измерения ценности одной и той же информации, которое подстраивается под **индивидуальные характеристики** ее получателя, то в зависимости от внешних и внутренних факторов действующих на получателя информации во времени и пространстве, результат оценки **будет разным**.

Однако это не означает, что ценность информации нельзя измерять количественно. Например, ценность информации может выражаться для потребителя в денежном эквиваленте, величина которого может формироваться рынком создания и потребления информации. Этот механизм работает весьма эффективно, и на его основе существуют целые отрасли создания и потребления информации (телевидение, печать, Интернет, ноу-хау и т.д.).

Критичность информации

Рассмотрим еще один аспект ценности информации, а именно **критичность информации**, т.е. термин, который часто используют руководители предприятий и организаций в определенной ситуации. В термине критичность информации присутствует определенное важное пороговое значение для принятия **решения** или для достижения **цели** ее **получателем**. Можно часто услышать следующую фразу, *что подтверждение достоверности этой информации нам весьма критично для принятия решения*.

Смысл этой фразы означает, что если получателю **важно** иметь определенную достоверную информацию в определенное время, в определенном месте, в определенных условиях для принятия решения или для достижения определенной цели, то такую информацию называют **критичной.**

Таким образом, критичность информации также является **субъективной** оценкой по отношению к ее получателям, т.е. одному получателю она может быть весьма важна, а для другого получателя она может означать информационный мусор. Поэтому ее величина ценности весьма индивидуальна. Этот вывод подтверждает то, что значение ценности любой информации **вариативно** по отношению к ее интеллектуальному приемнику.

Заключение

В настоящее время, во многих организациях и на предприятиях широко применяются информационные технологии, созданы корпоративные сети передачи и приема информации, базы и хранилища информации. По существу на этих предприятиях произошли определенные качественные изменения, которые означают, что без этих атрибутов информационных технологий эти предприятия и организации более работать не могут. В результате возникла задача по защите корпоративной информации предприятий и организаций. А какую информацию необходимо наиболее тщательно защищать, какую менее и т.д.?

Этот вопрос непосредственно связан с оценкой ценности корпоративной информации для предприятия или организации. Как мы сейчас понимаем, что эту оценку могут дать **только специалисты предприятия**, которые работают с соответствующей информацией. Но при условии, что они понимают, с какой целью используется или предполагается использовать конкретную информацию и, что произойдет, если эта информация **исчезнет**, **изменится** или будет **доступна конкуренту.**

Литература

1. Чернавский Д.С. Синергетика и информация,"Наука",2001г.

2. Стратонович Р.Л. Теория информации //М., Сов. радио, 1975, 424с

3. Бонгарт М. М. Проблема узнавания", 1967г.

4. Харкевич А. А. Избранные труды. Т. 1 и 3. - М.: Наука, 1973г.

5. Корогодин В.И. Информация как основа жизни. -Дубна:Феникс+,2000.-208с

Информация об авторе

Данилов А.Д. – Санкт-Петербург, Эксперт департамента по информатизации и телекоммуникациям Ленинградской области, Эксперт некоммерческого партнерства "Учебный и исследовательский центр "Протей", кандидат технических наук (Phd); e-mail: danilov@proteus-spb.ru, www.proteus-spb.ru

THE MAIN QUESTION OF THE INFORMATICS, 100 YEARS AFTER ITS POSEING

Stoyan Poryazov

Extended Abstract:

Paul Otlet (1905) defines term "documentation" (a predecessor of "Information Science" and "Informatics") as activity comprising **1. gathering**; **2. processing** (Handling); **3. storage**; **4. retrieval**; and **5. dissemination** (distribution) of documents. With this he posed the main question of the Informatics - what are the basic information activities?

In the long-term policy of the IFD (1937) documentation is defined as: gathering; storage; **6. classification** and selection; dissemination and **7. usage** of information of all kinds. (In this abstract, we'll numerate the new appearance only of the proposed information activities, in chronological order). Jacson (1954): Documentation is the art of **8. creation**; dissemination; and usage of documents. Mark & Taylor (1956): Documentation is a group of methods for: **9. ordered** presentation; **10. systematisation**; and **11. transmission** of recorded knowledge. Webster (1961): Documentation is gathering **12. coding** and dissemination of recorded knowledge. Thompson (1963) and Taylor (1962): Information Science includes studying of: creation; dissemination; gathering; **13. organization**; storage; retrieval; **14. interpretation**; and usage of information.

From the other hand from lehnkering (2005) we understand that: "The concept of logistics covers all activities relating to the procurement, transport, transhipment and storage of goods. Logistics as generally understood is concerned particularly with material flow (raw materials, interim and final products), but also involves providing companies with services and information". Something more: "Logistics is the art and science of managing and controlling the flow of goods, energy and information" (wikipedia (2005)). What is the difference between Informatics and Logistics? Our answer is:

The only activities, unique for Informatics are: **1. Creation of languages** (including designing of signs' material presentations and denotations); **2. Creation of information** (including creation of messages for observation models' presentation (model coding)); **3. Interpretation of information**; **4. Destroying of Information**. These very old activities are in the early stages of their scientific study in Informatics and mathematics.

Bibliography

IFD (1937). Outline of a long-term policy of the International Federation for Documentation. FID publ. 325. The Hague, FID, 1960.

Jacson (1954). E. B. Jackson. Inside documentation. "Special libraries", 1954, v. 45, No 4.

lehnkering (2005). www.lehnkering.com/leh/struktur/metatop/glossar/index,language=en.html

Mark & Taylor (1956). J. D. Mark, R. S. Taylor. A system of documentation terminology. In: A. Kent, J. W. Rerry (eds). documentation in action. New York, Reinhold, 1956.

Paul Otlet (1905). "L'Organisation Rationelle de l'Information et de Documentation en Matière économique: examen des moyens d'assurer aux services de renseignements des musées coloniaux et commerciaux, ainsi qu'aux offices de renseignements industriels et commerciaux indépendants une plus complète utilité au point de vue de l'expansion mondiale. Rapport présenté au Congrès International d'Expansion Economique Mondiale réuni à Mons les 24-28 septembre, 1905." *IIB Bulletin* 10 (1905): 5-48.

Webster (1961). Webster's Third new international dictionary. Cleveland-New York, 1961.

wikipedia (2005). en.wikipedia.org/wiki/Logistics

Author's Information

Stoyan Poryazov – Institute of Mathematics and Informatics, Bulgarian Academy of Science, Sofia – 1113, acad. G. Bontchev Str, block 8. Tel: (+359 2) 979 28 46, Fax: (+359 2) 971 36 49, e-mail: stoyan@cc.bas.bg

OBJECTS, FUNCTIONS AND SIGNS

Stoyan Poryazov

Extended Abstract:

The problem of sign definition is posed from the Greek Philosopher Plato (427-347 B.C.) and it is not solved yet.

Our approach to sign definition is based on the Black Box paradigm - every material entity is considered as a convertor (function) of the influences to the entity (input) to its reactions (output). We consider the following possible functions of the material objects, in case of their interaction with Subjects (e.g. humans):

1. Utilitarian function. The natural properties of the entity are used directly (e.g. water is used for drinking).
2. Non-information function. The properties of the entity cause change in the emotional state of the human (e.g. beautiful landscape, music) and/or physiological status of the Subject, e.g. mantras, yantras and mandalas.
3. Phenomenological (symptom). One or more values of the properties of the entity (symptom) are used for prediction of values of other properties of the same entity (e.g. one has temperature over 39 degrees by Celsius, consecutively he is ill). The Subject must have knowledge (a model) of dependencies between the values of the symptom and its meaning. No need of conventions with other persons exists, but previous experience or education is mandatory.
4. Modeling function. One or more values of the properties of the (A) entity are used for prediction of values of other properties of the other (B) entity. The Subject is needed of knowledge of the dependencies between the two sets of properties' values, and of a witness that A-entity is a model of the B-entity.
5. Symbolic (e.g. the flag of a country, Red Cross etc.). The symbol usually means belonging of one entity to the organization, idea, political party, etc. Some knowledge of the correspondence is necessary.
6. Sign-model. (e.g. "cuckoo-bird") This is a name that can remember the named entity (a limited modelling function is used). The correspondence between the properties of the sign and its meaning is not mandatory (the concrete cuckoo-bird may be mute).
7. Sign-name. Only a convention is necessary.
8. Sign-code. Sign-code is a sign-name used for denotation of the other sign. Only a convention is necessary.
9. Sign-element. The entity hasn't meaning. Its function is to be used for creation of signs, e.g. a pixel of a computer display.

Conclusions:

- Four types of function of the material entities are enough for characterizing all known usages of the entities:
  1. natural function,
  2. non-information function,
  3. modelling correspondence and
  4. convention correspondence.
- In the case of modelling correspondence a witness is necessary. In the case of convention correspondence, the same convention must be used in creation and interpretation of the signs.
- The carrier of the sign must be a real entity, because ideal entities (e.g. the mathematical objects) can't interact.
- The ideal objects may be (and are) denoted with signs, and their properties are modeled by rules (e.g. axioms in mathematics). Even mathematically non-existing objects may be denoted, e.g. non-empty subsets of the empty set.

Author's Information

Stoyan Poryazov – Institute of Mathematics and Informatics, Bulgarian Academy of Science, Sofia – 1113, acad. G. Bontchev Str, block 8. Tel: (+359 2) 979 28 46, Fax: (+359 2) 971 36 49, e-mail: stoyan@cc.bas.bg

7.3. The Intangible World

APPROACHING THE NOOSPHERE OF INTANGIBLE – ESOTERIC FROM MATERIALISTIC VIEWPOINT

Vitaliy Lozovskiy

***Abstract**: Exploration of intangible world is under the serious influence of esoteric. Mystics, religions, unrestrained use of metaphors, fairy tales, gossips, unverified and uncertified "facts" – all this needs accurate well-disposed, but sound scientific consideration. Our society really needs new ideas, new approaches and new paradigms. Technological civilization becomes more and more complicated, risky and ecologically critical. The current level of AI research cannot guarantee successful solution of societal control and management. Besides, the human being itself practically did not change its mental and psychological abilities for several hundred thousands years. We can lose control over our society and its technology, if we do not change cardinally ourselves. In this text, I tried to approach this problem – the problem of our noosphere from materialistic viewpoint.*

***Keywords**: philosophy, noosphere, esoteric, intangible world, beliefs, soul, God, materialism, idealism, egregors*

"Go there, don't know where, bring that, don't know what…"

Russian fairy tales folklore

"We believe that 2 * 2 = 4. Without this belief we would fail in our everyday calculations. We should be the true believers the whole yearlong. But once a year we ought to stop and think over this issue one more time…"

V.Lozovskiy, Wise Thoughts (unpublished)

Introduction

I was greatly embarrassed when decided at last to step forward with this theme. First, I am a newcomer to the field of esoteric, while this sphere of human activity is philosophized, theorized and practiced from the very beginning of humankind. Yes, the progress of human society started with the first inventions to support physical existence of prehistoric people. But no one society can exist without "ideal" world – semiotic systems which, of course, have material carriers, but whose sense is in representing some knowledge and communicating it between intellectual subjects.

It is very hard to handle existing information fund on esoteric: abundance of paradigms, approaches, term interpretations, metaphoric way of consideration (usually the authors never admit it, but go on in developing their personal viewpoints as if it was scientifically proven truth). Thus, for our current goals the idea to make substantial comparative presentation of this field was abandoned from the very beginning. And, again, it is unusually hard to make specific literature and author reference. I understand that in these circumstances one should leave the hope to pinpoint exact scientific priorities of all researchers in this field, including your obedient servant… Instead, we are aimed at much more important issue – try to find correct estimates of what is going on in our civilization, and how should we together handle all difficult problems, which life evolved around us. As a partial compensation for this liberty in approach, we shall try to give as precise definitions, as we can, for the nontrivial concepts, with which we are going to deal.

Amen!

Knowledge, Civilization, Noosphere, AI, Esoteric and the Future of the Humankind

Speaking about human beings, about society, we cannot bypass the question of human destination, goal, sense of living in this World. Of course, the answer varies among specific individuals, epochs, situations and the fundamental issue of humankind creation – was it done according to Creator's will, or was it completely natural process – including evolution, mutations, natural selection, or even extraterrestrial visitors interference. Looks like we can find the common denominator for all these variations and alternatives. I guess, it is knowledge.

Knowledge is the sound prerogative for any human deeds and actions irrelevantly to the prime cause: material, or ideal (In the beginning was the Word, and the Word was with God, and the Word was God [Bible]). Knowledge is obtained as a result of ***need, boredom, pleasure, curiosity*** and ***chance***. The universal method of obtaining knowledge is ***activity***. Life is activity. The natural frame of activity is ***adequacy***, or ***harmony***.

Of course, we cannot prove this thesis, but one can easily try to fit it over any imaginable situation – is it striving for life, scientific research, creating work of art. In any kind of human activity, we should obtain knowledge. New feelings are also new knowledge. Just imagine that due our inactivity or due completely routine activity one stops to acquire **new knowledge**. At that point life ceases its sense and we prefer to name it ***existence***.

From now, we will use the concept of Creator (God, Lord in Christianity). Even the orthodox materialist should agree, that such concept exists, and until now, science could not prove neither (His? Her? Its?) existence, nor absence. In our argumentation here we shall – not accept – but touch and from time to time refer to this concept. Assume, for the moment, existence of Creator who created World and humans. In all idealistic philosophies and religions, Creator is Omniscient and Almighty. In such case – what could be the reason for material World creation? What could be the sense for creation of humankind? The reader should agree that He created the whole World having no other way for obtaining new knowledge. Such ideal entity as Creator, by definition, could not be interested just in shuffling the matter. So, obtaining knowledge really is the sound candidate for the goal of human existence.

Humans became humans only due the creation of ***society***. Civilization is the product of social activity - communication, interaction of people, creation of crafts, engineering, science, art, religions, traditions. Civilization is the natural cradle for new knowledge creation. It became possible only because each human generation could absorb the knowledge of previous generations and, in its turn, lay its own knowledge strata for future generations to rest upon. Along this reasoning thread, we arrive at the concept of noosphere. It was formulated by V.I.Vernadsky [Vernadsky, 1943] and Père Pierre Teilhard de Chardin [Teilhard, 1947]. My own vision of noosphere in semiotic context could be found in [Lozovskiy, 2003]. The term «noosphere» (Ionian Greek "noos" = mind) is used for integrative designation of physical world realities together with the whole mankind knowledge.

One of the key milestones in the humankind development was when people tried to attack the fundamental concept of knowledge creation and handling. Information technologies (IT) and even artificial intelligence (AI) sprang into existence and were developed to the current level. Before the information age only humans (and, to some extent, higher animals) could generate, process and communicate knowledge. Now it could be done much more efficiently with the help of artificial means.

The stunning progress of IT and AI cannot conceal the problems encountered in this way of technological development of civilization. We encounter difficult problems in every step of knowledge acquisition, formalization, handling in computer systems. One of the main obstacles is a bad ergonomics of artificial systems. Many managers and applied specialists fail in efficient communication with expert systems and remain on primitive level of using technology of paper, pencil, phone, and maybe calculator. From the other side, human intelligence in its entirety and, in general, human performance, remains unreachable for current technical counterparts. We have very weak understanding of human intuition, emotions, flexibility of behavior, methods of handling imprecise, unreliable, incomplete information. The notorious human factor, intrinsic difficulties of management in organizational systems are still beyond the reach of contemporary knowledge engineers.

From the other side, our civilization is inevitably approaching its critical point. We have serious ecological problems, which have definite tendency for deterioration, we more and more depend upon technological world, created by us: the failure in power supply, communication, automated management and control systems can not only block the normal life process, but even lead to catastrophes and tragedies (Chernobyl, for example). This overcomplicated world requires much better level of understanding, interaction and management, than can

provide the human brain of contemporary people. And it is no wonder: human brain and its usage by humans practically did not change during the last hundreds of thousands years.

Of course, we cannot stop, reverse the civilization, return to caves and animal skins clothes – we ought to go on in our development, changing our ways, our mental and psychic abilities, harmonizing our interaction with Nature.

But does our brain, perceptional and psychic abilities work at their threshold? Neurophysiologists say that usually human brain works within the range of 3-6% of its potential. This phenomenon is not coincidental. It means that our life conditions until now did not stimulate its growth.

Everyone can name maybe several encounters in his/her life with persons demonstrating outstanding abilities – physical, mental, psychic. One elderly woman under the affect brought out of fire her huge trunk, which later could not lift four strong men. There are people who developed their memory to fantastic level – they memorize heaps of numbers, perform complicated calculations in mind. Some behave very soundly and efficiently under the stress situations, when others just got paralyzed. All these phenomena, though rare and frequently requiring tedious and systematic everyday training – do not seem to us marvelous: we understand, in principle, how these results could be achieved, and can give scientific explanation concerning their mechanisms.

Besides above mentioned phenomena, which can be explained by “materialistic” science, from ancient times we obtain evidences of something supernatural taking place – clairvoyance, foretelling, telepathy, telekinesis, levitation, nontraditional healing, bio-field, UFO, communication with Higher Intelligence, - all what is known as esoteric, or mystics. The number of books on this theme is overwhelming – it is hard to mark out the most fundamental ones – one usually adheres to some or other school according to personal taste or belief.

There are some specific problems with esoteric… First, “official science" calls it “pseudoscience”. Phenomena, pretending to be esoteric, rarely become objects of scientific investigation. In its turn, the number of esoteric paradigms is astonishing. As a rule, esoteric authors freely and without restraint use metaphoric presentation of their subject tacitly convincing the reader to buy it without any doubts. Personal and religious beliefs in this sphere play essential role. And again – authors usually avoid modest remarks of the type: “It seems to” or “I believe, that…”, so that the reader is left to his/her own judgment concerning the objectivity status of specific presentation.

My idea is that time is ripe for humankind to approach esoteric from the positions of scientific method. AI professionals should be among the researches in this field. And that is why. First, we should acknowledge the fact, that the modern technological society is approaching the crucial barrier of complexity, beyond which we can lose control over the civilization, we created. Our current understanding of noosphere structure, content and behavior may present only the tip of the iceberg and should be critically reconsidered. Even if majority of esoteric evidences will occur false, their remaining part may add substantially to our philosophy and practice. AIers are proficient in knowledge engineering and simulation of human behavior. Acquaintance with world of unconscious and other esoteric themes could enhance understanding of human mental and psychological characteristics and, as a result, lead to automation systems creation, which will be better adapted to the challenges of our growing and maturing society. In such research, we can expect several byproducts, for example, refining our philosophical background – handling this subtle border between material and ideal hypostases.

The proposed approach to esoteric requires refinement and elaboration of our basic definitions. Without such bureaucratic stage, we cannot attack such complicated humanitarian polyparadigmatic area.

Materialism and Idealism

Traditional approach to these paradigms is symmetric and frequently vulgar. They say that materialism does not accept in the world anything but substance and energy. Earthy materialists do not believe in good and evil, they deny the existence of soul, ethics, do not believe in God. They see only material interactions in the world and deny the existence of any higher will, intelligence governing the Universe. Contrarily to that, idealism is supposed to put forward “ideas” – abstract concepts, goals according to which sprang into existence, function and changes the physical world. That is why idealism is inevitably tied hard with the concept of God, Creator, whose goals and orders rule the Universe. Traditionally, consistent materialistic and idealistic paradigms are considered equally sound; i.e. one can adhere to any one of them, and do not be afraid of criticism from the opposite camp.

My argument is - that idealistic paradigm is weaker than materialistic one due the lack of experimental evidences. It is the routine approach to natural science theories: they all should be supported by practice. Speaking about

concepts, paradigms, ideas, language, phrases, plans, intentions, emotions, etc, we should inevitably postulate existence of the following components necessary for the success of communication act:

- subject generating and issuing these "ideal" assets (source of information);
- subject receiving, understanding and perceiving these information (reception of information);
- material data medium used for transmission and/or storage of information.

Both communicants should be intelligent enough, belong, at least, to the similar cultures, have common communication language, be involved in some common goal oriented activity – otherwise such communication act will fail. The state or activity of receiving subject should change somehow because of the communication being considered.

All three communication components are really indispensable. Thus, "information" cannot exist without some material substrate – media, bearing this information. So, material data medium should first be created, and only then, it could be used for handling "ideal" (information) entities. One of the main characteristics of an ideal entity is its semioticity.

From this trivial argumentation, one can infer the following significant corollaries.

1. The ideal entity creation should be preceded with obtaining the material bearer good for generation, storing or transmission of this ideal entity. Thus, the issue of priority is solved in the favor of matter.
2. Communication of ideal entities – information – is possible only between intellectual beings belonging to the same culture and of comparable intellectual level. So, it is impossible and senseless to give orders to inanimate objects, they cannot perceive and understand information.
3. Ideal entities exist in reality, but only within certain cultural world [Lozovskiy, 2003] and have sense and meaning only within this world – intellectual beings inhabiting it.

I feel that this is correct materialistic approach to the Nature, and one can rely upon it in the following investigations. As was shown above, our world, being materialistic in form, has means for representing ideal entities - information. One more remark here – concerning the origin of Universe, Solar system and human race. Here we, of course, cannot answer the question – whether our world is the result of natural physical and physiological processes and evolution, or it was the result of Creator's activity, or interference of some extraterrestrial civilization. All we can proclaim – it is the result of material processes.

Materialistic Approach to Exploration of Intangible World

Cognition of the World is the endless process. The whole Universe at each given moment of time can be divided into two subworlds: Tangible and Intangible.

Tangible World – is the part of the Universe, of which we are aware and have some positive knowledge about it. Adequateness of the paradigms used, theories created (and verified), their experimental confirmation varies within the broad range, though. From time to time this knowledge is changed to reflect better the new facts; sometimes old paradigms are replaced by the new ones within the course of scientific and technological revolutions. Knowledge (and folklore) about Tangible World as the part of cultural World [Lozovskiy, 2003] determine the current scope and state of the Tangible Noosphere.

Intangible World – is the collective term for all Unknown part of the Universe, of which we are unaware and have no rational verified knowledge.

Of course, this classification is fuzzy and provisional; there is no clear cut border between these worlds. In the course of scientific progress human cognition attacks, tries to explore the Intangible World gradually moving its fragments to Tangible one. The main obstacles on this road are fallacies, misbeliefs, fantasies which cannot be laid in the basement of natural science theories and usually become the part of human cultural World (fairy tales, religious beliefs, superstitions). If our destination really is to gain new knowledge, as we argued in the beginning of this communication, we ought to permanently attack the World of Intangible. But how? This situation reminds the classic one from the Russian fairy tales: go there, don't know where, bring that, don't know what… Of course, we ought to be armed with the technological approach, for example, new instruments for measuring radiation while exploring new areas of the Universe, but the most subtle and mysterious areas are tied with humans themselves, their inner world and relationship with Nature environment. By the way, the basic goal of

AI researches is to better understand "ideal" processes in human mind and psychics, try to simulate them and create better instruments to help humans solve their most important problems.

Frequently information about Intangible World researcher obtains from other humans in the form of their personal unusual experience or from witnesses of some extraordinary phenomena. These evidences, as a rule, have very troublesome features:

- extremely high level of dependence on specific person, their physical, mental and psychic state, character type, suggestibility, religiosity, innate features and personal training;
- high dependency on environment conditions: season, time of the day, weather, influence (usually interfering) and even silent disbelief or skepticism of listeners;
- aforementioned peculiarities frequently result in a bad recurrence of phenomena under investigation.

Esoteric is a very intriguing potential wonderland, which in my opinion quite deserve exploration, and that is why:

- it has a long history – thousands of years; esoteric beliefs accompanied human race from the very beginning of the history, have not vanished, but broaden and flourish;
- many outstanding persons were studying and practicing esoteric mysteries, and there is abundance of literature on this subject;
- there are a lot of esoteric schools, societies and communities – one can easily find one in one's own town – according to the taste and interests;
- esoteric is full of extremely rich promises, which are quite tempting and badly needed in contemporary human society in order to better correspond to requirements of our life, society, civilization – better physical, psychical and mental health, better control over own organism – even over its organs and subsystems, diagnostics and healing the diseases, which are caused by functional disorders or improper influence of the nerve system, mastering subconscious processes, development of extrasensory perception abilities and potential nontraditional ways of interpersonal communication and interaction.

Unbiased consideration of esoteric doctrines produces ambivalent impression. The lack of sound scientific theory and reliably confirmed facts positioned esoteric from viewpoint of Academy in the domain of pseudoscience. Situation is aggravated by very strong positions, which are held in many esoteric teachings by religions and mysticism. The time has come to investigate this problem using sound scientific method. Of course, the whole issue is exceedingly complicated, and here we will try, at most, approach it shyly not loosing solid ground beneath our feet.

So, the key source of information in esoteric is human evidence.

Evidences of Level 0. These are evidences of "folklore" character. Researcher obtains oblique information from unknown, unavailable or unreachable source: somebody had said something – no details, confirmations, trust. Most evidences unfortunately belong to this category and can be dismissed with light heart. This information probably deserves only to be put into protocol as possible hint for future, if it comes sometimes to relevant issues.

Evidences of Level 1. This sort of evidences is obtained directly from trustworthy subjects on the base of their personal experience, or experience of their close friends or relatives, from direct witnesses or participants in some action. Such evidences deserve the close examination. Their valuable feature is that it becomes possible to recur to the sources requiring detailed, probably documented and/or confirmed information; in the future it is possible to return to the issue under consideration, if new questions arise or new circumstances come to light or require clarification.

Sometimes, we may obtain distorted information, if the author of the evidence was in changed mind state: under hypnotic suggestion or in meditation. The serious problem within this Level 1 is that these evidences to high degree can be subjective.

Evidences of Level 2. Here we deal with the evidences obtained by researcher himself in the course of experiment or as a witness of some action or observation of natural phenomenon. In this case, it is possible not only to create a protocol of the experiment, but also actively participate in it on the stage of planning, execution and discussion of the results. Sometimes it becomes instructive to change the course of experiment "on line". The main objective in this situation is to discern the true nature phenomenon from the magician's tricks. Of course, the researcher should be exceedingly aware, if the experiment requires changing his/her own mind state. One should exclude chances of being drawn to such states unconsciously. Sometimes, the presence of qualified psychiatrist, whose task should be checking mind state of the participants, is required.

Evidences of Level 3. To this class belong evidences obtained as a result of personal training of the researcher (increasing sensitivity, developing extrasensory perception, special breathing techniques, experiments with changed mind states, meditation, perception of aura, biofield, etc.). Of course, it is the most complicated and time-consuming approach. Besides, it relies on the belief that extrasensorial abilities are not only inherited, but also can be trained by any dedicated person. Results, which could be obtained on this level by researcher stem, in greater part, from his/her introspection.

There are two main problems with esoteric. Firstly, these phenomena are closely related with brain activity on conscious and subconscious levels. And, secondly, we are lacking understanding sound physical backgrounds of these phenomena and, consequently, - instruments for obtaining objective measurements. It would be a pity to rest this domain to magicians, religious figures, naive people and businessmen. Such is the motivation of the current research.

Esoteric and its Strata

The whole domain of esoteric could be seen as consisting of four strata.

Stratum 1 - Personal (self-control and autosuggestion). Human organism is extremely complicated system. Our orthodox sciences (medicine, neurophysiology, biochemistry, psychology, psychiatry et al.) have until now very superficial understanding of life even on the level of distinct organs, phenomena and subsystems. When we get to the depths – the cell level, and, even more, if we are interested in integrated functioning of an organism, in gestalt effects, - our sciences really fail.

One of the big white spots is human ***subconsciousness***. Neurophysiologists fail to explain its main mechanisms. They say nothing about interfering with them. On the layman level, we understand, that subconscious work, from one side, as buffer, or long term memory, where are stored much greater volumes of data, past impressions, than we are aware, relying on our conscious sphere. It is well known fact that under hypnosis patient can recollect even events of his childhood or reconstruct forgotten relations between some events. The scope of subconscious memory is suspected to include probably genetic information – memory of generations. It could explain reminiscences of some humans about their previous incarnations.

Subconscious human sphere is not only memory – looks like it can process information in nontrivial way, finding correlations, rational solutions, of which people are completely unaware being awake. Psychologists call this phenomenon insight – as if solution comes from nowhere; but this could be the effect of processing data subconsciously and then just transferring it to daytime memory. The power and potential resources of subconscious memory exceed 90% of human brain. People from esoteric [Zykova] say, that we ought only to present our problems to our subconscious sphere, and solution found will be much better, than we would obtain “thinking” over them in ordinary way. Special methods exist which open the door to our subconscious processor: trances, meditations, which help us to lower our beta rhythm of vigil brain activity (15 - 30 and more Hz) down to alpha (7.5 – 13 Hz), teta (5 – 7 Hz) or even delta (0.5 – 4 Hz). Hypnosis, auto-training and meditation usually require subject to be brought, at least, to alpha rhythm. In this state various suggestions could be done, getting rid from negative habits and stress, solving different problems.

Human ***nerve system*** is the control system for the whole organism. We all know the saying: “All deceases are from nerves”. Evolution adapted us to the current life condition, our technological civilization, primitive medicine with its strong impact on separate organs, or just illness (“from cold”, “from stomach”, “from any pain”…). It led us to the status, where we lost control over our own body. We can discern several aspects of nerve system functioning.

- General psychic personality type and the current nerve system state. Sanguine, melancholic, phlegmatic, choleric character types; optimists and pessimists; gloomy, cheerful, shy types. All these not only determine one’s current psychic state: it has a strong influence on physical state of the organism. Psychical and physical spheres tend to be in harmony: in sound body – sound spirit and vice versa.
- Distal nerve and microcapillary systems work together near the border of the organism being sort of interface between it and environment. Ancient oriental medicine brought to us the theory of meridians and tsubo – small sensitive zones on the body surface [Serigawa]. According to oriental teaching, these zones “represent” internal organs, reflect their state and, from the other side, can be used to influence, correct the state of these organs.

- Nerve system direct control over organs and functional systems of the body. Contemporary humans, in general, lost this ability – it was obsolete due achievements of our technological civilization. However, looks like it did not vanish forever, but is now in the passive state. With the help of special training, we can control our heart, blood pressure, pulse, functioning of liver, kidney and other organs. Good news is that all these effects could be achieved not only by yogis or humans with inherent abilities – everyone can achieve results on this path, but, of course, results will differ. Really, a great artist can create masterpiece, but everyone can be taught to draw a house, a tree and a dog under it…

Extrasensorial perception is the ability for humans to perceive outer world information much more efficiently, than majority of people. It can be attributed to four sources.

- Heightened sensorial sensitivity.
- Broader frequency band.
- Using some "unusual" or yet unknown fields, energies, radiations: magnetic, electromagnetic, electrostatic fields, biofield.
- Perceptions of gestalt type, sometimes, named hypersensory (HSP), or intuition. A person with HSP is very observant and perceptive. They may be adept at reading body language or simply be more attentive to details than most people picking up subtle behavioral cues unconsciously, cues that are also unconsciously given. Strictly speaking, we deal here not with just perception, but with complicated process of general combined analysis of several stimuli on the basis of pattern recognition. For example, qualified and experienced physician can diagnose many organic disorders on the fly. Sometimes, he cannot even explain, how it is done – he simply feels, or it appears to him so. It means that these processes to the great degree are done subconsciously.

Nutrition and breath. These factors are exceedingly important not only for physical human body – it is self-evident. There are evidences, that breath do not only brings oxygen to the blood, but plays exceedingly important role in energetic processes in the organism and when interacting with environment, or even, Universe. This issue should be thoroughly studied by science.

Drawing a conclusion, we can state, that this personal stratum is materialistic foundation of esoteric and should be thoroughly studied, practiced and be developed further.

Stratum 2 – Bioenergetics. The issue of "bioenergetics" can be approached from different viewpoints. From physical point of view, each living organism radiates several types of energy: heat, acoustic waves, magnetic, electrostatic field and electromagnetic waves in rather broad frequency range. Correspondingly, these fields somehow influence the organism – on subconscious or biological level. Intensity of radiation is very low and drastically falls with distance. In esoteric it is postulated that there probably exists quite different type of energy – bioenergy, which is emanating from living organisms and form specific biofield around them. Sometimes, they speak about biofields of inorganic objects. The main problem with biofield is that until now we have no reliable physical methods and/or devices for measuring it. Even worse – there is no assurance, that biofield is really field – it may be bioplasma, or some other, maybe yet unknown fine material substance or radiation. Very queer circumstance is that these biofields can usually be detected and "measured" only by humans having specific sensitivity to them. Though, there exist methods pretending of making color photographs of aura – biofield around human body.

One of the most fascinating esoteric theories is the teaching about human chacras, auras and of the whole energetic system of the human organism, which support its functioning and play the role of interface gates between the organism and the rest of the Universe. Though these theories lack rigorous scientific confirmation, practically, one can develop personal sensitivity to the level, where some emanation from human body, or from other objects could be felt [for example, Bronnikov, 2005]. I have got training in Bronnikov's First Stage, starting from the zero sensitivity to biofield, and can present my subjective feelings and impression concerning my achievements. Really, after 10-days course and several months of training, I can feel "something" between my hands when I move them one towards another. This feeling in the palms is a mixture of elasticity – like soft children's air balloon, pricking, some heat, cold or rippling. At this stage, one can say, that, as there are no ***objective*** indications (measured with some physical device), - probably, it is psychological effect, suggestion, produced with neurons in my own brain, or between brain hemispheres. But things become more complicated when I feel the fields of other people, trees and even inorganic objects. Several successful sessions were carried, during which patient's headache, cramps and shoulder neuralgia were cured. The specific method used was of

Reiki type [Reiki, 2004] and procedures known as non-contact massage. Practice here evidently goes ahead of theory, but such nontrivial phenomena should not be just ignored due the lack of sound explanations.

Even more spectacular phenomenon in this bioenergetics stratum could have been the direct effects of bioenergy, for example, telekinesis. All I have in this vein now – evidences of the zero level – as majority of us.

Stratum 3 – Bioinformatics. Here we speak about "nontraditional" information transfer between humans, i.e. such information interaction, which cannot be attributed to optical, acoustical, electromagnetic methods of information communication. The most typical example of such phenomenon is telepathy. No sound, non-equivocal, reliable and repeatable confirmations of telepathy are present up to this day. Though, parapsychologists would argue that such evidences do exist. The time is high to solve this issue on the sound basis. Either yes/no experiment conducted rigorously with participants from the natural sciences will be held, or this problem will remain on fiction level for indefinite time ahead.

Stratum 4 - General noosphere. Absolute. This stratum is fundamental from philosophical viewpoint. In [Lozovskiy, 2003] I argued in the favour of cultural layer of the Earth's noosphere, which includes all knowledge of humankind in any form or domain – from science and religions to folklore and national habits. Esoteric goes much further insisting that "knowledge" per se does exist around us, and the sources of this knowledge are many, including non-biological objects on Earth, other galaxies and extraterrestrial civilisations. This general worldwide database, according to esoteric, not only keeps information about the past, but also about the future. At this point we should be prepared to accept the idea that such knowledge base is equivalent to the conception of Absolute, God, Creator, Higher Powers and so on. All people, according to this hypothesis, participate in creation of this database. Even the people, who have gone – their souls – can remain in this fine matter World.

In esoteric we can find even the exotic idea that human memory and even thought process are located not in the brain, but in that same global bio-informational field, and our brain function as a mere interface device between ourselves and Absolute. This hypothesis can be ignored until reliable experimental data in its support will be obtained.

The breathtaking revelation about existence of Absolute quite finely explains the principles of most marvellous esoteric deeds: clairvoyance, prophecies, voyages into the living organism cells, in the depths of matter and the Universe. And, of course, here we arrive to the concept of Creator – chief systems programmer, who made all the wheels go round…

So, the main problem here is to answer the question: does this Stratum 4 exist in reality, or it is just a beautiful metaphor. Usually, they say, that neither existence, nor absence of God cannot be proved. Sometimes this issue is tied with personal beliefs. We shall try to clarify this problem.

The Concept of Absolute

Let us start from the agreement, that, generally speaking, Absolute could be imagined as having two hypostases (contrarily to Christianity, where Trinity approach is adopted): mental (or cultural) and real. First, one cannot be empty, because we already have such ***concept***. As we stated earlier, this is an ideal entity dwelling either in somebody's brain (thoughts, beliefs, imagination, fantasies, hallucinations), or in some cultural layer. The most critical question concerns the second hypostasis: ***reality***. If Absolute is, at least, partially, real – it must manifest Itself (Himself, Herself) in reality, and thus, be liable to physical measurements. For example, if we arrived at the idea of general database, or information field existence, there should be possible to find out the material substance – carrier – and then methods of encoding used. At last, if it is not just passive database, but functioning control system, we ought to disclose the language of programming used and the programs themselves. Until now, there are no confirmations along these lines of reasoning. Either, we worked not enough diligently, or this hypostasis is empty. Remains the "God" within human brains: a system of beliefs, religions. Atheists have their own ethic-behavioral rules – their own "God".

We can imagine indirect methods of proving existence of real Absolute – for example, carefully studying the results of predictions, clairvoyance. If these phenomena do take place – it means that Absolute exists in reality.

Of course, one can say, that there can exist methods of protection, preventing humans inspect and interfere with "system programming" layers – as is done in computer operational systems. Then, the last argument can be drawn: there must exist, at least, some interaction with Absolute (analogy: application, calling some system function). If nothing of this sort will be produced, we should assume that the concept of Absolute is only a myth in

human brains. Situation reminds spy story: just imagine, that some country sends its spy as a resident to other country. This person legalizes there, marries, works somewhere – nobody can guess, that he is really a spy. From the other side, he does nothing contrary to law. So to say, the perfect spy… The question is: can we call him spy, while in reality he does not function as a spy?

Soul

Esoteric literature is full of references to human soul. Usually is postulated existence of soul independent of physical body. A lot of attention is paid to the incarnation hypothesis. All this is very interesting and exciting. But, as we already had got accustomed, we have no reliable experimental data on existence of soul as some ***separate*** entity from human body itself. Until then, I propose much more simple and natural concept of soul.

Soul is specific functionality of higher species, of their central nervous system. In humans this functionality is very rich and represents individual personality – with all spheres including intellect, emotions, ethics, social conduct, etc. The phenomenon of soul can be attributed to gestalt effect, where some complex entity show broader, or even new functionality than the sum of its parts have in common.

Take for example, car's engine. Being assembled, provided with gas, oil, water, air and electricity it demonstrates its gestalt functionality – its soul: it can rotate its output shaft. Of course, its "soul" is exceedingly primitive; human organism is much more complicated and perfect than our engine. But we are considering the fundamental principle. The engine works perfectly until some malfunction happens – its metal organism becomes ill. Eventually comes the day, when car mechanic says – alas, nothing could be done, it is impossible (or unreasonable) to recover the functionality of this engine… The engine is dead. It cannot accomplish its main functionality – rotating the output shaft. Its soul has gone. Where to? No one will be embarrassed with this question. The answer is self-evident: when parts cannot function together – functionality just vanishes. I can clap both my hands – and you will hear the sound. But with only one hand? We arrive at the same situation when dies human being. All his functionality disappears. Very frequently, it is done gradually – deteriorating short, long term memory, hearing, eyesight, coordination of movements, worsening brain's processing abilities… At last fail subsidiary systems: digestion, breath, blood circulation. Stops metabolism. Life has gone… And with it has gone human *individuality*. Knowledge, emotions, preferences, habits, beliefs… We may call all these mental and psychic abilities soul. Where does it go? Nowhere. It just disappears, ceases to exist.

Esoteric believers, who do not agree with this argumentation, should point out what soul features are missing here, certify that these features really exist and propose some candidate for their bearing media.

Beliefs

In esoteric the term "belief" is used, as a rule, in its religious sense, i.e. as belief in God. Sometimes it is said, that humans *should* believe, that those, who do not "believe" are somewhat inadequate persons. I can argue that normal human life is impossible without belief, but beliefs are many. Let us give the definition of belief.

Belief is some (model) entity in a knowledge representation system (KRS), which we consider to be in adequate correspondence with prototype entity in problem domain (PD). PD generally includes other KRS'es (reflexivity) and own KRS itself (selfreflexivity). Properly speaking, all model entities in any KRS are beliefs. Every belief should be accompanied with two indications: prototype status and adequacy of mapping.

Prototype status can be:

- real (we shall not discuss now objectiveness of reality while perceiving World through our senses);
- ideal (mental or cultural);
- indirect evidence - personal or cultural – (for example: "as my friend told me, we will reach town in two hours")

Adequacy of mapping is very delicate issue. Even when we measure physical parameters (weight, dimensions, temperature, etc), we do it with some accuracy. Our senses can mislead us. The mapping itself is not straightforward process: it includes, as a rule some processing (interpretation). That is why our beliefs could be so uncertain. Beliefs drastically vary among different people, religions, cultures.

Let us try to classify beliefs from the viewpoint of pragmatics.

- **Axiomatic beliefs.** They are assumptions taken by intellectual being without proof – on some pragmatic or aesthetic base, for example, axioms of Euclidean geometry. This belief is imaginary with absolute adequacy. Everyone agrees that axioms are purely ideal entities.
- **Beliefs-knowledge.** They can be theoretical and applied. Theoretical – predominantly mathematical – knowledge is obtained through inference from axioms in accordance with specified rules of inference. Applied knowledge is reflection of real world entities: "The Volga flows into the Caspian Sea", "Horses like oats". Of course, applied knowledge can be imprecise, unreliable, inadequate, based on the faulty paradigm or be result of senses mistake. This knowledge is frequently questionable, and it needs to be verified. If verification fails – it is a good reason to dismiss such belief as inadequate.
- **Cultural beliefs.** Ideal entities, created by humankind: sciences, arts, religions, etiquette, folklore, habits, customs, etc. Holding to specific cultural beliefs tends to consolidate people within corresponding social units.
- **"Pure" belief.** This belief usually concerns Intangible World, about which, according the definition, we have no positive knowledge. There is no use to ask pure believers foundations of their beliefs, motivations and confirmations. At most, you will receive reference to some "Holy" books or to authority of certain "known" people. Pure beliefs have nothing in common with World comprehension and can be thought of as some variety of cultural beliefs.
- **Autosuggestion beliefs.** They are beliefs in one's own mental constructions. This belief is of primary importance from the viewpoint of its influence on human personality, psychic and physical state. It correlates with cultural beliefs. We will consider these aspects closer in what follows.
- **Social beliefs**. The basic social unit is family. Members of normal family believe each other, believe their opinions. Intuitively, everyone tends to surround oneself and contact with people, which he could always count on in social aspects. These beliefs are verified and supported by mutuality. If you receive positive stimuli from your surroundings, it means that you are right in believing them. Social beliefs differ from cultural ones: here you tend to believe specific persons, and therefore – you believe to their ideas, views, advises.
- **Materializing beliefs and omens**. Esoteric people introduce these beliefs with the words: "We know that the thought is material". Heroes in fairy tales say: "According to jackfish will and to my wish…" – and requested action is performed in reality, or requested object materializes in the physical world. It should be said, that literally such effects are not confirmed rigorously. But such beliefs sometimes can have extraordinary positive material effects. But the leading factor here is psychology. In medicine is well known "placebo" effect, when some indifferent substance is given to the patient with accompanying words, that it is new very efficient remedy against his illness. The patient ***believes*** in it and mobilizes his nervous system, conscious and subconscious towards success. This effect can be obtained on the pure suggestion level – without material substance. And it supports our understanding that nerve system and psychological determination can really influence physiology of our organisms. In this aspect, such beliefs correspond to cultural, autosuggestion and social ones.

Autosuggestion and Cultural Beliefs, Egregors, Religions

The main principle, which we should, in my opinion, adhere to, is the well known philosophical maxim, Occam's razor: "Plurality should not be posited without necessity" [Occam]. It means that if we can explain some phenomenon with few simple arguments, it should be done unless we have the sound base to change our mind. In mathematics the good taste dictate choosing as simple, evident and small in number axioms, as is possible. This approach permits us filter out mysticism, metaphors, unrestricted imagination, which have no persuasive arguments in their support.

On the current level of our research, we can assert that Stratum 1 – personal – is really good entrance into the realm of bringing up superhuman race. This way is good for everyone. Only results and achievements will somewhat differ depending on genetics, psychological type, dedication to the method and trust to the teacher. Let us formulate several issues in this concern.

- You need to construct and then strictly adhere to positive psychological pattern: you are happy, healthy, optimistic, sociable, you explain your ideas freely and persuasively, you are always in good mood, smile to everybody and sincerely wish them all the best. The power of nerve system over the physical body and over your other abilities is so high, that the results are just miraculous. Besides, we can rely on the fact that in such complex system as our organism is sometimes cause and effect change places, and you can induce some situation if you really believe in its consequence. In medicine are known situations where patients cured themselves just with their "will power".
- As we argued earlier, our subconscious sphere has tremendous memory and processing power. We can apply to it with the help of special psychic practices, known as suggestion, autosuggestion, hypnosis, trance, meditations. There is abundance of literature on this theme; procedures depend on each specific person. Mastering these techniques personally is possible, but much easier success is achieved in groups and, of course, under the guidance of qualified teacher.
- Special exercises should be done to revitalize certain organism's functions - distal nerve and microcapillary systems, acuity of vision, biofield sensitivity.
- Nutrition, and especially, breathing is of vital importance not only for overall healthy state of the physical organism, but in developing extraordinary psychological abilities. We shall not accentuate importance of the general physical training and procedures of personal hygiene – it is self-evident.

Till now, we were speaking about personal psychological technique. Now we will turn to the social aspects. In the core of this issue always lies some community. It may be family, political party, ethnic, national, religious communities; sporting team, graduates of some university, etc. Such communities have certain features in common.

- Specific community is created around some idea, belief, sphere of common interests, or just on the basis of certain common features of their members.
- Frequently there is some documentation concerning the community. It may be unwritten law, ethnic traditions, statute of the party, "holy" book – in case of religion. Sometimes members of a community are formally registered and receive personal certificates confirming their membership.
- There exist special regulations concerning the membership. It may be free, or an applicant should satisfy some requirements (believe in some idea, have certain qualities, or be eager to obtain them, etc). So, there exist a procedure of initiation, enrollment into the community, and contrarily – procedure of disfellowship.
- Usually special symbolism is present. Official colors, banners, badges, carefully developed specific rituals: initiation, meeting, worshiping, special exercises, gestures, dances. Frequently community is organized as followers of some teaching, and thus there are real or imaginary persons in the past or even at present, which are greatly respected. Very popular are their portraits, exceedingly idealized, as a rule.
- Each member has a set of liabilities concerning the broad spectrum of what should and what should not be thought, said and done at proper circumstances. As a reward community's member receive sometimes material, but – what is much more important – moral, spiritual support, that wonderful feeling of participating in something great, true, even holy activity. Frequently, such community promises take a form of future happiness to come, maybe even after physical death. This issue has paramount importance and is probably determinative in the whole community–members' affair. This mechanism works on psychological level and has powerful suggestive effect.
- The "power" of the systems being considered heightens with the number of their members and with the system's age.

Community, complying to considerable extent with the issues given above, together with its members is named "***egregor***" (see, for example, [Bernstein]). Egregors could be thought as having two hypostases. First is social and psychological, and there is not a slight doubt that it is really existing ideal entity. The second hypostasis, from the current scientific point of view is mystical. Genuine esoteric considers that, besides socio-psychological foundation, egregors have some fine-material aspects, probably being the part of Stratum 4 – Absolute. This thesis today has no scientific confirmation.

Relying on the considerations just given, we can argue, that religions, Christianity, in the first place, are the typical representatives of egregors. This issue gives us the solid foundation for understanding religious activity in human societies.

Conclusion

Let us remember what was our pathos and determination when idea of creating AI sprang into existence. Yes, it was the dream about the future society, where AI, wise and helpful robots will help us in our intellectual work. Being carried away with this noble humane idea, we forgot about humans themselves. We thought that we know all principal mechanisms of brain activity, neglected sphere of unconscious and restricted our intellect simulation to known area. At the same time, already during several thousands of years flourished alternative approach to dealing with human intellect and psychics – religions, various beliefs, esoteric. Adepts of these disciplines suppose, that they possess much more general, powerful and complete knowledge about the World and humanity, but orthodox science usually avoid studying this domain, using the label: "pseudoscience". Benevolent consideration of this problem shows that, as always and as everywhere, the realm of esoteric encloses the broad spectrum of ideas, approaches, techniques, etc, with rather broad spectrum of utility. Some things look very unrealistic, even mystic, others could be positively considered if serious confirmations and verifications will be available. And still others look very rational and could be used in mental and psychological practice already today. Frankly saying, reality of esoteric experience should be carefully studied [JREF].

Besides that, until now we doing AI research used only hardware provided to us by engineers. It may happen, that "programming" human beings will give much more efficient results.

Acknowledgements

I am heartily grateful to Valeria Gribova from Vladivostok, Russia, who attracted my attention to the world of intangible, and presented her evidences and personal experience. I also had extensive exchange of opinions through e-mail with Elena Zykova (Juanita), parapsychologist from Irkutsk, who presented her understandings and feelings about the World. A low bow to you, Juanita for all efforts and time spent by you on your blunt pupil.

References

[Bible] The Gospel According to Saint John, 1:1, The Project Gutenberg Edition of the King James Bible, http://www.gutenberg.org/dirs/7/9/9/7999/7999-h/7999-h.htm

[Vernadsky] Владимир Иванович Вернадский, Несколько слов о ноосфере, 1943-1944, Из Архива В. И. Вернадского: http://vernadsky.lib.ru/ .

[Teilhard] Pierre Teilhard de Chardin, The Formation of the Noosphere, 1947, http://technoetic.com/noosphere/

[Lozovskiy] V.Lozovskiy, Towards the semiotics of Noosphere, International Journal "Information Theories & Applications", ISSN 1310-0513, Ed. in chief Krassimir Markov, 2003, Vol. 10, # 1, http://www.foibg.com, p. 29-36

[Bronnikov] V.M.Bronnikov, http://www.bronnikov.org/, http://galactic.org.ua/W-Bronnicov/Bk-1.htm , http://galactic.org.ua/W-Bronnicov/b_l_metod1.htm , First Stage

[Reiki] http://reiki.7gen.com/ , Reiki – Usui Shiki Ryoho: This is to certify that Vitaliy Lozovskiy has completed the First Degree in the Usui System of Natural Healing. You are initiated in the 8th generation of Mikao Usui, Chujiro Hayashi, Hawayo Takata, Phyllis Lei Furumoto, Fokke Brink, Hans Treffers by Victor Varchuk, 18.10.2004, Odessa

[Zykova] Elena Zykova, Irkutsk, personal communication, 2004 - 2005

[Serigawa] Katsusuke Serigawa, TSUBO. Vital Points for Oriental Therapy, Japan Publications, Inc, Tokyo, Japan, 1976, 256 p.

[Carroll] Robert Todd Carroll, The Skeptic's Dictionary: A Collection of Strange Beliefs, Amusing Deceptions, and Dangerous Delusions, http://www.skepdic.com/contents.html#O

[Occam] see [Carroll], William Occam, the medieval English philosopher and Franciscan monk, William of Ockham (ca. 1285-1349) http://www.skepdic.com/occam.html

[Bernstein] L.S. Bernstein, Egregor, Rosicrucian Archive, http://www.crcsite.org/egregor.htm

[JREF] One Million Dollar Paranormal Challenge, James Randi Educational Foundation, http://www.randi.org/research/index.html

Author's Information

Vitaliy Lozovskiy – Institute of Market Problems, Economic and Ecological Researches, National Acad. Sci. Ukraine, 29, Francuzskiy Boulevard, Odessa, 65044, Ukraine, Sen. Sci. Res,; e-mail: loz@loz.intes.odessa.ua

INFORMATICS, PSYCHOLOGY, SPIRITUAL LIFE

Larissa A. Kuzemina

Extended Abstract:

Informatics as a key link of our life is directly connected to a human perception with all its sides. Everything in our life rises in perception stems from it and develops.

To live means to perceive, to perceive means to develop oneself, to develop oneself means to live more and to perceive more [1].

The process of perception is connected to a psychological self-programming for selection, comprehension, accumulation and transmission of information. This brings us into the channel of a personal freedom psychology forming (ability of a person to control his development closely connected to self-consciousness, resourcefulness, openness, readiness to changes). In the course of self-consciousness development the range of the human choice and his freedom widens.

Freedom is considered as a form of activity characterized with three indicators: perceptiveness, instrumentality of "what for" value and controllability at any point. Respectively, deficiency of freedom may be related to misunderstanding of the forces acting on a subject, absence of clear value patterns and indecision, incapability to interfere in the course of the own life.

Under favourable conditions merging of freedom as a form of activity and responsibility as a form of regulation takes place. Under adverse conditions either own activity retreats to the background giving way to external requirements, situations, factors, or responsibility as a form of regulation doesn't regulate a manifestation of freedom.

Freedom is formed in the process of gaining by a person the internal right for activity and value patterns.

The general principle is expressed in the brilliant formula presented by Hegel: "Circumstances and motives have domination over a person only to such a degree that he himself allows them to have" [2]. So, freedom consists in rising regulation to a higher level.

Lest freedom be degenerated into tyranny it should have a value-semantic substantiation.

The same algorithm is present in a self-appraisal as a flowing value. The case in point is a strong personal appraisal moment (subjectivism of perception). As every individual exists in a medium (family, affiliation, society as a whole) then his perception by his associates and his self-perception is formed on the background of the existing standards and value patterns (general cultural, collective and individual ones). In brief, self-appraisal is a personal judgement about own value expressed in aims inherent in this individual.

There is an ingenious formula derived by James (1890) indicating two ways of rising a self-appraisal:

SELF-APPRAISAL = SUCCESS / CLAIMS.

In fact an individual can perfect his own appraisal either increasing the numerator of this fraction or decreasing its denominator as only relation of these indices is important for a self-appraisal. As James noted "our self-perception in this world depends purely on what we are going to become and what we are going to do".

In the system of multi-channel information the art is one of powerful means of action on perception, psychology and, ultimately, on the spiritual life of a person [3]. In this connection

1. purposeful action of all genres and aspects of art,
2. selection of coming information,
3. accumulation of positive information, development of art taste

are essential.

References

1. *Khol'nov S. Yu.* The Art of Perception or a Man without a Form – St-Petersburg: "Respeks", "Dilya", 2003. - 176p.
2. *Hegel G.W.F.* Works of different years. Moscow: "Mysl", 1971. V.2.
3. *Kuzemina L.* A. Music as the Source of Information Influence and Soul Education.

Author's Information

Kuzemina L.A. – Prof. of Kharkiv State University of Arts, kuzy@kture.kharkov.ua

INFORMATION SUPPORT OF PASSIONARIES

Alexander Ya. Kuzemin

Introduction

Noosphere. What is ahead? Can science give us understanding of the ideal of the future?!

It should be taken into consideration that among important problems facing the living ethnos [1] it is possible to single out some problems. What for, how do we live? Where and why can't we live better or how should we find the most effective way out of unfavorable situations. The answers to these questions can be found with a high degree of certainty on the basis of information about the past period of life of this ethnos, the data concerning the current state, their analytical processing and prediction of events movement in the future.

In this connection, **the problem of developing the concept of design of the PASSIONARIES information support arises.** The results of the development can make it possible to approach the answers to the questions above formulated. First, it should more objectively represent, understand and then prepare the decision-making and predict the ultimate situations' movement.

Solution of an isolated even very complicated problem for any subsystem in the sphere of human relations is sometimes cheerful, pleasant but always with no prospect with respect to objectively existing open for interaction with the external medium or other subsystems of the real system of human activity. Generally, solutions for analysis of the similar phenomena are found in the animate and inanimate nature. There are no complete analogs of the considered problems; so, a new problem should be solved anew. In some cases it seems that a new solution in the design work or investigation will be obtained if it is sufficient **to change a standpoint (an aspect) in the already performed investigation** or to add a new material or **to change an approximation degree** (for instance, to take a telescope or microscope instead of spectacles). However, on frequent occasions new efforts are needed which promise the same incomplete results as those ones without change of aspect in considering the problem. Such is the limit of the traditional methods of investigation.

In natural sciences, estimations of the events current state are irrelevant. First, the classification is necessary. In this case, it is important how can one obtain the needed classification. Thus, for example, zoologists classify as the earth ones sea (whales) and air (bats) animals as mammals because there is one correctly chosen criterion uniting them.

But not every generalization is fruitful for science. The concept "**humanity**" means **contraposition of the species** Homo sapience to all **other** animals [1]. The human body has a ***rigid relation to a feeding landscape.*** It is a **native land.** It is necessary to accommodate oneself to the use of the landscape resources in this case during a considerable time. One should adopt to the use of the landscape resources and it takes a considerable period of time. **Adaptation** (plasticity) of **generations** to unusual natural conditions is needed. **Assimilation by the descendants a set of traditions** necessary for future existence results eventually in transformation of **a native land into Motherland.** Forming of the human body is not limited by **imitation of ancestors**. "Eccentric" people - such **creative persons** as artists, poets, musicians giving the specific aspect to the human body appear and exist on their own in the human body

Combination of **adaptation of generations, imitation of ancestors and existence of the creative people** are **necessary** coordinates forming “ethnos”, which are characterized by the original stereotyped behavior and unique internal structure. **Ethnos** is a form of a Homo sapiens type. It is accepted to name generalization of all individuals of this anthroposphere or “ethnosphere”.

Sometimes from depths of subconsciousness of some people with a thin nervous organization the “**genetic memory”** shows itself (N.V. Timofejev-Recovsky named this phenomenon “an emergency gene”). Vague memoirs about not endured events occur to such people. The extremely seldom emerging gene transfers fragments of the information uniting mankind, which is represented by a mosaic of ethnoses. It is precisely the presence of genetic memory that unites the anthroposphere. Otherwise mankind would have scattered on some kinds and the racial theory would have triumphed.

Besides this the **combination of elements (people, families, generations) defines the system integrity of anthroposphere**. When generalizing the global history of mankind the legitimacy of the system approach is

obvious. **The landscapes** also comprise a part of a daily life of ethnoses as **elements**. This is "**ethnology**" – a combination of ethnography and natural scientific explanation of an origin and change of ethnic integrity.

From a point of view of the system approach, **ethnos unites** its members not only according to similarity of its members (common language, common religion, uniform authority). Application of the "similarity" concept leads to absurdity. The French is the ethnos, but they speak four languages. In 1937, L. Von Bertalonphy offered to consider a concept "**species**" as "**a complex of elements which are in interaction**" and named it "**an open type system**". A family (a husband, a wife, a mother-in-law, a son, a daughter, a shed, a draw-well, a cat) living in one house represents the known example of the system. Ethnos is not determined by borders of the state but by the interaction of its elements.

Not a single ethnos is eternal. Constant variability in time and space is a law of nature. Ethnoses develop according to the laws of irreversible entropy and lose the initial pulse caused them, the same as any movement fades due to resistance to the environment.

In 1908 V.I. Vernadsky paid attention to the information appeared in the newspaper about **a power phenomenon** of mass flight of the locust in an extremely great number, exceeding the stocks of all deposits of copper, zinc and tin on the Earth, from the blossoming valleys of Ethiopia to the Arabian desert to perdition. Eventually, V.A. Vernadsky came to an ingenious conclusion that **all living organisms posses some biochemical energy of living substance of biosphere** and completely not mystical energy but an ordinary, similar to electromagnetic, thermal, gravitational and mechanical ones. And in the case with the locust, the mechanical energy manifested itself. Mostly the biochemical energy of a living substance is in homeostasis i.e. an unstable balance, but sometimes its fluctuations i.e. the sharp rises and decays are observed.

The more complicated is the organism, the more factors define complication of its system integrity of ethnic systems and the more diverse is their manifestation in the apparent history. **The urgency of the development and use of a global, self-organizing information system is obvious, it would make it possible to preserve this integrity or, at least, to decrease the risk of a complete self-destruction, or to bring to nothing an opportunity of the origin of a negative situation.** There is a need in the information support of the current control, prediction, forecasting of a negative situation onset, analysis of the course and forecasting of the consequences of the ethnogenesis outbreak culmination, its end when the outbreak decays and movement passes into homeostasis of ethnic systems. It is known that all ethnoses passed the phases of rise, overheating, wretchedness and inertia, but every ethnos did it in its own way.

Laws of Dialectics

All types of energy are perceived by a human being not directly but through an observable effect, but to gain an effect the structure consisting of many elements is needed.

PASSIONARIES (individuals of power redundant type) are the individuals possessing an inherent ability to absorb more energy from an external environment than it is required only for personal and species self-preservation and to give out this energy as a purposeful work on modification of their environment. Judging about the increased passionarity of this or that person by the characteristic of his behavior and mentality (see. *Passionarity as the characteristics of behavior and mentality*). As L.N. Gumiliov showed at the ethnic level that the mass change of people behavior in the direction of increase of their activity is an effect of *energy of the biosphere living substance* we can speak about the reasons of the increased activity of a separate individual. In this case we should proceed from the assumption that the mental and intellectual activity requires expenses of energy precisely in the same way as the physical one, but this energy is in other form and it is more difficult to register and measure it.

Another aspect of the power redundant structure is its ability influence actively behavior and mental state of the associates, it is the so-called phenomenon of the *passionary induction* and this feature is inseparably linked with a high level of the general behavior activity.

L.N. Gumiliov made the historical description of the human bodies' activity change by the *ethnogenesis phases.* The author repeatedly addressed to the characteristics of the persons among outstanding PASSIONARIES showing that the frequency of occurrence of such individuals in different epochs is naturally associated with a general activity of the ethnic system (its *passionary tension*). Having analyzed the biographies of such figures as Napoleon, Alexander the Great king of Macedon, Sulla, Jan Hus, Jeanne d'Arc, arch-priest Avvacum, Hannibal, Genghis Khan and many others L.N.Gumiliov found out that all of them were united by a steady complex of

features. The core of them was an insuperable internal aspiration to the vigorous activity changing an environment. This aspiration was dictated neither by the conditions of environment, nor material boons. On the contrary, the unreasonable activity of many PASSIONARIES caused misunderstanding and condemnation of the associates and resulted in their deprivations and destruction. Moreover, the history knows many cases when people laid down their lives for the sake of an idea or a great common cause and not under the influence of a minute but deliberately. All these facts forced the author of the theory to single out the *passionarity* as a separate behavior pulse not reduced to any known biological instincts or mental properties of a person. Having described the *passionarity* on the popularized level (as the activity of ethnic bodies) and on the individual level (as the behavior pulse) L.N. Gumiliov offered the only inconsistent explanation of this phenomenon i.e. the passionarity theory as an effect of *energy of a living substance of biosphere* described by V.I. Vernadsky. As other explanations are not offered yet, we proceed in our definitions from this theory.

PASSIONARY INDUCTION – it is variation of moods and behavior of people at the presence of more passionary persons. *Passionaries* are able to impose their associates their own behavior aims, to impart to them the increased activity and enthusiasm not inherent to these people They begin to behave as if they were passionaries but as soon as a sufficient distance separates them from the *passionaries* they assume their own natural behavior and mental aspect. The phenomenon of Pasionary induction is the most appreciable during wars when passionary commanders succeed in leading behind themselves the armies consisting generally of the *harmonious people*. Thus, Napoleon and Suvorov in Italy succeeded in gaining brilliant victories, but as soon as the armies remained without their commanders the successes were replaced by defeats. It is impossible to explain this phenomenon only by the presence of a leader's talent. During a fight, the effect of a personal presence of a passionary stirring the army to attack is very important. It should be noticed that the passionary induction does not always proceed from the commanders. Very often private soldiers but not passionaries become its source. Passionary induction underlies successes of many famous public speakers whose speeches tremendously impressed the audience or threw them into fury. But when reading the texts of these orators, they don't produce a similar impression.

The cited examples are only the brightest demonstration of the passionary induction. As a whole it penetrates into all ethnic processes being a basis of all mass movements of people, initiators of such movements are passionaries carrying away less passionary people. This is the essence of political movements, great migrations, religious heresies.

Having observed a number of examples L.N. Gumiliov came to conclusion that the passionary induction acts upon the people belonging to the same ethnic group with passionaries (the sources of this induction) much more profoundly. All this clearly demonstrates the fact that passionary induction is one of major factors due to which the ethnos acts as a single unit. To explain the passionary induction phenomenon L.N. Gumiliov offered the hypothesis of the passionary field (see *A field in ethnology*).

The passionary pulses and passionaries themselves should have enough energy potentials for performance of their mission. One of the energy components inherent to passionaries is the information component. And this component can be substantially increased qualitatively and quantitatively. The purpose of the given work consists in the information support at the expense of the correct system analysis of the passionaries' purposes. The passionaries' purposes can make the basis of development of a simulation language and system simulation as a whole at the expense of system representations.

As a result of the system analysis, the information support of passionaries and passionary pulses should be linked with the activity of special situational centers.

The situational center (SC) for information support of the passionary activity can represent a set of management bodies and hardware-software means complex intended for information support of the administrative decision-making by the passionaries and practical support of situations management in real time under various conditions of the situation development:

- **at a normal development** when SC realizes collection, analysis and accumulation of the data circulating in the uniform information space of the ethnos existence and uses historical experience;
- **at a transitive stage** when the situation leaves the condition of a normal development, but does not reach a level menacing the safety of the state or ethnos when returning the situation to a normal development SC is used as the regulating tool;

- **at a crisis stage** when SC is used by the passionaries for returning the situation to a normal development by means of the united management in real time by all state bodies and structures, down to a separate executor.

The use of advanced information technologies allows to provide duly reception and exchange of the authentic information in the state bodies, reduction of terms of the decision-making by passionaries and brining decisions to the notice of executors, qualitative control of execution and reliable protection of the information being the state secret with a limited access. Moreover, the used technologies allow to supervise visually the situation in zones of risk, including on the border and in the zone of battle actions transmitting the information directly to the persons making decisions with the minimal expenses.

The situation management is based on the following principles:
- Independence and reliability of the information sources as the basis of the administrative decision-making;
- Relating of the information arriving, processed, stored and circulating in SC to coordinates of the territory related it;
- Creation of the uniform information space with a state databank allowing to trace the stages of situations origin and development, to store parameters describing a situation at all stages of its development;
- Simulation of separately taken situations development and their sets for estimation of the administrative decisions' adequacy to the situations' development stages using the technologies making it possible to estimate the condition situations under observation, to reveal the reasons of important changes, to predict development of processes and to ensure the automated search previously unknown laws in the uniform databank.

The basic purpose of SC consists in perfection of the state management mechanism based on the uniform information space and uniform system of situations development monitoring.

According to the specified purpose, the SC basic tasks are formulated as follows:
- continuous collection, processing and storage of information;
- support of situations development management including information support to the person making the state administrative decisions and representation of information about results of these decisions realization;
- formation of the appropriate information models of the situations development, revealing, estimation and forecasting of the destabilizing factors;
- perfection of mechanisms of threats neutralization and liquidation of their realization consequences;
- perfection of organizational structure and mechanism of the control system functioning.

The SC development and functioning for passionaries support intended for solving the global tasks of the ethnos management should be primarily associated with the problems of the ethnos economic safety. The importance of introduction into consideration of the given direction is primarily connected:
- with transient nature of economic processes, it borders in some cases unpredictability and errors in the forecasts concerning the subsequent economic development;
- prevailing of the globalization tendency in forming various processes both of specifically economic nature, and various spheres of scientific and technical development;
- with the extent of the extreme situations action on all spheres of human activity and necessity of passionaries' plans realization in the complete volume;
- with the possibility to consider unforeseen economic crises and decays as specific situational aspects of management affecting the efficiency and effectiveness of the decisions made by passionaries (mainly it is associated with the possibility of financing the measures directed on liquidation of emergency situations completely and with the highest economic efficiency).

In this case the economic safety, as the component of the situational center functioning should be understood:
- in a narrow sense – in terms of functioning of the center itself,
- and in a wide sense – in terms of the analysis and estimation of the general economic situation necessary for successful activity of a passionary.

But despite of the possibility to consider the economic safety from different standpoints the decisive meaning consists in its extended treatment. In other words, the most important is consideration of economic safety from the standpoint of the analysis and estimation of an arising general economic situation. It is connected with the fact that the economic safety criterion is the economy state estimation in terms of development of the major processes presenting the essence of the economic safety. Therefore, the criterion estimation of safety includes the estimates of:

- the resource potential and possibilities of its development;
- the level of efficiency of use of the resources, capital, work and its conformity to the level in the most advanced countries, and also the level where the threats of external and internal nature are reduced to a minimum;
- competitiveness of economy;
- integrity of territory and economic space;
- stability both conditions of prevention and solution of the social conflicts.

Here the most significant thing is to single out concrete tasks which should be solved in monitoring of the developing economic situation. Primarily, the following tasks should be singled out:

- the analysis of the financial flows movement direction of the different subjects of managing and their interaction within the limits of the so-called zones of Borderland (both inside of separately taken country, and at the interstate level). In this case, the special attention in the given analysis is paid to functioning of the bank system. As a separate task, we should single out prevention of the possibility of negative effect on the economic stability of the bank system. It can be achieved by means of the analysis and forecasting of economic and political conditions in the bank system. At the same time, such analysis should include definition of influence of certain political groupings in the bank sphere, dependence of the bank system on other branches of national economy being under control of various financial and industrial groups. Moreover, the analysis of economic safety of the bank system should include also consideration of dependence of the bank system subjects on political and economic forces of other states which can carry out a number of measures directed on ruining the bank system. As one of possible versions of using the given research results we may indicate: definition of priority directions of the financial flows movement between the countries of the Borderland to minimize the accompanying economic risk and, as a consequence, to increase the economic safety of the general control system of emergency situations;
- definition of the most probable economic risks rendering the greatest influence on efficiency of the emergency management system. Consideration of the degree of the risk-forming factors' effect on the financial stability of various subjects of managing aimed for warning and development of specific situational aspects of management associated with unforeseen economic consequences. In this case it is necessary to take into account that not the indices themselves but their threshold values, their limiting size are of importance for the economic safety. Violation of these values gives rise to the problems in the normal course of development of various elements of reproduction, results in formation of the negative, destructive tendencies in the field of economic safety. As an example (in relation to internal threats), it is possible to name a rate of unemployment, breaking in the incomes between most and least supplied groups of the population, rates of inflation. The approach to their extreme allowable value testifies to the increase in threats of the social-economic stability of the society, and the excess of the limiting or threshold values indicates to the fact that the society comes into the zone of instability and social conflicts. It is a real detriment of the economic safety. From the standpoint of external threats an extreme allowable level of the state debt, preservation or loss of positions in the world market, dependence of national economy and its major sectors (including the defense industry) on the import of foreign technique, furnishing products or raw material can serve the indicators;
- Substantiation and choice of the policy of finance administration of the situational center. That should provide diversification of the sources of attracting financial funds for solving various tasks of the situational center.

The not less important question in aspect of the considered directions of research is also the choice of formalization and analysis apparatus. As such an apparatus the theory of fuzzy sets can be used as just within the framework of the given theory it is possible to describe the processes occurring in economy and the procedures of emergency situations management associated with them.

Author's Information

Kuzemin A.Ya. – Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, (Ukraine), kuzy@kture.kharkov.ua

INDEX OF AUTHORS

Pavel P.	Antosyak	250
Irene L.	Artemjeva	132
Andrey	Averin	509
Ruslan A.	Bagriy	504
Andrii A.	Baklan	454
Vladimir B.	Berikov	286
Vyacheslav	Bikov	37
B.	Blyuher	265
Yevgeniy	Bodyanskiy	622
Igor A.	Bolshakov	328
Elena I.	Bolshakova	328
M.F.	Bondarenko	312
Frank M.	Brown	529,537,545,553,560
David	Burns	336
T.I.	Burtseva	517
Yuriy A.	Byelov	23,32
Dmitry	Cheremisinov	496
Liudmila	Cheremisinova	496
Andrey	Danilov	638,649
Dimiter	Dobrev	461
Volodymyr S.	Donchenko	218,223,404,605
Vladimir	Donskoy	289
Elena V.	Drobot	243
Olga V.	Dyakova	253
Pawel	Dymora	412
O.	Dytshenko	307
Yuliya	Dyulicheva	289
Alexander P.	Eremeev	272
Alexander E.	Ermakov	190,199
Richard	Fallon	336
D.	Fastova	307
Boris E.	Fedunov	446
M.	Fomina	76
Alla	Galinskaya	584
Tatiana	Gavrilova	127
Krasimira	Genova	279
Anatoliy Ya.	Gladun	158
Gleb S.	Gladun	395
Victor P.	Gladun	344,395
Grigoriy N.	Gnatienko	250
Lilia	Gnibeda	627
Anastasiya V.	Godunova	395
V.M.	Golovnya	237
Iryna	Golyayeva	265
Petro	Gopych	608
Yevgen	Gorshkov	622
T.	Goyvaerts	305
Valeriya	Gribova	153
Leonid L.	Gulyanitskiy	454
N.	Gusar	307
Vladimir D.	Gusev	53
Violeta	Gzhivach	389
Miroslaw	Hajder	412
Roman V.	Iamborak	23,32
Luke	Immes	355
Krassimira	Ivanova	631,638
Natalia	Ivanova	638
Yurii L.	Ivaskiv	395
Vladimir A.	Izotov	17
Grigorii V.	Jakemenko	395
Valeriy A.	Kamaev	171
Mykola F.	Kirichenko	404
Nadezhda N.	Kiselyova	364
Alexander S.	Kleshchev	132,147
Margarita A.	Knyazeva	147
Vitaliy	Kolodyazhniy	622
Anton	Kolotaev	165
Michael	Korbakov	567
May	Korniychuk	486
Aleksey P.	Kotlyarov	328
Todorka	Kovacheva	616
Valeriy N.	Koval	98,104,112
Sergey P.	Kovalyov	53
Alexey	Kravchenko	567
Anatoliy	Krissilov	265
Victor	Krissilov	262
Sergey	Krivoi	389,412
Larissa	Kuizemina	669
Yuriy V.	Kuk	104
A.	Kulikov	76
Michael	Kussul	584
Natalia	Kussul	567, 570
Alexander Ya.	Kuzemin	305,307,670
Alla	Lavrenyuk	627
Gennady	Lbov	60
Yu.G.	Lega	517
Natalja	Leshchenko	375
V.M.	Levykin	305,318
Phil	Lewis	336
Alexander	Lobunets	119
Vladimir	Lovitskii	336
Vitaliy	Lozovskiy	657

Tatyana	Luchsheva	88
Elena	Lyashenko	212
Igor	Lyashenko	212
Alexander	Makarenko	600
Nikolay N.	Malyar	247
Krassimir	Markov	631,638
Sergiy	Mashchenko	226
Lyudmyla	Matveyeva	389
Miroslaw	Mazurek	412
V.V.	Melnik	517
Nadiya	Mishchenko	347
Ilia	Mitov	631,638
Sergey	Mostovoi	256
Vasiliy	Mostovoi	256
Xenia	Naidenova	67,174,182,190
Andrey M.	Nalyotov	53
Svetlana	Nedel'ko	84
Victor	Nedel'ko	92
Olga	Nevzorova	351
Agris	Nikitenko	418
Vadim A.	Nitkin	199
Viktoriya N.	Omardibirova	223
Stuart	Owen	336
Alexander	Palagin	140
M.V.	Panchenko	237
A.N.	Papusha	517
Victor	Peretyatko	140
E.	Petrov	307
Nguyen Thanh	Phuong	465,567
Julia	Pisarenko	427
Valery	Pisarenko	427
Nicolaj	Pjatkin	351
Stoyan	Poryazov	655,656
Elina V.	Postol	395
Viktoriya	Prokopchuk	427
Ilya	Prokoshev	400
A.V.	Rabotyagov	312
Zinoviy L.	Rabynovych	1
Dmitri	Rachkovskij	522,578
Elena	Revunova	578
Yulia V.	Rogushina	158
Mikhail	Roshchin	171
Leonid V.	Rudoi	381
Sergey	Ryabchun	98,112
Victor	Rybin	433
Galina	Rybina	433
Natalia V.	Salomatina	53
Volodymyr	Savyak	98,112
Denis P.	Serbaev	404,605
Ivan	Sergienko	98
Fatma	Sevaee	479
Daria	Shabadash	262
Natalya	Shchegoleva	347
Andriy	Shelestov	567
S.	Skakun	570
Velina	Slavova	355
M.I.	Sliptshenko	312
Vitaliy	Snytyk	232
Artem	Sokolov	522
Alona	Soschen	355
Sergey	Sosnovsky	127
Yuri	Sotskov	375,381
Nadezhda	Sotskova	381
Inna	Sovtus	486
I.	Starikova	307
V.	Stepanov	265
Tatyana	Stupina	60
Vyacheslav	Suminov	400
Leonid A.	Svyatogor	371
Adil V.	Timofeev	442, 591
Sergiy V.	Tkachuk	23,32
Daniela	Toshkova	616
Yevgen	Tsaregradskyy	486
Vadim	Vagin	76,509
Yuriy	Valkman	37
Ivan	Varava	427
P.R.	Varshavsky	272
Vassil	Vassilev	279
Mariyana	Vassileva	279,297
Vitaly J.	Velichko	344,395
Olexy F.	Voloshin	205,237,247
Gennady S.	Voronkov	9,17
Karola	Witschurke	321
Anatoliy	Yakuba	98,112
Ekaterina	Yarovaya	627
Alla V.	Zaboleeva-Zotova	171
Nikolay G.	Zagoruiko	53
Arkadij D.	Zakrevskij	491
Alexandr V.	Zavertaylov	53
Yuri P.	Zaychenko	473,479
Julia	Zin'kina	351
Stepan G.	Zolkin	45

International Journal
"Information Theories and Applications"

International Conferences on Information Theories and Applications

supported and published by IJ ITA

Knowledge, Dialogue, Solutions

General Information Theory

Information Research, Applications and Education

Business Informatics

Multimedia Semantics

Digitisation of Cultural Heritage

General sponsor:

FOI BULGARIA

For more information:

www.foibg.com

© 2005 Markov, Ivanova, Mitov