



I T H E A



International Journal

INFORMATION **TECHNOLOGIES**
&
KNOWLEDGE



2011 Volume 5 Number 2



International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE

Volume 5 / 2011, Number 2

Editor in chief: Krassimir Markov (Bulgaria)

Victor Gladun (Ukraine)

Abdelmgeid Amin Ali	(Egypt)	Koen Vanhoof	(Belgium)
Adil Timofeev	(Russia)	Larissa Zaynutdinova	(Russia)
Aleksey Voloshin	(Ukraine)	Laura Ciocoiu	(Romania)
Alexander Kuzemin	(Ukraine)	Luis F. de Mingo	(Spain)
Alexander Lounev	(Russia)	Natalia Ivanova	(Russia)
Alexander Palagin	(Ukraine)	Nataliia Kussul	(Ukraine)
Alexey Petrovskiy	(Russia)	Nelly Maneva	(Bulgaria)
Alfredo Milani	(Italy)	Nikolay Lyutov	(Bulgaria)
Avram Eskenazi	(Bulgaria)	Orly Yadid-Pecht	(Israel)
Axel Lehmann	(Germany)	Radoslav Pavlov	(Bulgaria)
Darina Dicheva	(USA)	Rafael Yusupov	(Russia)
Ekaterina Solovyova	(Ukraine)	Rumyana Kirkova	(Bulgaria)
Eugene Nickolov	(Bulgaria)	Stefan Dodunekov	(Bulgaria)
George Totkov	(Bulgaria)	Stoyan Poryazov	(Bulgaria)
Hasmik Sahakyan	(Armenia)	Tatyana Gavrilova	(Russia)
Iliia Mitov	(Bulgaria)	Vadim Vagin	(Russia)
Irina Petrova	(Russia)	Vasil Sgurev	(Bulgaria)
Ivan Popchev	(Bulgaria)	Velina Slavova	(Bulgaria)
Jeanne Schreurs	(Belgium)	Vitaliy Lozovskiy	(Ukraine)
Juan Castellanos	(Spain)	Vladimir Ryazanov	(Russia)
Julita Vassileva	(Canada)	Martin P. Mintchev	(Canada)
Karola Witschurke	(Germany)	Zhili Sun	(UK)

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.

The rules for the papers for IJ ITK as well as the subscription fees are given on www.foibg.com.

Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the Consortium FOI Bulgaria (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.5, Number 2, 2011

Edited by the Institute of Information Theories and Applications FOI ITHEA, Bulgaria, in collaboration with:
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnic de Madrid, Spain.

Publisher ITHEA®

Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, www.foibg.com, e-mail: info@foibg.com

Printed in Bulgaria

Copyright © 2011 All rights reserved for the publisher and all authors.

© 2007-2011 "Information Technologies and Knowledge" is a trademark of Krassimir Markov

ISSN 1313-0455 (printed)

ISSN 1313-048X (online)

ISSN 1313-0501 (CD/DVD)

DATA ACQUISITION SYSTEMS FOR PRECISION FARMING

Oleksandr Palagin, Volodymyr Romanov, Igor Galelyuka, Vitalii Velychko,
Volodymyr Hrusha, Oksana Galelyuka

Abstract: *In the article it is described two structures of data acquisition systems, which are based on the family of portable devices "Floratest" and suitable for using in precision farming.*

Keywords: *Cautsky effect; chlorophyll; chlorophyll fluorescence induction; data acquisition system; fluorometer; portable device; precision farming.*

ACM Classification Keywords: *J.3 Life and Medical Sciences - Biology and Genetics.*

Introduction

Unforeseen changes of climate and difficulty of using statistical data demonstrated acute necessity to develop models for forecasting crop yield and climate influence on it. These models have to be based on live data, but not only statistical ones. Acquisition of input data for these models is urgent and important task. Earth remote monitoring data (e.g. space observations) and data from surface tools for plant state monitoring can be considered as such input data. It is important to note, that these data are very significant for precision farming technology, which also operates with models for forecasting crop yield and is used for minimization of costs (e.g. water, fertilizers etc.) and maximization of harvest.

Therefore, acquisition of plant cover state live and objective data in most cases is very important factor, which causes future strategy of keeping agricultural lands and proper decision making. Certainly, it would be ideal to obtain information about improvement or worsening of plant cover state beforehand, but not after the event. It lets to avoid increasing costs and save harvest from possible loss.

In the article there are is described tool for express-diagnostics of plant state, notably portable device "Floratest" [Romanov, 2007], and data acquisition systems, built on basis of these portable devices.

TOOL FOR MONITORING OF PLANT STATE

Photosynthetic processes are the processes which supply energy to the cells of plants. Chlorophyll is the main pigment of the cells of plants. One of the main features of the molecular of chlorophyll is ability of fluorescence. The intensity of chlorophyll fluorescence depends on photosynthetic activity. After irradiation of leaf the intensity of chlorophyll fluorescent signal is increasing at first and then slowly reduces. This effect is called as effect of Kautsky [Kautsky, 1931] or effect of chlorophyll fluorescent induction (CFI). The form of this curve is very sensitive to adverse environment.

It gave possibility to develop in the V.M. Glushkov Institute of Cybernetics of NAS of Ukraine the portable device "Floratest", which estimates in several seconds the plant state after drought, frosts, pollution, herbicides etc.

without plant damage. Like human cardiogram device builds CFI curve estimated photosynthesis process, which is the base of plant vital activity.

Device and relevant diagnostic methods refer to the area of biological object researches by detecting their biophysical properties, particularly native chlorophyll fluorescent induction. Device is defined as smart biosensor with fragment plant as sensing element.

Express-diagnostic of plant state is carried out by functional features and is based on using of features of separate specific sections of CFI curve, which refer to separate areas of photosynthesis chains as diagnostic features. By CFI curve form it is easily to detect influence of one or another stress factor on the plant state.

Application areas of portable device "Floratest":

- express-estimating plant vital activity after drought, frosts, sorts coupling, pesticide introduction;
- express-detection of optimal doses of chemical fertilizers and biological additives, what lets to optimize amount of fertilizers and additives and reduce nitrates content in vegetables and fruits;
- express-detection of level of pollution of water, soil and air by pesticides, heavy metals and superpoison;
- economy of energetic and water resources during man-made watering;
- developing precision farming technology for increasing the quality of agricultural products;
- using the device in the insurance agriculture to get predicted results of future yield;
- automation of researches in the plant physiology field.

During field experiments conducting on large areas the sequence of measurements in the time causes distortion of overall results of measuring. Irregularity of ground area causes internally irregularity of plant growth parameters and future crop. Complex and multistage agrotechnical process doesn't always correlate with changes of plant passing through phenological stages, what has place in conditions of unforeseen climate changes. For estimating state of plant photosynthetic apparatus it is used classical acetone method in different modifications. This method is characterized by limited plant samples and long-term tests. So, simultaneity of measuring on the field under the same conditions lets to implement necessary technological procedures, involve necessary agrotechnology, forecast and influence on future crop in time.

It is clear, that it is impossible to provide plant state low cost monitoring of large agricultural field in short time by some wired devices. To solve the problem we need to equip data acquisition system by family of portable devices "Floratest" with communication tools.

DISTRIBUTED DATA ACQUISITION SYSTEMS

Development of electronics and telecommunication technologies over last 10 years causes creating and implementation of electronic data acquisition systems with remote gathering of measured data in many areas of human activity. Such systems include smart sensor modules, communication lines, transmitters and receivers, central control stations, which gather and accumulate information.

As sensors in such system we use developed by us family of portable wireless devices "Floratest". Example of structure organization of developed data acquisition system is shown on Fig. 1.

In most cases the efficiency of data acquisition system work is defined by level of used technologies of data acquisition and data transmission medium. With increasing system scale the contribution of these components in

total efficiency of data acquisition system increases too. So, first of all, it is important to choose technology of data transmission. Analysis of data transmission technologies showed that it is most optimal to create such data acquisition system on basis of wireless network with mobile terminals, which use existing mobile communication systems (e.g. GSM/GPRS). In this case it is no necessity to issue the license on radiofrequency channel using and buy expensive transmission and receiving equipments. In addition, GSM network covers all territory of Ukraine and other countries.

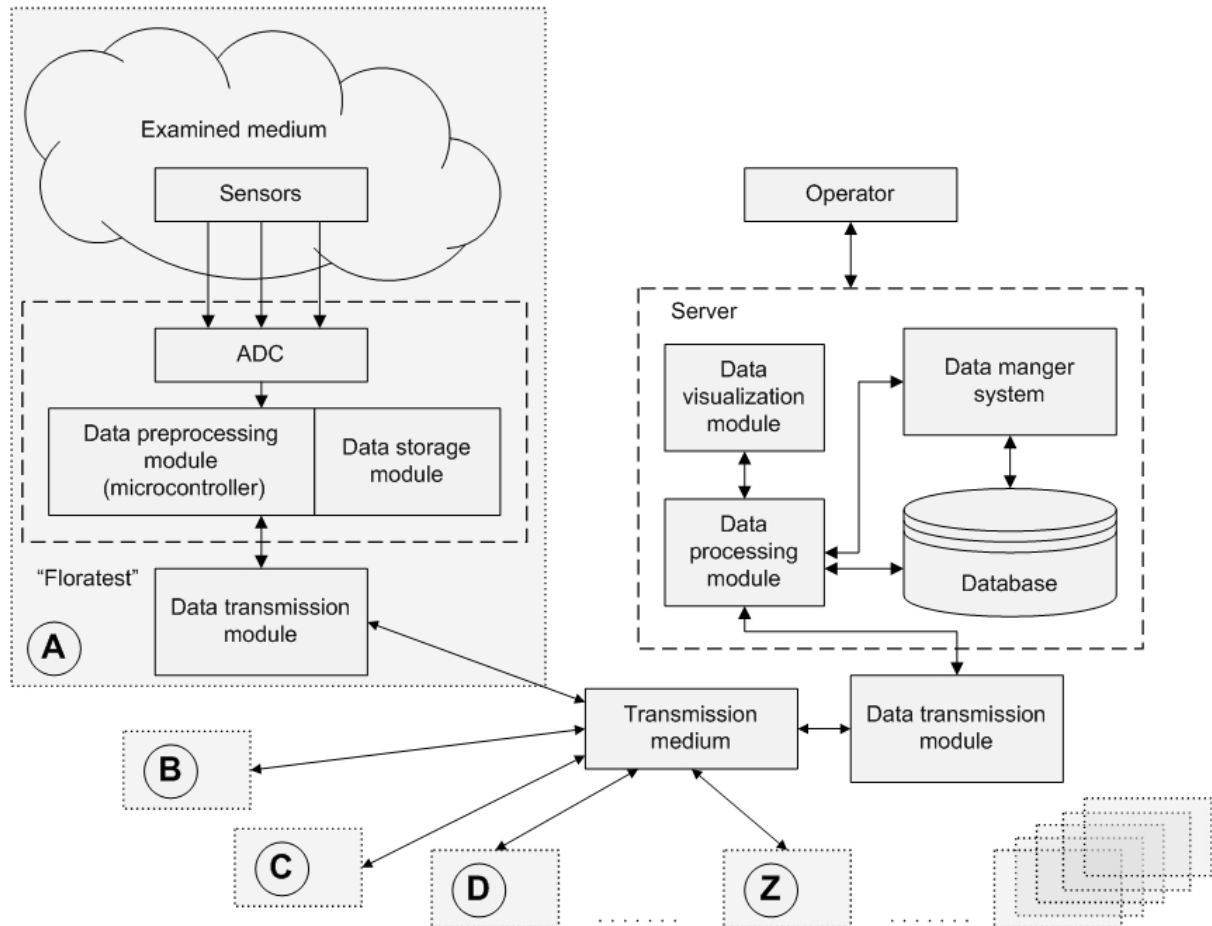


Fig. 1. Data acquisition system on the basis of device "Floratest" family

In developed data acquisition system smart portable devices are used as mobile tools for data acquisition in any country region. Measured data from portable devices are transmitted to central control station, where these data are processed, analyzed and generalized. Collected results are displayed in graphic, tabular or another form and then used for building generalized map of some region state. Map of state defines, for example, plant cover state of agricultural lands, ecological state, existence of plant diseases etc. on some territory or whole country.

In our opinion the main disadvantage of such system is next one: if you want to change or implement new applied calculation methods you need to reprogram all mobile portable devices "Floratest".

To remove this disadvantage we are developing new data acquisition system, which has significant differences from previous one. Principle of such system operation is next. Portable devices "Floratest" measure plant state

without any data processing. Then user defines the examined medium (e.g. plant sort) and "raw data" is transmitted to server. Taking into account the type of examined medium the server will process these "raw data" by means of specific calculation method. Having necessity to alter or develop new applied calculation methods we simple will put new applied calculation method on the server or replace existing one.

Taking into account technology of sensor networks the developing of new data acquisition system is very similar to developing of network element, so next tasks have to be solved:

- minimization of device dimensions;
- minimization of power consumption for long lifetime of batteries;
- polling of sensors and data transmission in digital format.

Unlike classical elements of sensor network our sensors don't support data retransmission from one sensor to another, but this feature will be examined in future.

Data acquisition system, as set of small smart sensors, which are places on the biological objects, needs, as rule, complex software on the server. Such software has to poll sensors and obtain data, control element network and set their operation modes, interpret, store and represent data from sensors.

Generally typical architecture of such system can be presented by next components: sensor module, communication unit with sensor modules, server and work terminal. Sensor module measures parameters of biological object state, in case of need partially processes measuring data, communicates with server and processes commands for work mode change. Communication unit with sensor modules provides two-side digital communication with sensor modules, saves all obtained data, transfers these data to server, retransmits control commands to sensor modules, on demand gives data from sensor modules. Server accumulates data from one or several communication units, saves all obtained data, sends saved data on demand of work terminal. Work terminal gives graphic interface to system users, provides representation, partial (in some cases full) processing and storing data for more detail processing, gives possibility to control system.

Presented architecture is classical for tasks of such type. Among system features one can note storing all data by sensor modules. It is made for increasing reliability of system in the case of server failure. In this case the data will be accumulated in the communication unit and then transmitted to the server. Communication unit with sensor modules performs minimum of intelligent functions and saves all data in "raw format". Data interpretation is made by next system components.

Typical architecture of such data acquisition system with applied calculation methods on server is shown on fig. 2.

New data acquisition system gives us significant advantages:

- 1) it is no necessity to alter or reprogram sensors (portable devices "Floratest") in the case of changing applied calculation methods;
- 2) there are no embedded applied methods, so it simplifies hardware and software of portable device "Floratest" and reduces the price;
- 3) automatic operation of portable device without specialist presence is possible;
- 4) it is large flexibility of the modernization of existing and developing new applied calculation methods.

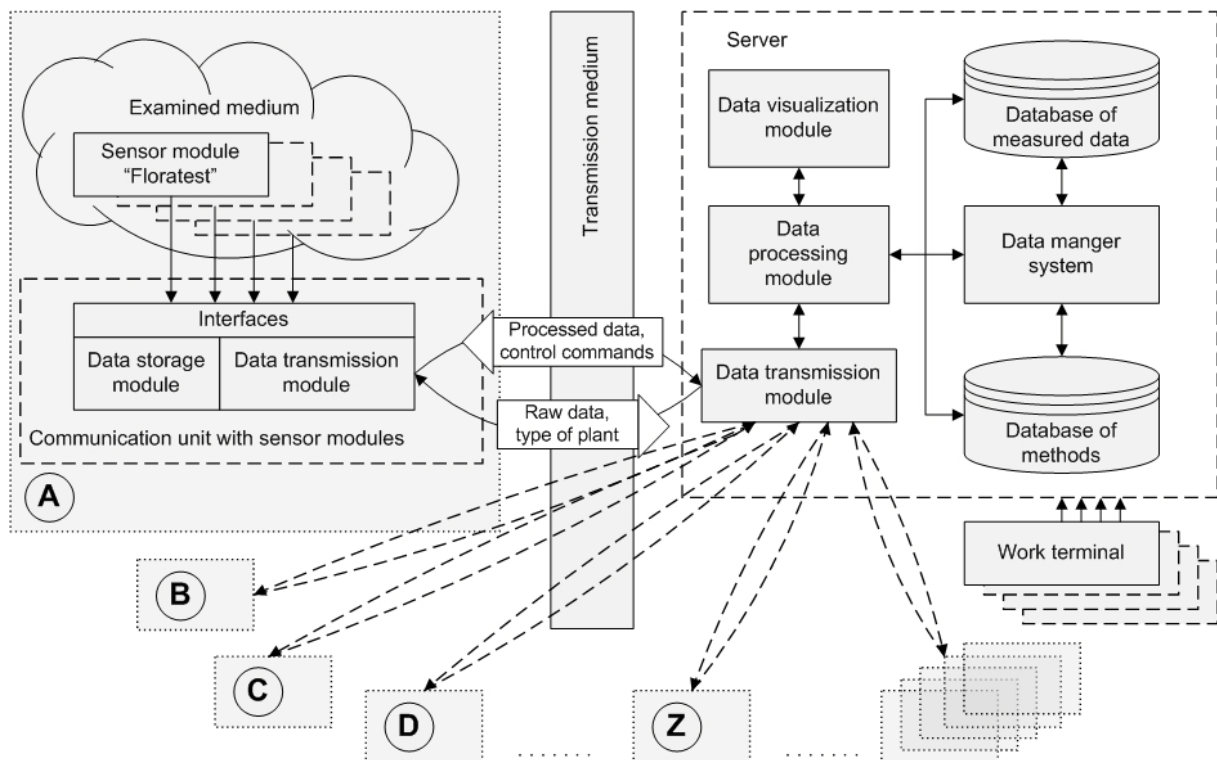


Fig. 2. Architecture of data acquisition system with applied calculation methods on server

DATA PROCESSING ON THE SERVER

For processing of measured data on the server and searching dependences and characteristic features of curve family we used the technology of growing pyramidal network (GPN) [Gladun, 2008]. GPN belongs to class of logic - linguistic information models, i.e. such models where the main elements are not numbers and calculations. The main elements in such models are names and logical connections. Logic - linguistic models can operate simultaneously with different-type data and be described adequately by means of natural language expressions.

The model of classes of the objects, used for the decision of tasks of classification, diagnostics and forecasting, should include all the most important attributes describing a class. The model also should display for this class the characteristic logical connections between essential attributes. Generalized logical multivariate models of objects classes are the *concepts* that in logic are usually defined as ideas that reflect essence of objects. [Voyshvillo, 1967]. The concept in GPN is a generalized logical attributive model of objects' class, and represents the belonging of objects to the target class in accordance with some specific combinations of attributes.

A GPN is an acyclic oriented graph having no vertexes with a single incoming arc. Vertexes having no incoming arcs are referred to as *receptors*. Other vertexes are named *conceptors*. Receptors correspond to values of attributes. When the network is building, the input information is represented by sets of attributes values describing some objects (materials, relations, actions, situations, names of properties, states of the equipment, illness etc.). In the task of searching characteristic features of curve ensemble the receptors correspond to intervals of values of measured data with time stamps. Conceptors correspond to descriptions of objects in general and to crossings of objects descriptions. Conceptors represent GPN vertexes. In this task conceptors correspond to CFI curves and crossings of CFI curves descriptions.

The result of network building is the formed logical expression contains logical relations, represented by allocation of *check vertices* [Gladun, 2008]. Logical expression describes the concepts in the network, defining different classes of objects. The system forms logical models of objects classes which allow taking into account influence on diagnosed or forecasting parameter of separate attributes and their various combinations as well. Besides, it is taken into account influence of "exclusive" attributes which are incompatible with diagnosed or forecasting value.

The analytical tasks, such as diagnostics or forecasting, can be reduced to the task of classification, i.e. to belonging the research object to a class of objects, with a set of properties significant for prognosis. Classification of new objects is performed by comparing the attribute descriptions of new objects with the concept, defining a class of forecasting or diagnosed objects. Objects can be classified by evaluating the value of the logical expressions that represent corresponding concepts. The variables, corresponding to the attribute values of the recognized object, set 1, other variable set 0. If the evaluated expression takes the value 1 it means that the object is included into the volume of concept.

By classification manner GPN is closest to the known methods of data mining as decision trees and propositional rule learning. The main characteristic of the pyramidal networks is the possibility to change their structure according to structure of the incoming information. Unlike the neural networks, the adaptation effect is reached without introduction of a priori network excess. Pyramidal networks are convenient for performing different operations of associative search. Hierarchical structure of the networks, which allows them to reflect the structure of composed objects and gender-species connections naturally, is an important property of pyramidal networks.

Pyramidal networks considerably decrease volumes of search operations that makes it possible to avoid the effect of "information explosion" when solving analytical problems on the basis of large-scale data. Certainly, the full advantages of pyramidal networks are appeared at their physical realization supposing parallel distribution of signals on a network. The important property of a network as means of storage of the information is that the opportunity of parallel distribution of signals is combined with an opportunity of parallel reception of signals on receptors.

Conclusion

- It is described two structures of data acquisition systems, which are based on the family of portable devices "Floratest" and suitable for using in precision farming.
- It is shown, that data acquisition systems with common applied methods on the server let to simplify smart sensor devices but shifting data processing to server and using universal taught method of logic – linguistic data processing.
- Developing and implementation of such systems let to considerably expand area of application of portable device "Floratest" family.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua)

Bibliography

[Romanov, 2007] Romanov V., Fedak V., Galelyuka I., Sarakhan Ye., Skrypnyk O. Portable Fluorometer for Express-Diagnostics of Photosynthesis: Principles of Operation and Results of Experimental Researches // Proceeding of the 4-th IEEE Workshop on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", IDAACS'2007. – Dortmund, Germany. – 2007, September 6–8. – P. 570–573.

[Kautsky, 1931] Kautsky H., Hirsch A. Neue Versuche zur Kohlenstoffassimilation // Naturwissenschaften. – 1931. – 19. – S. 964.

[Gladun, 2008] Gladun V., Velichko V., Ivaskiv Y. Selfstructured Systems // International Journal Information Theories and Applications, FOI ITHEA, Sofia, Vol.15, N.1, 2008. – P. 5-13.

[Voyshvillo, 1967] Voyshvillo E. The Concept. MGU-Moscow, 1967, 285 p. (in Russian)

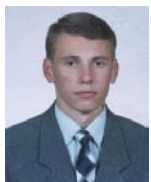
Authors' Information



Oleksandr Palagin – Depute-director of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Academician of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: palagin_a@ukr.net



Volodymyr Romanov – Head of department of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: VRomanov@i.ua; website: <http://www.dasd.com.ua>



Igor Galelyuka – Senior research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of technical science; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: galib@gala.net; website: <http://www.dasd.com.ua>



Vitalii Velychko – Senior research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of technical science; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: velychko@aduis.com.ua; website: <http://www.aduis.com.ua>



Volodymyr Hrusha – research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: vhrusha@gmail.com; website: <http://www.dasd.com.ua>



Oksana Galelyuka – Research fellow of Institute of encyclopedic researches of National Academy of Sciences of Ukraine; Tereshchenkivska str., 3, Kiev, 01004, Ukraine

TERMINOLOGICAL ANNOTATION OF THE DOCUMENT IN A RETRIEVAL CONTEXT ON THE BASIS OF TECHNOLOGIES OF SYSTEM "ONTOINTEGRATOR"

Olga Nevzorova, Vladimir Nevzorov

Abstract: In this article the method of terminological annotation of mathematical documents which is used in a context of text mining (in particular, for RDF network developing of a collection of mathematical documents) is considered. Terminological annotation of mathematical documents is carried out on the basis of universal design technology for applied problems solving developed in ontolinguistic system "OntoIntegrator" under control of system of ontological models.

Keywords: Natural language processing, ontological models, terminological annotation

ACM Classification Keywords: H.3.1.Information storage and retrieval: linguistic processing

Introduction

The idea of Semantic Web and the space of informational objects (a semantic net work of objects with an associate metadata) allows us deal with the information retrieval problem as the information retrieval problem in the space of informational objects. Such approach takes place in LinkedData project [Berners-Lee, 2011] where the space of informational objects represents as a PDF-network. Different semantic entities of the text documents such as structure elements of the document or terminological units or segments with different semantic might be represented as informational objects.

The information retrieval problem in the space of informational objects is essentially different from the searching methods and relevance assessment used nowadays. As distinct from keyword searching in the space of objects (in general keywords are represented by random symbolic sequences) implies searching in the space of names, relations and properties and could be realized on database meta-language. Therefore the important task is to generate the space of informational objects, to separate semantic objects of the text which are conceptual units of the semantic space out.

This paper is dedicated to the method of terminological annotation of mathematic documents, which is used in tasks of extraction semantic objects from the text (particularly in case of generating PDF-network of mathematic documents) and terminological indexation of collection of documents.

Terminological annotation of the document

It supposed that the quality of searching might be better if we use the terminological annotation of a source text and make up a terminological index.

The main problem of this approach is lack of accessible terminological sources in different areas of knowledge, on base of which we could realize the terminological annotation.

The alternative solution of the problem is an automatically extraction of terminology from the text. Nowadays there are 3 methods of terminology extraction: linguistic methods, statistic methods and combined methods. Linguistic methods are based on lexical-syntactic models of one-word and multi-words terminology and a system of filters when non-terminology is shifted out.

Statistic methods are based on idea of a terminological frequency. A term combination usually correspond with n-gramms (binomial, trinomial and tetratomic word combinations), which are characterized with the high level of steadiness. In case of evaluating the steadiness of word combinations in text we usually use *MI-score*, *t-score*, *Log-Likelihood*, *C-value*, χ^2 *criterion* and other.

Combined methods of terminology analysis suppose using both lexico-grammatical models, methods of terms word combination generation, the system of filters and statistic's instruments [Loukashevich et al., 2010].

The main problem of all methods is a filtration of generated terms word-combinations (a candidates for terms).

The other important task occurs in terminology searching is an evaluation of term relevance. In the information retrieval we use MI statistic measure (mutual information) and its modification (t-score measure).

The formula of *MI* measure is the following: $MI(ab) = \log \frac{N * freq(ab)}{freq(a) * freq(b)}$

where *freq ()* – the frequency of words and word-combinations, *N* – the number of words in collection. *MI* measure shows the difference between using the word in word- combination and the word's separate using.

The quality of extracted term word combinations might be realized on base of *AvP* measure (an average precision) [Ageev et al., 2004]. The precision *PrecTerm* of term word-combination extraction from a list of *n* word-combinations is defined as $PrecTerm = \frac{T}{n}$, where *T* is a number of terms in a list.

In the arranged list we can get the precision on the level of *n*-terms $PrecTerm(n)$, which is determine as the quantity of relevant terms among the first set of *n* in the distributed list divided by *n*.

In this case the average precision is counted according to the formula: $AvP = \frac{1}{k} \sum_i PrecTerm(i)$,

where *k* is the number of terms in the list of *n*-elements.

For example, if we have three elements in a list (*N=3*) and two terms (*k=2*) occur in the first and the third position of the list, we will get $AvP=(1/2)(1+2/3)=5/6$.

It should be made clear that when we figure out the *MI* measure the word order and their correlation (the relevance of syntactic structure of terms) doesn't take into account.

The syntactic structure of multiword term determines types of syntactic relations between components of the word-combination.

Let introduce the notion of precision of term extraction with the specified syntactic structure *R* - ($PrecTerm_R$), which defines as a quantity of relevant terms with the specified syntactic structure *R* taken from the list of an *n* word-combinations, in other words $PrecTerm_R = \frac{T_R}{n}$, where *T_R* is a number of terms with the specified *R* syntactic structure in the list.

It supposed that the relevant documents should contain word-combinations with the same syntactic structure as the structure of the word combinations in request. For example, if there is a *rank of an Abelian group* word-combination in the request, the documents which contain a *torsion-free Abelian groups of rank one* word-combination should be recognized as irrelevant, because the main word of the *rank of an Abelian group* word-combination is *rank* since in the *torsion-free Abelian groups of rank one* the main word *rank* has a dependent position (attributive relation).

Thus a new approach of the terminology searching improvement grounds on the idea, that the searching based on terminological annotation should rely on semantic-syntactic equivalence of the document models with the searching request, what is different to a traditional keyword searching.

Let us overview the syntactical relations of Russian language might be used for the term word combinations generation. The general syntactical model for term word combination is a nominal group that includes the main word (a noun) and the modifier (dependent word). The structure of nominal group determines by syntactical relations between the main and dependent words. The syntactical relations found on correlation between lexical meaning of the words and their grammatical forms.

There are five main types of syntactical relations of word-combinations in Russian language: an attributive, an objective, a subjective, an adverbial and a completive.

Attributive relations. We can talk about attribute relations when a noun (with the general lexical meaning of subject) correlates with the word of attributive meaning that might be coordinated and uncoordinated with the noun. A formal model of the attribute relations could be represented as:

1) $Atr \cap_N N$ – is a model with a coordinated attribute *Atr* (it's coordinated by the whole set of grammatical categories), for example, in Russian *конечная группа* (a finite group).

2) $N + Atr$ – is a model with an uncoordinated attribute (a parataxis), for example, in Russian *группа без кручения* (a torsion-free group), in Russian *группа с обычной арифметической операцией умножения* (a group with standard arithmetic multiplication operator).

Objective relations. We can talk about objective relations when a verb (a participle or an adverb) correlates with a noun or more rare infinitive. These word combinations are semantically bounded, because the main word has lexical meaning of action, sense, perception, since the dependent word means an object of this action, sense, perception (it's a direct object mainly). The noun of action takes an acting structure after the verb (in Russian *решить уравнение* – решение уравнения).

A formal model of the objective relations could be represented as:

$N_V + N_{p2}$, where N_V is a verbal noun, N_{p2} is a genitive case noun. Particular model of objective relations rely on a model of government of the verb.

Subjective relations characterize the word-combinations with a verb or participle in a passive voice. The dependent word in this case shows an actor (an instrumental case). For example, in Russian *предложенный автором* (метод). A formal model of the subjective relations could be represented as: $V^2 / A_V^2 + N_{p5}$, where V^2 / A_V^2 is a verb or participle in a passive voice and, N_{p5} is a noun in instrumental case.

Adverbial relations characterize a verb word-combination and rely on the lexical meaning of the process. The adverbial relations are specified as the adverbial relations of attribute, time, place, cause and purpose.

The example of an adverbial relation of place is in Russian группа параллельных переносов в линейном пространстве (a group of parallel transfers in the linear space), compare with строка в таблице (in Russian) (a row in the table). The latter is an attributive model.

A formal model of adverbial relations could be represented as: $N_v + PPNP$, where N_v is a verbal noun, $PPNP$ is prepositional phrase (more often in an adverbial meaning).

Completive relations appear in idiomatic word-combinations. A formal model of the completive relations might be represented as a list of the corresponding word combinations.

In general, a multi-word term might be generated as a superposition of the aforementioned relations. As an example, in Russian абелева группа без кручения первого ранга (*torsion-free Abelian groups of rank one*) word-combination is a superposition of the word-combinations which are ((абелева группа) без кручения первого ранга), (группа (параллельных [переносов в линейном пространстве])), where round brackets mark out the attributive relations and square brackets indicate the adverbial relations.

A semantic structure of multi-word terms word-combinations might be represented in a structure of the terminological annotation, in which we can distinguish a type of relation, a main word and a dependent word. The correlation of the elements are similar to the subordination relations of compound sentences.

The terminological annotation might be organized as an XML notation with the following characteristics:

- the type of relationship is represented by the value of a LINK attribute;
- the Holder attribute indicates the main element of a word-combination;
- the Dependent attribute defines the dependent element of a word-combination according to a syntactical relation.

The announcement of attributes in XML format are designated as:

- <!ATTLIST TERM Link (attributive | objective | subjective | adverbial | completive)
- <!ATTLIST TERM Holder CDATA >
- <!ATTLIST TERM Dependent CDATA >

Thus, the terminological annotation of the 'Abelian group' word-combination in XML format could be represented as: <TERM Link="attributive" Holder="группа" Dependent="абелева">

```

абелева группа
</TERM>

```

Design technology for applied problems solving in "OntoIntegrator" system

The "OntoIntegrator" system is an ontolinguistic research software development kit for the solution of applied problems, connected with an automatic text processing. The "OntoIntegrator" system contains the following functional subsystems, which are [Nevzorova, 2007]:

- an "Integrator" subsystem;
- an "OntoEditor+" subsystem of ontology modeling;

- a "Text analyzer" subsystem;
- a subsystem of dealing with external linguistic recourses;
- a subsystem of ontological models.

The "OntoEditor+" [Nevzorova, 2006] subsystem of ontology modeling provides the main table functions for dealing with an ontology (addition, modification, deletion, automatic correction; keeping of more than one or compound ontologies, in other words with the general lists of relations, classes, text equivalents and others; an import of the ontologies with the different formats of data; a filtration of ontology; keeping of statistics automatically, searching for chains of relations and others). The functions of the visualization unit support different graphic modes of system, including the graphic mode of the ontology modeling.

The "Text analyzer" subsystem includes linguistic tools are useful for solution the problem of the morphology analysis, of the ontological markup, of the polysemy resolution, of segmentation and the applied linguistics modeling. The subsystem of dealing with the external linguistic recourses supports keeping the basic linguistic recourses that contain a grammatical dictionary and a set of specialized linguistic data bases. The Integrator subsystem provides integrated base of the applied linguistic problem solution and control under the applied problem solution generation.

Let us overview the process of generation the solution of applied problems in the "OntoIntegrator" system by the example of the terminological annotation task. The process of the solution is realized under control of the ontological models system with a reflective kernel (the system is represented by the relational data bases). The system of the ontological models includes different types of ontologies; there are applied ontologies (domain ontology and the relations for inference in it), the ontology of models and the ontology of problems [Nevzorova et. al., 2011].

In terms of structure the system of ontological models represents a ternary associative system. The components of the system are the semantic networks (the ontological subsystems); there are the ontology of problems, the ontology of models and the applied ontology. The system allows us the interpreting of the applied ontology as a set of ontologies of different domains, which can be external, attached by a user, and internal, integrated into "Ontointegrator" system (with the possibility to spread-out, to edit and the support of calculation) with the purpose of the applied problems solution.

In order to build a solution for the linguistic problem it is necessary to decompose it to structural elements represented in problem ontology. Then, structural elements of solution for the linguistic problem are mapped to the set of model ontology structures. For easier interpretation all model-concepts are split to conditional groups supported by complex visualization mechanisms:

- Basic models providing the minimal functionality of system of ontological models;
- Syntactical models reflecting T-models (text models);
- Structure-semantic models creating adequate to applied problem solution structure and interpretation of its results;
- User models which are dynamically created by user.

Meaningfully, models are used for implementation of solution to problem-concept and allow to set (to interpret) solution components as a problem of setting the property, or as a problem of identification of the relation or the

problem with known evaluation algorithm. The solution for applied task is composed on the basis of text fragments (T-models) which are derived from the models from model ontology (S-models) by model identification procedure. Syntactic models, which are the basis for the extraction of terminological word-combinations, are based on the rules of syntactic analysis of NP. Basic syntactic models of NP structure are distinct by type of syntactic relations between the noun and modifiers. Also, arbitrary number of modifiers and combinations of any types of syntactic relations (attributive, objective, subjective and adverbial) are allowed.

Another developed method allows the extraction of NP from constructions with conjunctive reduction. The inverse problem of extraction of construction's potential components for their recognition as independent terminological entities is solved through terminological analysis of conjunctive syntactic constructions of certain types. For example, the components "*natural numbers addition operator*" and "*natural numbers multiplication operator*", which are recognized as independent terminological entities, are extracted from the syntactic construction "*natural numbers addition and multiplication operators*". The extraction of components from conjunctive constructions is made on the basis of special rules, which take the phenomenon of "semantic homogeneity" into consideration. Semantic homogeneity assumes the composition of syntactic construction with semantically homogeneous parts, in other words, all members of homogeneous constructions must belong to the same semantic class. During the phase of construction of rules, two main semantic classes - the concrete and abstract entities - are separated. The semantic homogeneity principle allows the composition of conjunctive construction for either concrete or abstract entities. For example, the constructions of type "*Eigen vectors and values of matrix*" (concrete homogeneity) or "*number addition and multiplication*" (abstract homogeneity) are acceptable. Likewise, the semantic homogeneity of classes of attributes is required for composition of conjunctive attributive set (conjunction by attributes). For example, the conjunction by attribute "number sign" ("*positive numbers and negative numbers*") is presented in the "*positive and negative numbers*" construction; but, attributes from different semantic classes are defined in the "*nondegenerate and symmetric matrices*" construction and, therefore, this construction is treated as attributive. The rules of extraction of components from conjunctive constructions are well-defined in [Nevzorova et. al., 2009].

Recognition of terminology in text on the basis of main syntactic NP and conjunctive reduction models is performed in "OntoIntegrator" system on the basis of applied ontological resource - the ontology of mathematical knowledge (theory of groups section). The basic list of mathematical terms (prop terms) from this section is used for experiments. Syntactic link type of the prop term and its syntactic role (main or dependent word) is determined for prop term containing terminological word-combination selected in text. The corresponding data is represented via XML notation.

In general case, the technology of problem preparation and solving in "OntoIntegrator" system consists of following phases:

1. Preparation of linguistic resources. Basically, this phase involves the supplement of system dictionaries with new terminology of given problem field. The selection mechanisms of candidates to terms are provided in the system.
2. System configuration. Besides the ability of creating new ontology system or connection one developed earlier, the possibility of combining the ontology of models and ontology of problems from different systems into applied ontology system is provided in the "OntoIntegrator" system. Adding new concepts to ontology of models and the redefinition of them in the reflexive kernel as corresponding relations in ontology of problems and applied

ontology is sufficient in order to accomplish this. Also, dynamic change of applied concepts (instances) classification on the basis of model-concepts (classes) is possible through connection of different classification, which was developed earlier.

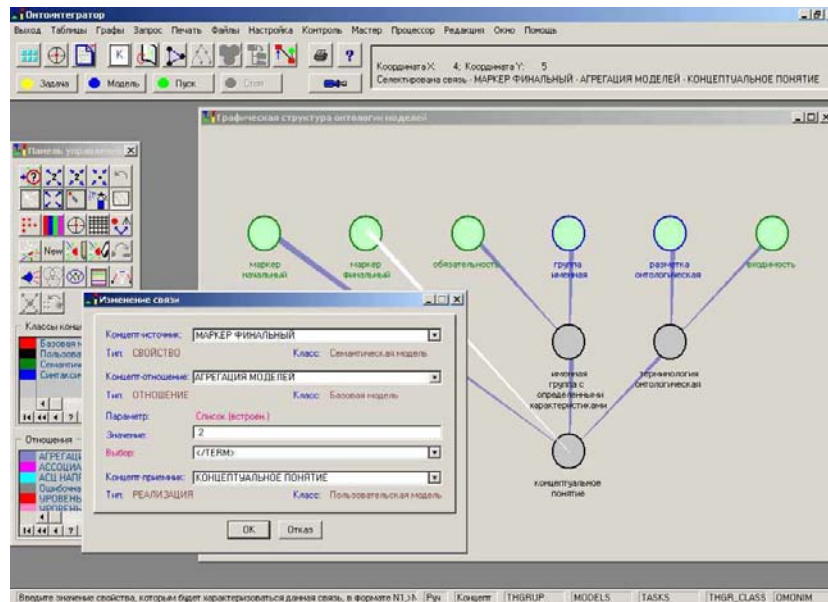


Fig.1. Construction of model-concepts of "Implementation" type

3. Construction of problem-concepts of "Implementation" type. In the context of this article, this means the construction of solution process of abstract problem "Document tagging on the basis of problem model" structure in the ontology of problems basis using inclusion relationship. Problem-concept of type "Implementation" is meaningfully a link (node) to semantic subnet, which describes the sequence (structure) of solution process.

4. Construction of model-concepts of "Implementation" type. Several such models could be used simultaneously during the problem solving, but they have to be combined in the semantic subnet with the node (link) as model-concept of type "Implementation" that would be treated as problem model, using model aggregation relationship. The problem model "Abstract concept", which is used for terminological tagging of document, is presented on fig.1. This concept has the "Final marker" property which is updated through parameter of link with </TERM> value, chosen from integrated ontology of text markers. Integration of chosen models is performed on the base of aggregation relationship.

5. Supplement of integrated ontologies. In problem under discussion, markers <TERM Link="link relationship" Holder="main element" Dependent="dependent element"> and </TERM> should be added to integrated ontology of text markers.

6. Preparation of applied ontology. During problem solving, it is sufficient to supplement applied ontology with basic list of terms (prop terms) without identifying the relationships between them and to configure its linguistic shell, which would provide the ontological tagging of document being processed.

7. Classification of applied concepts. For problem under discussion this phase is performed by default, because the model "Conceptual idea" does not contain model-concepts of type "Implementation" that identified in text through their instances (concepts of applied ontology).

8. Choice of problem-concept and model-concept. Text selection for processing and/or building-up the natural language query is viewed as part of the solution to problem of tagging.

9. Startup of solution process (problem-concept).

The result of terminological tagging of the document is presented in OMDOC format in the Internet-browser window on fig.2. For example, the tagging of terminological group *finite p-groups* includes indication of syntactic relationship type (attributive word-combination), selection of head (*p-group*) and the dependent word (*finite*). Prefix (“\$\$\$”) denotes the form (*p*).

```

<?xml version="1.0" ?>
<!-- This OMDoc document is generated from an sTeX-encoded one via LaTeXML, you may want to reconsider editing it. -->
<omdoc xmlns="http://omdoc.org/ns" xmlns:stex="http://kvarc.info/ns/sTeX" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:om="http://www.openmath.org/OpenMath" xmlns:m="http://www.w3.org/1998/Math/MathML" stex:srcref="at
solvable.tex; line 24 col 17">
<metadata>
<dc:creator>E. M. Колесова</dc:creator>
<dc:language>ru</dc:language>
<dc:title>
Вполне
<TERM Link="attributive" Holder="$$$-группы" Dependent="разложимые, абелевы, без, кручения">
разложимые абелевы
<om:OMOBJ>
<om:ONLY name="End" />
</om:OMOBJ>
- группы без кручения
</TERM>
</dc:title>
<dc:description />
</metadata>
<omgroup layout="sectioning" stex:srcref="at solvable.tex; line 34 col 27" xml:id="intro">
<metadata>
<dc:title stex:srcref="at solvable.tex; line 34 col 27" />
</metadata>
<context stex:srcref="at solvable.tex; line 35 col 39" type="introduction" xml:id="intro.p1">
<CMP stex:srcref="at solvable.tex; line 35 col 33" xml:id="intro.p1.p1">
<cp stex:srcref="at solvable.tex; line 35 col 33" xml:id="intro.p1.p1.p1">
В
<TERM Link="attributive" Holder="теории" Dependent="абелевых, групп">теории абелевых групп</TERM>
важной задачей является изучение связи между
<TERM Link="attributive" Holder="группой" Dependent="абелевой">абелевой группой</TERM>
и ее
<TERM Link="attributive" Holder="группой" Dependent="эндоморфизмов">группой эндоморфизмов</TERM>
. Возникает вопрос, при каких условиях
<TERM Link="attributive" Holder="группа" Dependent="абелева">абелева группа</TERM>

```

Fig.2. The result of terminological tagging in OMDOC format

Experiments

Experiments for terminological annotation of mathematical articles are done on the experimental collection of group theory articles and on the original terminological list of group theory concepts from the corresponding section of DBPedia.

Content processing of elements of mathematical documents includes text segmentation to sentences, text objects (extraction of formulas, number sequences, words, punctuation marks, abbreviations etc.) recognition, recognition of NP containing terms from applied ontology, recognition of complex syntactic constructions (conjunctive reduction groups) and other procedures (for example, homonyms extraction and classification).

In articles being considered, single- and multiword NP are selected on the basis of corresponding syntactic models (*finite group*, *Sylow 2nd subgroup of group*, *Klein group*, *dihedral group*, *nilpotent non-Abelian group* etc.). It is necessary to note that mathematical texts contain a large amount of words with prefix-formulas (*p-subgroup*, *2-subgroup*) and postfix-formulas (*group G*, *subgroup K*). Arbitrary formulas and expressions could be used as prefixes. These objects are not contained in system dictionary and are processed with special methods, which are separating left prefix-formulas and are working on the basis of right word-part syntactic model. Postfix-formula containing words are processed on the basis of NP with abbreviation syntactic model (*group G*, *subgroup K*).

Conclusion

New ideas of consideration of syntactic structure of terminological word-combination in search problem context and the "OntoIntegrator" system solution technology of problem of terminological annotating are considered in this article. Main technological phases of problem solution preparation are described on the basis of concrete applied problem.

Proposed technology allows the unification of process of solution composition of wide range of applied problems that are oriented on ontology usage and methods of automated text processing. All phases of applied problem solution composition are supported by convenient graphical interfaces and specialized graphical editors. Detailed description of conceptual and technological solutions made in "OntoIntegrator" system could be found in bibliographic links. At the present time, the developed technology of applied problems solution are probed in concrete applications linked to terminological and structural annotation of mathematical documents, basic linguistic problems, such as ontological tagging of text, context rule-based homonymy disambiguation etc.

Acknowledgements

The work has been completed with partial support of Russian Foundation of Basic Research (grant № 11-07-00507).

Bibliography

[Berners-Lee, 2011] T. Berners-Lee. Linked Data - Design Issues. At <http://www.w3.org/DesignIssues/LinkedData.html/>. (Accessed on January 18, 2011).

[Ageev et al., 2004] Ageev M., Kuraleonok I. Official metrics of ROMIP'2004. In Proceedings of the second Russian seminar on the evaluation of methods of information retrieval, Saint-Petersburg, pp. 142-150 (2004). In Russian.

[Loukashevich et al., 2010] Loukashevich N.V., Logachev A.M. Attribute combination for automated term extraction. Computational methods and programming (11), 108-116 (2010). Available at: <http://num-meth.srcc.msu.su/>. In Russian.

[Nevzorova, 2006] Nevzorova O. Instrumental system "OntoEditor+" for visual design of ontologies in linguistic applications. KSTU named after A.N.Tupolev newsletter, (3), 56-60, (2006). In Russian.

[Nevzorova et. al., 2011] Nevzorova O., Nevzorov V. Multi-layer ontological system for applied problems solution planning. In Proceedings of international conference "Open Semantic Technologies for Intelligent Systems" (OSTIS'2011), pp. 323-330, (2011). In Russian.

[Nevzorova, 2007] Nevzorova O. Ontolingvistic systems: Technologies of applied ontology interaction. Kazan State University scientific notes, Physic-mathematical sciences, 149 (2), pp.105-115 (2007). In Russian.

[Nevzorova et al., 2009] Nevzorova O., Nevzorov V. Ontological analysis of the domain: Automated methods of term extraction in "OntoIntegrator" system. In "Modelling methods" symposium, Kazan, pp.196-208 (2009). In Russian.

Authors' Information

Olga Nevzorova – Chebotarev Research Institute of Mathematics and Mechanics, Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Russia; e-mail: olga.nevzorova@ksu.ru

Vladimir Nevzorov – Kazan State Technical University, Kazan, Russia; e-mail: nevzorov@mi.ru

TOWARDS LINGUISTICS ANALYSIS OF THE BULGARIAN FOLKLORE DOMAIN

Galina Bogdanova, Konstantin Rangochev,
Desislava Paneva-Marinova, Nikolay Noev

Abstract: *This paper presents an investigation of the lexical structure of the Bulgarian folklore, made during the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage"¹ project. This is the first attempt for computational lexical analysis of the Bulgarian folklore and its constituents. Based on this research some linguistic components, aiming to realize different types of analysis of text folk objects are implemented in the Bulgarian folklore digital library. Thus, we lay the foundation of the linguistic analysis services in digital libraries aiding the research of kinds, number and frequency of the lexical units that constitute various folk objects.*

Keywords: *multimedia digital libraries, frequency and concordance dictionaries, systems issues, user issues, online information services, folklore rubrics.*

ACM Classification Keywords: *H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues*

Linguistics Research and Analysis of the Bulgarian Folklore

The research of the lexical structure of the Bulgarian folklore is very important task for different science domains such as folkloristic, ethnology, linguistics, computational linguistics, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what the lexical structure of Bulgarian folklore works is. During the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" project [Bogdanova et al., 2006] [Paneva-Marinova et al., 2010] [Todorov, 2007] we lay the foundation of the computational lexical analysis of the Bulgarian folklore and its constituents. Our attention was directed to these researches in order to enrich both the content and functionality of the developed multimedia digital library of Bulgarian folklore² (also called Bulgarian Folklore Digital Library or BFDL, <http://folkknow.cc.bas.bg/>)

¹ The "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" is a national research project of the Institute of Mathematics and Informatics, supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No IO-03/2006. Its main goal is to build a multimedia digital library with a set of various objects/collections (homogeneous and heterogeneous), selected from the fund of the Institute for Folklore of the Bulgarian Academy of Science. This research aims to correspond to the European and world requirements for such activities, and to be consistent with the specifics of the presented artefacts [Bogdanova et al., 2008][Berger et al., 2008].

² The Bulgarian folklore digital library is built during the "Development of Digital Libraries and Information Portal with Virtual Exposition 'Bulgarian Folklore Heritage'" module of the national research project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage". This

[Pavlov et al., 2010] [Paneva-Marinova et al., 2010] [Rangochev et al., 2007]. Thus we aim to expand the target group of potential users of the library, covering not only those who are interested in Bulgarian folk music, but also narrow specialists in different fields of humanities (folklore, ethnology, linguistics, text linguistics, structural linguistics, etc.). The Bulgarian folklore digital library has a flexible structure that involves additional linguistic components in order to provide real observation and analysis of text folk objects. Digital library with similar analyzing services are presented at [Pavlov et al., 2007] [Pavlova-Draganova et al., 2007] [Pavlov et al., 2006].

As a basis of our research we took the analysis of folklore lexical structure so called main component of the linguistic research of the Bulgarian folklore. We try to answer to the questions: How many and what token it contains? Is there and what is the domination or the lack of some groups of tokens, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what the real lexical structure of Bulgarian folklore works is. With a few exception (for Bulgarian heroic epic [Rangochev, 1994] and for "Veda Slovena", See <http://www.bultreebank.org/veda/index.html>) lexical analysis for the Bulgarian folklore and its constituents is missing, the regional characteristics of the folklore lexical structure is unknown. Unfortunately, in 2011 the Bulgarian linguistics, folklore, ethnology, etc. cannot answer the question what are the lexical components of Bulgarian folklore (number, frequency, word forms, etc.) and so far, this type of research is carried out systematically and with a purpose.

This paper presents the basic components of the linguistics research – the different types of dictionaries, frequency dictionaries, concordance dictionaries, terminological dictionaries, valence dictionaries, etc. The paper also describes the BDFL linguistics components for frequency analysis that manipulate the sets of folklore objects of text media type. Finally, the project of a dictionary – concordances of songs, prose, interviews, etc. is outlined.

Classification of Basic Components of the Linguistics Research

The basic components of the linguistics research are the dictionaries. According to accepted definitions every dictionary is a list of words and their meanings in alphabetical order. It is also an alphabetically arranged publication containing information about words, meanings, derivations, spelling, pronunciation, syllabication and usage. (See <http://www.web-ezy.com/cit/main/webzglos.htm>). The dictionary could give information for pronunciation, grammar, derivatives, history and etymology of the basic word, as well as recommendations for usage, examples, phraseological expressions, examples. Dictionaries are usually in the form of books, but recently electronic dictionaries are more and more recognized.

Qualification of dictionaries is based on different criteria. Many qualifications exist in different lexicographic and lexicological papers [Hartmann, 1993][Svensen, 1993]. Usually dictionaries are combined, which makes them more effective, but this makes their differentiation in categories more difficult.

- By form
 - Traditional dictionaries – they are made with the help of a computer, but their end form is on a paper.

Internet-based environment is a place where folklore objects (mainly from the Funds of the Institute for Folklore at the Bulgarian Academy of Sciences) of different kinds and origins were documented, classified, and „exhibited“ in order to be widely accessible to both professional researchers and the wide audience.

-
-
- Digital dictionaries – online (web-based) or local (desktop) dictionaries.
 - By their purpose
 - Descriptive dictionaries – descriptive dictionary for the meaning of words according to a common convention;
 - Grammar dictionaries – includes definitions and grammar rules;
 - Dictionary of synonyms – unilingual, includes words with similar meanings;
 - Valence dictionary – dictionaries for the variations of one language, for example – British, Bulgarian, American;
 - Dictionaries of etymology – they trace the development of a language's words in time, giving historical examples, to show the origin and the changes afterwards;
 - Phrase logical dictionaries – dictionaries that present phraseological units of one language. They contain: the most used phraseologies in colloquial speech besides literary units, jargon units, folklore units, vulgarisms and civisms, which are typical for the speech of young people;
 - Frequency dictionaries – gives information on how frequently a word, phrase is used in a particular corpus of texts;
 - Translation dictionaries – bilingual dictionaries, which are used for translation from one language to another;
 - Concordance dictionaries – dictionaries that shows the lexeme with/ in her context.
 - Specialized dictionaries – contain words (terms), that are used by a particular group of people in a professional environment;
 - Terminological dictionaries – contain the most frequently used words with detailed description for each of them;
 - By the number and type of languages
 - Unilingual (mono-lingual) dictionaries – unilingual is the dictionary, in which words are described in the same language;
 - Bilingual – dictionaries which contain translation of words in two languages;
 - Multilanguage dictionaries – dictionaries, which contain translations of words in more than two languages.

Frequency Dictionaries and Concordance Dictionaries for Bulgarian Folklore

For the folklore domain more suitable dictionaries are the frequency dictionary and the concordance dictionary [Rangochev et al., 2010]. The frequency dictionary presents the frequency of the lexemes in a definite corpus of texts. It is considered that the facts in one frequency dictionary are reliable enough if there are minimum 20 000 lexical units in it. The frequency dictionaries gave versatile information: presence/ absence of definite lexemes or group of lexemes in comparison with a standard frequency dictionary of the Bulgarian speech [Radovanova, 1968]; frequency of verbs (the so called "verb temperature" [Gerganov et al., 1978] (for the Bulgarian speech at least 21 % verbs in the examined corpus of texts); investigating of the paradigmatic relations in the vocabulary of the text corpus (river- stream- brook- rill...). The domination of group lexemes and respectively small number or absence of other group reveals the constituent characteristics of the text type and its originators [Rangochev, 1994].

- A general frequency dictionary – it contains the all lexical units which are in the BFDL (songs, proverb and descriptions of the rites...);
- A regional frequency dictionary – it contains all the text units which come of a definite folklore region or of a concrete settlement (if there are enough texts). Practically, this is a dialect dictionary of the region/ settlement as far as the folklore regions coincides with the dialect areas.
- A functional frequency dictionary – it contains all the text units which have identical functions: descriptions of the rites, various types of songs, narratives etc. This kind of dictionary would describe some genre specifics of the different parts of the Bulgarian folklore;
- Another dictionary – by user's wish.

The advantage of creating of frequency dictionaries is the possibility to make comparisons between the different types of texts and it can be also followed the tendencies in the dynamics of the lexis – presence/absence of various group of lexemes, etc.

The following table 1 illustrates the comparison of the Bulgarian folklore and spoken languages based on data available in frequency dictionaries.

Concordance dictionaries are these which show the lexeme with/ in her context – it is present the previous one (or more than one) lexeme and the following lexeme according to the examined lexeme. Example: "Fifty heroes are drinking wine" – the underlined lexeme is the examined and the lexemes in italic are her context. Of course, about the songs this could be concordance dictionary of their verses, about the narrative texts (descriptions of the rituals, etc.) – sentences in which they are contained (from point to point...). The creating and using of concordance dictionaries of the texts from BFDL would give good possibilities for folklorists and ethnologists to solve a series of problematic areas as presence/ absence of formulas in the folklore songs and epics, the structure of the folklore text, etc.

Rank list			
Bulgarian spoken language ³		Bulgarian heroic епос ⁴	
1. <u>съм</u> – 4 041	14. <u>си</u> – 1065	1. <u>съм</u> – 1342	14. <u>го</u> – 338
2. <u>и</u> – 3764	15. <u>казвам</u> – 1 045	2. <u>да</u> – 1 247	15. <u>му</u> – 320
3. <u>да</u> – 3 148	16. <u>тя</u> – 1044	3. <u>си</u> – 548	16. <u>че</u> – 318
4. <u>аз</u> – 2 433	17. <u>викам</u> – 1 031	4. <u>Марко</u> – 1 036	17. <u>а</u> – 286
5. <u>той</u> – 2 288	18. <u>те</u> – 1014	5. <u>се</u> – 828	18. <u>кон</u> – 276
6. <u>не</u> – 1 956	19. <u>какъв</u> – 938	6. <u>на</u> – 801	19. <u>от</u> – 272
7. <u>се</u> – 1 928	20. <u>за</u> – 913	7. <u>и</u> – 796	20. <u>ми</u> – 233
8. <u>този</u> – 1 701	21. <u>че</u> – 874	8. <u>па</u> – 657	21. <u>ти</u> – 225
9. <u>на</u> – 1 669	22. <u>с</u> – 809	9. <u>у</u> – 582	22. <u>що</u> – 222
10. <u>ти</u> – 1 249	23. <u>имам</u> – 768	10. <u>я</u> – 553	23. <u>по</u> – 218
11. <u>ще</u> – 1 183	24. <u>така</u> – 742	11. <u>та</u> – 526	24. <u>добър</u> – 201
12. <u>един</u> – 1 131	25. <u>от</u> – 731	12. <u>не</u> – 412	25. <u>три</u> – 201
13. <u>в</u> – 1 099		13. <u>юнак</u> – 396	

Table 1: Comparison of the Bulgarian folklore and spoken languages

³ The frequency dictionary is made of texts of the Bulgarian spoken language and the corpus contains 100000 lexemes [Nikolova, 1987].

⁴ The frequency dictionary is made of 100 song from [Romanska, 1971] and the texts of the songs contains 7871 verses while there are in it 40042 lexemes.

A Conceptual Framework of a Linguistics Components in the Bulgarian Folklore Digital Library

In the process of the primary testing of BFDL come into being the necessity of insurance of resources for linguistic analysis of the folklore knowledge [Rangochev et al., 2010]. For this aim it was projected and worked out a frequency dictionary with the following functional specification:

- Linguistic analysis of the available multitude of folklore objects of text media type in BFDB;
- Determination of the frequency of meeting the lexemes in text folklore objects;
- Creating of lists of the lexemes,
 - in frequency order
 - in alphabetical order.
- Taking the number of the lexical units;
- Taking the number of the repeats of the lexical units.

Figure 1 depicts the sequence of actions that has to be executed in order to be generated a frequency dictionary. Standard step is the passing through BFDL search service and its sub-functions: 1) user searches by some criteria; 1.1) service performs search in metadata repository, 1.1.1) service gets media data for the found objects, 1.1.1.1) service returns all found media objects by the search criteria, and 1.1.1.1.1) result sent to user. When the result set is generated the user could choose to generate a functional dictionary (step 2). Dictionary generation is performed and the result could be shown by frequency or alphabetically.

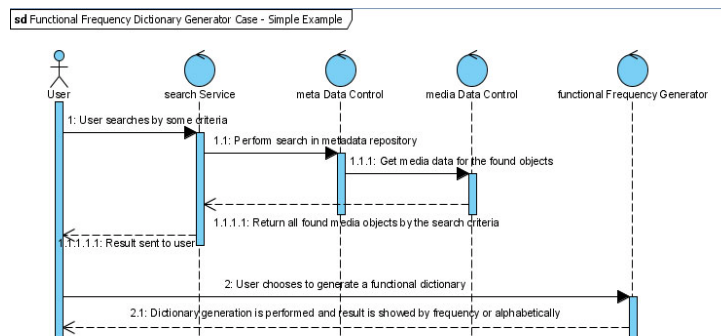


Figure 1: Sequence Diagram

Figure 2 depicts analysis class diagram for the BFDL linguistic component.

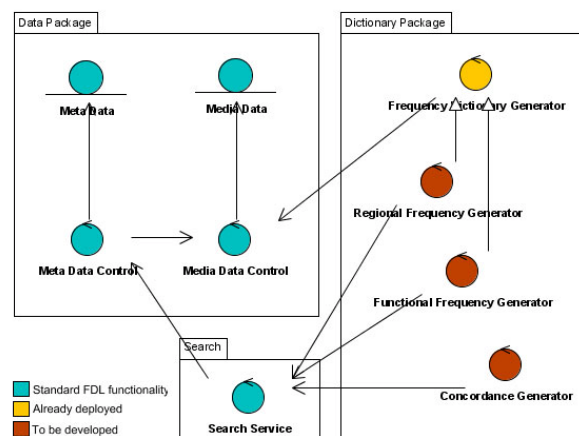


Figure 2: Analysis class diagram

The diagram shows the relations between the data package, the dictionary package and the search service. In the dictionary package there are clearly illustrated different types of generators for frequency dictionary, regional frequency dictionary, functional frequency dictionary and dictionary-concordance [Rangochev et al., 2010].

The Frequency Dictionary Project

The main objective of this project is to build frequency dictionary for texts with folklore themes. The dictionary provides information on how often a particular word or phrase is used in a particular corpus of texts. For the project aims a special hierarchical dataset and WEB interface have been created. The system allows full text search of big corpuses of texts. The dictionary uses rules and concepts in the field of Bulgarian folklore that filter the words/phrases (figure 4). The words/phrases are representatives of 20 different folklore rubrics (thematic headings).

The chosen folklore rubrics are: 1) Village information; 2) Rituals and feasts; 3) Songs; 4) Instrumental music (descriptions); 5) Dance folklore (descriptions); 6) Children folklore; 7) Prose; 8) Proverb, saying; 9) National beliefs and knowledge; 10) National medicine; 11) Magic; 12) Fortune-telling; 13) Dreams; 14) Clothing and adornment; 15) Belongings; 16) National art; 17) Architecture, monuments; 18) Food and feeding; 19) Festivals, gatherings and reviews; 20) Others.

The dictionary serves two types of users: administrator and ordinary user. The administrative area is composed of sections of the main operations of the data modeling. The section for adding of texts allows addition of text and a place for uploading source file. The system has an option for subscription of information from a file. It can upload it to a server as a useful source of reference that can be use by other applications. User area allows the user to search for a word in different sections, the system returns a complete answer on how many times and where the word contains.

The **administrative part** contains the following sections for the main operations on data modifying:

- Adding (presented on figure 3):
 - of a text: here the application has a text field, that enables addition of text and a field that enables upload of the source file.
 - of a rubric: the application is simplified to the limit and the administrator chooses the level on which he wants to add a rubric and gives only its name.

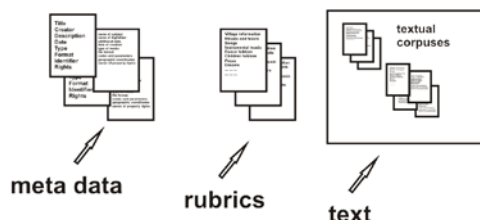


Figure 3: Adding data

Change of a rubric: The application gives an option for a change of the name of the rubric and the unique key is the same. The administrator has an option to choose the rubric, which he wants to change. After the choice is

made the text is saved in a field that can be modified. The query the data base is simplified to the highest degree. All needed parameters are given by drop down menus and all rubrics are a part of that menu, which contains their respective identification numbers. There are two types of the query:

```
"UPDATE rubrics SET rubric_name="" + TextBox1.Text + "" WHERE id = " +
Convert.ToInt32(DropDownList1.SelectedValue);
```

- Deletion: After an object is chosen to be deleted at the chosen level, the system deletes cascade all lower levels. The following source code is the query:

```
int sid = Convert.ToInt32(RadioButtonList1.SelectedValue);
if (sid == 1) {SqlDataSource1.DeleteCommand = "DELETE FROM rubrics WHERE id = " +
Convert.ToInt32(DropDownList1.SelectedValue);    SqlDataSource1.Delete();}
```

User part is composed of the search form that allows for selecting a desired item, the level and the corresponding text. The results appeared on the screen in which information rubric, how many files and how the words are distributed.



Figure 4: Full-text search

An example of search query is:

```
try {
string query = "SELECT info_text, path FROM rubric_info WHERE Contains(info_text,@text) AND rid =
@rid AND table_id = @tblId";
SqlCommand cmd = new SqlCommand(query, cn);
cmd.Parameters.AddWithValue("@text", TextBox1.Text);
cmd.Parameters.AddWithValue("@tblId", tblId);
cmd.Parameters.AddWithValue("@rid", rid);
cn.Open();
```

```

SqlDataReader dr = cmd.ExecuteReader();
int count = 0;
int fileNumber = 1;
while (dr.Read()) {
string text = dr["info_text"].ToString();
string path = dr["path"].ToString();
count += CountWords(text, TextBox1.Text, i, fileNumber, path);
fileNumber++;
}
fileNumber -= 1;
Label3.Visible = true;
lbAllFiles.Text = "Number of selected files by this rubric is: " + fileNumber + ".";
lbNumWords.Text = "Number of words is: " + count + ".";
dr.Close();
dr.Dispose();
}
finally {
cn.Close();
}

```

Because of the nature of the task, the usage of the following additional function is needed. It counts words in the respective texts.

```

private int CountWords(string text, string word, int i, int fNum, string fileName) {
char[] delims = new char[] { ' ', ':', ';', ',', '!', '@', '#', '$', '%', '&', '\t', '\n', '\0' };
foreach(string s in text.Split(delims)){
if (s.ToLower() == word.ToLower()) i++;
}
return i;
}

```

In result there is shown information how many files there are in every rubric, how words are divided, etc.

MsSQL, Visual Studio, HTML, CSS, JavaScript are used for the creation of the dictionary. A hierarchical structure of data (tree) is used for organization of data. The hierarchical structure of data has tables included for administration of rubrics (categories) and growing of the tree structure is allowed in volume and depth.

The system offers uses an easy and fast search system, due to the hierarchy of the data. It enables introduction of many different rubrics and nevertheless they don't influence the speed of searching. The individual tables contain only the names of rubrics, as well as their keys for organization the hierarchy. The help table contains all texts of all rubrics, organized with the help of indexes, which enables a fast access to the relevant texts and rubrics. There is an option for construction of a dynamic growing of the tree of tables in depth.

Acknowledgements

This work is supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant №IO-03-02/2006 "Development, Annotation and Protection of a Digital Archive "Bulgarian Folklore Heritage"" and №IO-03-03/2006 "Development of Digital Libraries and Information Portal with Virtual Exposition "Bulgarian Folklore Heritage"" from the project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage".

Bibliography

[Berger et al., 2008] Berger, T., Todorov, T., Improving the Watermarking Process with Usage of Block Error-Correcting Codes, *Serdica Journal of Computing*, 2008, Vol. 2, pp. 163-180.

[Bogdanova et al., 2006] Bogdanova, G., Pavlov, R., Todorov, G., Mateeva, V., Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage, *Advances in Bulgarian Science Knowledge*, National Center for Information and Documentation, 2006, Vol. 3, pp. 7-15.

[Bogdanova et al., 2008] Bogdanova, G., Todorov, T., Georgieva, Ts., New approaches for development, analyzing and security of multimedia archive of folklore objects, *Computer Science Journal of Moldova*, 2008, Vol. 16, 2(47), pp.183-208.

[Gerganov et al., 1978] Gerganov, E., Mateeva, A., Experimental Research of the Frequency of the Bulgarian Language, In the Proceedings of the national conference "Contemporary problems of the native language education", Sofia, Bulgaria, 1978.

[Hartmann, 1993] Hartmann, R.R.K., *Lexicography. Principles and Practice*, Applied Language Studies) London/New York: Academic Press, 1993.

[Nikolova, 1987] Nikolova, C., *A frequency dictionary of the Bulgarian spoken language*, Sofia, Bulgaria, 1987.

[Paneva-Marinova et al., 2010] Paneva-Marinova, D., Pavlov, R., Rangochev, K., Digital Library for Bulgarian Traditional Culture and Folklore, In the Proceedings of the 3-rd International Conference dedicated on Digital Heritage (EuroMed 2010), Lymassol, Cyprus, 2010, Published by ARCHAEOLOGIA, pp. 167-172.

[Pavlov et al., 2006] Pavlov R., Pavlova-Draganova, L., Draganov, L., Paneva, D., e-Presentation of East-Christian Icon Art, In the Proceedings of the Open Workshop "Semantic Web and Knowledge Technologies Applications", Varna, Bulgaria, 2006, pp. 42-48.

[Pavlov et al., 2007] Pavlov R., Paneva, D., Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, *Cybernetics and Information Technologies*, 2007, Vol. 6, № 3, pp. 51-62.

[Pavlov et al., 2010] Pavlov, R., Paneva-Marinova, D., Rangochev, K., Goynov, M., Luchev, D., Towards Online Accessibility of Valuable Phenomena of the Bulgarian Folklore Heritage, In the Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech'10), Sofia, Bulgaria, 2010, ACM ICPS Vol. 471, pp. 329-334.

[Pavlova-Draganova et al., 2007] Pavlova-Draganova L., Georgiev, V., Draganov, L., Virtual Encyclopaedia of Bulgarian Iconography, *Information Technologies and Knowledge*, 2007, Vol.1, №3, pp. 267-271.

[Radovanova, 1968] Radovanova, V., „Representative frequency dictionary of text with length 500 000 tokens“, Master thesis, University of Sofia 'St. Kl. Ohridski', Sofia, Bulgaria, 1968.

[Rangochev et al., 2007] Rangochev K., Paneva, D., Luchev, D., Bulgarian Folklore Digital Library, In the Proceedings of the International Conference on Mathematical and Computational Linguistics „30 years Department of Mathematical Linguistics”, Sofia, Bulgaria, 2007, pp. 119-124.

[Rangochev et al., 2010] Rangochev, K., Goynov, M., Paneva-Marinnova, D., Luchev, D., Linguistics Research and Analysis of the Bulgarian Folklore, Experimental Implementation of Linguistic Components in Bulgarian Folklore Digital Library, In the Proceedings of the International Conference „Classification, Forecasting, Data Mining” (CFMD 2010), Varna, Bulgaria, 2010, pp. 131-137.

[Rangochev, 1994] Rangochev, K., “Structural particularities of the epic text (using material of the Bulgarian heroic epos)”, PhD Thesis, Sofia University “St. Kliment Ohridski, Sofia, Bulgaria, 1994.

[Romanska, 1971] Romanska Cv. (Ed.), "Bulgarian heroic epos", Col. 53, Sofia, Bulgaria, 1971.

[Svensen, 1993] Svensen, Bo, Practical Lexicography: Principles and Methods of Dictionary-Making. Oxford: Oxford University Press, 1993.

[Todorov, 2007] Todorov, T., Performance of an error correction scheme for image watermarking, Proc. of the International Workshop on Optimal codes and Related Topics, White Lagoon, Bulgaria, 2007, pp. 233-236.

Authors' Information



Galina Bogdanova – PhD in Informatics, Associated Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: galina@math.bas.bg

Major Fields of Scientific Research: Information Society Technologies, Multimedia Digital Archives, Data Mining, Information Society Technologies, Steganographia, Coding Theory, Computer Science, Algorithms, Knowledge Technologies and Applications.



Konstantin Rangochev – PhD in Philology, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: krangochev@yahoo.com

Major Fields of Scientific Research: Ethnology, Folklore studies, Culture Anthropology, Linguistics, Computational Linguistics, Digital Libraries.



Desislava Paneva-Marinova – PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: dessi@cc.bas.bg

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.



Nikolay Noev – PhD student in Informatics, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: nickey.noev@gmail.com

Major Fields of Scientific Research: Multimedia Digital Archives, Internet Technologies, Computer Science, Knowledge Technologies and Applications.

ENVIRONMENTAL RISK ASSESSMENT USING GEOSPATIAL DATA AND INTELLIGENT METHODS

Nataliia Kussul, Sergii Skakun, Oleksii Kravchenko

Abstract: *In this paper, we describe intelligent methods and technologies for environmental risks assessment using geospatial data. The risk assessment process is based on fusion of data acquired from different sources: models, in-situ observations and remote sensing instruments. The ensemble approach is used for data processing. Several real-world applications are described to demonstrate efficiency of the proposed approach, namely numeral weather prediction (NWP), land biodiversity assessment, vegetation state assessment, fire monitoring and flood mapping. These applications are being implemented within international projects within the UN-SPIDER Regional Support Office (RSO) in Ukraine.*

Keywords: *intelligent methods, risk assessment, remote sensing from space, satellite data processing, environmental monitoring, vegetation state assessment, fire monitoring, UN-SPIDER.*

ACM Classification Keywords: *D.2.12 [Software Engineering] Interoperability; Information Systems; H.1.1 [Models and Principles] Systems and Information Theory; H.3.5 [Information Storage and Retrieval] Online Information Services; I.4.8 [Image Processing and Computer Vision] Scene Analysis - Sensor Fusion.*

Introduction

At present, global climate changes on the Earth made rational land use, environmental monitoring, and prediction of natural and technological disasters the tasks of great importance. The basis for the solution of these crucial problems lies in integrated use of multisource data of different nature, in particular modelling data, in-situ measurements and observations, and indirect observations such as airborne and spaceborne remote sensing data [GEOSS, 2005].

In particular, models can be used to fill in gaps in data by extrapolating and estimating necessary parameters to the site of interest, to better understand and predict different processes occurring in the atmosphere, land, ocean and sea. The models can also help to interpret measurements and to design new observing systems. In-situ measurements are often used for calibration and validation of modelling and remote sensing data, and usually assimilated into models. Satellite observations have an advantage of acquiring data for large and hard-to-reach territories, as well as providing continuous and human-independent measurements. Many important applications such as environmental monitoring, agriculture monitoring, monitoring and predictions of natural disasters heavily rely on the use of Earth observation (EO) data from space. For example, both spaceborne microwave and optical data can provide means to detect drought conditions, estimate drought extent and assess the damage caused by the drought events [Kogan et al, 2004; Wagner et al, 2007]. To assess vegetation health/stress, which is extremely important for agriculture applications, optical remote sensing data can be used to derive biophysical and biochemical variables such as pigment concentration, leaf structure, water content at leaf level and leaf area index (LAI), fraction of photosynthetically active radiation absorbed by vegetation (FPAR) at canopy level [Liang,

2004]. The satellite-derived flood extent [Kussul et al, 2011] is very important for calibration and validation of hydraulic models to reconstruct what happened during the flood and determine what caused the water to go where it did [Horritt, 2006]. Information on flood extent provided in the near real-time (NRT) can also be used for damage assessment and risk management, and can benefit to rescuers during flooding.

The EO domain is characterized by the large volumes of data that should be processed, catalogued, and archived [Shelestov et al, 2006]. The processing of satellite data is carried out not by the single application with a monolithic code, but by the distributed applications. This process can be viewed as a complex workflow that is composed of many tasks: geometric and radiometric calibration, filtration, reprojection, composites construction, classification, products development, post-processing, visualization, etc. Dealing with EO data, we have to also consider the security issues regarding satellite data policy, the need for processing in NRT for fast response within international programs and initiatives, in particular the International Charter "Space and Major Disasters" and the International Federation of Red Cross. It should be also noted that the same EO data sets and derived products can be used for a number of applications. For example, information on land use/change, soil properties, and meteorological conditions is important for droughts identification, vegetation state assessment and floods. Therefore, once we develop interfaces to discover and access the required data and products, they can be used in a uniform way for different purposes and applications. This represents one of the important tasks that are being solved within the development of the Global Earth Observation System of Systems [GEOSS, 2005] and European initiative Global Monitoring for Environment and Security [GMES, 2008]. Services and models that are common for different EO applications (e.g. flood monitoring and crop yield prediction) are shown in Figure 1.

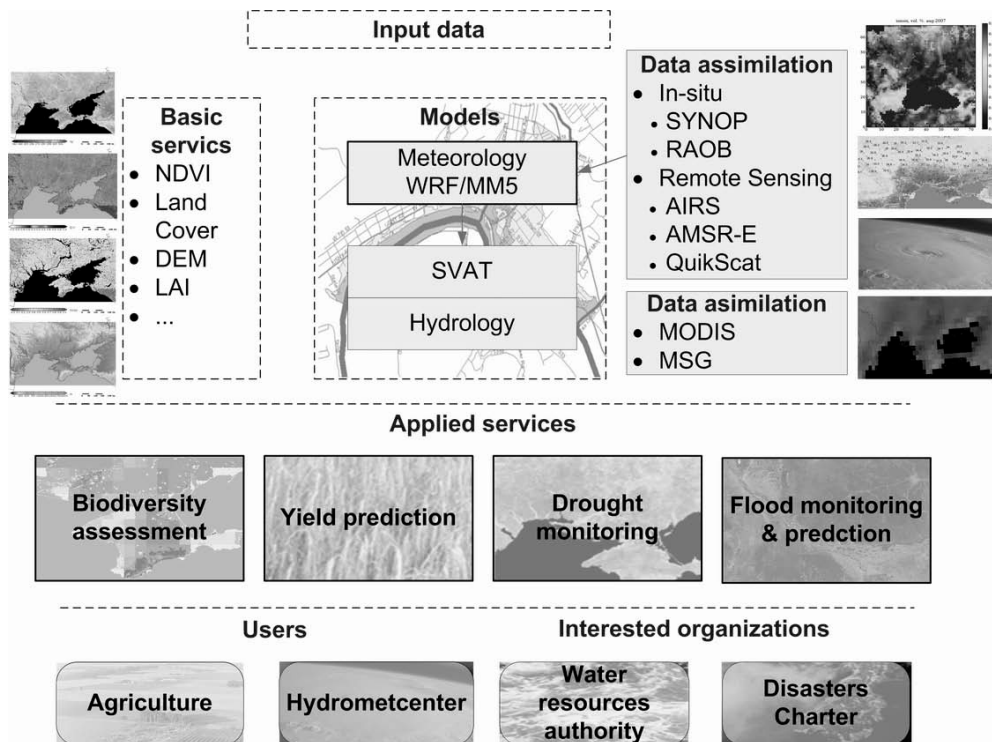


Figure 1. Common services and models for different applications

A considerable need therefore exists for intelligent methods and appropriate technologies that will enable the integrated and operational use of multi-source heterogeneous data for different application domains, and in particular environmental risk assessment.

In this paper, we describe intelligent methods and technologies for environmental risks assessment using geospatial data. The risk assessment process involves a fusion of data acquired from different sources: models, in-situ observations and remote sensing instruments. Several real-world applications are described to demonstrate efficiency of the proposed approach, namely *numeral weather prediction (NWP)*, *land biodiversity assessment*, *vegetation state assessment*, *fire monitoring* and *flood mapping*. Most of these applications are being implemented within international projects within the UN-SPIDER Regional Support Office (RSO) in Ukraine (<http://un-spider.ikd.kiev.ua>).

Environmental Risk Assessment using Geospatial Information

Usually, risk represented as a combination of the likelihood of an occurrence of a hazardous event or exposure(s) and the severity of injury or ill health that can be caused by the event or exposure(s) [OHSAS, 2007]. Mathematically, risk R often simply defined as a function f of disaster probability and expected loss (hazards): $R = f(\text{probability}, \text{loss})$.

Event probability could be estimated using a neural network (forecast) model [Haykin, 1999] based on data acquired from remote and in-situ observations (data fusion approach) [Kussul et al, 2009]. To identify the neural forecast model we use risk functional minimization theory developed within a theoretical framework known as computational learning theory [Bishop, 2006] or statistical learning theory [Vapnik, 1998]. Within this approach there are three types of empirical risk minimization problems: classification problem, regression retrieval problem and problem of indirect experiments interpretation. For each of the problems a specific loss function is determined.

To estimate event probability density function information from different sources is integrated (Figure 2).

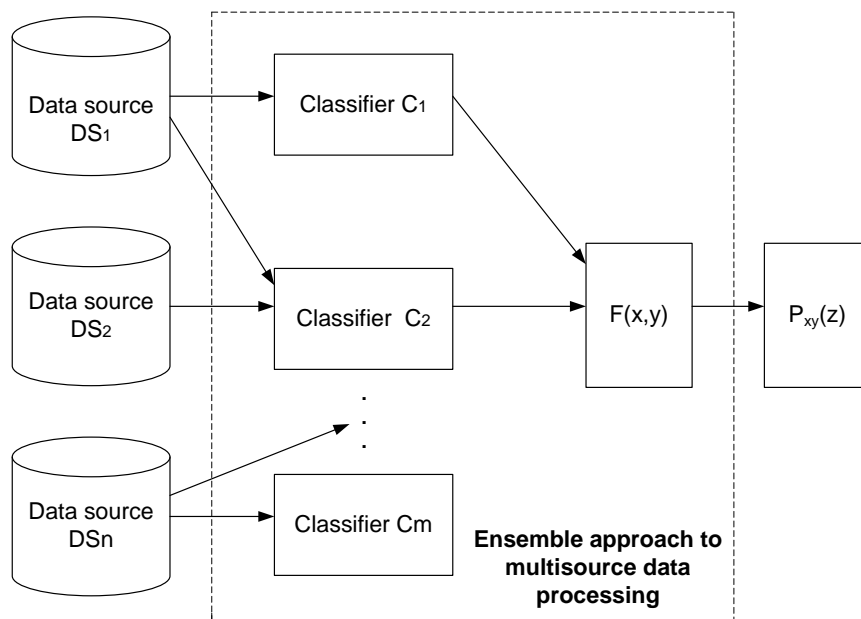


Figure 2. Event probability density estimation from multisource data using ensemble approach

Each classifier (can also be referred as an *expert*) provides an opinion on the event using corresponding data source (geospatial information, point observations). Their outputs are combined through a generalized rule F . Such a framework is known as a *mixture-of-experts model* [Jacobs et al, 1991]. In the following sections we describe how multisource data are combined using this approach for applied problems solving in different domains.

Applications

Numerical Weather Modelling (NWP). Prediction of meteorological parameters represents one of the core services for a number of applications (e.g. floods, droughts, agriculture, etc). Currently, we run the Weather Research and Forecasting model (WRF) [Michalakes et al, 2004] in operational mode for the territory of Ukraine. The meteorological forecasts are generated every 6 hours with a spatial resolution of 10 km. Forecast range is for 72 hours in advance. The horizontal grid dimension is 200 by 200 cells with 31 vertical levels. We use forecasts from the Global Forecasting System (NCEP GFS) for boundary conditions. This data is available via Internet through the National Operational Model Archive & Distribution System (NOMADS).

The workflow of the model run is composed of the following steps: data acquisition; data pre-processing, computation of forecasts using WRF model and data post-processing; visualization of the predicted parameters.

To run WRF model, it is necessary to obtain boundary and initial conditions for the territory of Ukraine. This data can be extracted from the GFS model forecasts. To get the required data, the dedicated script was developed. This script downloads global forecasts every 6 hours. To decrease the data volume, our script uses a special Web-service capable of selecting subsets of the GFS data for the territory of Ukraine. The acquired data is transferred to the storage subsystem and marked as unprocessed (i.e. it has to be processed by the WRF model). After the GFS data has been downloaded, the Karajan script initialises a workflow for data pre-processing, WRF run, and data post-processing.

Data pre-processing step is intended to transform the downloaded data into the format that is used to run the WRF model. GFS data is delivered in the GRIB format in geographical projection. This data is transformed into the internal WRF format by the `grib_prep.exe` command, warped into the Lambert Conformal Conic projection (by executing `hinterp.exe` command) and vertically interpolated using the `vinterp.exe` command. These utilities (`grib_prep.exe`, `hinterp.exe` and `vinterp.exe`) are tools from the WRF Standard Initialization (SI) package. The results of these transformations are stored in the netCDF format. After that, the `real.exe` command is used to produce initial and boundary conditions for WRF model run. The inputs to `real.exe` command are GFS data in netCDF format and WRF configuration file (`namelist.input`).

Data processing step consists in running WRF model using `wrf.exe` command. The outputs of the command are forecasts of the meteorological parameters. This is the most computationally intensive task. After WRF model run, post-processing step is carried out. For specified weather parameters and for each forecast frame (3 hours), a graphic representation (in PNG format) of spatial distribution is created. Additionally, special files containing georeferencing information are created (files with `*.wld` extension). The results of the post-processing phase are used to visualize the WRF forecasts via the mapping service. This service provides to the users animations of the weather forecasts (Figure 3). The service provides tools to select a forecast time, forecast frames (up to 72 hours in advance), and weather parameters to display. Selected by the user information is packed into the

request to the server. To process the request, all required data (in PNG and WLD formats) is retrieved from the storage subsystem and passed to the mapping server in order to create the maps. Maps are further processed by the script to generate weather animation in GIF format. Finally this animation is presented at user side.

We have also tested the performance of the WRF model in dependence of the number of computational nodes of the supercomputer SCIT-3. For test purposes, we used the WRF model version 2.2 with a model domain identical to those used in operational NWP service (200x200x31 gridpoints with horizontal spatial resolution 10 km). We observed almost linear productivity growth within increasing number of computation nodes. For instance, 8 nodes of the SCIT-3 cluster gave the performance increase in 7.09 times (of 8.0 theoretically possible) when compared to the single node. The use of 64 nodes increases the performance 43.6 times [Kussul et al, 2009].

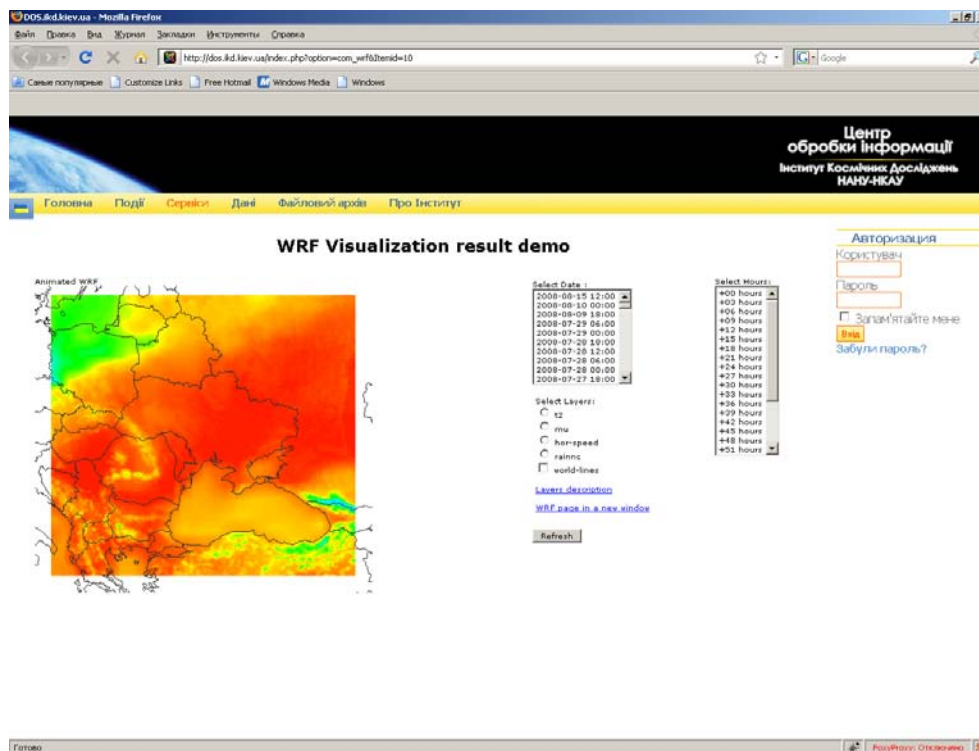


Figure 3. Example of land temperature forecasts using WRF model

Land biodiversity assessment. We have developed a Web service for biodiversity assessment for the Pre-Black Sea region of Ukraine using EOS data products [Popov et al, 2008]. Biodiversity is associated with a number of abiotic and biological factors that can be identified using remote sensing data. These factors include: landscape types, geographical latitude/altitude, climate conditions (such as mean daily temperatures, humidity, etc), structure and primary productivity of a vegetation mantle [Hansen and Rotella, 1999]. These factors can be

estimated using EO data from space [Popov et al, 2008]. The workflow for biodiversity estimation consists of the following steps: data acquisition, data processing, and visualization. Figure 4 shows the overall architecture of the service with information flows and integration modules.

Special system was developed in order to acquire multisource satellite data on a regular basis. This system operationally monitors for the new products and provides automatic data acquisition from different sources: Level 1 and Atmosphere Archive and Distribution System (LAADS), Land Processes Distributed Active Archive Center (LP DAAC) and National Snow and Ice Data Center (NSIDC). The acquired data are stored in the data archive of Space Research Institute.

After the required data has been acquired, the data is re-projected to a conical Albers projection and scaled to the spatial resolution of 250 m. Since we use data from multiple sources different tools were applied for the re-projection and scaling purposes. In particular, we used MODIS Swath Reprojection Tool, MODIS Reprojection Tool, and GDAL library (Geospatial Data Abstraction Layer, <http://www.gdal.org>). Since biodiversity index represents a parameter that is estimated for the time range, it is required to calculate average values for the parameters influencing biodiversity. For this purpose, average composites of images were created. Using these composites and solar irradiation acquired from SRTM DEM v2, we estimated the biodiversity index using a fuzzy model [Popov et al, 2008]. The resulting product is a georeferenced file in GeoTIFF format showing biodiversity index over the given region. The workflow of the data processing step is controlled by the Karajan engine while the data are processed on the computational resources of the Grid system using the GRAM service [Shelestov et al, 2006; Kussul et al, 2009; Hluchy et al, 2010].

The proposed Web service is implemented on the basis of OGC standards, Web Map Service 1.1.1 (<http://www.opengeospatial.org/standards/wms>) and Web Coverage Service 1.0 (<http://www.opengeospatial.org/standards/wcs>). The developed Web service is accessible via Internet through the address <http://inform.ikd.kiev.ua/biodiv/> (Figure 5). It represents current distribution of the potential biodiversity and allows monitoring each of the factors that influence biodiversity.

Vegetation state assessment. A cascade of models is used to vegetation state assessment (Figure 6). This includes a regional NWP model WRF that was described in the previous subsection and comprehensive land surface model (Noah). Remote sensing observations along with ground measurements are assimilated into these models to derive meteorological parameters (temperature, rainfall), land and soil parameters (moisture and temperature). Additionally, satellite-based products (for example, vegetation indices) are used monitor vegetation state.

Such an approach was used to monitor severe droughts that hit Ukraine in spring-summer 2007. Consequences were catastrophic: 1,4 million ha of crops totally destroyed, 8,5 million ha of crops damaged, 100 million of U.S. dollars losses. The use of the proposed approach allowed us to identify regions that were mostly affected by the disaster, and estimate potential losses. Figure 7 shows comparison of vegetation index of 2007 and 2006.

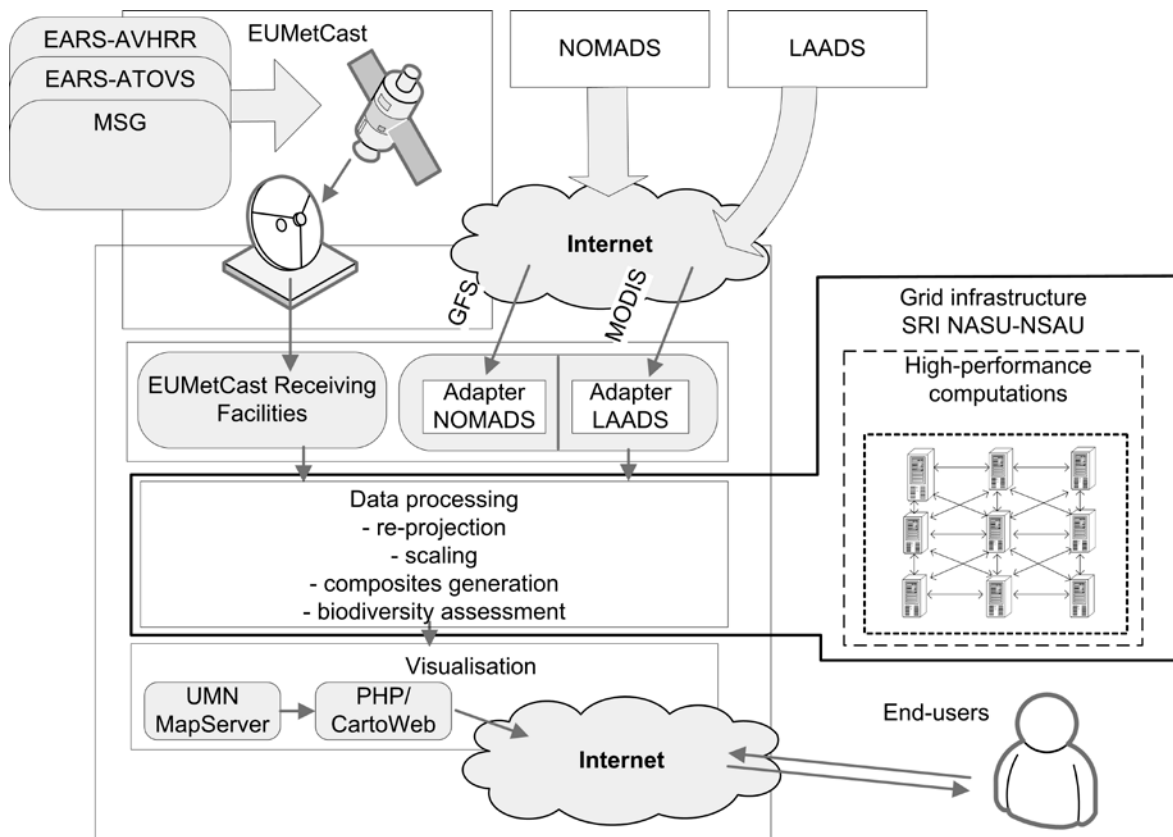


Figure 4. Overall architecture of the service with information flows

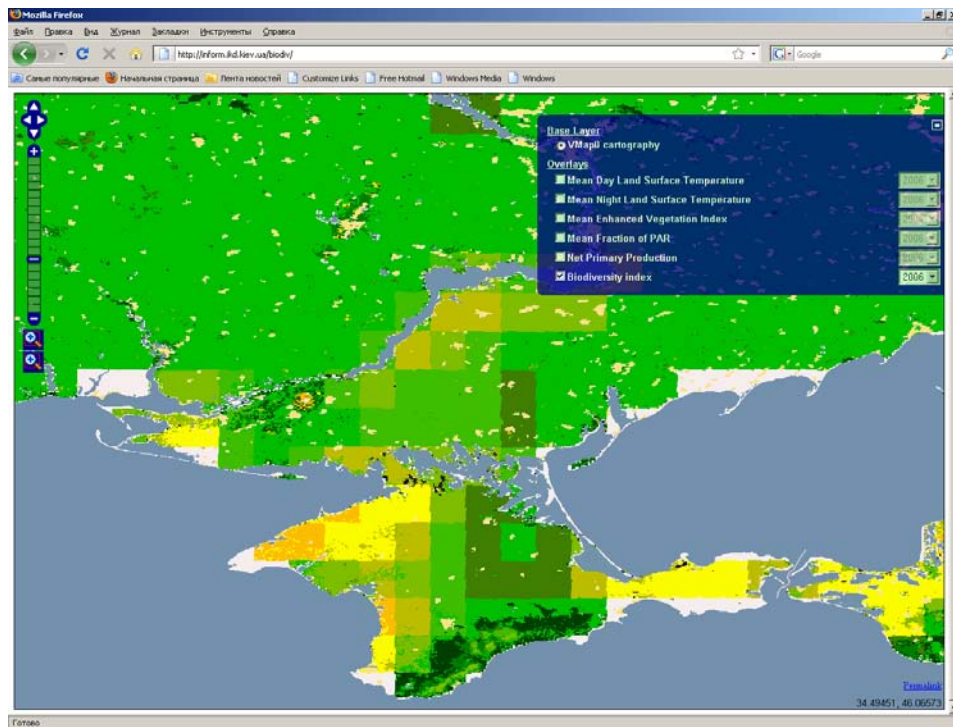


Figure 5. Demonstration of Web service for biodiversity assessment using EOS data products for the Pre-Black Sea region of Ukraine

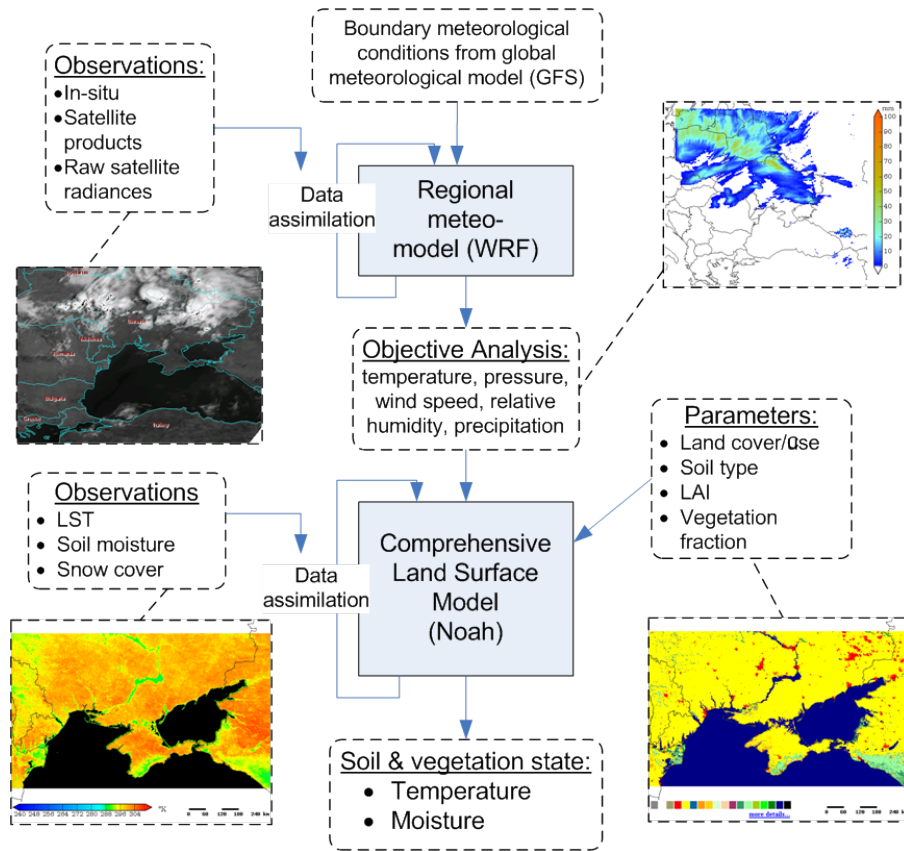


Figure 6. Modelling cascade for drought monitoring in Ukraine

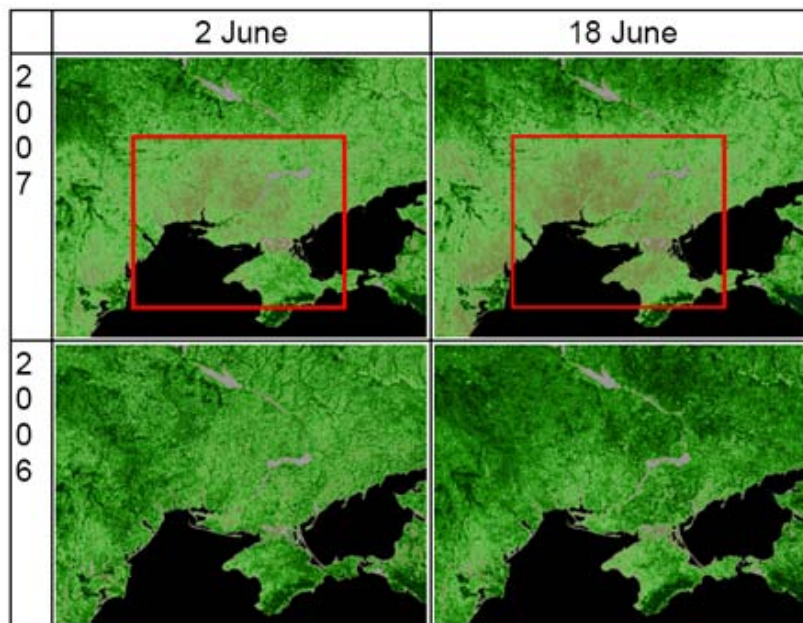


Figure 7. Evolution of Enhanced Vegetation Index (EVI) in the 2006 and 2007 vegetation seasons. Drought affected territories are highlighted by a rectangle

Fire monitoring. In July-August, 2010, Ukraine suffered from fires due to extremely high temperature: +35-39 C in Eastern regions and +40-42 C in South regions. On average 200 fires per day were detected. There was high risk of forest fires and fires approaching ammunition depots. Operational monitoring of fires was carried out using the following datasets:

- EO-1/ALI data acquired through Sensor Web prototype (date: 14.08.2010 08:15UTC)
- Landat-5/TM (date: 02.08.2010 08:15UTC)
- ZKI Fire Service that is available on daily basis and is using MODIS instrument onboard Terra & Aqua satellite.

The data products were extracted specifically for the territory of Ukraine. MODIS products were operationally delivered twice per day while other products were delivered on demand for the regions with the highest risks of fires. Cross-validation of MODIS and Landsat-5 products was done and showed good correspondence between data (Figure 8).

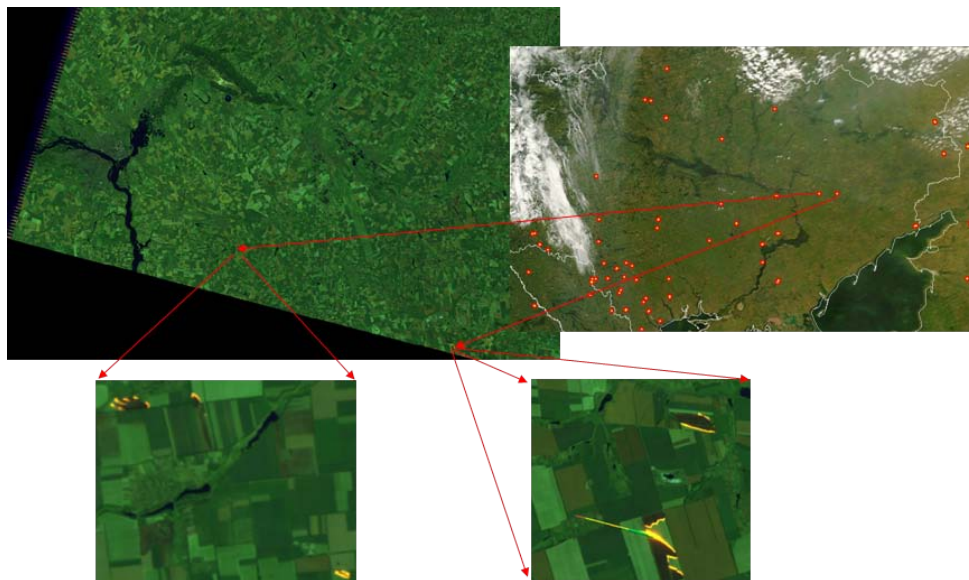


Figure 8. Cross-validation of fire products from MODIS and Landsat-5

International projects within UN-SPIDER RSO in Ukraine

UN-SPIDER is the United Nations Platform for Space-based Information for Disaster Management and Emergency Response and aims at providing universal access to all types of space-based information and services relevant to disaster management. The UN-SPIDER Regional Support Office (RSO) in Ukraine was established on basis of Space Research Institute in 2010. The RSO in Ukraine provides expertise in satellite data processing and product generation, operational delivery of services in case of emergency situations, and training activities. The RSO in Ukraine is actively involved in international projects. One of such a project is the Namibian Pilot on integrated flood management and water related vector borne disease modelling. Within this project one of the main tasks is flood risk assessment based on heterogeneous data.

These data are (Figure 9):

- Satellite imagery: synthetic-aperture radar (Envisat/ASAR, Radarsat-2), optical (EO-1, MODIS, Landsat-5), TRMM
- Modelling data: meteorological data (numerical weather prediction), hydrological data (river catchments).
- In-situ observations and river gauges: rainfall and river flow rate
- Statistical data: statistical information on floods for previous years.

The integration of different products is shown in Figure 10.

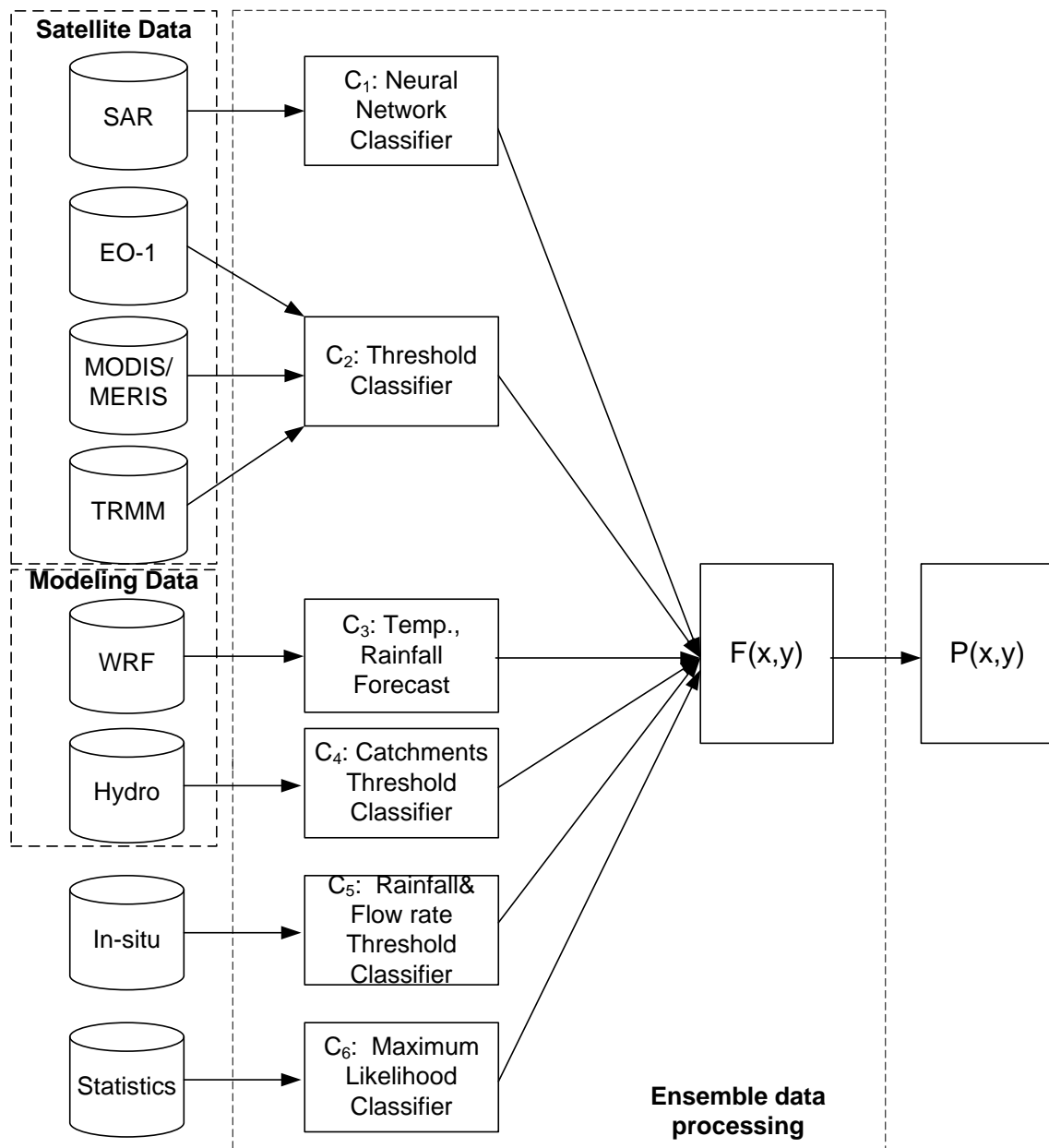


Figure 9. Integration of multisource data to flood risk assessment for the Namibian project

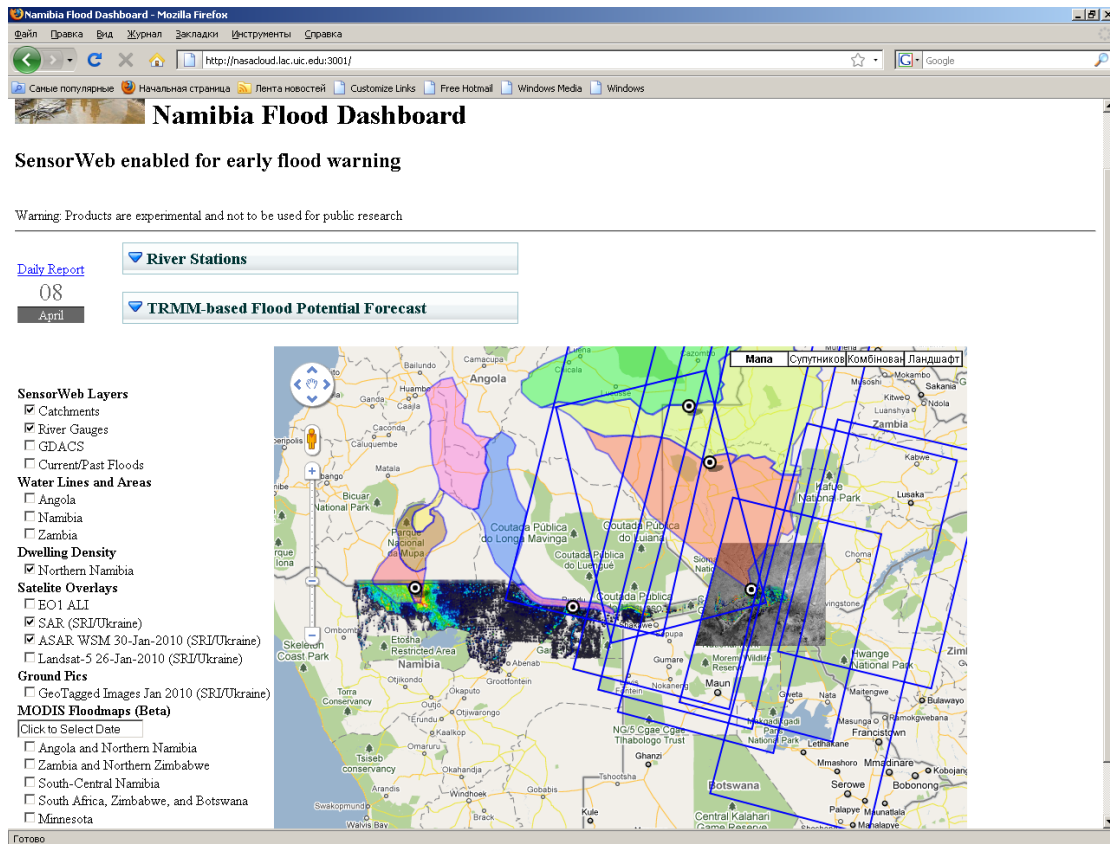


Figure 10. Namibian pilot project portal

Conclusions

In this paper we presented intelligent methods and corresponding technologies for environmental risk assessment. The risk assessment process is based on fusion of data acquired from different sources: models, in-situ observations and remote sensing instruments. The concept where the same data sets are applied for different applications is used. Therefore, once interfaces to discover and access the required data and products are developed, they can be used in a uniform way for different purposes and applications. This provides a basis for effective and operational exploitation of data.

The mixtures-of-experts concept for environmental risk assessment is introduced. Different experts provide a partial decision on the event using corresponding data, and their opinions are combined through some generalized rule. This allows for the problem to be broken into smaller sub-problems, and these sub-problems might be easier to solve than the overall problem.

Several real-world applications are described to demonstrate efficiency of the proposed approach, namely *numeral weather prediction (NWP)*, *land biodiversity assessment*, *vegetation state assessment*, *fire monitoring* and *flood mapping*. Most of these applications are being implemented within international projects within the UN-SPIDER Regional Support Office (RSO) in Ukraine (<http://un-spider.ikd.kiev.ua>).

Bibliography

- [Bishop, 2006] C.M. Bishop. Pattern Recognition and Machine Learning. New York: Springer Science+Business Media, 2006. 738 p.
- [GEOSS, 2005] GEOSS, Global Earth Observation System of Systems. 10-Year Implementation Plan: Reference Document. ESA Publication Division, Netherlands, 2005.
- [GMES, 2008] Global Monitoring for Environment and Security (GMES):we care for a safer planet. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Brussels, COM (2008) 748 final.
- [Hansen and Rotella, 1999] A.J. Hansen, J.J. Rotella. Abiotic factors. In: Hunter ML (ed) Maintaining biodiversity in forest ecosystems, Cambridge: Cambridge University Press, pp 161-209, 1999.
- [Haykin, 1999] S. Haykin. Neural Networks. A comprehensive Foundation. — New Jersey: Prentice Hall, 1999.
- [Hluchy et al, 2010] L. Hluchy, N. Kussul, A. Shelestov, S. Skakun, O. Kravchenko, Y. Gripich, P. Kopp, E. Lupian. The Data Fusion Grid Infrastructure: Project Objectives and Achievements. Computing and Informatics, 29(2), pp. 319-334, 2010.
- [Horritt, 2006] M.S. Horritt. A methodology for the validation of uncertain flood inundation models. J. of Hydrology, 326, pp. 153-165, 2006.
- [Jacobs et al, 1991] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton. Adaptive mixtures of local experts. Neural Computation, Vol. 3, pp. 79-87, 1991.
- [Kogan et al, 2004] F. Kogan, R. Stark, A. Gitelson, E. Adar, L. Jargalsaikhan, C. Dugrajav, S. Tsooj. Derivation of Pasture Biomass in Mongolia from AVHRR-based Vegetation Health Indices. Int. J. Remote Sens, 25(14), 2889-2896, 2004.
- [Kussul et al, 2009] N. Kussul, A. Shelestov, S. Skakun. Grid and sensor web technologies for environmental monitoring. Earth Science Informatics, 2(1-2), pp. 37-51, 2009.
- [Kussul et al, 2011] N. Kussul, A. Shelestov, S. Skakun. Flood Monitoring on the Basis of SAR Data. In F. Kogan, A. Powell, O. Fedorov (Eds.) "Use of Satellite and In-Situ Data to Improve Sustainability". - NATO Science for Peace and Security Series C: Environmental Security, pp. 19-29, 2011.
- [Liang, 2004] Liang S. Quantitative Remote Sensing of Land Surfaces. Wiley, 2004.
- [Michalakes et al, 2004] J. Michalakes, J. Dudhia, D. Gill, T. Henderson, J. Klemp, W. Skamarock, W. Wang. The Weather Research and Forecast Model: Software Architecture and Performance. In: Proc of the 11th ECMWF Workshop on the Use of High Performance Computing In Meteorology (25-29 October 2004, Reading U.K), 2004.
- [OHSAS, 2007] OHSAS 18001:2007 - Occupational Health and Safety Management Systems Requirements Standard, 2007.
- [Popov et al, 2008] M. Popov, N. Kussul, S. Stankevich, A. Kozlova, A. Shelestov, O. Kravchenko, M. Korbakov, S. Skakun. Web Service for Biodiversity Estimation Using Remote Sensing Data. Int. J. of Digital Earth, 1(4), pp. 367-376, 2008.
- [Shelestov et al, 2006] A.Yu. Shelestov, N.N. Kussul, S.V. Skakun. Grid Technologies in Monitoring Systems Based on Satellite Data. J. of Automation and Information Science, 38(3), pp. 69-80, 2006.

[Shelestov, 2008] A. Shelestov. Workflow Modelling in Grid System for Satellite Data Processing. International Journal on Information Theory and Applications, 15(3), pp. 225-231, 2008.

[Vapnik, 1998] V. Vapnik. Statistical Learning Theory. New York: Wiley, 1998.

[Wagner et al, 2007] W. Wagner, C. Pathe, D. Sabel, A. Bartsch, C. Kuenzer, K. Scipal. Experimental 1 km soil moisture products from ENVISAT ASAR for Southern Africa. ENVISAT & ERS Symposium, Montreux, Switzerland, 2007.

Authors' Information

Nataliia Kussul – Deputy Director, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: inform@ikd.kiev.ua

Sergii Skakun – Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: serhiy.skakun@ikd.kiev.ua

Oleksii Kravchenko – Scientific Researcher, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: oleksiy.kravchenko@gmail.com

SELF-ORGANIZING ROUTING ALGORITHM FOR WIRELESS SENSORS NETWORKS (WSN) USING ANT COLONY OPTIMIZATION (ACO) WITH TINYOS

Nuria Gómez Blas, Luis F. de Mingo, Levon Aslanyan, Vladimir Ryazanov

Abstract: *This paper describes the basic tools to work with wireless sensors. TinyOS has a component-based architecture which enables rapid innovation and implementation while minimizing code size as required by the severe memory constraints inherent in sensor networks. TinyOS's component library includes network protocols, distributed services, sensor drivers, and data acquisition tools – all of which can be used as-is or be further refined for a custom application. TinyOS was originally developed as a research project at the University of California Berkeley, but has since grown to have an international community of developers and users. Some algorithms concerning packet routing are shown. In-car entertainment systems can be based on wireless sensors in order to obtain information from Internet, but routing protocols must be implemented in order to avoid bottleneck problems. Ant Colony algorithms are really useful in such cases, therefore they can be embedded into the sensors to perform such routing task.*

Keywords: *Mobile ad-hoc networks, ant colony optimization, routing protocols, simulation, TinyOS.*

ACM Classification Keywords: *A.0 General Literature - Conference proceedings, I.6. Simulation and Modelling, I.2.8 Problem Solving, I.2.11 Distributed Artificial Intelligence, H.5.2 Information interfaces and presentation: Miscellaneous.*

Preliminaries

Ant Colony Optimization (ACO) [19, 20] technique is an optimization technique to solve optimization problem. It has been developed for combinatorial optimization problems. ACO are multi-agent system in which the behaviour of each single agent, called ant, is inspired by the behaviour of real ants. ACO has been successfully employed to combinatorial optimization problems such as maximum loadability in voltage control study, loss minimization in distribution networks, unit commitment problem, multiobjective reactive power compensation, and complex multi-stage decision problem.

Figure 1 shows the behavior of an ant colony. There is a path, along which ants are walking, for example from food source A to the nest E, and vice versa. Suddenly an obstacle appears and the path is cut off. So at position B the ants walking from A to E (or at position D those walking in the opposite direction) have to decide whether to turn right or left. The path of ants followed is shown in figure 1. Because path BCD is shorter than BHD, the first ant following it will reach D before the first ant following path BHD which is shown in figure 1.c. The result is that an ant returning from E to D will find a stronger trail on path DCB, caused by the half of all the ants that by chance decided to approach the obstacle via DCBA and by the already arrived ones coming via BCD: they will therefore prefer (in probability) path DCB to path DHB. As a consequence, the number of ants following path BCD per unit of time will be higher than the number of ants following EHD. This causes the quantity of

pheromone on the shorter path to grow faster than on the longer one, and therefore the probability with which any single ant chooses the path to follow is quickly biased toward the shorter one. The final result is that very quickly all ants will choose the shorter path.

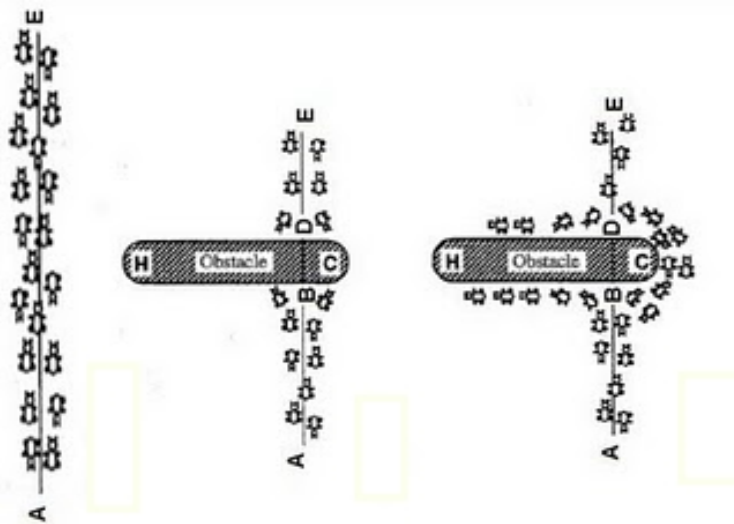


Figure 1.- Ant Colony Behaviour.

Ants use the environment as a medium of communication. They exchange information indirectly by depositing pheromones, all detailing the status of their "work". The information exchanged has a local scope, only an ant located where the pheromones were left has a notion of them. This system is called "Stigmergy" and occurs in many social animal societies (it has been studied in the case of the construction of pillars in the nests of termites). The mechanism to solve a problem too complex to be addressed by single ants is a good example of a self-organized system.

The algorithm is derived from the study of real ant colonies. Therefore the system is called as Ant System (AS) [20] and the algorithms as Ant algorithms. The use of artificial ant colonies as an optimization tool, will have some major differences with a real (natural ant) one:

1. artificial ants will have some memory,
2. they will not be completely blind,
3. they will live in an environment where time is discrete.

Ant colony optimization is a major breakthrough in Engineering as well as non engineering applications.

Wireless Sensors Networks Tools

Wireless Sensors Networks (WSN) consists of spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants, and to cooperatively pass their data through the network. The more modern networks are bi-directional, enabling also to control the activity of the sensors. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance; today such networks are used in many industrial and consumer application, such as industrial process monitoring and control, machine health monitoring, environment and habitat monitoring, healthcare applications, home automation, and traffic control.

The WSN is built of "nodes" - from a few to several hundreds or even thousands, where each node is connected to one (or sometimes several) sensors. Each such sensor network node has typically several parts, see figure 2: a radio transceiver with an internal antenna or connection to an external antenna, a microcontroller, an electronic

circuit for interfacing with the sensors and an energy source, usually a battery. A sensor node might vary in size from that of a shoebox down to the size of a grain of dust, although functioning "motest" of genuine microscopic dimensions have yet to be created. The cost of sensor nodes is similarly variable, ranging from hundreds of dollars to a few pennies, depending on the complexity of the individual sensor nodes. Size and cost constraints on sensor nodes result in corresponding constraints on resources such as energy, memory, computational speed and communications bandwidth. The topology of the WSNs can vary from a simple star network to an advanced multi-hop wireless mesh network. The propagation technique between the hops of the network can be routing or flooding [9, 10, 11].

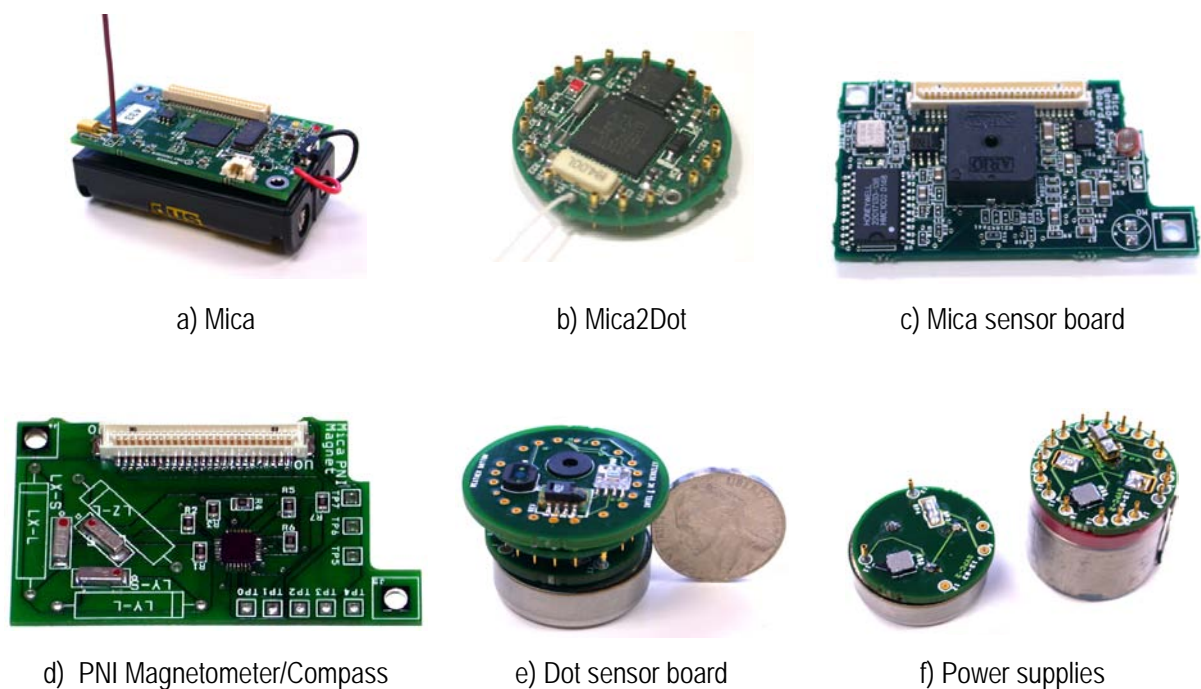


Figure 2.-Hardware implementation of different motest and available sensors and power supplies to connect to.

TinyOS

TinyOS[1, 16] is an event-driven operating system designed for sensor network nodes that have very limited resources. TinyOS, is used, for example, on the MICA motest (see figure 2), which are small wireless sensor nodes. TinyOS has extensive networking support, and this support includes technically excellent protocol designs which have become de facto standards, or in some cases, parts of Internet standards. This support has been in part due to TinyOS's use as a platform by many leading low-power wireless research groups, who have then released their code for general use and supported it well. The TinyOS net2 Working Group is responsible for adding, improving, and maintaining TinyOS's network protocols.

TinyOS supports low duty cycle operation through low-power link layers. Rather than keep the radio always on, TinyOS turns the radio on periodically (e.g., every few hundred ms) to check if there is a packet to receive. This enables the network to appear "always on" yet support sub-1% duty cycles: the basic tradeoff is that

communication has higher latency. TinyOS supports multihop, network-wide sub-millisecond time synchronization through the Flooding Time Synchronization Protocol, developed by researchers at Vanderbilt University.

Data collection protocols build a self-organizing, self-repairing routing topology to data collection points known as "roots." Typically these roots are connected to a PC or other device, such that the data collected can be stored in a database for later use. Collection protocols send data in only one direction (towards a root): they do not support messages to arbitrary nodes in the network. TinyOS's standard collection protocol, the Collection Tree Protocol (CTP), is highly efficient and robust: it continues to deliver data even after large numbers of node failures and has emerged as the gold standard against which other routing protocols are measured.

Data dissemination protocols reliably deliver a piece of data to every node in a network. TinyOS supports three dissemination protocols: Drip, DIP, and DHV [13, 14, 15]. These three protocols represent a gradual evolution towards more efficient algorithms. Generally speaking, applications should use DHV.

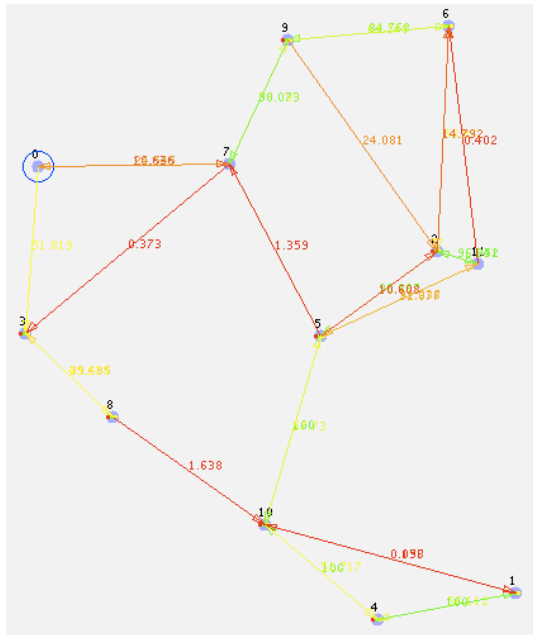
TinyOS includes support for reprogramming a multihop wireless network over the air with the Deluge protocol. A Deluge-enabled network supports having multiple binaries in the network at once: a command line tool can instruct the network to change programs. This operation takes a short while as the nodes reprogram themselves.

All of the above protocols are subjects of a long literature of research and publications, such that there is extensive information in how they work. They are all designed to work on top of low power link layers [12].

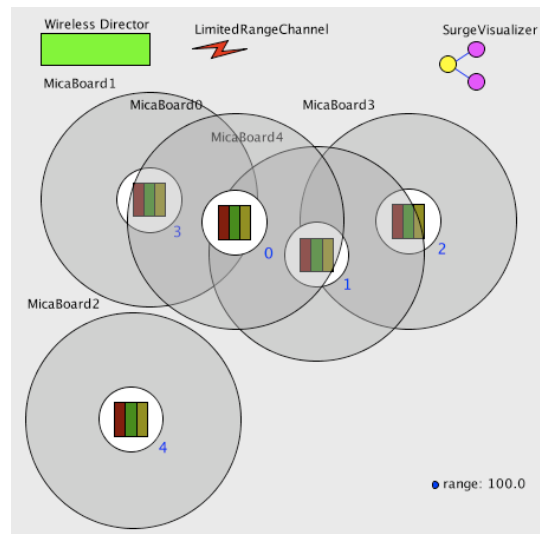
Tinyviz, Viptos and Ptolemy II

Viptos (Visual Ptolemy and TinyOS) [2, 4] is an integrated graphical development and simulation environment for TinyOS-based wireless sensor networks. Viptos allows developers to create block and arrow diagrams, see figure 3, to construct TinyOS [1] programs from any standard library of nesC/TinyOS components. The tool automatically transforms the diagram into a nesC program that can be compiled and downloaded from within the graphical environment onto any TinyOS-supported target hardware. Viptos is based on TOSSIM [1] and Ptolemy II [3]. TOSSIM is an interrupt-level simulator for TinyOS programs. It runs actual TinyOS code but provides software replacements for the simulated hardware and models network interaction at the bit or packet level. Ptolemy II [7, 8] is a graphical software system for modelling, simulation, and design of concurrent, real-time, embedded systems. Ptolemy II focuses on assembly of concurrent components with well-defined models of computation that govern the interaction between components. While TOSSIM only allows simulation of homogeneous networks where each node runs the same program, Viptos supports simulation of heterogeneous networks where each node may run a different program. Viptos simulations may also include non-TinyOS-based wireless nodes. The developer can easily switch to different channel models and change other parts of the simulated environment, such as creating models to generate simulated traffic on the wireless network.

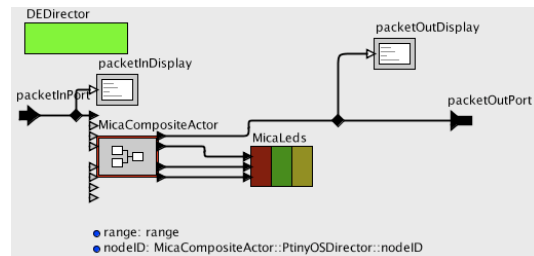
Viptos inherits the actor-oriented modelling environment of Ptolemy II [5, 6], which allows the developer to use different models of computation at each level of simulation. At the lowest level, Viptos uses the discrete-event scheduler of TOSSIM to model the interaction between the CPU and TinyOS code that runs on it. At the next highest level, Viptos uses the discrete-event scheduler of Ptolemy II to model interaction with mote hardware, such as the radio and sensors. This level is then embedded within VisualSense to allow modelling of the wireless channels to simulate packet loss, corruption, delay, etc. The user can also model and simulate other aspects of the physical environment including those detected by the sensors (e.g., light, temperature, etc.), terrain, etc.



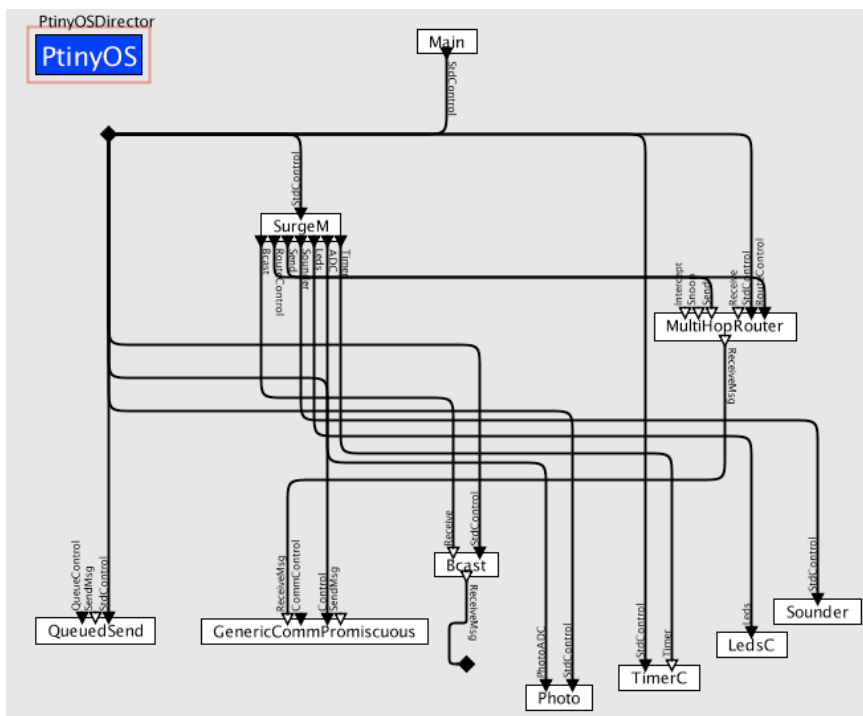
a) TiniViz simulation of motes showing radio link power.



b) Distribution of motes in space.



c) Mote behaviour.



d) Mote components and interfaces.

Figure 3.- TiniViz simulation of motes with radio links a), and Viptos simulation at different abstraction levels b), c) and d).

TinyViz [1] is a Java-based GUI that allows you to visualize and control the simulation as it runs, see figure 3.a), inspecting debug messages, radio and UART packets, and so forth. The simulation provides several mechanisms for interacting with the network; packet traffic can be monitored, packets can be statically or dynamically injected into the network.

Ant colony self-organizing routing algorithms

Ant colony algorithms were first proposed by Dorigo et al as a multi-agent approach to difficult combinatorial optimization problems like the traveling salesman problem (TSP), figure 4, and the quadratic assignment problem (QAP), and later introduced the ACO meta-heuristic.



Figure 4.- Solved TSP problem using ant colony optimization.

There are two types of ants applied in the algorithm, forward ants and backward ants. Forward ants, whose main actions are exploring the path and collecting the information from the source nodes to destination node, have the same number as the source nodes [21]. The paths that forward ants travel will construct a tree when they merge into each other or reach the destination and data is transmitted along the tree paths, see figure 5.

POSANT Routing Algorithm [19, 20] is ant colony optimization based routing algorithm which uses location information to improve its efficiency. POSANT is able to find optimum or nearly optimum routes when a given network contains nodes. *Zone based Routing Algorithm using Cluster*. Concept of clustering needs grouping of nodes in the network. This grouping depends upon transmission range and number of hop in a group. Each node group will have a group head called Cluster head having the responsibility of communication among its member nodes and other cluster heads. Cluster head should contain address of its member nodes as well as that of other

cluster heads. Member nodes need to store address information of their cluster head and neighbor nodes. When information needs to pass from one node to another, member node sends this information to its corresponding cluster head, which decides whether the destination is a member or not.

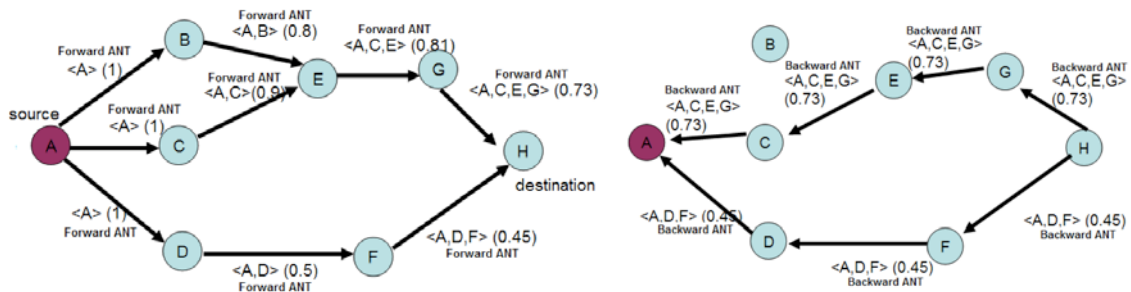


Figure 5.- Ant colony routing in mobile ad-hoc networks (MANET) protocol [17].

Ant Colony Routing Algorithm with Zones [18]. Concept of Ant Colony algorithm is merged with zone based (clustering) algorithm to form ant colony routing algorithm with zones. This algorithm will provide advantage of both ant colony and zone based algorithm. Like ant colony algorithm, here we need not store large routing tables in nodes, we need to store only neighboring node information and previous traversed node information. As nodes in mobile ad-hoc network will have memory of small storage capacity, it would be tough to store large routing table inside each node.

A local routing, instead of storing the whole network graph, will be more suitable in order to keep track of the information going to a destination node.

The local routing table in every node/mote keeps the following information:

A list of neighbourhood nodes/motes that have internet connection, see figure 6. This table is build using a discovery ant that every node will run, when the ant reaches the internet sink a backward ant will be sent back to the source node that updates the probability and lookup table of nodes. These discovery ants start if there is no internet connection at regular time intervals.

In order to be able to sent back data packets, the MAC address or ID of the source node must be keep in the path of the route to the internet sink. That is, every node/mote stores the pair:

- the MAC/ID of the source of every transmitted packet (to be able to sent data back to the source),
- and the MAC/ID of the connected node/mote of transmitted packet (to sent data back).

Data packets are sent using the Internet lookup table, according to the propability of the node. When the echo information pass a node and reaches the source, then the probabilities are also updated.

The probability of internet lookup table at node i that has a radio link with node j's updated using the following equation:

$$P_{ij} = \frac{\tau_j \alpha_{ij} \beta_j}{\sum_j \tau_j \alpha_{ij} \beta_j}$$

where:

τ_j is the pheromone information updated by backwards ants

α_{ij} is the radio link power between nodes

β_j is the node power status

This approach does not take into account low power consumption since this algorithm will be implemented at in-car entertainment systems (ICE) and does not take into account memory limitations, this can be solved using a circular table in order to remove low probabilities. Real-time information is not need and backward information could also use a circular table when there are a lot of nodes.

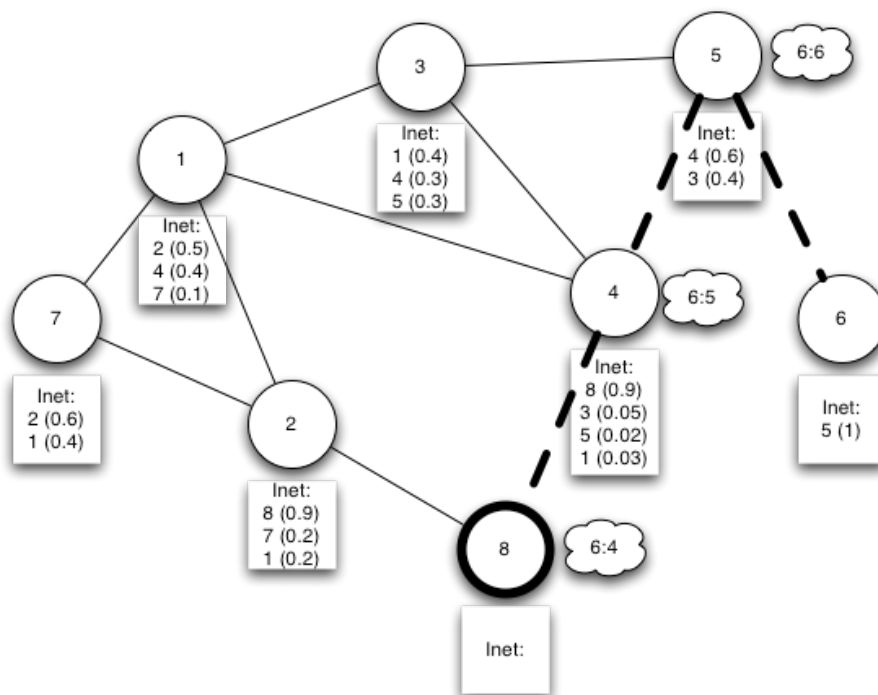


Figure 6.- WSN topology with internet routing table and backward information.

Future remarks

In-car entertainment systems, sometimes referred to as ICE, is a collection of hardware devices installed into automobiles, or other forms of transportation, to provide audio and/or audio/visual entertainment, as well as automotive navigation systems (SatNav). This includes playing media such as CDs, DVDs, Freeview/TV, USB and/or other optional surround sound, or DSP systems. Also increasingly common in ICE installs are the incorporation of video game consoles into the vehicle. In Car Entertainment systems have been featured TV shows such as MTV's Pimp My Ride. In Car Entertainment has become more widely available due to

reduced costs of devices such as LCD screen/monitors, and the reducing cost to the consumer of the converging media playable technologies. Single hardware units are capable of playing CD, MP3, WMA, DVD.

MIT's CarTel project is investigating how cars themselves could be used as ubiquitous, highly reliable mobile sensors. At the Association for Computing Machinery's sixth annual Workshop on Foundations of Mobile Computing, members of the CarTel team presented a new algorithm that would optimize the dissemination of data through a network of cars with wireless connections. Researchers at Ford are already testing the new algorithm for possible inclusion in future versions of Sync, the in-car communications and entertainment system developed by Ford and Microsoft.

Described algorithms could be embedded into ICE in order to improve the routing algorithm since there is no need of real-time information retrieval and, in some cases, no need of low power consumption.

Bibliography

- [1] TinyOs web page. <http://www.tinyos.net>
- [2] Vptos web page. <http://ptolemy.eecs.berkeley.edu/vptos/>
- [3] Ptolemy II web page. <http://ptolemy.eecs.berkeley.edu/ptolemyII/>
- [4] [Elaine Cheong](#), [Edward A. Lee](#), [Yang Zhao](#): Vptos: a graphical development and simulation environment for tinyOS-based wireless sensor networks. [SenSys 2005:302](#)
- [5] [Christopher X. Brooks](#), [Edward A. Lee](#), [Stavros Tripakis](#): Exploring models of computation with ptolemy II. [CODES ISSS 2010:331-332](#)
- [6] [Yang Zhao](#), [Yuhong Xiong](#), [Edward A. Lee](#), [Xiaojun Liu](#), [Lizhi C. Zhong](#): The design and application of structured types in Ptolemy II. [Int. J. Intell. Syst. \(IJIS\) 25\(2\):118-136 \(2010\)](#)
- [7] [Kyungmin Bae](#), [Peter Csaba Ölveczky](#): Extending the Real-Time Maude Semantics of Ptolemy to Hierarchical DE Models [RTRTS 2010:46-66](#)
- [8] [Martin Schoeberl](#), [Christopher X. Brooks](#), [Edward A. Lee](#): Code Generation for Embedded Java with Ptolemy. [SEUS 2010:155-166](#)
- [9] [Rajkumar K. Raval](#), [Carlos Fernández](#), [Chris J. Bleakley](#): Low-power TinyOS tuned processor platform for wireless sensor network motes. [ACM Trans. Design Autom. Electr. Syst. \(TODAES\) 15\(3\) \(2010\)](#)
- [10] [A. Sivagami](#), [K. Pavai](#), [D. Sridharan](#), [S. A. V. Satya Murty](#): Energy and Link Quality Based Routing for Data Gathering Tree in Wireless Sensor Networks Under TINYOS - 2.X [CoRR abs/1005.1739 \(2010\)](#)
- [11] [Swarup Kumar Mitra](#), [Ayon Chakraborty](#), [Subhajit Mandal](#), [Mrinal K. Naskar](#): Simulation of Wireless Sensor Networks Using TinyOS- A case Study [CoRR abs/1004.4154 \(2010\)](#)
- [12] [Doina Bucur](#), [Marta Z. Kwiatkowska](#): Software verification for TinyOS. [IPSN 2010:400-40151EESimon Kellner](#): Flexible Online Energy Accounting in TinyOS. [REALWSN 2010:62-73](#)
- [13] [Muhammad Hamad Alizai](#), [Bernhard Kirchen](#), [Jó Ágila Bitsch Link](#), [Hanno Wirtz](#), [Klaus Wehrle](#): TinyOS meets wireless mesh networks. [SenSys 2010:429-430](#)
- [14] [Nelson I. Dopico](#), [Carlos Gil-Soriano](#), [Inigo Arrazola](#), [Santiago Zazo](#): Analysis of IEEE 802.15.4 Throughput in Beaconless Mode on micaZ under TinyOS 2. [VTC Fall 2010:1-5](#)
- [15] [Werner Backes](#), [Jared Cordasco](#): MoteAODV - An AODV Implementation for TinyOS 2.0. [WISTP 2010:154-169](#)

-
-
- [16] [Angelo Paolo Castellani](#), [Paolo Casari](#), [Michele Zorzi](#): TinyNET - a tiny network framework for TinyOS: description, implementation, and experimentation. [Wireless Communications and Mobile Computing \(WICOMM\) 10\(1\):101-114 \(2010\)](#)
- [17] Ajay C Solai Jawahar: Ant colony optimization for mobile ad-hoc networks. ECE-572 - Parallel and Distributed Computing. Rutgers University.
- [18] Bandyopadhyay, M. and Bhaumik, P.: Zone Based Ant Colony Routing In Mobile Ad-hoc Network. Communication Systems and Networks (COMSNETS), 2010, ISBN: 978-1-4244-5487-7, pp.: 1-10. (2010).
- [19] M. Dorigo, Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italie, 1992.
- [20] A. Coloni, M. Dorigo et V. Maniezzo, Distributed Optimization by Ant Colonies, actes de la première conférence européenne sur la vie artificielle, Paris, France, Elsevier Publishing, 134-142, 1991.
- [21] K.Saleem, N.Fisal, S.Hafizah, S.Kamilah and Rozeha A. Rashid; Ant based Self-organized Routing Protocol for Wireless Sensor Networks . International Journal of Communication Networks and Information Security (IJCNIS), Vol. 1 (2). 2009.

Acknowledgment

This work has been partially supported by the Spanish Research Projects:

TRA2010-15645. "COMUNICACIONES EN MALLA PARA VEHICULOS E INFRAESTRUCTURAS INTELIGENTES" (Mesh communication with intelligent vehicles). (2010)

TEC2010-21303-C04-02. "ESTRUCTURAS RESONANTES PARA APLICACIONES DE SEÑAL FOTONICA DE BANDA ANCHA ". (2010).

Authors' Information

Nuria Gómez Blas – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: ngomez@eui.upm.es Research: DNA computing, Membrane computing, Education on Applied Mathematics and Informatics

Luis F. de Mingo – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: lfmingo@eui.upm.es Research: Artificial Intelligence, Social Intelligence, Education on Applied Mathematics and Informatics

Levon Aslanyan– Head of Department, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: lasl@sci.am

Vladimir Ryazanov– Russian Academy of Science, Russian Federation; e-mail: rvccas@mail.ru

SAFETY OPERATIONS OF THE COMPLEX ENGINEERING OBJECTS

Nataliya Pankratova

Abstract: *The safety operations of the complex engineering objects on the basis of system control is realized. The essence of such control is systemically coordinated evaluation and adjustment of the operational survivability and safety during the functioning process of an object. The diagnostic unit, which is the basis of a safety control algorithm for complex objects in abnormal situations, is developed as an information platform of engineering diagnostics. By force of systematic and continuous evaluation of critical parameters of object's functioning in the real time mode, the reasons, which could potentially cause the object' tolerance failure of the functioning in the normal mode, are timely revealed.*

Keywords: *survivability, risks, abnormal mode, safety, information platform for engineering diagnostics*

ACM Classification Keywords: *H.4.2. INFORMATION SYSTEM APPLICATION: type of system strategy*

Conference topic: *Applied Program Systems*

Introduction

The practice of the last decades of the last century suggests that the risks of man-made and natural disasters with the consequences of regional, national and global scale are continuously increasing [1], that is due to various objective and subjective conditions and factors [2]. Analysis of accidents and catastrophes can identify the most important causes and weaknesses of control principles for survivability and safety of complex engineering objects (CEO). One of such reasons is the peculiarities of the functioning of the diagnostic systems aimed to identify failures and malfunctions. This approach to safety precludes a possibility of a priori prevention of abnormal modes and as a consequence, there is the possibility of its subsequent transition into an accident and catastrophe.

Therefore, it is necessary to develop a new strategy to solve safety problems of modern CEO for various purposes. Here we propose a strategy that is based on the conceptual foundations of systems analysis, multicriteria estimation and risk forecasting [3]. The essence of the proposed concept is the replacement of a standard principle of identifying the transition from operational state of the object into inoperable one on the basis of detection of failures, malfunctions, defects, and forecasting the reliability of an object by a qualitatively new principle. The essence of this principle is the timely detection and elimination of the causes of an eventual transition from operational state of the object into inoperable one on the basis of systems analysis of multifactorial risk of abnormal situations, a reliable estimation of margin of permissible risk of different modes of complex engineering objects operation, and forecast the key indicators of the object survivability in a given period of its operation.

The processes of CEO functioning and processes of ensuring their safety are principally different. The first is focused on achieving the main production target of complex engineering systems, so they are focused on at all

stages of a product's life cycle. The second is regarded as secondary by the defined category of specialists, because in their view, all the major issues of efficiency and reliability and, consequently, the safety of the products are resolved at the stages of its development, refinement, handling, testing. As a result, there are precedents when the developments of goals, objectives and requirements for safety and, above all, for an engineering diagnostics system have not proper justification. As a consequence, it turns out that the figures and properties of the created safety system do not correspond to real necessities of complex objects, which they must satisfy.

Thus, there is a practical necessity to qualitatively change the principles and the structure of operational-capability controls and the safety of modern engineering systems in real conditions of multifactor risk influence. First of all, the control of complex objects should be systemized which means that there should be system coordination of operability control and safety control not merely by the corresponding goals, tasks, resources, and expected results but also, importantly, by the immediacy and effectiveness of interaction in real conditions of abnormal situations. Such coordination should provide immediate and effective interaction between the mentioned control systems. On the one hand, an effectiveness of the safety system should be provided for timely detection of abnormal situations, evaluation of risk degree and level, and the definition of an permissible risk margin during the process of forming recommendations about immediate actions given to the decision maker. On the other hand, the system of operational capability control after receiving a signal about abnormal situations should, in an effective and operative manner, make a complex object ready for an emergency transition to an offline state and should make it possible to effect this transition within the limits of permissible risk. This can be achieved only under the condition that the system of engineering diagnostics fully complies with the timeliness and efficiency of personnel actions in case of emergencies. Namely: Diagnosis should provide such level of completeness, accuracy and timeliness of information about the state and changing of technologically hazardous processes, which will allow staff to prevent the transition of abnormal situation to an accident and catastrophe in time.

It must be noted that the requirement of timeliness is a priority, as the most accurate, most reliable information becomes unnecessary when it comes to staff after an accident or catastrophe. So there is a practical need of systemic coherence of diagnostic rates with the pace of work processes in different modes of complex engineering systems operation. Such coherence can be one of the most important conditions for ensuring the guaranteed safety for the objects with increasing the risk [4].

1. Mathematical Formulation of Complex Object System Control Problem

Let us show the mathematical formulation of this problem with a priori set variation intervals of main indicators of the system in the normal mode and predefined permissible boundes of the influence of external factors. It is known that system functioning is characterized by the following sequence of complex system states: E_1, E_2, \dots, E_k . Every state E is characterized by specified indicators of system function processes (Y_k, X_k, U_k) and specified indicators of external environmental influence and risk factors Ξ_k :

$$E_k = \{(Y_k \in Y) \wedge (X_k \in X) \wedge (U_k \in U) \wedge (\Xi_k \in \Xi)\},$$

where the meaning of indicators at the moment $T_k \in T^\pm$ is defined by the following relations:

$$Y_k = \hat{Y}[T_k]; X_k = \hat{X}[T_k]; U_k = \hat{U}[T_k]; \Xi_k = \hat{\Xi}[T_k];$$

$$T_k = \{t_k | t_k > t_{k-1}\}; T_k \in T^\pm; T^\pm = \{t | t^- \leq t \leq t^+\}; Y = (Y_i | i = \overline{1, m}); X = (X_j | j = \overline{1, n});$$

$$U = (U_q | q = \overline{1, Q}); \Xi = (\Xi_p | p = \overline{1, P}).$$

Here Y is a set of external parameters Y_i that includes technical, economic, and other indicators of system-function quality; X is a set of internal parameters X_j that includes constructional, technological, and other indicators; U is a set of control parameters U_q ; Ξ is a set of external environmental influence parameters and parameters of risk factor influence Ξ_p ; $\hat{Y}[T_k]$, $\hat{X}[T_k]$, $\hat{U}[T_k]$ and $\hat{\Xi}[T_k]$ are sets of meanings of appropriate parameters at the moment T_k ; and T^\pm is a specified or predicted complex object functioning period. Required: determine in the moment $T_i \in T^\pm$ such values of degrees η_i and levels W_i of risk, as well as a margin of permissible risk T_{ar} , which provides, during the abnormal mode, the possibility of transition from the mode \bar{R}_{tr}^+ during the period \tilde{T}_{tr}^\pm to the normal mode till the critical moment T_{cr} of transition of abnormal mode becomes an accident or catastrophe. Here, the mode \bar{R}_{tr}^+ is controlled functioning mode conditioned by the control influence U_{tr} of a safety control system. During the time period \tilde{T}_{tr}^\pm this mode leads to the reduction of the abnormal mode R_{an} to the normal mode R_{nm} . The Mode \bar{R}_{tr}^+ is characterized by the following functional:

$$\bar{R}_{tr}^+ : R_{an} \xrightarrow{U_{tr}} R_{nm}$$

which defines the process of the reduction of the abnormal mode R_{an} to the normal mode R_{nm} under the influence of the control system. The main system property is an operational capability characterized by given quality indicators defined by the set Y . System safety will be considered as an ability to timely prevent a consecutive transfer from a normal mode to an accident or a catastrophe on the basis of timely detection of essential risk factors and elimination of the possibility of their conversion into catastrophic risk factors. Safety is characterized by the following indicators: degree of risk η_i , level of risk W_i and the margin of permissible risk T_{pr} of an abnormal mode; the margin of permissible risk T_{as} of an accident; and the margin of permissible risk T_{cr} of a catastrophe. The quantitative values of safety indicators are defined on the basis of the general problem of multifactor risk analysis, with mathematical definition is described in [4].

2. Strategy for Solving the Problem of System Control of Complex Objects

The main goal of the proposed strategy is to guarantee a rationally justified reserve of survivability of a complex system in real conditions of fundamentally irremovable information and time restrictions.

The main idea of the strategy is to ensure the timely and credible detection, recognition, and estimation of risk factors, forecasting their development during a definite period of operation in real conditions of a complex objects operation, and on this basis ensuring timely elimination of risk causes before the occurrence of failures and other undesirable consequences.

The main approaches and principles of the strategy for providing guaranteed safety of complex systems should be formed on the basis of the following principles [5]:

- system coordination according to the goals, tasks, resources, and expected results of measures aimed at ensuring the safety of a complex system;
- mutual coordination of goals, tasks, resources, and expected results of control of serviceability and safety of a complex system;
- timely detection, guaranteed recognition, and system diagnosis of factors and situations of risk;
- efficient forecasting and credible estimation of abnormal and critical situations;
- timely formation and efficient realization of decisions of safety control in the process of prevention of abnormal and critical situations.

Therefore, the most important and obligatory requirement of the strategy is system coordination of decisions and actions at all stages of a product's life cycle according to its goals, tasks, terms, resources, and expected results. The coordination must be provided simultaneously from the position of guaranteeing both the required indicators of safety and survivability and the required indicators of serviceability during the given period of operation [5].

In particular, the consistency of the diagnosis and control are especially important for transport systems, where there principally cannot be an emergency stop in conditions of unexpected effect of catastrophic risk factors. Such systems include all categories and all types of aircraft.

First, note the principal differences between the given problem and typical control problems. The main difference is that the initial information about a complex object contains only a small part of information about its state, properties, functioning processes, and operational capability characteristics. This information represents only the state and work characteristics of such objects in normal mode. Undoubtedly, this information is enough for decision making during the complex object control only on the condition that the normal mode continue for a long time. However, in real objects in view of existing technical diagnosis systems, oriented toward failure and malfunction detection, it is impossible to ensure that a malfunction or a failure will not appear within the next 5–10min. It is a priori unknown how much time it will take to repair a malfunction. It may take from a few minutes up to several hours or even days and months. And, consequently, the possible damage is a priori unknown, and thus the safety control system is, essentially, a recorder of information about facts and damage. A fundamentally different approach can be realized on the basis of the system control of complex objects. The essence of such control is systemically coordinated evaluation and adjustment of the operational capability and safety during the functioning process of an object.

The general strategy of such an approach is shown in Fig. 1. There is incomplete, fuzzy information about the object functioning state at the moment $T_k \in T^\pm$. This information is not enough for decision making. This implies the significant property of such an approach. This property means that the situation analysis and decision making are provided not only in typical conditions of exact recognition of a normal or an abnormal system mode, but also in conditions where there is only fuzzy, incomplete information about a situation. It is significant that this strategy, in conditions of fuzzy information about a situation, allows one, if necessary, to make a timely decision on emergency stop of the system operation. In the following control strategy in blocks 1–3 there are realized

procedures of complex object functioning diagnostics and analysis. In block 4, on the basis of the results of the execution of the procedures of blocks 2 and 3, of a normal functioning mode occurs. During this process, three possible variants of a complex object state are analyzed: the normal functioning mode remains (transition of a control to block 5.0); signs of a normal mode violation appears that make it possible to reveal that at time $T_k \in T^\pm$ the situation is abnormal (transition of a control to block 5.1); or at time $T_k \in T^\pm$ the situation becomes undefined (transition of a control to block 5.2).

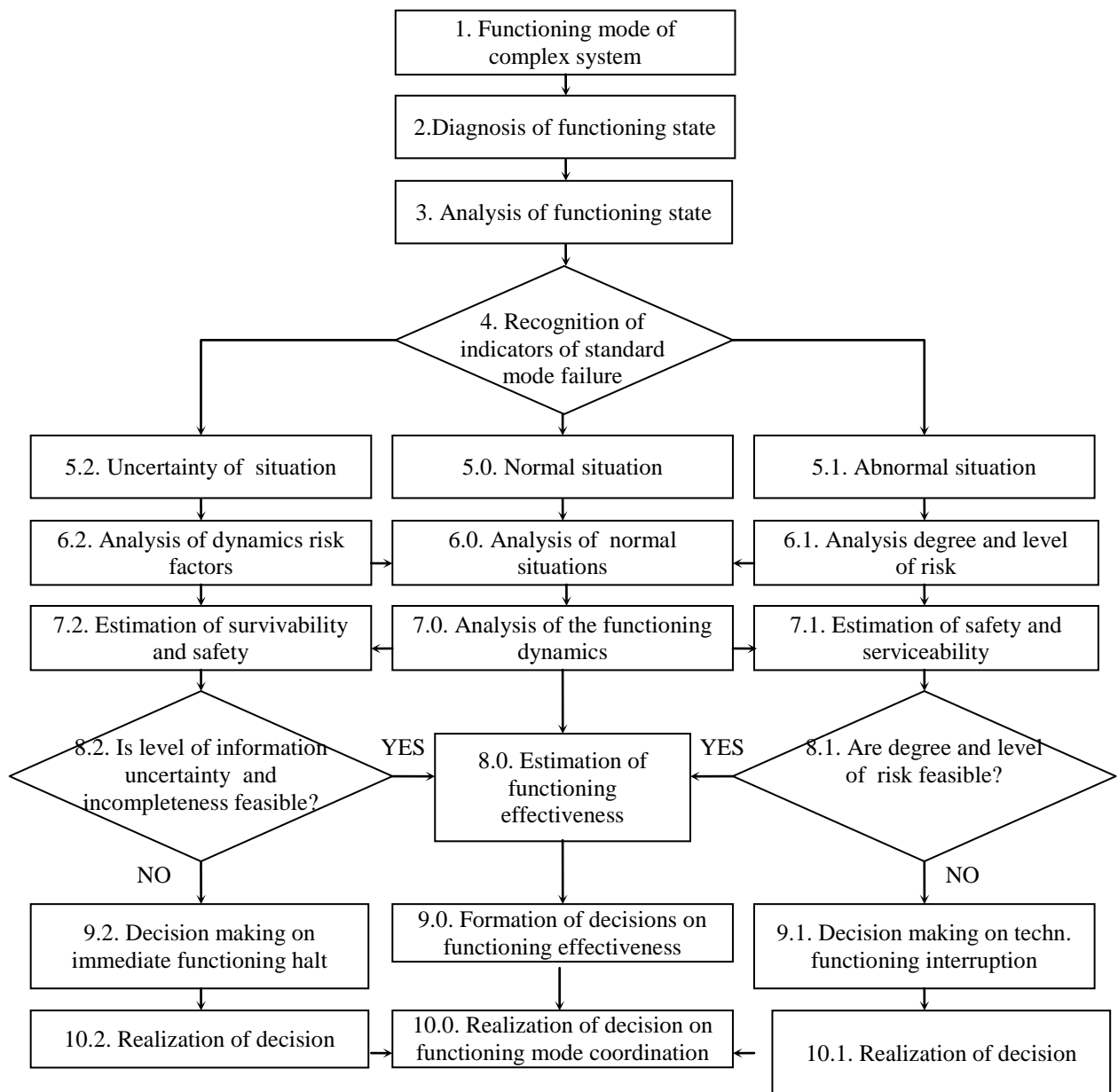


Figure 1. Structural scheme of the system control strategy of a complex objects' serviceability and safety

In the first case the system operates in normal mode, and quality control is executed (6.0–10.0 units). In the second variant on the basis of the sequence of abnormal situations the following actions are realized. The risk degree and level of an abnormal situation sequence is analyzed, and the safety and operational capability of complex objects are evaluated (6.1–7.1 units) and a decision is made regarding the scheduled stop of the complex object functioning (transition of the control to the blocks 8.1–10.1) or a decision is made regarding the continuation of the complex object functioning if the values of risk degree and level are acceptable (transition of the control to block 8.0). In the third variant, an evaluation of survivability and safety of the system in conditions of uncertain information about abnormal situations is made. For this, the following sequence of operations is realized. An analysis is made of risk factors sequence of abnormal situations, on the basis of which the complex object survivability and safety are evaluated (6.2–7.2 units). If a certain uncertainty level and incompleteness level are acceptable, then the decision about the continuing functioning of an object is made (transition of the control to block 8.0). Otherwise, the decision on emergency stop of an object functioning is made (transition of the control to the blocks of 9.2, 10.2).

The strategy of system control of complex objects serviceability and safety is realized as an information platform of engineering diagnostics of the complex objects.

3. Information platform for engineering diagnostics of CES operation

The diagnostic unit, which is the basis of a safety control algorithm for complex objects in abnormal situations, is developed as an information platform [6] that contains the following modules:

- acquisition and processing of the initial information during the CEO operation;
- recovery of functional dependences (FDs) from empirical discrete samples;
- quantization of the discrete numerical values;
- identification of sensors failure;
- timely diagnosis of abnormal situations;
- forecast of nonstationary processes;
- generation of the process of engineering diagnostics.

Let us detail these modules of the information platform of engineering diagnostics (IPED).

Acquisition and processing of the initial information during the CEO operation. By a CEO we mean an complex engineering object consisting of several multi-type subsystems that are system-consistent in tasks, problems, resources, and expected results. Each subsystem has functionally interdependent parameters measured with sensors. To this end, groups of sensors are connected to each subsystem, each having different parameters (time sampling, resolution, etc.), depending on what its nature is.

The engineering diagnostics during the CEO operation requires samples of size N_{01} and N_{02} , where $N_{01}(N_{01} \gg 200)$ is the total sample size during the CEO real-mode operation; $N_{02}(N_{02} \ll N_{01}; N_{02} = 40 \div 70)$ is the size of the basic sample required for estimation the FDs. The initial information is reduced to a standard form, which makes it possible to form FDs from discrete samples. In view of the proposed methodology, biased Chebyshev polynomials are taken as basic approximating functions, which normalizes all the initial information to the interval $[0, 1]$.

Recovery of FDs based on Discrete Samples. The approximating functions are formed as a hierarchical multilevel system of models [6]. We will use the Chebyshev criterion and biased Chebyshev polynomials $T_{j_s p}(x_{j_s p}) \in [0,1]$. Such an approach reduces the procedure of forming the approximating functions to a sequence of Chebyshev approximation problems for inconsistent systems of linear equations [5].

Due to the properties of Chebyshev polynomials, the approach of formation the functional dependences makes it possible to extrapolate the approximating functions set up for the intervals $[\hat{d}_{j_s}^-, \hat{d}_{j_s}^+]$ to wider intervals $[d_{j_s}^-, d_{j_s}^+]$, which allows forecasting the analyzed properties of a product outside the test intervals.

Quantization of Discrete Numerical Values. The quantization is applied in order to reduce the influence of the measurement error of various parameters on the reliability of the solution being formed. The procedure of quantization of discrete numerical values is implemented as follows. As the base reference statistic for each variable $x_1, \dots, x_n, y_1, \dots, y_m$, the statistic of random samples in these variables of size $N_{01} \geq 200$ is taken. As the base dynamic statistic in the same variables, the statistic of the sample of the dynamics of the object for the last N_{02} measurements is taken. Therefore, the very first measurement of the original sample should be rejected and measurements should be renumbered in the next measurement $N_{02} + N_2$. Figure 2 schematizes the sample for the instant of time $t = t_0, N_{02} = 40$ and $t = t_0 + \Delta t$ ($t = 1, 2, 3, \dots, t_k, \dots, T$).

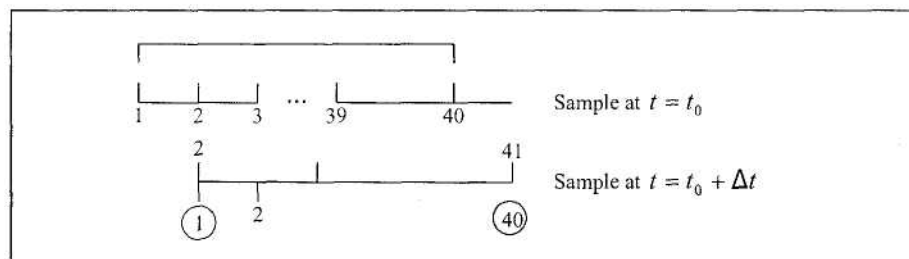


Figure 2. Sample at $t = t_0$ and $t = t_0 + \Delta t$

For the current dynamic parameters, we take the statistics of samples of size $N_{02} + N_2$ biased by N_2 with respect to the statistics of samples of size N_{02} .

Identification of sensors failure. Functioning of CEO involves monitoring the state of this system using various equipments, sensors, measuring devices. In this case, the recorded figures are not checked for validity in most cases. Often indicators of the transition system in abnormal or emergency mode of operation may be false. Thus, in this situation it is expedient to introduce procedures of identification possible failure of sensors.

A procedure of identification possible failure of sensors is based on the following thoughts. If the sensor functions are normal, each of his indications is not out of the threshold level. Any indication can be confirmed by previous and subsequent values. This is firstly connected with the nature of monitored processes: the majority of changes in the status process are not instantaneous. Therefore, the abrupt change in the sensor readings can be taken as evidence of failure of measuring devices. This approach is realized in the following way. At each step the

arithmetic mean P of previous y_{i-1} and subsequent y_{i+1} measurements is calculated. Then we compare this value P with current value y_i : $\Delta = P - y_i$. If deviation Δ exceed of threshold level, then operator displays a message about a possible sensor's failure.

Also failure of the sensors operation can be monitored by comparing the forecasted and actual results of measurements. Since the forecast follows the general behavior of the system, based on recent measurements, the deviation of the actual one may indicate the failure of sensors. Therefore, in operation of CEO a regular comparison of forecasts and their corresponding recovered values are implemented. As in the previous case, the deviation, which is greater than a threshold level, gives the message about the possible failure of the sensor.

Timely Diagnosis of Abnormal Situations. Timely diagnosis of abnormal situation is an important aspect of complex system. The possibility of prevention not only the result of abnormal situation, but also abnormal situation may reduce risk of disruption to a minimum. Therefore, several ways of work with abnormal situation was provided.

First, each value is obtained in a result of the recovery of functional dependence, compared with the threshold level of abnormal and emergency situation. In this case, having reached any of the indicators of such a boundary, a warning about the occurrence of the abnormal situation is appeared on the operator scoreboard. The reason of this situation and the current level of risk in the system are also indicated.

Also the reason of this situation and the current level of risk in the system are indicated. This approach allows us to monitor the immediate developments in the monitored system. From the standpoint of the algorithm it is described as fig. 3:

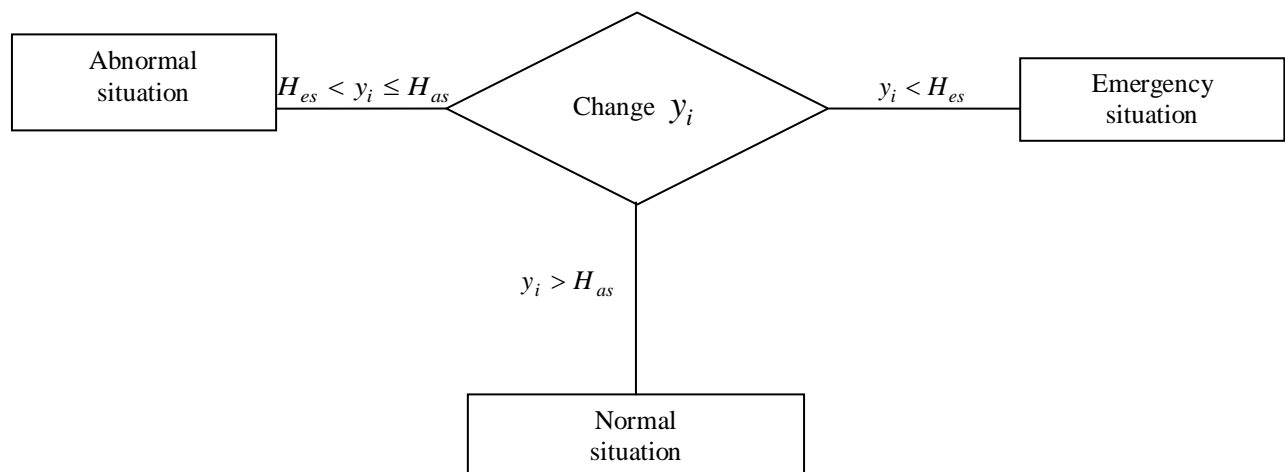


Fig. 3. Structural algorithm of timely diagnosis of abnormal situations

Secondly, for early detection of the possibility of the abnormal situation becoming a similar approach is used for the forecasted results. That is, after each step of forecasting, the obtained values are compared with the boundary values of the abnormal or emergency situation. The operator also receives all the information on the scoreboard. This allows to pass ahead of the abnormal situation for a few steps, and if possible to prevent it.

Thirdly, the operator can monitor the development of the system and to respond to negative trends in the process. For it the operator scoreboard displays the level of danger in the system and the current and predicted risk. The level of danger 7 means an emergency situation, levels 4, 5 and 6 indicates the abnormal one. Accordingly, levels 0–3 mean the normal functioning. In this case, as can be seen, the higher the risk, the closer the system to the emergency situation. If it is at danger level 3, the operator must be prepared for a possible deterioration of the process development and the transition to abnormal mode of operation. Using this approach operator gets several points of monitoring the system state. He can determine in advance the approximation of the emergency, he receives a warning when it is directly approaching, and he will be warned when the abnormal situation is happened, if it is not able to avoid it in time.

Forecast of Nonstationary Processes. Forecast of the critical parameters is done by using the method of coordinate descent and the degree of approximation to the sensors indicators according the algorithm:

- based on a number of values of input parameters currently using alternating-variable descent method, their predictive value in the future was calculated, system of equations has been solved

$$\begin{pmatrix} t^p & t^{p-1} & \dots & t^0 \\ t^p & t^{p-1} & \dots & t^0 \\ t^p & t^{p-1} & \dots & t^0 \\ t^p & t^{p-1} & \dots & t^0 \end{pmatrix} \cdot \begin{pmatrix} x_{1pr}^0 \\ x_{1pr}^1 \\ \dots \\ x_{1pr}^p \end{pmatrix} = \begin{pmatrix} x_1^0 \\ x_1^1 \\ \dots \\ x_1^p \end{pmatrix};$$

- with the found vector $(x_{1pr}^0 \ x_{1pr}^1 \ \dots \ x_{1pr}^p)^T$ predicted value index of the k-th sensor at time t_{i+1} is

$$\text{given by } x_{kpr}(t_{i+1}) = \begin{pmatrix} t^p & t^{p-1} & \dots & t^0 \end{pmatrix} \cdot \begin{pmatrix} x_{1pr}^0 \\ x_{1pr}^1 \\ \dots \\ x_{1pr}^p \end{pmatrix} = \sum_{j=0}^p t^j x_{1pr}^{p-j};$$

- the functional dependences were recovered, which was the desired predicted values of objective functions.

The forecast was realised on 10 steps ahead, as forecasts for more of them are not appropriate since it is accumulated the error of solving method of equations system. Functional dependences for the values of incoming parameters recovered at each step to ensure the most accurate forecast in order not to miss a moment of possible becoming abnormal situation.

Setting up the Process of Engineering Diagnostics. We will use the system of CES operation models to describe the normal operation mode of the object under the following assumptions and statements.

- Each stage of CEO operation is characterized by the duration and by the initial and final values of each parameter y_i determined at the beginning and the end of the stage, respectively. The variations of y_i within the stage are determined by the corresponding model.
- All the parameters y_i are dynamically synchronous and in phase in the sense that they are simultaneously (without a time delay) increased or decreased under risk factors.

- The control $U = (U_j | j = \overline{1, m})$ is inertialess, i.e., there is no time delay between the control action and the object's response.
- The risk factors $\rho_{q_k}^\tau | q_k = \overline{1, n_k^\tau}$ change the effect on the object in time; the risk increases or decreases with time.
- The control can slow down the influences of risk factors or stop their negative influence on the controlled object if the rate of control exceeds the rate of increase in the influence of risk factors. The negative influence of risk factors is terminated if the decision is made and is implemented prior to the critical time T_{cr} . At this moment the risk factors cause negative consequences such as an accident or a catastrophe.

To analyze an abnormal mode, let us introduce additional assumptions as to the formation of the model and conditions of recognition of an abnormal situation.

- The risk factors $\rho_{q_k}^\tau | q_k = \overline{1, n_k^\tau}$ are independent and randomly vary in time with a priori unknown distribution.
- The risk factors can influence several or all of the parameters y_i simultaneously. A situation of the influence of risk factors is abnormal if at least two parameters y_i simultaneously change, without a control, their values synchronously and in phase during several measurements (in time).
- The influence of risk factors will be described as a relative change of the level of control. The values of each risk factor vary discretely and randomly.

We will recognize risk situations by successively comparing $\tilde{y}_i[t_k]$ for $\tilde{y}_i[t_k]$ several successive values of $t_k, k = \overline{1, k_0}$, where $k_0 = 3 \div 7$. As follows of the assumptions, the condition of a normal situation is synchronous and in phase changes of \tilde{y}_i for several (in the general case, for all) parameters, whence follows a formula for different instants of time t_k for all of the values of i and for the same instants of time t_k for different values of i (different parameters):

$$\text{sign}\Delta\tilde{y}_i[t_1, t_2] = \dots = \text{sign}\Delta\tilde{y}_i[t_k, t_{k+1}] = \dots = \text{sign}\Delta\tilde{y}_i[t_{k_0-1}, t_{k_0}], \quad (1)$$

$$\text{sign}\Delta\tilde{y}_1[t_k, t_{k+1}] = \dots = \text{sign}\Delta\tilde{y}_i[t_k, t_{k+1}] = \dots = \text{sign}\Delta\tilde{y}_n[t_k, t_{k+1}], i = \overline{1, n}. \quad (2)$$

As follows from (1) and (2), given an abnormal situation on the interval $[t_1, t_{k_0}]$, the following inequalities hold simultaneously:

- the inequality of the signs of increment $\Delta\tilde{y}_i$ for all the adjacent intervals $[t_k, t_{k+1}]$ for $k = \overline{1, k_0}$ for each parameter $\tilde{y}_i, i = \overline{1, n}$;
- the inequality of the signs of increment $\tilde{y}_i, i = \overline{1, n}$, for all of the parameters \tilde{y}_i for each interval $[t_k, t_{k+1}], k = \overline{1, k_0}$.

4. Diagnostic of reanimobile's functioning

Contentive Statement of a Problem. The work of reanimobile, which moves in the operational mode, i.e. with the patient on board, is considered. Patient's life is provided with medical equipment, which is powered from the reanimobile's onboard electrical.

Basic equipment includes:

- ICE1 — basic internal combustion engine (ICE), which causes the car to move and rotate the main generator of G1;
- G1 — the main generator, with the capacity of 1.1KW that generates electricity when the angular velocity of crankshaft rotation is above 220 rad/sec (when the speed is above 220 rad/sec generator is switched on, when falls down 210 rad / s is off);
- TGB — transmission — gearbox (gear ratio: 1 — 4.05; 2 — 2.34; 3 — 1.39; 4 — 1; 5 — 0.85; main transmission — 5.125);
- ICE2 and T2 — auxiliary engine with a generator power of 1.1kW, which is used in emergency situations to provide power (standby ICE2 consumes fuel ICE2 0.5liters / hour);
- RB — rechargeable battery that provides power to the equipment when the generators do not generate electricity;
- PD — power distribution unit, which provides: battery charge, users' power from one of the generators, or from the battery, or the combination mode.

Tension in the on-board network depends on the generators and the level of battery charge. In the normal mode all equipment power is provided from the main generator and RB.

The main consumers, which are considered during the simulation:

- medical equipment, which consumes about 500 watts;
- illumination of the main cabin — 120 W;
- outdoor lighting (lights) — 110 W;
- car's own needs — 100 W.

Charge current is limited at the level that corresponds to the power extracted from the generator, equal to 200 watts. Reanimobile must travel a distance of 70 km. with a specific schedule of speed, which is formed by road situation.

It is required to ensure electric power for medical equipment, which is located in the main cabin. Since the motion is carried out at night, it is needed to provide additional coverage of the inner and outer. Kinematics parameters approximately correspond to the ambulances, based on GAZ.

Depending on the speed transmission, ratio is changed, therefore, the frequency of crankshaft rotation of the main internal combustion engine (ICE1) is changed. At the beginning of the way there are 47liters of fuel in the tank. Nutrition ICE1 and ICE 2 are from the same tank. In normal situation, the car safely drives patient for 11,700

seconds (3 hours and 15 minutes). In this case, the battery voltage does not decrease less than 11.85V. At the end of the way there are 4.1liters of fuel in the tank.

Transition into abnormal mode is caused by malfunction of the charger, voltage sensor RB. It is assumed that the sensor gives out false information that the battery is fully charged. Since recharging RB is not done, then with the lapse of time the battery is discharged, and, consequently, the voltage on-board network on the intervals of generator outages (when switching gears, ICE1 idling) will also be decreased. Due to deep discharge the mode is occurred when the output voltage RB is not enough to maintain the medical equipment operability and this is an emergency situation.

The Recognition of an Abnormal Situation. The recognition of an abnormal situation occurs in accordance with prescribed critical values.

1) For stress in the on-board network: abnormal is 11.7V, emergency is 10.5V

2) For the amount of fuel: abnormal is 21, and emergency is 11.

3) For the voltage at the rechargeable battery: an abnormal situation is 11.5V. Thus, while reducing the value of the function below one of the set values, the operation of reanobile goes to an abnormal mode of functioning.

In other words, if $Y_t < H$ critical exists, at the moment of time t CES functioning goes to an abnormal mode.

Where Y_t is a predicted value for the recovered functional dependence. On the diagrams, this process can be observed in the form of decreasing a prediction level (pink curve) below the threshold of the abnormal mode (blue line).

Critical variables:

- Board voltage (depending on the parameters of the RB, the generators condition, the load current). This option could lead directly to an emergency, if the board voltage drops below trip level of medical equipment.
- Fuel level. Depends on the power, which is taken off from the main engine (made in proportion to rotation speed). Decline below a certain point can lead to abnormal (when you can call another car or refueling, and catering equipment from RB) or emergency mode (when the car made a stop for a long time without charging).
- Voltage RB (depending on the generators condition, the total electricity consumption).

Real-time monitoring of the technical diagnostics is conducted in the reanimobile operation process with the purpose of timely exposure of potentially possible abnormal situations and guaranteeing the survivability of the system's functioning. In compliance with the developed methodology of the guaranteed CTO functioning safety at the starting phase $t = t_0$, functional recovery $y_i = f_i(x_1, \dots, x_j, \dots)$ is performed using $N_{02} = 50$ given discrete samples of values y_1, y_2, y_3 and their arguments. Here $y_1 = Y_1(x_{11}, x_{12}, x_{13}, x_{14}), y_2 = Y_2(x_{21}, x_{22}),$ and $y_3 = Y_3(x_{31}, x_{32}, x_{33}),$ where x_{11} is the measured voltage RB; x_{12} is the velocity of crankshaft rotation; x_{13} is power, which is provided by auxiliary generator; x_{14} is the total power consumption; x_{21} is the velocity of crankshaft rotation; x_{22} is power, which is provided by auxiliary generator; x_{31} is the velocity of crankshaft

rotation; x_{32} is power, which is provided by auxiliary generator; x_{33} is the total power consumption. All data on the variables $Y_i, i=1,2,3$ and their arguments $x_i, i=1,2,3$ are given as samples during the reanimobile's motion within 50000 seconds.

In this case, the voltage sensor gives false information about the voltage RB. When the voltage drops below 11.7V the diagnostic system provides a driver with the signal about an abnormal situation which can be developed into an emergency. The driver stops the car ($t=7323s$), switches on a standby generator ($t=7414s$) and eliminates the failure ($t=7863s$). Having recharged the battery from a standby generator when $t=8533s$, the driver turns off the standby generator and resumes the motion ($t=8623s$). Due to low battery, voltage at its terminals starts to decrease rapidly. The diagnostic system warns about abnormal situation again, to solve the problem the driver forcefully supports ICE1 speed at 250 rad/s, thus ensuring continued operation of the main generator.

As a result, fuel consumption is increased, which leads to the abnormal situation ($t=13000s$) when the amount of fuel is reduced to 1liter. At this moment of time the car is forcibly stopped by the signal of the diagnostics system (before reaching their destination) and a standby generator is switched on to provide the electric power supply (one liter of fuel is enough for 2 hours operation of standby generator that allows refuel the car or call for help).

The Risk Detection Procedure. Taking into account the specifics of operation of the system, following risk detection procedures were constructed. When reanimobile is functioning, possibility of abnormal situation is calculated with the formula

$$F(\rho_k) = 1 - (1 - \rho_{Gv})(1 - \rho_{Av})(1 - \rho_F),$$

where ρ_{Gv} is the probability that the board voltage drops below the emergency level; ρ_{Av} is the probability that the battery voltage drops below the emergency level; ρ_F is a probability that the fuel level drops below the emergency level. ρ_{Gv}, ρ_{Av} and ρ_F are calculated in the following way:

$$\rho_{Gv} = 1 - \left| (H_{1an} - y_{1pr}) / |1,75 * (H_{1an} - H_{1e})| \right|; \quad H_{1an} \neq H_{1e},$$

$$\rho_F = 1 - \left| (H_{2an} - y_{2pr}) / |1,75 * (H_{2an} - H_{2e})| \right|; \quad H_{2an} \neq H_{2e},$$

$$\rho_{Av} = 1 - \left| (H_{3an} - y_{3pr}) / |1,75 * (H_{3an} - H_{3e})| \right|; \quad H_{3an} \neq H_{3e},$$

where H_{1an} is board voltage in abnormal situations ($Y_{1r} \Leftarrow 11.7V$); y_{1pr} is the current board voltage (recovery functional dependence using forecast); H_{1e} is board voltage in an emergency ($Y_{1r} \Leftarrow 10.5V$); H_{2an} is the level of fuel in abnormal situations ($Y_{2r} \Leftarrow 1L$); y_{2pr} is the current value of the fuel (recovery functional dependence using forecast);

H_{2e} is the level of fuel in an emergency ($Y_{2r} = 0$); H_{3am} is a battery voltage in the abnormal mode ($Y_{3r} \Leftarrow 11.7B$); y_{3pr} is the current battery voltage ((recovery functional dependence using forecast); H_{3e} is a board voltage in an emergency ($Y_{3r} \Leftarrow 10.5V$).

This structure of risk was taken on the basis of the normalization behavior of the process in the interval (0,1). Design formula repelled by conditions: the risk during the emergency must be equal to 1, the risk at the border of abnormal mode should be equal to 0,4. In the result, the risks on all fronts are taken into account. The overall risk is 1 during the damage 0,5–0,6 at the border of the abnormal mode.

Some results of reanimobile's functioning during the first 7000 sec. are shown in Fig. 4 as the diagrams of stress distribution of the on-board network, the amount of fuel in the tank, the rechargeable battery voltage. The transition into abnormal mode happens due to failure of the sensor battery voltage. So far as the battery recharging is not implemented, the battery is discharged with the lapse of time and, consequently, the voltage in the on-board network in the period of 6500-7400 sec. is also decreased and transits into abnormal mode. The fuel level, which depends on the capacity of the internal combustion engine, is also reduced.

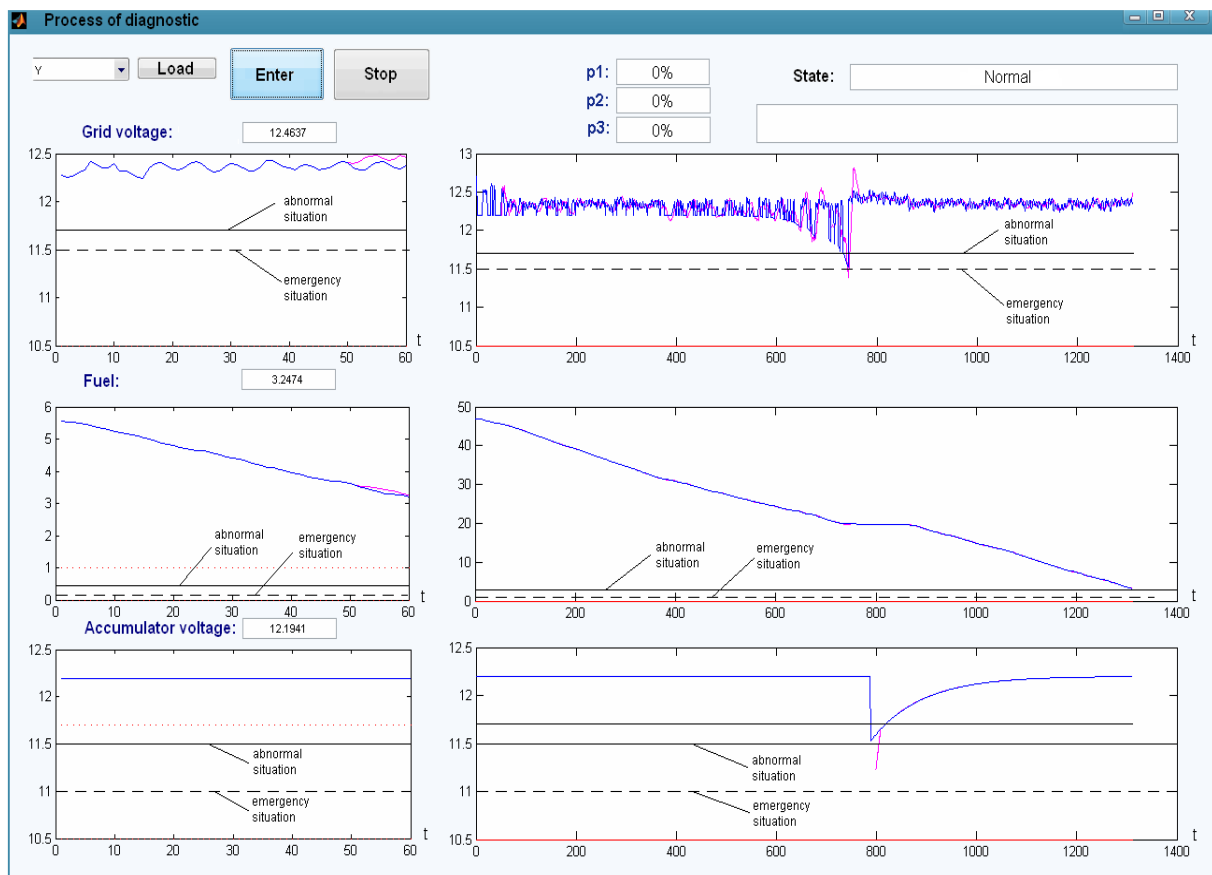


Figure 4. Stress distribution of the on-board network, the amount of fuel in the tank, the rechargeable battery voltage in accordance of time t sec.

At any time of the program operation user has the ability to look at the operator scoreboard, which displays a series of indicators that reflects the character of the state of CEO of the reanimobile functioning. This is such indicators as: indicators of sensors accumulator battery voltage, fuel quantity in the tank, the voltage on-board network, the state of the system, the risk of the damage, the causes of the abnormal or emergency mode, as well as the indicator of the danger level of the system operation and possible failure of sensors.

4. Conclusion

System coordination of survivability and safety control by the goals, objectives, resources and expected results, as well as by efficiency and effectiveness of interaction in the real conditions of abnormal situations allows to provide the effective and efficient interaction of these control systems. On the one hand, it is ensured the efficiency and effectiveness of safety systems according to timely detection of abnormal situations, estimation of its degree and level of risk, definition of the margin of permissible risk in the process of forming the recommendations for the prompt actions of the DM. On the other hand, the survivability control system must effectively and efficiently operate after receiving a signal about the abnormal situation to ensure the availability of a complex object for the emergency transition into abnormal mode and provide its realization within a margin of permissible risk.

The proposed strategy of system coordination of survivability and safety engineering objects operation, implemented as a tool of information platform of engineering diagnostics of the complex objects, ensures the prevention of inoperability and the danger of object's functioning. By force of systematic and continuous evaluation of critical parameters of object's functioning in the real time mode, the reasons, which could potentially cause the object' tolerance failure of the functioning in the normal mode, are timely revealed. For situations, development of which leads to possible deviations of parameters from the normal mode of the object's functioning, it is possible timely to make a decision about the change of the operation mode of the object, or an artificial correction of the parameters to prevent the transition from the normal mode into the abnormal one, accident and catastrophe.

The principles, which are included in the implementation of the guaranteed safety of CEO operation strategy, provide a flexible approach to timely detection, identification, forecasting and system diagnosis of factors and risk situations, formation and implementation of sustainable solutions during the acceptable time within the fatal time limit.

Bibliography

- [1] Frolov K. V. (gen. ed.), Catastrophe Mechanics [in Russian], Intern. Inst. for Safety of Complex Eng. Syst., Moscow —1995. —389 p.
- [2] Troshchenko V. T. (exec, ed.), Resistance of Materials to Deformation and Fracture: A Reference Book, Pts. 1, 2 [in Russian], Naukova Dumka, Kyiv. —1993, 1994. —702 p.

-
- [3] Pankratova N. and Kurilin B., Conceptual foundations of the system analysis of risks in dynamics of control of complex system safety. P. 1: Basic statements and substantiation of approach // J. Autom. Inform. Sci. —2001. —33, №. 2. —P. 15-31.
- [4] Zgurovsky M. Z., Pankratova N. D., System Analysis: Theory and Applications, Springer, Berlin. —2007. — 475 p.
- [5] Pankratova N.D., "System strategy for guaranteed safety of complex engineering systems", Cybernetics and Systems Analysis 46, 2 (2010): 243-251.
- [6] Pankratova N.D., "System approach to estimation of guaranteed safe operation of complex engineering systems", International Book Series «Information science&computing». –New Trends in Information Technologies. ITHEA. SOFIA (2010):115-128.

Authors' Information



Nataliya Pankratova – DTs, Professor, Depute director of Institute for applied system analysis, National Technical University of Ukraine "KPI", Av. Pobedy 37, Kiev 03056, Ukraine; e-mail: natalidmp@gmail.com

Major Fields of Scientific Research: System analysis, Theory of risk, Applied mathematics, , Applied mechanics, Foresight, Scenarios, Strategic planning, information technology

INDIRECT APPROACH OF DETERMINATION OF COLLECTIVE ALTERNATIVE RANKING ON THE BASIS OF FUZZY EXPERT JUDGEMENTS

Pavlo P. Antosiak, Oleksij F. Voloshin

Annotation: The article suggests methods for determining the collective ranking based on the indirect approach. We consider the case of fuzzy expert preferences given in the form of matrices of fuzzy tournaments and also the case of ordinal fuzzy expert assessments. For aggregation used method for calculating the linguistically quantized speech as well as OWA operator. The first method makes it possible to do without the complex optimization problems that arise in group decision making. Another method can be used for direct ranking of alternatives by experts.

Keywords: fuzzy expert judgements, group decision making, group ranking, indirect approach.

ACM Classification Keywords: H.4.2 Information Systems Applications: Types of Systems: Decision Support.

Introduction

The methods of group decision making which were called as the collective expert judgement are increasingly frequently used in the applied mathematics and different spheres of human activity. The peculiarity of collective expert judgement as the scientific tool for solving of complex slightly structured problems is fuzziness which is appropriate to the expert judgements.

Setting of the problem

The problem of collective ordinal expert judgment is considered in the following setting [Тоценко, 2006].

Given: finite set of alternatives $A = \{a_1, \dots, a_{n_A}\}$; qualitative criterion of alternatives judgment (K); normalized

coefficients $\alpha = \{\alpha_1, \dots, \alpha_{n_E}\}$, $l \in N_E = \{1, \dots, n_E\}$, $\sum_{l=1}^{n_E} \alpha_l = 1$, of expert competence concerning the subject of expertise.

To find: collective alternative ranking of set A according to criterion K , which generalizes the opinion of all experts in the best way and is agreed taking into consideration the expert competence.

Indirect approach provides at least two stages: the stage of expert information aggregation and the stage of decision making. The aggregated (collective, group, social, agreed, etc.) fuzzy P_C is formed on the aggregation stage. The best alternative or resulting alternative ranking is determined on the stage of decision making on the

basis of agreed judgement P_C . It is possible to realize the aggregation stage in different ways according to the kind of individual judgements.

Method of fuzzy collective ordinal judgement determination of alternatives on the basis of fuzzy expert matrixes of paired comparisons

On the aggregation stage of fuzzy expert information from fuzzy individual preferences, given by the experts in the form of paired comparison matrixes, the fuzzy collective preference P_C in the form of matrix with elements $\mu_{ij}^{(C)}$, $i, j = 1, \dots, n_A$ is built. Each value $\mu_{ij}^{(C)} \in [0, 1]$ expresses the level of confidence about preference of alternative a_i over alternative a_j of expert group in general.

For aggregation of individual preferences we use an approach which is based on the fuzzy majority [Kacprzyk, 1985i, Kacprzyk, 1985ii], according to which we have:

$$\mu_{ij}^{(C)} = \mu_Q \left(\frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} \right) \quad (1)$$

where $\mu_Q(\cdot)$ is membership function of fuzzy quantifier Q , $\mu_{ij}^{(l)}$ is confidence degree about the alternative preference a_i over a_j in the opinion of the l st expert.

Statement 1. Let all individual preferences are fuzzy tournaments. If Q is non-decreasing linguistic quantifier with such data as (a, b) , that $a + b = 1$, then collective preference built according to the rule (1), is also fuzzy tournament.

Proof. Let's choose the arbitrary indexes $i, j \in N_A$.

Let $\frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} \leq a$, then $\mu_{ij}^{(C)} = 0$.

$$\begin{aligned} \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} \leq a &\Rightarrow \frac{1}{n_E} \sum_{l=1}^{n_E} (1 - \mu_{ji}^{(l)}) = 1 - \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} \leq a \Rightarrow \\ \Rightarrow 1 - a &\leq \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} \Rightarrow b \leq \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} \Rightarrow \mu_{ji}^{(C)} = 1 \end{aligned} \quad (2)$$

Suppose now $a < \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} < b$. Taking into consideration the equivalent transformations in (2), we have the following implications:

$$\begin{aligned} a < 1 - \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} < b &\Leftrightarrow a - 1 < -\frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} < b - 1 \Leftrightarrow \\ \Leftrightarrow 1 - b < \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} < 1 - a &\Leftrightarrow a < \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} < b \end{aligned} \quad (3)$$

Further, taking into consideration (3), we have:

$$\mu_{ij}^{(Q)} + \mu_{ji}^{(Q)} = \frac{\frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} - a}{b-a} + \frac{\frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ji}^{(l)} - a}{b-a} = \frac{\frac{1}{n_E} \sum_{l=1}^{n_E} (\mu_{ij}^{(l)} + \mu_{ji}^{(l)}) - 2a}{b-a} = \frac{1-2a}{b-a} = 1.$$

If $\frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ij}^{(l)} \geq b$, then analogous to the previous case one can make sure that this statement is correct. \square

Another approach to aggregation of individual preferences can be the use of various kinds of operators of information aggregation. Let's examine the most reasonable and often used in practice family of aggregation operators. Ordered weighted averaging operator (OWA operator) was suggested by R. Yager in the work [Yager, 1988] and later more studied and characterized in [Yager, 1993]. OWA operator is commutative, idempotent, continuous, steady, neutral, equilibrated and stable relatively to lineal transformations but in general case is nonassociative. OWA operator accepts values from interval between the values of operators $Min(\cdot)$ and $Max(\cdot)$. Fundamental aspect of OWA operator is reordering of arguments in accordance with an importance (signification) of their values. R. Yager [Yager, 1993] defined the induced ordered weighted averaging operator (OWA operator) as the generalization of OWA operator for a case when information about competence of experts in form of crisp weight number is available. The same fuzzy principle of majority [Yager, 1994] is suggested to be used for calculation of weight numbers of OWA operator.

When using OWA operator for aggregation of individual expert preferences the following result is fair.

Statement 2 [Chiclana, 2003, p. 74]. Let Q is non-decreasing linguistic quantifier with such data as (a, b) that $a + b = 1$. Then OWA operator managed with the quantifier Q retains the property of additivity.

For building an indirect collective ranging of alternatives on the basis of fuzzy collective preference we use the method suggested in the work [Скофенко, 1983]. It consists in the fact that on the basis of the available matrix of preferences $P = (\mu_{ij})_{i,j=1,\dots,n_A}$ it is possible to define the judgements of truth (assurance) of more difficult propositions concerning the alternatives, that is of the following propositions ω_{ik} , $k = 0, \dots, n_A - 1$, $\forall i \in N_A$:

$$\omega_{ik} = \text{"alternative } a_i \text{ and is better than } k \text{ alternative from set } A\text{"}.$$

If in the quality of t -norm, t -conorm and investor occur $\min\{\cdot, \cdot\}$, $\max\{\cdot, \cdot\}$, $1 - \cdot$ then determination of truth degree of proposition ω_{ik} comes to the following rule [Скофенко, 1983]:

$$\mu(\omega_{ik}) = \begin{cases} 1 - \mu_{ip_1}, & \text{if } k = 0, \\ \min\{\mu_{ip_k}, 1 - \mu_{ip_{k+1}}\}, & \text{if } 0 < k \leq n_A - 2, \\ \mu_{ip_{n_A-1}}, & \text{if } k = n_A - 1, \end{cases}$$

where $\mu_{ip_k} \geq \mu_{ip_{k+1}}$, $k = 1, \dots, n_A - 1$.

Method of fuzzy collective ordinal judgement determination on the basis of fuzzy expert ordinal judgements

Rather wide spread procedure of expert information gaining is direct ranking of alternatives. Expert is proposed all set of alternatives for judgement and he is proposed to put them in order according to preference. Direct ranking of alternatives can be realized by different ways [Литвак, 1996]. But in general case ordinal judgements given by an expert can be fuzzy. In a quality of fuzzy ordinal judgement can be the following fuzzy propositions of an expert:

- place (rank) of alternative a_i is nearly r_i ;
- a_i is nearly within the limits from $r_i^{(1)}$ to $r_i^{(2)}$;
- fuzzy propositions which contain linguistic variable "rank".

We formalize for our problem first two fuzzy judgements in form of triangular fuzzy number and trapezoidal fuzzy number accordingly.

$$\mu_r(x) = \begin{cases} 0, & \text{if } x < r - \frac{n_A}{10}, \\ 10 \cdot \frac{x-r}{n_A} + 1, & \text{if } r - \frac{n_A}{10} \leq x < r, \\ 10 \cdot \frac{r-x}{n_A} + 1, & \text{if } r \leq x < r + \frac{n_A}{10}, \\ 0, & \text{if } x \geq r + \frac{n_A}{10}. \end{cases}$$

$$\mu_{r_1 r_2}(x) = \begin{cases} 0, & \text{if } x < r_1 - \frac{n_A}{10}, \\ 10 \cdot \frac{x-r_1}{n_A} + 1, & \text{if } r_1 - \frac{n_A}{10} \leq x < r_1, \\ 1, & \text{if } r_1 \leq x < r_2, \\ 10 \cdot \frac{r_2-x}{n_A} + 1, & \text{if } r_2 \leq x < r_2 + \frac{n_A}{10}, \\ 0, & \text{if } x \geq r_2 + \frac{n_A}{10}. \end{cases}$$

As it is noted in the work [Рыжов, 1998], it is easier for specialists in the applied problems where the expert judgements are widely used to formulate them in the terms of natural language. Such propositions of an expert are possible to formalize through the linguistic variable which is described by the tuple $\langle X, T(X), U, G, M \rangle$. Here X is the name of linguistic variable which reflects some object; $T(X)$ is the set of values or terms of this variable which are the names of fuzzy variables; U is a set, which is the branch of terms definition; G is syntactic procedure (grammar), which describes the process of creation of set elements $T(X)$ of new values of linguistic variable; M is semantic procedure which allows to ascribe to each new meaning of linguistic variable some semantics by means of formation of corresponding fuzzy set. For our case: $X = \text{"rank"}$; $T(X) = \{\text{"high", "middle", "low"}\}$; $U = [1, n_A]$; $G = \{\text{"very", "more or less", "not", "and", "or"}\}$; as the semantic rules we use the above mentioned rules for logic connection and negation. Membership functions of the corresponding terms can be defined in the following way:

$$\mu_{high}(x) = \begin{cases} 0, & \text{if } x < 1, \\ 1, & \text{if } 1 \leq x < \frac{n_A}{10} + 1, \\ 10 \cdot \frac{n_A - 3x + 3}{7n_A}, & \text{if } \frac{n_A}{10} + 1 \leq x < \frac{n_A}{3} + 1, \\ 0, & \text{if } x \geq \frac{n_A}{3} + 1. \end{cases}$$

$$\mu_{middle}(x) = \begin{cases} 0, & \text{if } x < \frac{n_A}{4}, \\ \frac{4x - n_A}{n_A}, & \text{if } \frac{n_A}{4} \leq x < \frac{n_A}{2}, \\ \frac{3n_A - 4x}{n_A}, & \text{if } \frac{n_A}{2} \leq x < \frac{3}{4}n_A, \\ 0, & \text{if } x \geq \frac{3}{4}n_A. \end{cases}$$

$$\mu_{low}(x) = \begin{cases} 0, & \text{if } x < \frac{2}{3}n_A, \\ 10 \cdot \frac{3x - 2n_A}{7n_A}, & \text{if } \frac{2}{3}n_A \leq x < \frac{9}{10}n_A, \\ 1, & \text{if } \frac{9}{10}n_A \leq x \leq n_A, \\ 0, & \text{if } x > n_A. \end{cases}$$

Thus an alternative of paired comparison of alternatives (in literature such method of expert judgements giving got the name of giving "object-object"[Cook, 1983]) can be the method "object-rank" of expert judgements giving. As a result of such approach the experts evidently or implicitly form their individual judgements in the form of matrixes $P_l = (\mu_{ik}^{(l)})_{i,k=1,\dots,n_A}$, $l = 1, \dots, n_E$, elements of which show the truth degree of proposition

$$\omega_{ik} = \text{"alternative } a_i \text{ has rank } k \text{"}.$$

On the stage of expert information aggregation on the basis of available fuzzy ordinal individual judgements we define the truth degree of the following fuzzy proposition:

$$\omega_{ik}^{(C)} = \text{"most experts consider that alternative } a_i \text{ has rank } k \text{"}.$$

Truth degree of such fuzzy proposition is calculated according to the following rule:

$$\mu_{ik}^{(C)} = \begin{cases} 0, & \text{if } \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ik}^{(l)} \leq 0.3, \\ \frac{2}{n_E} \sum_{l=1}^{n_E} \mu_{ik}^{(l)} - 0.6, & \text{if } 0.3 < \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ik}^{(l)} < 0.8, \\ 1, & \text{if } 0.8 \leq \frac{1}{n_E} \sum_{l=1}^{n_E} \mu_{ik}^{(l)} \leq 1, \end{cases}$$

where $\mu_{ik}^{(l)} = \mu_l(\omega_{ik})$, $\mu_l(\cdot)$ is the membership function of corresponding fuzzy rank, given by the l st expert.

By collective fuzzy ranking of the set of alternatives A we shall understand the set of all fuzzy subsets A_i , $i = 1, \dots, n_A$, which are determined by the values $\mu_{ik}^{(C)}$, $k = 1, \dots, n_A$, and correspondingly with the following membership functions [Скофенко, 1983]:

$$\mu_{A_j}(k) = \frac{\mu_{ik}^{(C)}}{\max_{k=1, \dots, n_A} \mu_{ik}^{(C)}}. \quad (4)$$

Approaches to the definition of strict collective ranking of alternatives on the basis of fuzzy collective ordinal judgement

It is known [Скофенко, 1983], that in the case when the matrix of fuzzy preference is the matrix of fuzzy tournament then the corresponding fuzzy ranking has the property of prominence in the sense of fuzzy set promonence. At the cut A_j , which comes into fuzzy ranking according to all values of membership degree, the single segment in the set of crisp ranks will be put in correspondence to each alternative. If the matter is about the fuzzy individual rankings and if it is impossible to define crisp resulting rankings on the basis of α -cut, the experts are proposed to overview their judgements, after which the above described approach is used again.

One can use the following approach for definition of crisp strict ranking which is in some sense "the closest" to the fuzzy collective ranking. To the crisp ranking of alternative by giving "object-rank" expert judgements evidently corresponds matrix $X = (x_{ik})_{i,k=1, \dots, n_A}$, elements of which satisfy the conditions $x_{ik} \in \{0,1\}$,

$\sum_{i=1}^{n_A} x_{ik} = \sum_{k=1}^{n_A} x_{ik} = 1$, $i, k = 1, \dots, n_A$. If as the proximity measure between fuzzy rankings Hamming distance is taken between the corresponding matrixes, then the following arrangement is justified:

$$\sum_{i=1}^{n_A} \sum_{k=1}^{n_A} |x_{ik} - \mu_{ik}^{(C)}| \rightarrow \min,$$

$$\sum_{i=1}^{n_A} x_{ik} = \sum_{k=1}^{n_A} x_{ik} = 1, \quad x_{ik} \in \{0,1\}, \quad i, k = 1, \dots, n_A.$$

Example

Let each of the expert group $\{e_1, e_2, e_3, e_4\}$ makes direct ranking of seven alternatives of set A . Result of carried out judgements is shown in table 1.

Table 1. Fuzzy ordinal expert judgements

	e_1	e_2	e_3	e_4
a_1	high	nearly 1	average	nearly 2
a_2	nearly 3	high	not low and not very high	average
a_3	average or low	not very high	low	very low
a_4	not low	small	nearly 2	Very high
a_5	Not very high	average	not very low	nearly 4
a_6	Within limit of [5,7]	not high	high	not very high
a_7	low	nearly 6	average	Not high

We calculate value $\mu_{ik}^{(C)}$ for $i, k = 1, \dots, 7$ on the basis of expert rankings (table 1) and put them into the table 2.

Table 2. Collective fuzzy ordinal judgement

	1	2	3	4	5	6	7
a_1	0.543	0.522	0.002	0	0	0	0
a_2	0	0.189	0.838	0.4	0	0	0
a_3	0	0	0.236	0.257	0.125	1	1
a_4	0.543	0.733	0.064	0	0	0	0
a_5	0	0.138	0.879	1	0.593	0.067	0
a_6	0	0.067	0.379	0.543	0.9	0.9	0.9
a_7	0	0	0.155	0.257	0.216	0.808	0.543

Then according to the equation (4) we calculate value $\mu_{A_i}(k)$, $i, k = 1, \dots, 7$. Fuzzy ranking, given in the form of membership function is showed on the figure 1.

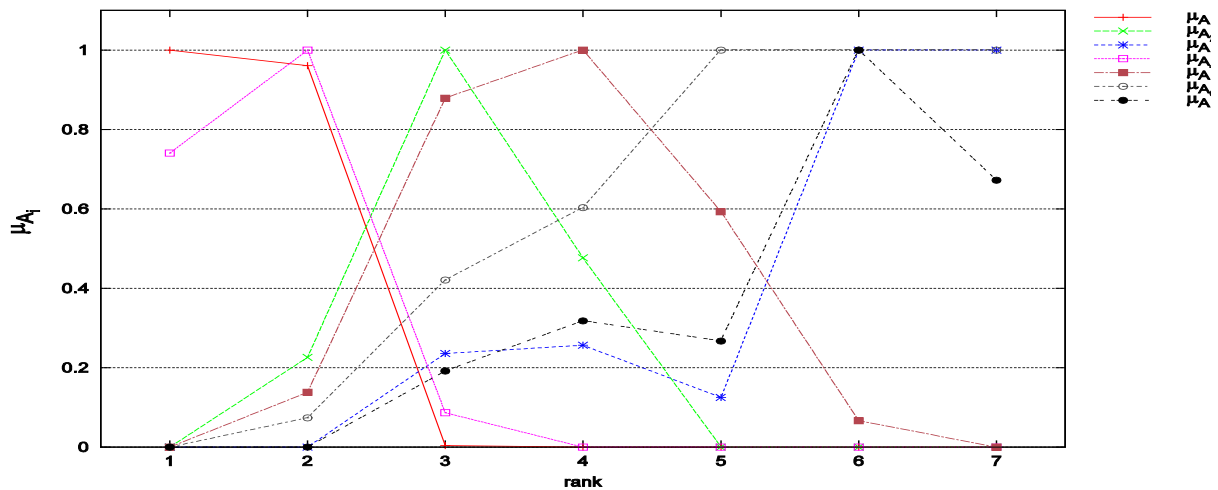


Figure 1. Fuzzy collective ranking of the set of alternatives A

On the basis of cut of fuzzy collective ranking according to the highest membership degree as the solution of our model problem we get crisp strict ranking of alternatives to which corresponds permutation of alternative indexes (1,4,2,5,6,7,3).

Conclusions

Indirect approach developed in this work can serve as an alternative to the direct approach, realized by the authors before in the work [Антосяк, 2010].

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA Bulgaria www.ithea.org and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine www.aduis.com.ua.

Bibliography

- [Тоценко, 2006] Тоценко В. Г. Об унификации алгоритмов организации экспертиз // Проблемы правовой информатизации, 2006, 2 (12). – С. 96-101.
- [Kacprzyk, 1985i] Kacprzyk J. Group decision making with a fuzzy majority via linguistic quantifiers. part i: A consensory-like pooling // Cybernetics and Systems, 1985, 16. – P. 119-129.
- [Kacprzyk, 1985ii] Kacprzyk J. Group decision making with a fuzzy majority via linguistic quantifiers. part ii: A competitive-like pooling // Cybernetics and Systems, 1985, 16. – P. 131-144.
- [Yager, 1988] Yager R. R. On ordered weighted averaging aggregation operators in multicriteria decision making // Systems, Man and and Cybernetics, Part A: Systems and Humans, 1988, 18. – P. 183-190.
- [Yager, 1993] Yager R. R. Families of owa operators // Fuzzy Sets and Systems, 1993, 59, 2. – P. 125-148.
- [Yager, 1994] Yager R. R. Quantifier guided aggregation using owa operators // International Journal of Intelligent Systems, 1994, 11. – P. 49-73.
- [Chiclana, 2003] F. Chiclana, F. Herrera, E. Herrera-Viedma, L. Martinez. Anote on the reciprocity in the aggregation of fuzzy preference relations using owa operators // Fuzzy Sets and Systems, 2003, 137. – P. 71-83.
- [Скофенко, 1983] Скофенко А. В. Применение нечеткой логики при ранжировании объектов методом парных сравнений // Кибернетика, 1983, 3. – С. 116-118.
- [Литвак, 1996] Литвак Б. Г. Экспертные оценки и принятие решений. И-во «Патент», Москва 1996. – 271 с.
- [Рыжов, 1998] Рыжов А. П. Элементы теории нечетких множеств и измерения нечеткости. И-во «Диалог-МГУ», 1998. – 116 с.
- [Cook, 1983] W. D. Cook, L. Seiford, S. Warner. Preference ranking models: Conditions for equivalence // Journal of Mathematical Sociology, 1983, 9. – P. 125-127.
- [Антосяк, 2010] Антосяк П. П. Узагальнення медіанного підходу на випадок нечітких індивідуальних переваг // Вісник Київського університету. Серія: фіз.-мат. науки, 2010, 2. – С. 81-86.

Authors' Information



Oleksij Voloshin – professor, TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV, Faculty of Cybernetics. Kyiv, Ukraine; e-mail: ovoloshin@unicyb.kiev.ua
 Major Fields of Scientific Research: decision theory, optimization methods, Mathematical Economics, decision support system, expert systems.



Pavlo Antosiak – assistant lecturer, National University of Uzhgorod, Department of Mathematics. Uzhgorod, Ukraine; e-mail: antosp@ukr.net
 Major Fields of Scientific Research: group decision making, methods of discrete optimization, soft computing.

SELECTIVE EVOLUTION CONTROL METHOD FOR EVOLUTION STRATEGIES WITH NEURAL NETWORK METAMODELS

Pavel Afonin

***Abstract.** This paper presents a new evolution control method to reduce the number of computationally expensive simulations for evolution strategies with fitness function models. A feedforward neural network is used as a fitness model and constructed with the help of some previously evaluated solutions in the search space. On-line learning is implemented during searching process. In the evolution strategy with the proposed method the number of controlled individuals is changed during optimization and the choice of parents for the next generation is always made out of controlled individuals. The results of the evolution strategy implementation with the selective evolution control method for three standard test functions in comparison with other known evolutionary strategies are presented.*

***Keywords:** evolution strategy, neural network, metamodel, evolution control*

***ACM Classification Keywords:** I.2 Artificial Intelligence: I.2.6 Learning: Connectionism and neural nets*

***Conference topic:** Neural Networks*

Introduction

During the last several years evolutionary algorithms have found wide application for solving a great number of design optimization problems, simulation optimization problems as well as other complex problems demanding applying global optimization methods. However, in the most cases the large number of function evaluations are required for a evolutionary algorithms to converge a near-optimal solution.

One of the ways of solving the problem given is using approximate models (metamodels) instead of computationally expensive fitness function evaluations. The polynomial models [1, 8], artificial neural networks include multilayer perceptrons [4, 6], radial-basis-function networks [8] and support vector machines [9] as well as the kriging models [2, 3] can be employed as fitness function models in evolutionary algorithms (for example, evolution strategies and genetic algorithms).

In this paper a selective evolution control method for evolution strategies based on metamodels are proposed and investigated. A multilayer perceptron is used as a metamodel and on-line learning is implemented during searching process.

The remaining part of the paper is devoted to: A brief review of the methods described in the literature of approximate model incorporation into evolutionary algorithms is presented in section II. Section III introduces the selective evolution control method proposed. Section IV presents experimental results from simulations on three benchmarks. Section V summarizes paper conclusion and planning for future research.

Related works

There are two approaches to integrate metamodels into evolutionary algorithms: surrogate approach and evolution control.

In the surrogate approach at first the optimum of the metamodel is determined and after that evaluated on the real fitness function. The new evaluation is used to update the model, and the process of the metamodel improving is repeated.

In concept of evolution control propose two methods [6]: controlled individuals and controlled generations. In generation-based evolution control, all individuals in population is evaluated on either the metamodel or the fitness function. In individual-based control, part of the individuals in current population are chosen and evaluated with the real fitness function. Remaining individuals are evaluated with the approximate model. Individuals that are evaluated on the fitness function call as controlled individuals.

The main issue in the individual-based evolution control is to define which individuals in the each generation are evaluated with the real fitness function and which with the approximate model [1, 3, 6, 9].

Then describe two main individual-based evolution control methods: the best strategy and pre-selection strategy.

In the best strategy [6], $\lambda' = \lambda$ offspring are estimated with the fitness model and the λ^* best ones are evaluated with the real fitness function. After model construction the remaining $\lambda' - \lambda^*$ individuals are evaluated again with the fitness model. The μ best individuals from the λ individuals become parents of the next generation.

In the pre-selection strategy [9], $\lambda' > \lambda$ offspring are generated out of μ parents through recombination and mutation and after that estimated with the fitness model. The $\lambda^* = \lambda$ best individuals are pre-selected from the λ' offspring and re-evaluated with the real fitness function.

The important conclusion in [4] is that the stability of the evolution strategy with individual based evolution control might be improved if the parents for the next generation are selected out of controlled individuals, as it is done in the pre-selection strategy.

However, at present there remains actual one question: how many and which individuals must be controlled?

Selective evolution control method

The model fidelity can be changed from one generation to the next one due to the change of the region where population is located as well as data change for model construction. Therefore, selection quality of the model evaluated in the current generation could be invalid for predicting the number of controlled individuals in the next generation.

One of the quality criteria of the model is the rank correlation prank [7], which in turn depends on the difference between the rank of the offspring individual based on the real fitness function and on the approximate model.

It may be expected, that if at first we could evaluate the model quality in current generation and after that employ this evaluation for determining controlled individuals in the same generation than this method might prove to be effective for correct selection of parents to the next generation.

The main idea of approach is that controlled individuals should be chosen from the current generation depending on the quality of the model which should be evaluated in the same generation by means of evaluating rank difference for some individuals taken as small separate units.

For this purpose number of controlled individuals (η), which must have the best rank among all λ individuals of current generation is introduced. In this case the number of controlled individuals λ^* for each generation may be from η to λ .

For more stable work of the selection operator the condition $\eta \geq \mu$ is introduced, which means that the selection of μ parents for the next generation must be made out of controlled individuals.

If model quality is lower then high probability exists that the first η individuals will change their rank. In this case the number of controlled individuals is increased. If model quality is higher then low probability exists that the first η individuals will change their rank and so the number of controlled individuals is decreased.

Let us consider graphical presentation of evolution strategy with the selective evolution control (figure 1). At first, all λ offspring is generated out of μ parents of the current generation by means of recombination and mutation. After that all λ offspring estimated with the model. Further individuals are evaluated with the real fitness function. At the Step 1 the η best individuals from the λ offspring are evaluated. The first individual changing rank 1 for rank 2 remains within the η best individuals and the second one goes out of η best individuals changing rank 2 for rank 6. At the Step 2 the only one non-controlled offspring from η best individuals is evaluated, thus 1-st individual changes rank 1 for rank 4 and the 2-nd individual of rank 2 for rank 1. At the Step 3 only the 2-nd individual is evaluated and as a result it changes rank 2 for rank 1 and the 1-st individual does from rank 1 for rank 2. So all η best individuals are controlled. Further the model is updated taking into consideration the λ^* controlled individuals in current generation. In conclusion the μ best individuals are selected only from the η controlled individuals.

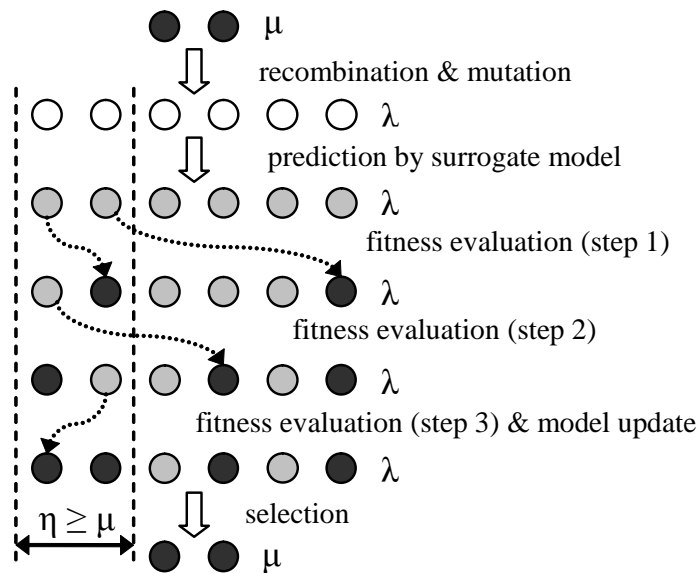


Figure 1. Evolution strategy with selective evolution control

Thus, in the evolution strategy with the proposed evolution control method the number of controlled individuals λ^* for each generation depending on the quality of the model for same generation and the choice of μ parents for the next generation is always made out of controlled individuals.

Experiments on benchmarks

The (3, 12) evolution strategy with covariance matrix adaptation [5] is taken in this work. Three test functions: 12D Ackley function, 12D Rosenbrock function and 12D Schwefel function are carried out for the investigation. The evolution strategy proposed (Sel ES) is compared with the three evolution strategies: pure evolution strategy (pure ES), pre-selection strategy (PreSel ES) and best strategy (BS ES).

The values of λ' and λ are based on recommendations [4, 9] equals for pre-selection strategy: $\lambda'=12$; $\lambda=24$ and for best strategy: $\lambda'=6$; $\lambda=12$. The main parameter of selective evolution control method is equal 3 ($\eta = \mu$). The neural network consists of 12 inputs, one hidden layer with 8 hidden neurons and one output. According to the recommendations [4, 9] for achieving good approximation in the current local region only data from 4λ to 5λ of last fitness evaluations are used for training neural network.

Program realization of the algorithms and research are made with Matlab 7.1. In Figures 2, 3, 4 (for three test functions) the median of the best fitness in each generation over 25 runs are showed against the number of fitness evaluations.

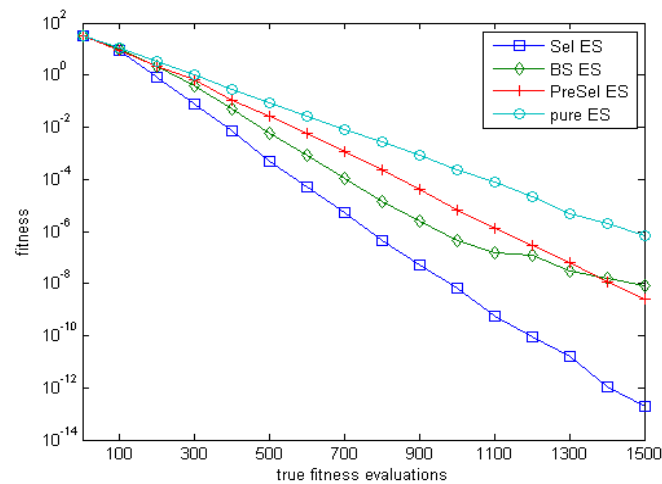


Figure 2. Results for the Sphere function

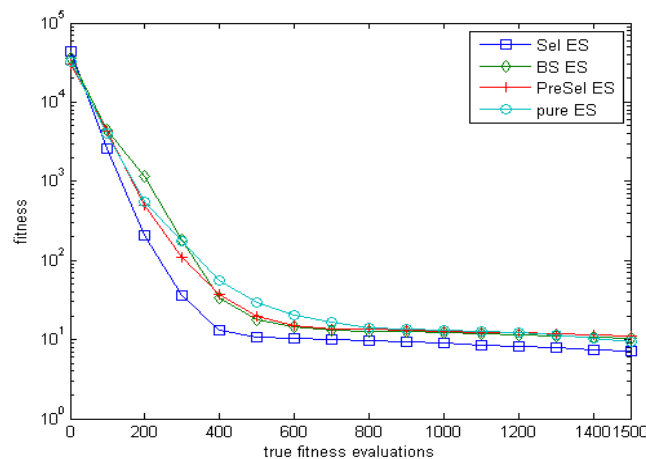


Figure 3. Results for the Rosenbrock function

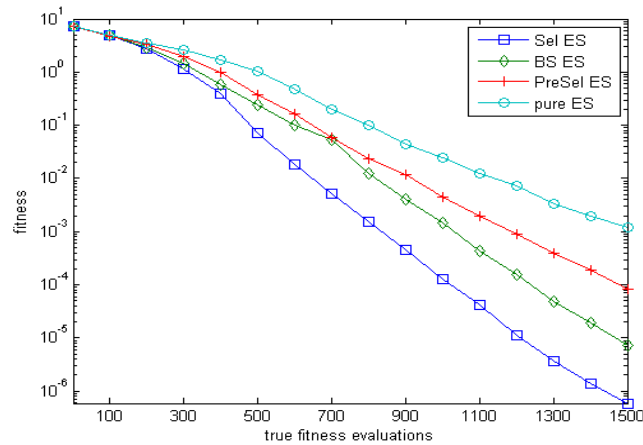


Figure 4. Results for the Ackley function

The number of generations for each number of controlled individuals from η to λ for strategy proposed is shown in Figures 5, 6, 7 (for three functions investigated).

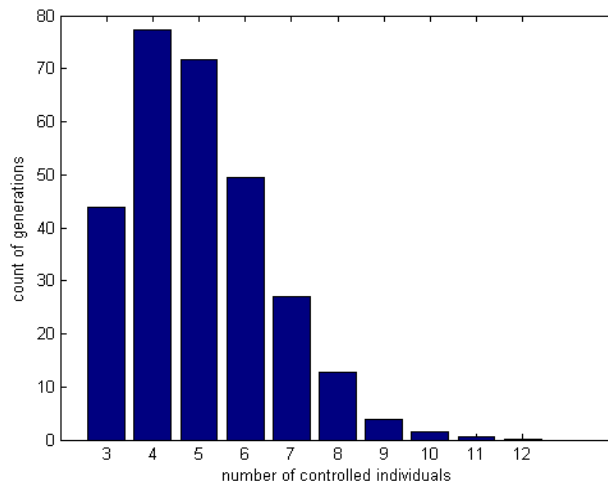


Figure 5. Results for the Sphere function

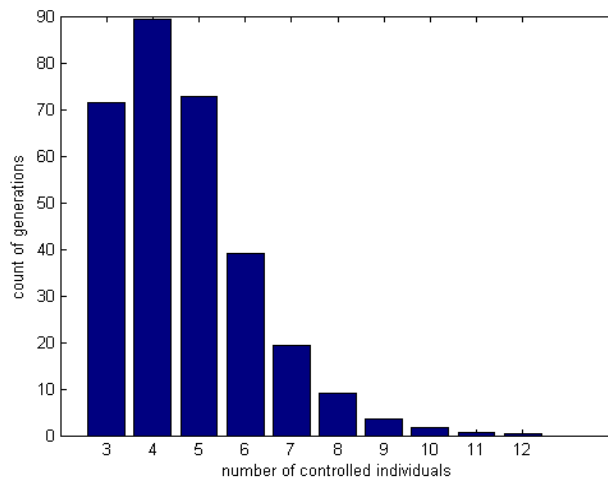


Figure 6. Results for the Rosenbrock function

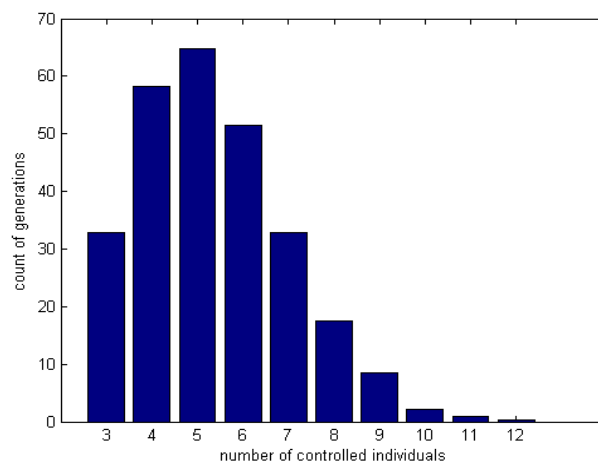


Figure 7. Results for the Ackley function

It can be seen, from the results that the evolution strategies with the selective evolution control method has better performance than the others. For Sphere and Rosenbrock functions the number of controlled individuals is less than half of the population size and this number is enough for effective algorithm convergence. The number of controlled individuals for Ackley function happened to be more than for Sphere and Rosenbrock functions. Probably, this fact may be explained that Ackley function has more complex landscape and as a consequence the model can have a greater error during the optimization process.

Conclusion

In this work a new evolution control method for evolution strategies with metamodels are proposed and investigated. This method can be used to reduce the number of computationally expensive fitness function evaluations in complex optimization problems solving. From the results, we showed that the evolution strategy with the proposed evolution control method has better performance than the others investigated strategies for common benchmarks. Future work is planned to implementation and investigation the evolution strategy with the selective evolution control method for the several real-world optimization problems.

Acknowledgement

The paper is supported by the Russian Foundation of Basic Research (project №11-07-00780).

Bibliography

- J. Branke and C. Schmidt. Fast convergence by means of fitness estimation. *Soft Computing Journal*, 9(1):13-20, 2005.
- D. Bueche, N.N. Schraudolph, and P. Koumoutsakos. Accelerating evolutionary algorithms with Gaussian process fitness function models. *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, 35(2), pp. 183-194, 2005.
- M. Emmerich, A. Giotis, M. Özdenir, T. Bäck, and K. Giannakoglou. Metamodel-assisted evolution strategies. In *Parallel Problem Solving from Nature*, number 2439 of *Lecture Notes in Computer Science*, pp. 362-370, 2002.

L. Gräning, Y. Jin, and B. Sendhoff. Individual-based Management of Meta-models for Evolutionary Optimization with Application to Three-Dimensional Blade Optimization. *Evolutionary Computation in Dynamic and Uncertain Environments*, pp. 225-250, 2007.

N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2), pp. 159–196, 2001.

Y. Jin, M. Olhofer, and B. Sendhoff. A framework for evolutionary optimization with approximate fitness functions. *IEEE Transactions on Evolutionary Computation*, 6(5), pp. 481-494, 2002.

Y. Jin, M. Huesken, and B. Sendhoff. Quality measures for approximate models in evolutionary computation. In Alwyn M. Barry, editor, *GECCO 2003: Proceedings of the Bird of a Feather Workshops, Genetic and Evolutionary Computation Conference*, pp. 170–173, Chigaco, 11 July 2003. AAAI.

D. Lim, Y.S. Ong, Y. Jin, and B. Sendhoff: A study on metamodeling techniques, ensembles, and multi-surrogates in evolutionary computation. In *Genetic and Evolutionary Computation Conference*, pp. 1288-1295, 2007.

H. Ulmer, F. Streichert, and A. Zell. Evolution strategies with controlled model assistance. In *Congress on Evolutionary Computation*, pp. 1569--1576, 2004.

Authors' Information

Pavel Afonin – *ph.d.*; *Bauman Moscow State Technical University; 2-d Baumanskaya st.,5, Moscow, Russia;*
e-mail: pavlafon@yandex.ru

COMPUTER SIMULATION OF MIMA ALGORITHM FOR INPUT BUFFERED CROSSBAR SWITCH

Tasho Tashev, Tatiana Atanasova

Abstract: The investigations for throughput of a new algorithm for computing of non-conflict schedule in crossbar switch node are presented in this paper. By means of Generalized nets (GN) a model of MiMa (Minimum from Maximal Matching) algorithm is synthesized. The results of computer simulation of a GN-model performing uniform load traffic are presented. Evaluated throughput of the switch mode tends near the 100% and no regions of instability are detected.

Keywords: *Modeling, Generalized Nets, Communication Node, Crossbar Switch, Algorithm, Simulation.*

ACM Classification Keywords: *B.4.4 Performance Analysis and Design Aids, C.2.1 Network Architecture and Design, C.4 Performance of Systems*

Introduction

At present, digital telecommunications streams are based on the exchange of packets. In information exchange networks the essential nodes are commutation nodes called switches and routers. Crossbar packet switches route traffic from input to output where a message packet is transmitted from the source to the destination.

The randomly incoming traffic must be controlled and scheduled to eliminate conflict at the crossbar switch. The goal of the traffic-scheduling for the crossbar switches is to minimize packet blocking probability and packet waiting time and to maximize the throughput of packet through a switch [Elhanany, 2007]. So achieving a maximum throughput of the switch depends on the calculation of non-conflict plan for switching incoming packets.

The problem for calculating of non-conflict schedule is NP-complete [Chen et al, 1990]. Algorithms are suggested which solve the problem partially. Constantly rising levels of traffic communication require developing of new, more efficient algorithms for calculating the schedule.

The origin of a series of parallel algorithms is the PIM-algorithm (Parallel Iterative Matching) [Anderson et al, 1993]. One of the research directions is working on modifications to PIM-algorithm, relying on input buffering with virtual output queuing (VOQ) [Guannan Qu et al., 2010]. Other studies are directed to the use of inputs and intermediate buffering (CICQ) [Dinil Mon Divakaran et al., 2010]. The approach with an intermediate load balancing also attracts attention. Of course, research is also directed towards a fully optical switching [Lin Liu et al., 2010].

Cellular automata, neural networks, etc. are used as formal means to describe and study the characteristics of crossbar switch nodes. In this investigation the apparatus of Generalized Nets (GN) are used as a powerful modern tool for formal modeling of parallel processes. Generalized nets (GN) [Atanassov, 1991, Atanassov, 1997] are a contemporary formal tool created to make detailed representation of connections between the

structure and temporal dependencies in parallel processes. They are used in different fields of application, telecommunication is one of them [Gochev, 2008], [Tashev and Gochev, 2009]. The apparatus of GN in this research is applied to synthesize a model of one new algorithm for computing of non-conflict schedule in the crossbar switch node.

In this paper we presented the investigations on the proposed new algorithm for crossbar switch - MiMa (Minimum from Maximal Matching) algorithm. The algorithm is design to calculate a non-conflict schedule in crossbar switch node with VOQ. It is based on a new criterion for selecting of non-conflict solutions. Checking of its applicability is done by computer simulation of switching of these non-conflict solutions through synthesized Generalized-Nets based model of the MiMa algorithm. Its assessment is based on the modeling firstly of the throughput in the presence of uniform distributed incoming traffic. For this purpose, four templates are used to simulate uniform demand traffic and the evaluation of the performance of the MiMa algorithm has been obtained.

Algorithms of Non-Conflicts Schedule for Commutation

The requests for transmission through switching $n \times n$ line switch node is presented by an $n \times n$ matrix T , named traffic matrix (n is integer). Every element t_{ij} ($t_{ij} \in [0, 1, 2, \dots]$) of the traffic matrix represents a request for packet from input i to output j . For example $t_{ij} = 2$ means that two packets from the i^{th} input line have to be send to j^{th} output line of the switch node, etc.

It is assumed that a conflict situation is formed when in any row of the T matrix the number of requests is more than 1 – this corresponds to the case when one source declares connection with more than one receiver. If a column of the matrix T hosts more than one digit 1, it indicates a conflict situation. Avoiding conflicts is related to the switch node efficiency [Elhanany, 2007].

In our previous investigations algorithms for computing of non-conflict schedule are modeled by Generalized nets based on the principle of sequent-random choice [Tashev and Vorobiov, 2008].

The new developed MiMa algorithm also is based on the principle of sequent-random choice, but it uses a new criteria. Its informal description is as the following:

Matrix T is introduced. A vector-column, which consists of the number of conflicts in which row (conflict weights) is calculated. A vector-row, which consists of the number of conflicts in each column (column weight), is calculated too. In the vector-row we choose the maximal element (the column with the most conflicts). In the vector-column we choose the maximal element (the row with the most conflicts). If there is a request in the place of intersection in T we take it as an element of the non-conflict matrix Q_k . If not – we choose the element in the vector-column following the maximal element. We check if there is a request, etc. As a result for the chosen column of T we will choose a request (if indeed it exists) and we will reduce the weight of conflict of the corresponding row and column. We take the element following the maximal element in the column and we search for a request for it using previously described criteria. As a result the first matrix Q_1 will consists of elements (requests) with maximal weight of conflicts in T . For the last matrix Q_k it will be left only non-conflict requests in T .

Generalized Net Model of MiMa-Algorithm

The algorithm MiMa can be described formally by the means of Generalized Nets. The model is developed for switch node with n inputs and n outputs. Its graphic form is shown on Figure 1.

The model has possibilities to provide information about the number of switching in crossbar matrix, as well as about the average number of packets transmitted by one switch. Analysis of the model proves receiving a non-conflict schedule. Calculation complexity of the solution depends on the power of four of the dimension n of the matrix T ($O(n^4)$). Numerical modeling should provide us with the answer to the question: do we have a better solution with this algorithm or not in comparison with existing ones?

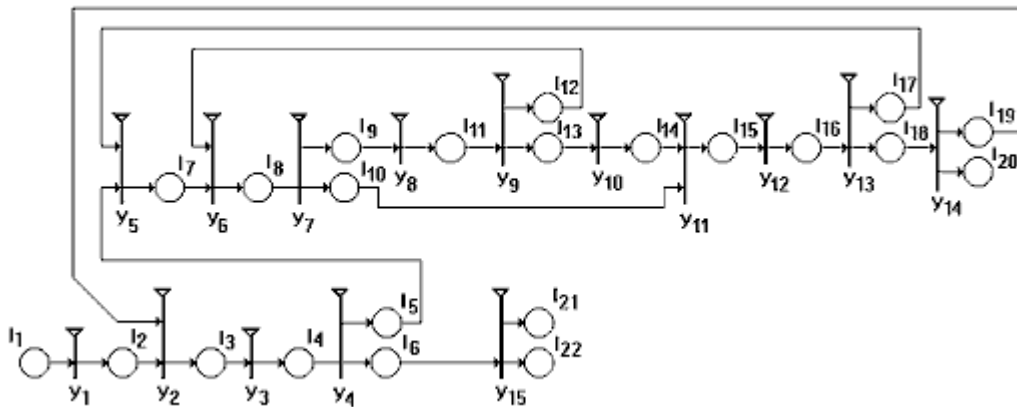


Figure 1: Graphical form of GN-model of MiMa-algorithm

Computer Simulation

The transition from GN-model to executive program is performed as in [Tashev and Vorobiov, 2007]. The program package Vfort of Institute of mathematical modeling of Russian Academy of Sciences is used [Vfort]. The source code has been tested and then compiled by the PC with Intel E8400 2x3.0 GHz, 2 GB RAM. The resulting executive code is executed in the DOS-console under Windows XP SP2. Main restriction for the choice of parameters in simulation (dimension n and type of load traffic) is the time for execution of the program.

Achieving maximal throughput of crossbar switch node depends on creation of non-conflict schedule for packet commutation. The first step while checking their efficiency is throughput modeling of the switch by uniform demand traffic. The matrix T defines a uniform traffic demand matrix if the total number of packets in each row and that of each column are equal [Gupta and McKeown, 1999].

The uniform demand traffic matrix is called in the investigation as *Pattern_i*. The index i shows values of element in the traffic matrix. All elements in the traffic matrix are equal. This allows calculating the throughput because in this case an optimal solution is known. The throughput is computed by dividing the result of optimal solution on the result of the simulated solution. The result of algorithm is a number of non-conflict matrices. Their sum is equal to T , as number of matrices shows number of commutations.

Figure 2 presents the used input data – *uniform matrix T*, defined by us. The first type of the matrix is called *Pattern₁*. Its specification is shown on the left of the figure 2. The optimal schedule requires n switching of

crossbar matrix for $n \times n$ switch. The second type of the matrix is called $Pattern_i$. Its specification is shown on figure 2 (right). The optimal schedule requires $(i \times n)$ switching of crossbar matrix for $n \times n$ switch.

$$T = \begin{matrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ 2 \times 2 \end{matrix} \begin{matrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\ 3 \times 3 \end{matrix} \dots \begin{matrix} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \\ k \times k \end{matrix} \quad T = \begin{matrix} \begin{bmatrix} i & i \\ i & i \end{bmatrix} \\ 2 \times 2 \end{matrix} \begin{matrix} \begin{bmatrix} i & i & i \\ i & i & i \\ i & i & i \end{bmatrix} \\ 3 \times 3 \end{matrix} \dots \begin{matrix} \begin{bmatrix} i & \dots & i \\ \vdots & \ddots & \vdots \\ i & \dots & i \end{bmatrix} \\ k \times k \end{matrix}$$

Figure 2: Types of the uniform traffic matrix T

The results from the computer simulation of the MiMa-algorithm with input data $Pattern_1$ and $Pattern_5$ are displayed on figure 3 and 4. The crossbar matrixes of the size 2×2 up to 130×130 are simulated. The results of simulations with input data $Pattern_{10}$ and $Pattern_{50}$ are demonstrated on figure 5 and 6. It can be seen that when the size of $Pattern$ and dimension of switch field increases, the throughput asymptotically tends to 100%.

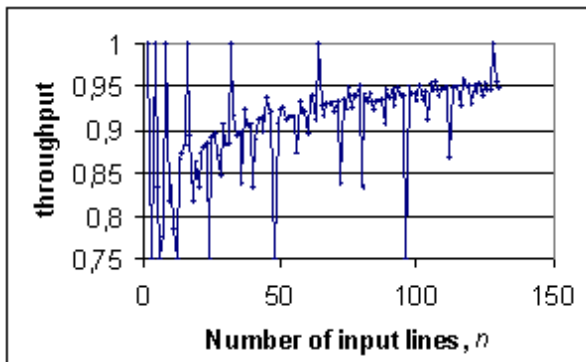


Figure 3: Throughput with $Pattern_1$

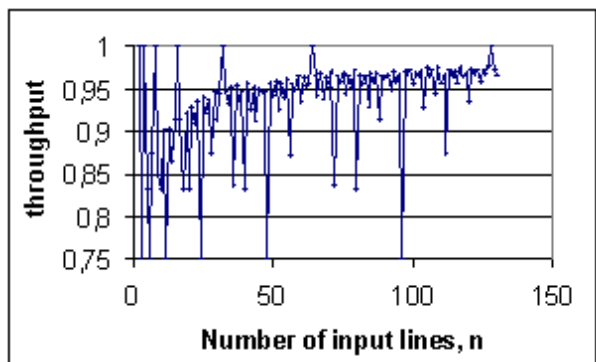


Figure 4: Throughput with $Pattern_5$

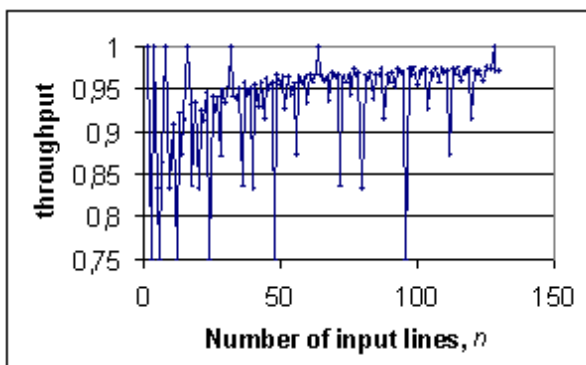


Figure 5: Throughput with $Pattern_{10}$

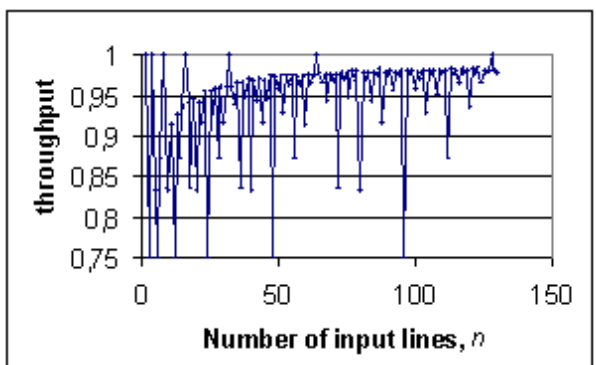


Figure 6: Throughput with $Pattern_{50}$

Evaluation of the simulation results illustrates that MiMa algorithm has low sensitivity to the increasing of input buffer. Figure 7 gives evidence for the difference between throughput in cases of $Pattern_{50}$ and $Pattern_1$ – when

input buffer increases in 50 times. The difference between throughput in cases of $Pattern_{50}$ and $Pattern_{10}$ is shown in figure 9 (increasing 5 times).

For checking existence of regions of instability we used a version of $Pattern_1$, in which the main diagonal consists of elements equal to zero (called $Pattern_{1-0}$). Figure 9 shows throughput for this case. The differences between throughput from figure 1 and figure 9 tend to zero. This result is summarized on figure 10.

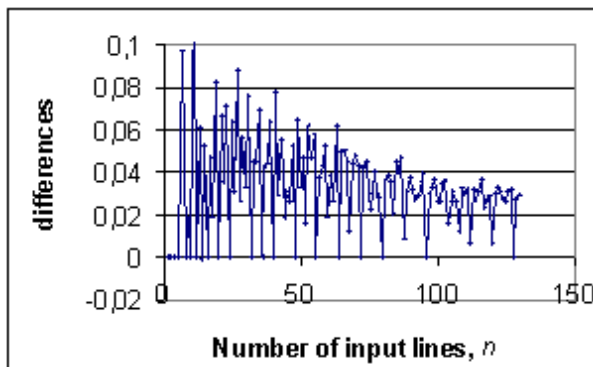


Figure 7: Differences between throughput $P_{50}-P_1$

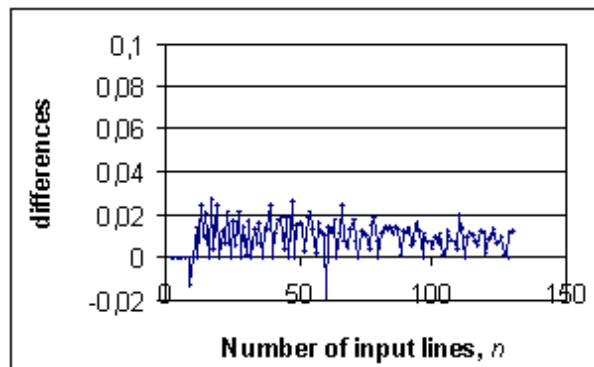


Figure 8: Differences throughput $P_{50}-P_{10}$

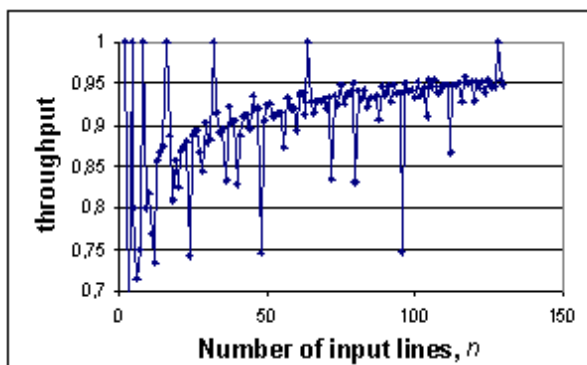


Figure 9: Throughput with $Pattern_{1-0}$

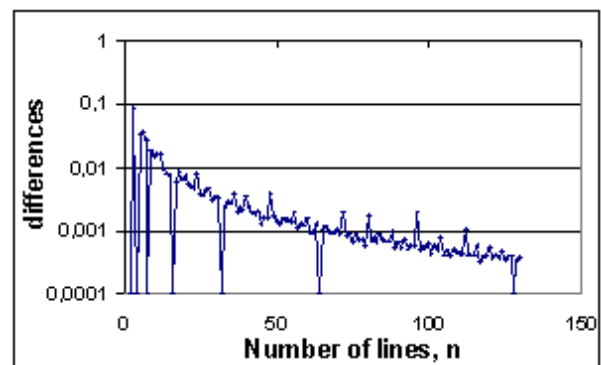


Figure 10: Differences between throughput P_1-P_{1-0}

We specified the family of variants of $Pattern_i$, in which the main diagonal consists of elements equal to zero (called $Pattern_{i-0}$). The cases of $Pattern_{5-0}$, $Pattern_{10-0}$ and $Pattern_{50-0}$ were also checked. The results are similar – they show absence of instability.

The promising results from the simulation on the MiMa algorithm lead us to an idea of conducting of large-scale simulations for non-uniform demand traffic that can be direction for the future work.

Conclusion

In this paper the investigations on new algorithm for calculating a non-conflict schedule for crossbar switch node are presented. Computer simulations of a Generalized Nets-based model of MiMa-algorithm performing uniform load traffic have been carried out. The results of simulation are evaluated. The simulations illustrate that MiMa

algorithm has low sensitivity to the increasing of input buffer. It is shown that throughput of switch mode asymptotically tends to 100% under uniform traffic and no regions of instability are detected.

Future work should be directed to carrying out large-scaled computer simulation to study the throughput for large dimensions of the switching field and a wide range of incoming demand traffic. The opportunities to parallelize the proposed algorithm could be searching for.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

- [Elhanany, 2007] Elhanany, I., Hamdi, M. High-performance packet switching architectures. Springer-Verlag London, 2007.
- [Chen et al, 1990], W., Mavor, J., Denyer, P., and Renshaw, D. Traffic routing algorithm for serial superchip system customisation. IEE Proc 137:[E]1, 1990.
- [Anderson et al, 1993] Anderson T., Owicki S., Saxe J., and Thacker C. High speed switch scheduling for local area networks. ACM Trans. Comput. Syst., vol. 11, no.4, 1993, pp.319-352.
- [Guannan Qu et al., 2010] Guannan Qu, Hyung Jae Chang, Jianping Wang, Zhiyi Fang, Si-Qing Zheng, Designing Fully Distributed Scheduling Algorithms for Contention-Tolerant Crossbar Switches, Proceedings of IEEE Conference on High-Performance Switching and Routing (HPSR 2010), pp. 69 - 74
- [Dinil Mon Divakaran et al., 2010] Dinil Mon Divakaran, Fabienne Anhalt, Eitan Altman, Pascale Vicat-Blanc Primet, Size-Based Flow Scheduling in a CICQ Switch, Proceedings of IEEE Conference on High-Performance Switching and Routing (HPSR 2010) , pp. 57 - 62
- [Lin Liu et al., 2010] Lin Liu, Zhenghao Zhang, Yuanyuan Yang, Packet Scheduling in a Low Latency Optical Packet Switch, Proceedings of IEEE Conference on High-Performance Switching and Routing (HPSR 2010), 13-16 June 2010, Richardson, TX, pp. 63-68
- [Atanassov, 1991] Atanassov K. Generalized Nets. World Scientific, Sing., N.J., London, 1991
- [Atanassov, 1997] Atanassov K. Generalized Nets and System Theory. Acad. Press "Prof.M.Drinov", Sofia, Bulgaria, 1997.
- [Gochev, 2008] Gochev V., Modeling of telecommunication traffic with Generalized nets. Proc. of National conference with international participation "Elektronika'2008", 29-30 May 2008, Sofia, Bulgaria. pp. 88-93 (in Bulgarian).
- [Tashev and Gochev, 2009] Tashev T., Gochev V., One Generalized Net Model for Estimation of Decisions for PIM-Algorithm in Crossbar Switch Node. Proc. of the Tenth Int. Workshop on Generalized Nets, 5 December 2009. Sofia, Bulgaria. pp. 51-58.

[Tashev and Vorobiov, 2008] Tashev T.D., Vorobiov V.M. Computer modeling of one algorithm of the frictionless schedule in the switching site. Proc. of Int. Workshop "Distributed Computer and Communication Networks – DCCN'2008" October 20-23 2008, Sofia, Bulgaria. Moscow, Russia, 2008. pp. 95-100.

[Tashev and Vorobiov, 2007] Tashev, T., Vorobiov, V.: Generalized Net Model for Non-Conflict Switch in Communication Node. Proc. of Int. Workshop "DCCN'2007" Sept 10-12, 2007, Moscow, Russia. IPPI Publ., Moskow, 2007. Pp.158-163.

[Vfort] <http://www.imamod.ru/~vab/vfort/download.html>

[Gupta and McKeown, 1999] Gupta, P., McKeown, N. Designing and Implementing a Fast Crossbar Scheduler. IEEE Micro, Jan-Feb 1999, pp. 20-28.

Authors' Information



Tasho Tashev, Department "Optimization and Decision Making", Institute of Information and Communication Technologies – Bulgarian Academy of Sciences, "Acad. G. Bonchev" bl. 2 Sofia 1000, Bulgaria; e-mail: ttashev@iit.bas.bg

Major Fields of Scientific Research: Distributed Information Systems Design, Methods and tools for net models researches



Tatiana Atanasova, Department "Optimization and Decision Making", Institute of Information and Communication Technologies – Bulgarian Academy of Sciences, "Acad. G. Bonchev" bl. 2 Sofia 1000, Bulgaria; e-mail: atanasova@iit.bas.bg

Major Fields of Scientific Research: Complex Control Systems, Learning Structures, Distributed Information Systems

PREDICTION OF EDUCATIONAL DATA MINING BY MEANS OF A POSTPROCESSOR TOOL

Oktay Kir, Irina Zheliazkova

Abstract: *A methodology for application of several linear methods of prediction of correct, missing, and wrong knowledge using a postprocessor tool is presented.*

Keywords: *teacher, learner, moving average methods, error of prediction, methodology, prediction skill, sessionscript*

ACM Classification Keywords: *Computer and Information Science Education, Knowledge Representation*

Introduction

In a recent exhaustive survey of Romero Cr., Ventura S., 2010 [3] prediction has been pointed out as one of the oldest *Educational Data Mining (EDM)* task. The variables more often predicted are the learner's performance, knowledge, scores, and mark and the techniques most commonly applied are neural and Bayesian networks, rule-based systems, regression, and correlation analysis.

In a previous authors' paper [2] a teacher's tool for the *EDM* called postprocessor was reported from design, implementation, and user's points of view. The term "postprocessor" stands for processing standardized output data sets after each session, e.g. test, lecture or exercise from a corresponding task-oriented environment. For ensuring the tool's intelligence and its adaptation to the teacher a power and expressive script language called *SessionScript* was implemented. Programming of descriptive statistics, visualization, and correlation analysis techniques was demonstrated using two output data sets respectively from environment for knowledge testing and for exercise task performing.

In business modeling the classical linear and non-linear methods of prediction are referred to as a temporal data mining technique for estimation of unknown values of an observed variable [5]. According to [1] the base line, e.g. time seria with the observed numerical, continuous or discrete values has to meet four important requirements:

a) The results of observations have to be sorted from the earliest to the last one; b) All time periods have to be of equal length; c) The observations have to be fixed at the same time in each period; d) Missing even a single observation is not desirable and missing data has to be complete with estimated ones. If a given base line does not meet any of these requirements it is likely the prediction error to be unacceptable.

The computationally complex techniques listed in the above-mentioned survey are proved as successful for the tasks concerning mainly mediate- and long-term prediction. In this paper studying simple linear methods for the *L*'s short-term prediction of correct, missing, and wrong knowledge of testing is reported using the postprocessor. Firstly the pedagogical experiment carried out for gathering the input data set is described. The focus of the

paper is on implementation of four methods called moving average in three cases, e.g. correct, missing, and wrong knowledge of testing. A comparative analysis of the corresponding errors of prediction by means of the tool is made to choose the most precise method. Conclusions summarize the methodology proposed for the application of the considered methods using the postprocessor.

Pedagogical Experiment Description

The data set for implementation of linear methods in the postprocessor for short-term prediction was gathered in the framework of an experiment carried out the academic 2008/2009 year. Four groups of bachelor degree regular students (3-rd year, 1-st semester) specialty "Computer Systems and Technologies" at Rouse University were involved in it (63 students in total). The test session was carried out within the framework of the course in Software Engineering and covered 30 hours taught lecture material. The test was created by the lecturer of the course as an intelligent posttest in order to evaluate the correct, missing, and wrong knowledge, as well as the time undertaken for the test performance.


Each student was accessible through a common device to a template Microsoft Word document. The number of questions was 30 with total scores $P_{max} = 352$ and planned time $T_{max} = 120$ min. The questions types were four, namely: multiple choice, unordered keywords, ordered keywords, and unordered pairs [4] and answering was reducing to filling an empty edit field in correspondence with a simple syntax. Depending on its type each question brought different number of scores $p_{max,j}$. A question "no" answer or subanswer was interpreted as missing knowledge, and incorrect answers or subanswers as wrong knowledge. After the test performance the student had to upload the fulfilled document back on the common device. The students were also told that the time for the test performance actually is unlimited and together with wrong and missing knowledge will be used as assessment indicators only for research purpose. Later the lecturer manually calculated the questions correct, missing, and wrong knowledge, their total scores for each student and his/her final mark in the traditional six-based scale: $0 \leq P \leq 0.4 * P_{max} - "2"; 0.4 * P_{max} < P \leq 0.55 * P_{max} - "3"; 0.55 * P_{max} < P \leq 0.70 * P_{max} - "4"; 0.70 * P_{max} < P \leq 0.85 * P_{max} - "5"; 0.85 * P_{max} < P \leq 1.0 * P_{max} - "6"$. The experience accumulated during the last decade by Zheliazkova's research group has pointed out that such non-linear scale is acceptable by both teachers and students [6]. The Word document of a "good" student, e.g. received test mark "4" is shown on fig. 1. This experiment confirmed the fact that the Ls go to such intelligent test performance only if they were preliminary self assessed at least with the mark "3". The students also were well motivated and stated that were waiting for an objective and precise test assessment. As a result a tendency of shifting the average test mark from "good" to "very good" also was monitored.

Tool's Description

In order to solve a new task a new user has to familiar with the data mining techniques, tasks classification, as well as with SESSIONSCRIPT language. The full specification of its first version can be found in [2]. By means of a standard text editor the user can review the script of the programs for related tasks.

A new free-text formulated problem has to be clear, precise, and compact. Although the problem solving is presented as a sequence of steps in practice some steps can be omitted others repeated or interpreted as subproblem solving. To perform each step from the technological scheme the teacher has to know the syntax

and semantics of the corresponding group of commands. In order to enhance the SESSIONSCRIPT language learning the following color coding scheme has been accepted: the correct commands names and symbols for operations in *Aqua*; table, row and column names in *Yellow*; values in *Pink*; unknown keywords and current values of program variables in *Dark grey*; and messages in *Grey*.



РУ "Ангел Кънчев"
кафедра "Компютърни Системи и Технологии"

Многоцелеви групов тест върху лекционния материал
по дисциплината "СОФТУЕРНО ИНЖЕНЕРСТВО" (СИ) за студенти редовно обучение 2008/2009

Авторски колектив:

1. Ангел Иванов, студент-бакалавър, 3-ти курс, спец. КСТ
2. Юсуф Хасанов, студент-бакалавър, 3-ти курс, спец. КСТ
3. Юмер Юмер, студент-бакалавър, 3-ти курс, спец. КСТ
4. Денис Ислям, студент-бакалавър, 3-ти курс, спец. КСТ
5. доц. д-р Ирина Желязкова, преподавател по СИ, кат. КСТ

Цели:

1. Оценка на изходното ниво знанията;
2. Диагностика на пропуските в лекционния материал;
3. Оценка качеството на теста.

Скала за оценка:
от 000 до 127 т. – 2
от 128 до 174 т. – 3
от 175 до 221 т. – 4
от 222 до 268 т. – 5
от 269 до 315 т. – 6

Очаквано време за изпълнение: 120 мин.
Студент: Ана Данчева Панова **Фак. номер:** 063178 **Група** 20 б

Тема 1: ПРОЦЕДУРЕН ПОДХОД	7 въпроса	65 т.
----------------------------------	------------------	--------------

ВЪПРОС 1: Подредени ключови думи
Попълнете липсващите ключови думи: "Различават два вида ... между модулите: ... и вътрешна. Една програма трябва да се разбие на ... така, че да съществува ... връзка между отделните модули и обратно -... връзки в модулите."
Отговор: връзки > външни > модули > слаби > силни
Параметри: L=2; Q_i=10; C_p=0.50
Препратка: 1.1. Същност на процедурния подход
Знания: 8,0,2
...

ВЪПРОС 18: Неподредени двойки
Укажете съответствието между: 1) INTERFACE, 2) IMPLEMENTATION на статична библиотека с име StatDLL за нуждите на програмата от фиг. 5 и други програми, използващи ПП от една и съща DLL.
Отговор: 1>a; 1>c ;2>b ;2>d
Параметри: L=2; Q_i=12; C_p=0.66
Препратка: 3.4. Създаване на интерфейсен модул за DLL
Знания: 10,0,0
...

ВЪПРОС 31: Считате ли, че получената оценка е обективна и точна? Да/Не
ВЪПРОС 32: За колко минути попълнихте теста?
Резултат: 218,66,31,106,4

РЕЧНИК				
библиотеки	инициализиращите	портове	AStud	MEML
библиотеката	интерфейсната	празна	Begin	new(PPrezident)
библиотеките	клавиатура	прекомпилирани	Build	PORTW
връзките	код	принтер	call	program
външни	константи	програмата	const AStud: TStud =	self
вътрешна	масиви	програмните	End.	Total
главната	методите	променливи	export	Total:=0;
глобални	минимална	реализационната	exports	TPrezident
датата	модул	свързаност	External	^TPrezident;
двубайтови	модули	сегмент	FAR	Unit StatDLL;
деклариране	нови	силни	index	USES
декларацията	обекта	статична	inline	Value : word;
диаграми	обектен тип	стека	library	var
директиви	обектна	съществуващи	Make	VAR Total : word;
дискон файл	отместване	текущия	MEM	with
еднобайтови	периферните	фрагменти	MEMW	{SF+}
запис	полетата	APrezident^		

Дата: 28.01.09 ЖЕЛАЕМ ВИ УСПЕХ!

Fig. 1. The word document of a student's test

The left side window (fig. 2) contains the script code of the program written by the teacher, and the right side window the table with the input data set. The description of the problem solved in a free text format can be seen switching from "Table" to the "Description" tabs (not shown here). The teacher is also recommended to save it in a standard text file serving as a common catalogue of the problems already solved by the members of the course team.

To facilitate the visualization the transposition of the input table is recommended. Each row of the table on fig. 2 corresponds to a test question from 1 to 30, and its first three columns the input data set, e.g. respectively the base line with correct (RA), missing (EA), and wrong (WA) knowledge. The next three columns contain their normalization values (P_RA), (P_EA), and (P_WA) respectively to the maximal scores $p_{\max j}$ the corresponding question. The teacher can choose also the menu-command Window|Variables to view the names and values of the system variables and the program variables. The visualization allows choosing different kinds of diagrams, such as bar, pie, line, and point viewed in separated windows. In order to view the corresponding diagram the table name has to be chosen from the menu-item View|Bars. For the needs of this kind of tasks a power command for visualization of a family of lines in a common coordinate system had been implemented in the postprocessor: `PLOT(x1,x2,...xn, 'propertyName1 = propertyValue1'; y1,y2,...yn, 'propertyName2 = propertyValue2'; ...) propertyName and propertyValue => { TITLE = "<string>", LINSTYLE = DOT | SOLID, LINEWIDTH = <integer>, MARKVISIBLE = TRUE | FALSE, POINTVISIBLE = TRUE | FALSE, POINTSIZE = <integer>, POINTSTYLE = SQUARE | TRIANGLE | CIRCLE | DOWNTRIANGLE | CROSS | DIAGONAL | STAR | DIAMOND, COLOR = CL<COLOR_NAME>}`

The screenshot shows the software interface with two main windows. The left window is a Command Prompt containing a script for calculating normalized scores and plotting. The right window is a table titled 'STUDENT_063178' with columns RA, EA, WA, P_RA, P_EA, and P_WA. The table contains 30 rows of data.

	RA	EA	WA	P_RA	P_EA	P_WA
1	8	0	2	0.8	0	0.2
2	15	0	0	1	0	0
3	6	2	2	0.6	0.2	0.2
4	7	0	0	1	0	0
5	4	1	0	0.8	0.2	0
6	11	0	2	0.8462	0	0.154
7	4	1	0	0.8	0.2	0
8	16	0	0	1	0	0
9	10	0	0	1	0	0
10	10	0	0	1	0	0
11	1	0	0	1	0	0
12	4	2	4	0.4	0.2	0.4
13	8	2	0	0.8	0.2	0
14	15	0	0	1	0	0
15	4	0	0	1	0	0
16	10	2	0	0.8333	0.167	0
17	6	2	2	0.6	0.2	0.2
18	10	0	0	1	0	0
19	12	2	0	0.8571	0.143	0
20	6	0	4	0.6	0	0.4
21	4	4	2	0.4	0.4	0.2
22	16	0	0	1	0	0
23	6	4	0	0.6	0.4	0
24	10	0	0	1	0	0
25	10	0	0	1	0	0
26	11	0	0	1	0	0
27	0	8	4	0	0.667	0.333
28	8	0	2	0.8	0	0.2
29	0	15	0	0	1	0
30	4	4	2	0.4	0.4	0.2

Fig. 2. The tool's screen with the script and table windows

Correct, Wrong, and Missing Knowledge Prediction

Hereinafter three linear methods called Moving Average (MA) are remained from the reviewed INTERNET literature [7,8,9,10] where the term *MA* stands for the mean value for a certain period of time.

➤ *Simple Moving Average (SMA)* uses average demand for a fixed sequence of periods and is good for a stable demand with no pronounced behavioral samples. The calculated formula for the step averaging procedure

is $F_{t+1} = \sum_{i=t-N+1}^t F_i$, where F_{t+1} is the prediction for the $(t+1)^{th}$ period of time; F_i is the actual value at the i^{th}

period; N is the number of observed periods. The main disadvantage of the method is losing several predicted values which number is equal to the number of the time periods for the MA.

➤ *Weighted Moving Average (WMA)* allows placing a greater emphasis on more recent data in order to reflect changes in demand samples. The weights used are based on the experience of the human predictor. In practice the weighting factors are often chosen to give more weight to the most recent data in the time series and less

weight to older one. The corresponding formula is $F_{t+1} = \sum_{i=t-N+1}^t w_i F_i$, $\{w_1, w_2, \dots, w_n\}$ the vector of weights

such that $\sum_{i=1}^N w_i = 1$. Note, that this method does not avoid the disadvantage of the *SMA* and requires a more

complex calculation at each step of averaging procedure. Additionally, if the data from each step are not available for analysis, it can be difficult if impossible to reconstruct a changing signal accurately. However, if the number of missing steps is known, the weighting of values in the *MA* can be adjusted to give equal weight to all missing samples to avoid this issue.

➤ *Exponential Moving Average (EMA)* is a better trend indicator than the *SMA* as it puts greater weight to most recent data than older ones. Unlike *SMA* and *WMA* the older data never goes away in the calculation of *EMA*. In this case the step calculation formula has the following form $F_{t+1} = (1 - \alpha)F_t + \alpha Y_t = F_t + \alpha(Y_t - F_t), t > 1; \alpha = 2/(M + 1); 0 < \alpha < 1$, where the coefficient α is called *smoothing factor* and M is called *length*.

The results of applying the above-mentioned methods of prediction on the first base line, e.g. for the correct knowledge are shown on fig. 3. Hereinafter the following color scheme is used for visualization: the observation values of the base line in red, line with prediction values for the *SMA* method (5 periods of time) in green, for the same method but with 15 periods in blue, line with prediction values for the *WMA* method with 5 periods in black, and line with prediction values for the *EMA* method with 5 periods in pink. For the three cases, e.g. correct, wrong, and missing answers the weights for the *WMA* method, e.g. were calculated as follows: $W_1 = 1 / (5 + 7 * 1.61803398)$, $W_2 = W_1 * 61803398$, $W_3 = W_1 + W_2$, $W_4 = W_2 + W_3$, $W_5 = W_3 + W_4$, where W_i is the weight of the i th period ($i = 1, \dots, 5$). The *WMA* attaches more value to the latest data.

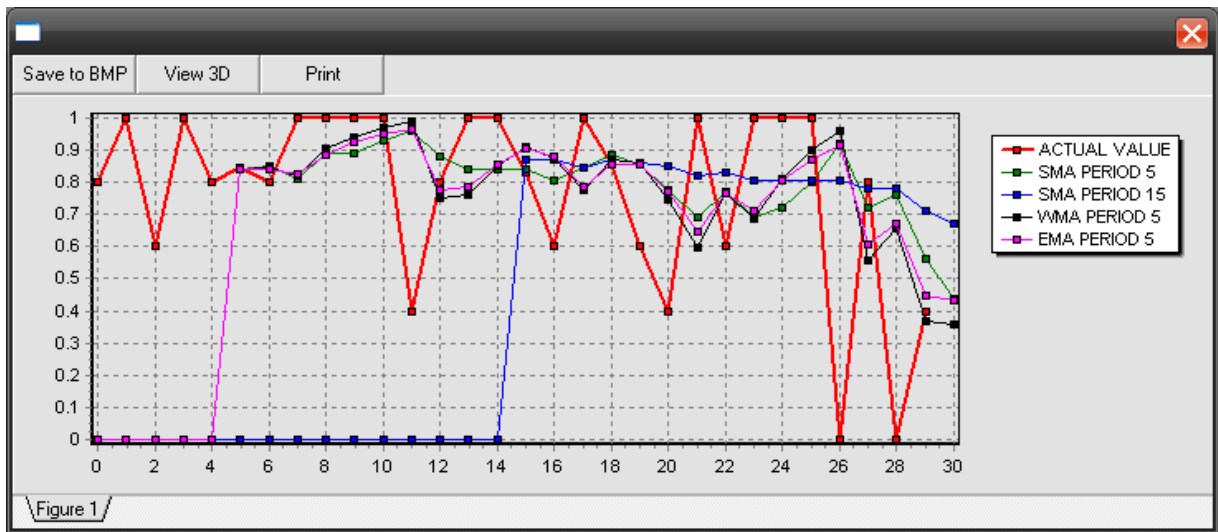


Fig. 3. The window for visualization of three prediction methods together with the base line for correct knowledge

On fig. 4 the results of programming the same four methods of prediction for the wrong knowledge are shown.

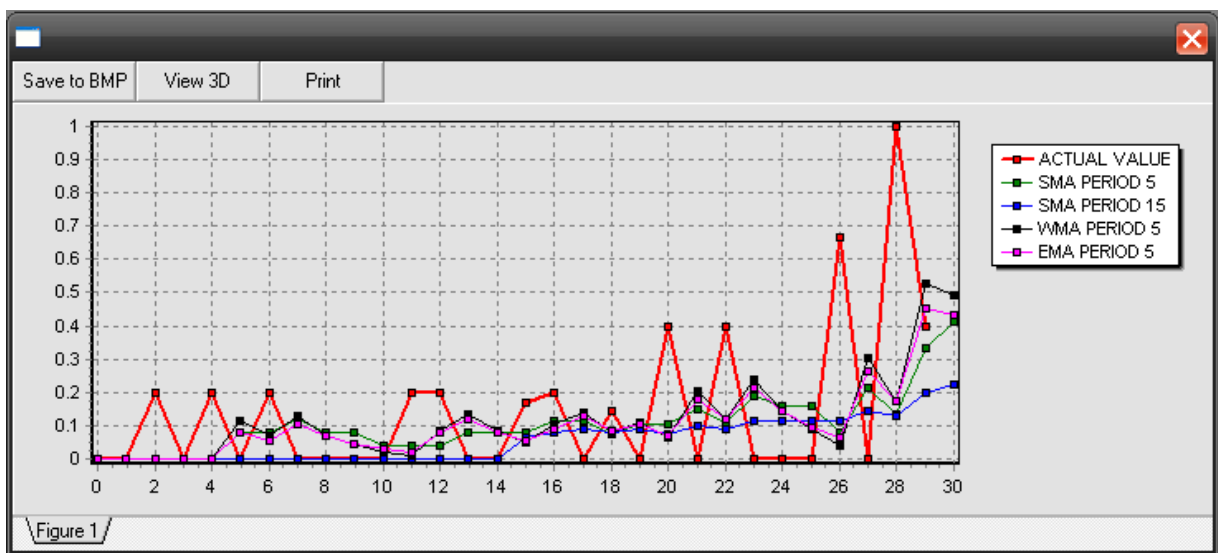


Fig. 4. The window for visualization of three prediction methods together with the base line for wrong knowledge

The results of programming the above-mentioned methods of prediction from the base line for the missing knowledge are shown on fig. 5.

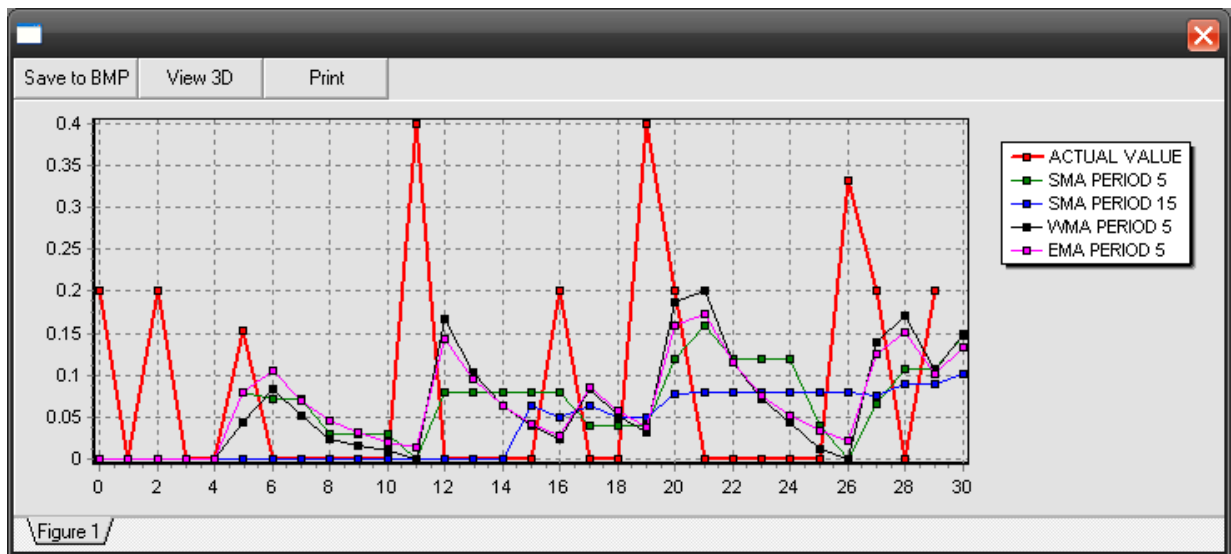


Fig. 5. The window for visualization of three prediction methods together with the base line for missing knowledge

Fig. 3, 4, and 5 illustrate the specific feature of the *SMA*, e.g. absence of several initial prediction values which number is equal to the number of time periods for calculation of the *MA* (for the first and second methods their number is 5 and 15 respectively). In all considered cases the *SMA* with 5 periods works so well as the much more difficult for calculation *EMA* with the same, e.g. 5 number of periods. This finding also is in line with the theoretical basis of prediction. The only case where the corresponding *MA* considerably diverges from each other is when the weight coefficients, assigned to the latest data, are different. The question which type of the *MA* is the best choice has no correct answer. As a rule, the *EMA* is more sensitive to changes than the *SMA*, but less than the *WMA*. However, the answer depends on the specifics of the base line, as well as on the prediction error.

Error and Skill Analysis

Prediction error E_t for the t^{th} calculation period is the difference between the actual value Y_t and the predicted one F_t , e.g. $E_t = |Y_t - F_t|$. The evaluation of the prediction can be equally performed through analyzing certain measures of the aggregate error that could be one of the following:

➤ *Mean Absolute Error (MAE)* is the average of the difference between predicted and actual values in all test

cases, e.g. it is the average prediction error and is calculated as $MAE = \frac{1}{N} \sum_{t=1}^N |E_t|$.

➤ *Mean Absolute Percentage Error (MAPE)* is the mean or average of the absolute percentage errors of

predictions as it is calculated as $MAPE = \frac{1}{N} \sum_{t=1}^N |E_t / Y_t|$.

➤ *Mean Absolute Deviation Percentage (MADP)* is calculated refer to the proportion

$$MADP = \left(\sum_{t=1}^N |E_t| \right) / \sum_{t=1}^N |Y_t|.$$

➤ *Mean Squared Error (MSE)* is the average loss, e.g. the expectation of the squared deviations of the arguments from their respective target value. It is calculated as $MSE = \frac{1}{N} \sum_{t=1}^N E_t^2$.

➤ *Root Mean Squared Error (RMSE)* is one of the most commonly used measures of success for numerical prediction and is computed by taking the average of the squared differences between each predicted value and its corresponding correct value, e.g. $RMSE = \sqrt{MSE} = \sqrt{\left(\sum_{t=1}^N E_t^2 \right) / N}$.

➤ *Skill in Prediction (SP)* is defined a root mean squared error as scaled representation of prediction error that relates the prediction accuracy of a particular prediction model to some reference one. If \bar{Y} is the prediction for the period t then the *SP* is calculated as:

$$SP = 1 - \frac{MSE_f}{MSE_c}; \quad MSE_f = \frac{1}{N} \sum_{t=1}^N E_t^2; \quad MSE_c = \frac{1}{N} \sum_{t=1}^N (\bar{Y} - Y_t)^2; \quad \bar{Y} = \frac{1}{N} \sum_{t=1}^N Y_t$$

From the formula of the MAE type of error follows that the prediction is perfect if it is equal to 0. Obviously, this error will decrease with the MA decreasing. The calculated MAE error for the fourth methods in case of correct knowledge was: 0.229, 0.279, 0.228, and 0.222 respectively. In percentage that approximately means 23% that is unacceptable having in mind that the length of the scoring intervals for the marks different from "2" is 15%. In case of missing knowledge the precise values of the MAE were: 0.178, 0.221, 0.189, and 0.1787 respectively. In percentage that means approximately 19% that is closer to the same error of the SMA with 5 periods than to that one of the SMA with 15 periods of time. The calculated results of the MAE error in case of wrong knowledge are: 0.116, 0.119, 0.115, and 0.115 respectively. In percentage that approximately means 12% 15 % that is approximately two times lower than in case of correct knowledge. The graphical interpretation for the MAE analysis is given on fig. 6.

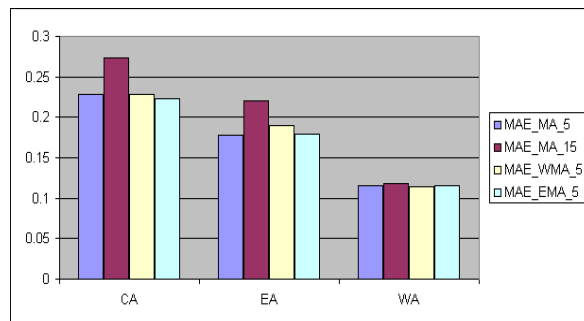


Fig. 6. Graphical comparison of the MAE for all cases

From the formula for the *MAPE* follows that this type of error can't be calculated when the base line contains a zero value. According to the formula for *SP* a perfect prediction has a *SP* equal to 1.0, a prediction with similar skill to the reference prediction would be equal to 0.0, and a prediction which is less skillful than the reference prediction will have negative values. The picture on fig. 6 is slightly changed when the most precise indicator, e.g. the *SP* is used (fig. 7). In case of correct knowledge for all four methods its value is negative and very close to zero (-0.1, -0.14, -.011, and -0.62 respectively). In practice that means none of the methods is acceptable. In case of the missing knowledge the values of *SP* are positive (0.84, 0.69, 0.83, and 0.88 respectively) and close to a "good" prediction. The prediction is "excellent" only in case of wrong knowledge, as the values of *SP* (0.95, 0.94, 0.95, and 0.85 respectively) are positive and very close to 1.00.

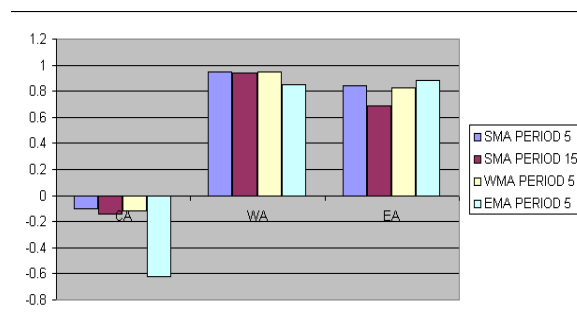


Fig. 7. Graphical comparison of the *SP* for all cases

The actual results can significantly differ from the predicted ones, which can be due to some side effects and/or external factors not taken into consideration. Regarding learners, they could be: attempt for fraud, unfamiliar type of tasks, insufficient attention or low motivation. Other external factors, related to the teacher, could be: poor session planning, organization, and/or delivery.

Conclusions

Application of four simple for computation methods of prediction, e.g. simple moving average with 5 and 15 periods of time, weighted moving average, and exponential moving average has been illustrated for the short-term task for prediction of correct, missing, and wrong knowledge of testing.

The findings from this study are more likely not to be valid for other individual students as the process of testing as the process of learning depends to great degree on the L's attitudes. In connection with this the following methodology for the test performance prediction for other Ls is recommended using the same tool for data mining: 1) Constructing the input table with the rows equal to the test questions, and columns to the number of predicted cases, e.g. correct, missing, and wrong knowledge; 2) Adding new columns with the normalized values of the base lines; 3) Programming formulas of prediction for all chosen methods; 4) Generation of a table with the *MAE* error for the all methods; 5) Generation of a table with the *SP* values; 6) Drawing the first base line with the corresponding predicted lines by using command **PLOT**; 7) Repeating step 4,5 and 6 for all chosen for comparison methods; 8) Making decision about preferable method for prediction on this base respectively to the 15% length of the six-scale intervals.

Bibliography

- [1] Carlberg C., Business Analysis with Microsoft Excel, Sofia, "SoftPress", 2003.
- [2] Kir O., Zheliazkova I, Teodorov G., Educational Data Mining by Means of a Power Instructor's Tool, Proceedings of International Conference on Entrepreneurship, Innovation, and Regional Development (ICEIRD), 2011 (Accepted).
- [3] Romero Cr., Ventura S. Educational Data Mining: A Review of the State of the Art, IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews 2010, No. 40/6, pp. 601-18.
- [4] Zheliazkova I. I., Andreeva M. H., An Intelligent Multimedia Environment for Knowledge Testing, E- learning and the Knowledge Society, Gent & Brussels, Belgium, 6-8 September 2004, pp. 3.13.1-3.13.24.
- [5] Zheliazkova I. I., Atanasov A. Z., An Approach to Teaching Systems for Decisions Making Support in Company Management, Bulgarian Journal "Automatics and Informatics", No. 1, 2008, pp. 62-68 (in Bulgarian).
- [6] Zheliazkova I. I., Kolev R. T., Task Results Processing for the Needs of Task-Oriented Design Environments, Int. J. Computers & Education, vol. 51, 2008, pp. 86-96.
- [7] <http://www.amstat.org/publications/jse/v11n1/datasets.hays.html>.
- [8] <http://www.smetoolkit.org/smetoolkit/en/content/en/416/Demand-Forecasting>.
- [9] http://en.wikipedia.org/wiki/Exponential_smoothing
- [10] http://en.wikipedia.org/wiki/Forecast_skill.

Authors' Information

Okta Kir – PhD student, University of Rousse, Studentska street 8, Rousse 7017, Bulgaria;

e-mail: kir.oktay@gmail.com

Irina Zheliazkova – Associate Professor; University of Rousse, Studentska street 8, Rousse 7017, Bulgaria;

e-mail: irina@ecs.ru.acad.bg

TABLE OF CONTENTS

Data Acquisition Systems for Precision Farming	
Oleksandr Palagin, Volodymyr Romanov, Igor Galelyuka, Vitalii Velychko, Volodymyr Hrusha, Oksana Galelyuka.....	103
Terminological Annotation of the Document in a Retrieval Context on the Basis of Technologies of System "Ontointegrator"	
Olga Nevzorova, Vladimir Nevzorov.....	110
Towards Linguistics Analysis of the Bulgarian Folklore Domain	
Galina Bogdanova, Konstantin Rangochev, Desislava Paneva-Marinova, Nikolay Noev	119
Environmental Risk Assessment Using Geospatial Data and Intelligent Methods	
Nataliia Kussul, Sergii Skakun, Oleksii Kravchenko	129
Self-Organizing Routing Algorithm for Wireless Sensors Networks (WSN) Using Ant Colony Optimization (ACO) With Tinyos	
Nuria Gómez Blas, Luis F. de Mingo, Levon Aslanyan, Vladimir Ryazanov.....	142
Safety Operations of the Complex Engineering Objects	
Nataliya Pankratova.....	152
Indirect Approach of Determination of Collective Alternative Ranking on the Basis of Fuzzy Expert Judgements	
Pavlo P. Antosiak, Oleksij F. Voloshin.....	168
Selective Evolution Control Method for Evolution Strategies with Neural Network Metamodels	
Pavel Afonin	176
Computer Simulation of MiMa Algorithm for Input Buffered Crossbar Switch	
Tasho Tashev, Tatiana Atanasova	183
Prediction of Educational Data Mining by Means of a Postprocessor Tool	
Oktay Kir, Irina Zheliazkova.....	190
Table of Contents.....	200