



I T H E A



International Journal

**INFORMATION THEORIES
&
APPLICATIONS**



2006 Volume 13 Number 2



**International Journal
INFORMATION THEORIES & APPLICATIONS**

ISSN 1310-0513

Volume 13 / 2006, Number 2

Editor in chief: Krassimir Markov (Bulgaria)

International Editorial Staff

Chairman: Victor Gladun (Ukraine)

Adil Timofeev	(Russia)	Larissa Zainutdinova	(Russia)
Alexander Eremeev	(Russia)	Levon Aslanian	(Armenia)
Alexander Kleshchev	(Russia)	Luis F. de Mingo	(Spain)
Alexander Kuzemin	(Ukraine)	Martin P. Mintchev	(Canada)
Alexander Palagin	(Ukraine)	Milena Dobрева	(Bulgaria)
Alexey Voloshin	(Ukraine)	Laura Ciocoiu	(Romania)
Alfredo Milani	(Italy)	Natalia Ivanova	(Russia)
Anatoliy Shevchenko	(Ukraine)	Neonila Vashchenko	(Ukraine)
Arkadij Zakrevskij	(Belarus)	Nikolay Zagorujko	(Russia)
Avram Eskenazi	(Bulgaria)	Petar Barnev	(Bulgaria)
Boicho Kokinov	(Bulgaria)	Peter Stanchev	(Bulgaria)
Constantine Gaidric	(Moldavia)	Plamen Mateev	(Bulgaria)
Eugenia Velikova-Bandova	(Bulgaria)	Radoslav Pavlov	(Bulgaria)
Frank Brown	(USA)	Rumyana Kirkova	(Bulgaria)
Galina Rybina	(Russia)	Stefan Dodunekov	(Bulgaria)
Georgi Gluhchev	(Bulgaria)	Tatyana Gavriloва	(Russia)
Iliа Mitov	(Bulgaria)	Valery Koval	(Ukraine)
Jan Vorachek	(Finland)	Vasil Sgurev	(Bulgaria)
Juan Castellanos	(Spain)	Vitaliy Lozovskiy	(Ukraine)
Koen Vanhoof	(Belgium)	Vladimir Jotsov	(Bulgaria)
Krassimira Ivanova	(Bulgaria)	Zinoviy Rabinovich	(Ukraine)

IJ ITA is official publisher of the scientific papers of the members of
the Association of Developers and Users of Intellectualized Systems (ADUIS).

IJ ITA welcomes scientific papers connected with any information theory or its application.

Original and non-standard ideas will be published with preferences.

IJ ITA rules for preparing the manuscripts are compulsory.

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

The camera-ready copy of the paper should be received by e-mail: foi@nlcv.net.

Responsibility for papers published in IJ ITA belongs to authors.

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

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.13, Number 2, 2006

Printed in Bulgaria

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

Publisher: FOI-COMMERCE - Sofia, 1000, P.O.B. 775, Bulgaria. www.foibg.com, e-mail: foi@nlcv.net

© "Information Theories and Applications" is a trademark of Krassimir Markov

Copyright © 1993-2006 FOI-COMMERCE, Publisher

Copyright © 2006 For all authors in the issue.

All rights reserved.

ISSN 1310-0513

A MULTICRITERIA DECISION SUPPORT SYSTEM *MULTIDECISION-1*¹

Vassil Vassilev, Krasimira Genova, Mariyana Vassileva

Abstract: The present paper describes some basic elements of the software system developed (called *MultiDecision-1*), which consists of two separate parts (the systems *MKA-1* and *MKO-1*) and which is designed to support decision makers in solving different multicriteria analysis and multicriteria optimization problems. The class of the problems solved, the system structure, the operation with the interface modules for input data entry and the information about DM's local preferences, as well as the operation with the interface modules for visualization of the current and final solutions for the two systems *MKA-1* and *MKO-1* are discussed.

Author Keywords: Multicriteria analysis, Multicriteria optimization, Multicriteria decision support system.

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

Introduction

Multicriteria decision making problems can be divided [Vincke, 1992] into two separate classes depending on their formal statement. In the first class of problems, a finite number of alternatives are explicitly given in a tabular form. These problems are called discrete multicriteria decision making problems or multicriteria analysis problems. In the second class, a finite number of explicitly set constraints in the form of functions define an infinite number of feasible alternatives. These problems are called continuous multicriteria decision making problems or multicriteria optimization problems.

Different methods have been developed to solve multicriteria analysis problems, which can be divided into several groups. A great number of the methods developed up to now can be grouped in three separate classes: weighting methods, outranking methods and interactive methods. Each one of these methods has its advantages and shortcomings, connected mostly with the ways of deriving information by the decision maker (DM) regarding his/her local and global preferences. The main element in the weighting methods is the way of determining the criteria weights, which reflect DM's preferences to the highest degree. Many methods for criteria weighting have been developed. A value tradeoff method is proposed in [Keeney and Raiffa, 1976]. Several versions of the analytic hierarchy process (AHP method) are developed in [Saaty, 1980], [Saaty, 1994], using pair-wise criteria comparison. A direct ranking and rating method is proposed in [Von Winterfeldt and Edwards, 1986], in which the DM first ranks all the criteria according to their importance. The weighting methods use a DM's preference model, which does not allow the existence of incomparable alternatives and the preference information obtained by the DM (different types of criteria comparison) is sufficient to determine whether one of the alternatives must be preferred or whether the two alternatives are equal for the DM. The outranking methods use a DM's preference model, which allows the existence of incomparable alternatives and the preference information obtained by the DM may be insufficient to determine whether one of the alternatives is to be preferred or whether the two alternatives are equal for the DM. The criteria and the alternatives are not compared by the DM in these methods,

¹ This paper is partially supported by the National Science Fund of Bulgarian Ministry of Education and Science under contract № I-1401\2004 "Interactive Algorithms and Software Systems Supporting Multicriteria Decision Making."

but he/she has to provide the so-called inter- and intra-criteria information. Some of the well-known representatives of the outranking methods are TACTIC method [Vansnick, 1986], PROMETHEE I-II methods [Brans and Mareschal, 1990], ELECTRE I-IV methods [Roy, 1991] and others. In order to solve multicriteria analysis problems with a large number of alternatives and a small number of criteria, the "optimizationally motivated" interactive methods have been suggested [Korhonen, 1988], [Sun and Steuer, 1996], [Narula et al., 2003].

One of the most developed and widespread methods for solving multicriteria optimization problems are the interactive methods [Gardiner and Vanderpooten, 1997], [Miettinen, 1999]. This is due to their basic advantages – a small part of the Pareto optimal solutions must be generated and evaluated by the DM; in the process of solving the multicriteria problem, the DM is able to learn with respect to the problem; the DM can change his/her preferences in the process of problem solution. The interactive methods of the reference point (direction) and the classification-oriented interactive methods [Miettinen, 1999] are the most widely spread interactive algorithms solving multicriteria optimization problems. Though the interactive methods of the reference point are still dominating, the classification-oriented interactive methods enable the better solution of some chief problems in the dialogue with the DM, relating to his/her preferences defining, and also concerning the time of waiting for new non-dominated solutions that are evaluated and selected.

The software systems supporting the solution of multicriteria analysis and multicriteria optimization problems can be divided in two classes – software systems with general purpose and problem-oriented software systems. The general-purpose software systems aid the solution of different multicriteria analysis and multicriteria optimization problems by different decision makers. One method or several methods from one and the same group are usually realized in them for solving multicriteria analysis and multicriteria optimization problems. This is due to the following two reasons:

- in the methods from the different groups, different types of procedures are used to get information from the DM, which leads to considerable difficulties in the realization of appropriate user's interface modules in the software systems;
- the designers of the software systems are usually interested in the realization of their own method (methods) or have distinct preferences towards methods from one and the same group.

The problem-oriented multicriteria analysis systems are included in other information-control systems and serve to support the solution of one or several types of specific multicriteria analysis problems. Hence, some simplified user's interface modules are usually realized in them. That is why methods from different groups of multicriteria analysis methods are included in some of these systems.

Well-known general-purpose software systems supporting the solution of multicriteria analysis problem are the following systems VIMDA [Korhonen, 1988], ELECTRE III-IV [Roy, 1991], Expert Choice [Saaty, 1994], HIVIEW [Peterson, 1994], PROMCALC and GAIA [Brans and Mareschal, 1994], Decision Lab [Brans and Mareschal, 2000], Web-HIPRE [Mustajoki and Hamalainen, 2000]. One representative of the problem-oriented systems, called Agland Decision Tool, is discussed in [Parsons, 2002].

Some well-known general-purpose software systems, which solve problems of multicriteria optimization, are the following systems VIG [Korhonen, 1987], CAMOS [Osyczka, 1988], DIDAS [Lewandowski and Wierzbicki, 1989], DINAS [Ogryczak et al., 1992], MOLP-16 [Vassilev et al., 1993], MONP-16 [Vassilev et al., 1993], LBS [Jaskiewicz and Slowinski, 1994], MOIP [Vassilev et al., 1997], NIMBUS [Miettinen and Makela, 2000]. The first type comprises the interactive algorithms of the reference point and of the reference direction [Wierzbicki, 1980], [Korhonen, 1987]. These are systems such as DIDAS, VIG, CAMOS, DINAS and LBS. The second type of interactive algorithms includes the classification-oriented algorithms [Benayoun et al, 1971], [Narula and Vassilev,

1994], [Miettinen, 1999], [Vassileva et al., 2001]. These interactive algorithms are built in the systems NIMBUS, MOLP-16, MONP-16 and MOIP.

The present paper describes some basic elements of the software system developed (called *MultiDecision-1*), which consists of two separate parts (the systems *MKA-1* and *MKO-1*) and which is designed to support decision makers in solving different multicriteria analysis and multicriteria optimization problems. The class of the problems solved, the system structure, the operation with the interface modules for input data entry and the information about DM's local preferences, as well as the operation with the interface modules for visualization of the current and final solutions for the two systems *MKA-1* and *MKO-1* are discussed.

Functions, Structure and User's Interface of *MultiDecision-1* System

The system *MKA-1*, the first part of the system *MultiDecision-1*, is designed to support decision makers in solving different multicriteria analysis problems. In *MKA-1* system an attempt has been made to realize three methods – a weighting method, an outranking method and an interactive method. These methods are respectively AHP method [Saaty, 1994], PROMETHEE II method [Brans and Mareschal, 1990] and CBIM method [Narula et al., 2003]. They are the most often used methods in the three groups of methods. The interface modules in the system allow the successful realization of different types of procedures for obtaining information by the DM and also for the entry of different types of criteria – quantitative, qualitative and ranking criteria.

The system *MKO-1*, the second part of the system *MultiDecision-1*, is designed to support decision makers in solving linear and linear integer problems of multicriteria optimization. Three classification-oriented interactive algorithms [Vassilev et al., 2003], [Vassileva, 2004] are included in *MKO-1* system, which enable the DM to define not only desired and acceptable levels of the criteria (as in reference point interactive algorithms), but also desired and acceptable intervals and directions of alteration in the values of the separate criteria. The first interactive algorithm, called GAMMA-L is intended to solve linear problems of multicriteria optimization. The second and the third algorithms, called GAMMA-I1 and GAMMA-I2 respectively, are designed to solve linear integer multicriteria optimization problems. In solving integer problems of multicriteria optimization, the dialogue with the DM is influenced largely by the time, during which he/she is expecting new non-dominated solutions for evaluation and choice. This is so, because the single-criterion integer problems [Nemhauser and Wolsey, 1988], solved at a given iteration, are NP-problems and the time for their exact solution is an exponential function of their dimension. When the solution time proves to be much longer, the DM may lose patience and interrupt the dialogue, refusing to look for a new solution. The classification-oriented interactive algorithms GAMMA-I1 and GAMMA-I2 allow the solving of single-criterion problems at each iteration, with the following two basic properties: a known initial feasible solution and a comparatively "narrow" feasible region. The properties of this type of single-criterion problems, above indicated, facilitate their solution, and also enable the use of approximate single-criterion algorithms. There exists at that high probability that the solutions found will be close to or coincide with the non-dominated solutions of the multicriteria problem.

The system *MKA-1* consists of solving modules, interface modules and internal-system modules. This modularity enables greater flexibility when including new methods or new interface realizations. The *MKA-1* system contains three solving modules. Every module encloses a software realization of one of the three methods - AHP method, PROMETHEE II method and CBIM method and help procedures for each method as well.

The system modules contain all global definitions of variables, functions and procedures of general purpose. The object possibilities of Visual Basic are utilized in *MKA-1* system, creating several classes with respect to internal system structures. They are: a class for messages, which capsules the output of error messages; dynamic context help information and registering of events in the debug window; a class matrix with some specific

procedures, necessary for AHP method; a class for storing the information specific for the criteria in PROMETHEE method and a class for storing system site. The renewal function starts the installation procedure.

PROMETHEE II Method

Evaluation Table

	Cost	Target	Duration	Efficiency	Manpower
News	60	900	22	Average(Fair)	8
Herald	30	520	31	Essential bad(low)	1
Panels	40	650	20	Good(High)	2
Mailing	92	750	60	Bad(Low)	3
CMM	52	780	58	Exceptionally good(high)	1
NCB	80	920	4	Very good(high)	6

Properties of criterion: Cost

Criterion Type: Quantitative

Min/Max: Minimum

Weight: 1

Preference Function: V-Shape with indiff.

Indifference Threshold: 10 Max Val - Min Val

Preference Threshold: 40 62

Gaussian Threshold:

Threshold Unit: Absolute Percent

Average Performance: 59

Unit: 1 000 USD

Legend

Min value(rating)

Max value(rating)

Quantitative's Scale

1- Exceptionally bad(low)	6- Good(High)
2- Essential bad(low)	7- Very good(high)
3- Very bad(low)	8- Essential good(high)
4- Bad(Low)	9- Exceptionally good(high)
5- Average(Fair)	

Previous Set Values Solve

Fig. 1. MKA-1 system PROMETHEE solving windows

The interface modules ensure the interaction between *MKA-1* system, the DM and the operating system. This interaction includes the entry of the data for the multicriteria problems, the entry of information specific for every method, information about DM's preferences, visualization of the current results and of the final result, graphical presentation of the solutions, print out, reading and storing of files, multi-language support, etc. Fig. 1 shows a window with DM's preferences in operation with PROMETHEE II method for one real multicriteria analysis problem, concerning the selection of an appropriate marketing action for advertising of bicycle manufacturing company products [Brans and Mareschal, 2000].

The interface with the DM is realized on the principle of an adviser – a sequence of windows (steps), each one with a distinctly expressed function, which considerably assists and facilitates DM's work. The DM has the possibility to move forward to a following step and also backward; returning for some corrections to the information already entered. The windows, which must be accessible in more than one stage of DM's operation with *MKA-1* system, are included in the menu or in the instruments band. *MKA-1* system possesses dynamic context help information. It gives a brief description of every visual component just by dragging the mouse over it. In addition to this, a debug window is used, that outputs service information about the system internal processes. It can be printed out or stored in a text file. This allows the obtaining of exact debug information when an error occurs. *MKA-1* enables the storing in a file of the input data for every multicriteria problem and of the data about the solution process. Thus, the solution process of a multicriteria problem can be interrupted at any stage and activated from the place of its interruption at any time. *MKA-1* system has comparatively rich printing functions – every piece of the data (entered or computed) may be printed. In this way the entire process of decision making

is documented – you can review the input data of the multicriteria problem being solved, the DM's preferences entered, the current values obtained, and the final result also, which on its turn can be printed out in the form of values or graphics.

MKO-1 system consists of the following three main parts: a control program, optimization modules and interface modules. The control program is an integrated software environment for creating, processing and saving of files associated with *MKO-1* system (ending by *“.mko”* extension) and also for linking and executing different types of software modules. The basic functional possibilities of the control program can be divided into three groups. The first group includes possibilities to use the standard for MS Windows applications menus and system functions – “File”, “Edit”, “View”, “Window”, “Help” and others in system own environment. The second group of control program facilities includes the control of the interaction between the modules realizing: creating, modification and saving of *“.mko”* files associated with *MKO-1* system, which contain input data and data concerning the process and the results from solving multicriteria linear and linear integer problems; interactive solution of the multicriteria linear and linear integer problems which have been entered; localization and identification of the errors occurring during the system operation. The third group of the functional features of the control program includes possibilities for visualization of important information, concerning the DM and the system operation as a whole.

The optimization modules realize three classifications oriented interactive algorithms GAMMA-L, GAMMA-I1 and GAMMA-I2, and also exact and approximate single-criterion algorithms solving problems of linear and linear integer programming.

The interface modules realize the dialogue between the DM and *MKO-1* system during the entry and correction of the input data necessary for the multicriteria problems during the interactive process of these problems solution, and also for the dynamic visualization of the main parameters of the process. An editing module serves to enter, alter and store the descriptions of the criteria, of the constraints, and also of the type and bounds of variables alteration. Another interface module enables the setting of DM's local preferences for alteration in the values of the separate criteria. A third interface module realizes two types of graphic presentation of the information about the values of the criteria at different steps and the possibilities for comparison. Dynamic Help is provided, which outputs specific information about the purpose and way of use of the fields and radio buttons in a separate window.

MKO-1 system is working under MS Windows. It can be added to *Programs* group and/or with a *Desktop* icon, from where it can be started. The system registers the *“.mko”* extension and associates it. Thus, at double clicking on a valid *“.mko”* file, the system will be started and this file will be loaded. There is a menu in the main window with the standard for MS Windows drop-down menus and commands. With their help, the operation of a new file is started or an existing *“.mko”* file is loaded and the operation may continue with the information stored in it.

The entry and correction of the problem criteria and constraints is realized in *“MKO-1 Editor”* window. Every criterion and every constraint is entered separately in the respective text field for edition. Syntax check is accomplished when they are added to the data already entered. The syntax accepted is similar to the mathematical record of this class of optimization problems. The type of the optimum looked for is entered first – “min” or “max”. After that, the digital coefficient with its sign is entered, followed by the variable name it refers to. The variables names can be an arbitrary set of letters and numbers. Each one of these elements is separated by a space. The constraints have similar syntax – digital coefficients and variables names are successively entered. The type of the constraints is defined by some of the symbols “<=”, “>=” or “=”. By double clicking on the constraint or criterion already entered, they are transferred to the editing field again, if subsequent corrections are necessary.

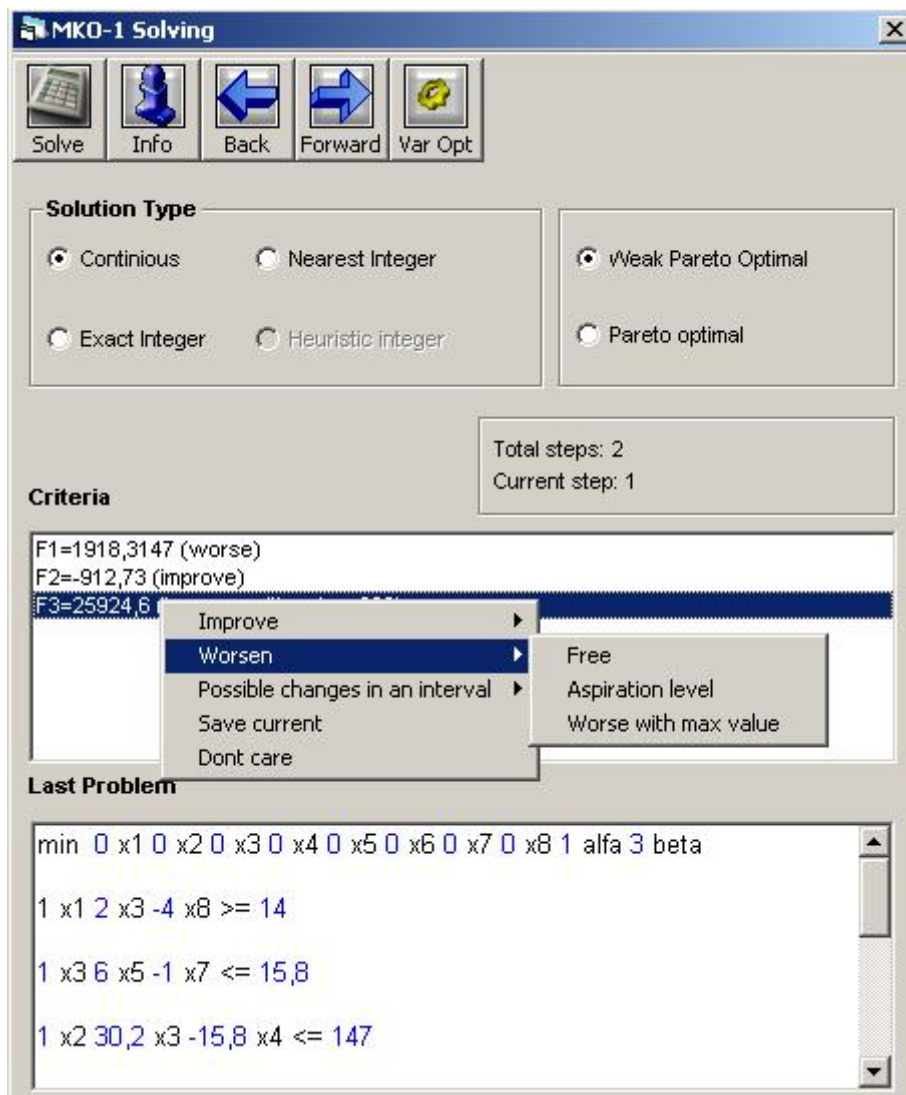


Fig. 2 MKO-1 Solving window

The interactive problems solving is realized in “*MKO-1 Solving*” window. “*MKO-1 Solving*” window is divided into several zones (Fig. 2). Its upper part contains a band with buttons that realize the main functions of the process for interactive solution of multicriteria linear and linear integer problems.

The next field of “*MKO-1 Solving*” window contains radio buttons for setup of the *MKO-1* solution looked for: continuous, integer, approximate integer, the closest integer, as well as weak Pareto optimal or Pareto optimal. Below them information is found about the time of the system operation for the current problem in seconds, the number of the step being currently considered and the total number of the executed steps.

Two text fields follow. The first one outputs successively the values of the criteria obtained at the current step. It is an operating field where DM’s preferences relating to the search of the next solution are set. After marking each one of the criteria, a context field is opened with the help of the mouse right button, where the DM sets the desired alteration in the value of this criterion at the following iteration. In case the selection is connected with the necessity to enter a particular value, *MKO-1* system opens an additional dialogue window and waits for the entry of the corresponding digital information.

When interactive algorithms are used for multicriteria optimization problems solving, it is an advantage to present information not only about the last solution found, but also about the process of search, as well as about all the previous steps. Given that some significant solutions are made on the basis of these results, it is important for the DM to be able to "testify" how he has reached this solution. That is why the information about the interactive process of multicriteria optimization problem considered, which consists of the problem input data, the solutions obtained at each step, the preferences set by the DM for a new search and the constructed scalarizing problems, saved in "*.mko" files, associated with *MKO-1* system serve not only for restarting an interrupted solution process, but also for documentation. "Print" command from the main menu can be used for selective print of the type of information chosen by the DM.

Conclusion

MultiDecision-1 system is designed to support the DM in solving different multicriteria analysis and multicriteria optimization problems. *MKA-1* system is designed to support the DM in modeling and solving problems of multicriteria ranking and multicriteria choice. The integrating of three different types of methods expands DM's possibilities to set his/her preferences about the quality of the most preferred solution. *MKO-1* system is designed to model and solve linear and linear integer problems of multicriteria optimization. The interactive classification-oriented algorithm included in the system offers to the DM wide possibilities to set his/her preferences about the qualities of the most preferred solution. The user-friendly interface of *MKA-1* system and *MKO-1* system facilitates the operation of DMs with different qualification level relating to the analysis and the optimization methods and software tools. *MKA-1* and *MKO-1* systems can be used for the purposes of education and for experimental and research problems solving as well.

Bibliography

- [Benayoun, et al., 1971] R.Benayoun, J.Montgolfier, J.Tergny, O.Laritchev. Linear Programming with Multiple Objective Functions: Step Method (STEM). *Mathematical Programming*, 1, 136-375, 1971
- [Brans and Mareschal, 1990] J.P.Brans and B.Mareschal. The Promethee Methods for MCDM: the Promcale, Gaia and Bankadviser Software. In: *Readings in Multiple Criteria Decision Aid*. Ed. C.A.Bana e Costa. Springer-Verlag, Berlin, 216-252, 1990.
- [Brans and Mareschal, 1994] J.P.Brans and B. Mareschal. The PROMCALC & GAIA Decision Support System for Multicriteria Decision Aid. *Decision Support System*, 12, 297-310, 1994.
- [Brans and Mareschal, 2000] J.P.Brans and B.Mareschal. How to Decide with PROMETHEE? Available on the Internet: <http://www.visualdecision.com>, 2000.
- [Gardiner and Vanderpooten, 1997] L.R.Gardiner and D.Vanderpooten. Interactive Multiple Criteria Procedures: Some Reflections. In: *Multicriteria Analysis*. Ed. J. Climaco. Springer-Verlag, Berlin, 290-301, 1997.
- [Jaszkievicz and Slowinski, 1994] A.Jaszkievicz and R.Slowinski. The Light Beam Search Over a Non-Dominated Surface of a Multiple-Objective Programming Problem. In: *Multiple Criteria Decision Making*. Ed. G.H.Tzeng, H.F.Wang, U.P. Wen and P.L.Yu. Springer-Verlag, Berlin, 87-99, 1994.
- [Keeney and Raiffa, 1976] R.Keeney and H.Raiffa. *Decisions with Multiple Objectives, Preferences and Value Trade Offs*. John Wiley & Sons, New York, 1976.
- [Korhonen, 1987] P.Korhonen. VIG - A Visual Interactive Support System for Multiple Criteria Decision Making. *Belgian Journal of Operations Research, Statistics and Computer Science*, 27(1), 3-15, 1987.
- [Korhonen, 1988] P.Korhonen. A Visual Reference Direction Approach to Solving Discrete Multiple Criteria Problems. *European Journal of Operational Research*, 34, 152-159, 1988.

- [Lewandowski and Wierzbicki, 1989] A.Lewandowski and A.P.Wierzbicki. Aspiration Based Decision Support Systems. Lecture Notes in Economics and Mathematical Systems, 331. Springer – Verlag, Berlin, 1989.
- [Miettinen, 1999] K.Miettinen. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston, 1999.
- [Miettinen and Makela, 2000] K.Miettinen and M.Makela. Interactive Multiobjective Optimization System WWW-NIMBUS on the Internet. Computer and Operation Research, 27, 709-723, 2000.
- [Mustajoki and Hamalainen, 2000] J.Mustajoki and R.P.Hamalainen. Web-HIPRE: Global Decision Support by Value Tree and AHP Analysis. INFOR, 38, 208-220, 2000.
- [Narula and Vassilev, 1994] S.C.Narula and V.Vassilev. An Interactive Algorithm for Solving Multiple Objective Integer Linear Programming Problems. European Journal of Operational Research, 79, 443-450, 1994.
- [Narula et al., 2003] S.C.Narula, V.Vassilev, K.Genova and M.Vassileva. A Partition-Based Interactive Method to Solve Discrete Multicriteria Choice Problems. Cybernetics and Information Technologies, 2, 55-66, 2003.
- [Nemhauser and Wolsey, 1988] G.L.Nemhauser and L.Wolsey. Integer and Combinatorial Optimization. Wiley, New York, 1988.
- [Ogryczak et al., 1992] W.Ogryczak, K.Stuchinski and K.Zorychta. DINAS: A Computer-Assisted Analysis System for Multiobjective Transshipment Problems with Facility Location. Computers and Operations Research, 19, 637-648, 1992.
- [Osyczka, 1988] A.Osyczka. Computer Aided Multicriterion Optimization System. In: Discretization Methods and Structural Optimization – Procedures and Applications. Ed. H.A.Eschenauer and G.Thierauf. Springer-Verlag, 263-270, 1988.
- [Parsons, 2002] J.Parsons. Agland Decision Tool: A Multicriteria Decision Support System for Agricultural Property. In: EMSs 2002, Integrated Assessment and Decision Support, Proceedings, Vol. 3, 181-187. Available on the Internet: <http://www.iemss.org/iemss2002/>, 2002.
- [Peterson, 1994] C.R.Peterson. HIVIEW – Rate and Weight to Evaluate Options. OR/MS Today, April, 1994.
- [Roy, 1991] B.Roy. The Outranking Approach and the Foundations of ELECTRE Methods. Theory and Decision, 31, 49-73, 1991.
- [Saaty, 1980] T.S.Saaty. The Analytic Hierarchy Process. McGraw-Hill, New York, 1980.
- [Saaty, 1994] T.S.Saaty. Highlights and Critical Points in the Theory and Application of the Analytic Hierarchy Process. European Journal of Operational Research, 74, 426-447, 1994.
- [Sun and Steuer, 1996] M.Sun and R.Steuer. InterQuad: An Interactive Quad Free Based Procedure for Solving the Discrete Alternative Multiple Criteria Problem. European Journal of Operational Research, 89, 462-472, 1996.
- [Vansnick, 1986] J.C.Vansnick. On the Problem of Weights in Multiple Criteria Decision Making (the Noncompensatory Approach). European Journal of Operational Research, 24, 288-294, 1986.
- [Vassilev et al., 1993] V.Vassilev, A.Atanassov, V.Sgurev, M.Kichovitch, A.Deianov and L.Kirilov. Software Tools for Multi-Criteria Programming. In: User-Oriented Methodology and Techniques of Decision Analysis and Support. Ed. J.Wessels and A.Wierzbicki. Springer- Verlag, Berlin, 247-257, 1993.
- [Vassilev et al., 1997] V.Vassilev, S.Narula, P.Vladimirov and V.Djambov. MOIP: A DSS for Multiple Objective Integer Programming Problems. In: Multicriteria Analysis. Ed. J. Climaco. Springer-Verlag, Berlin, 259-268, 1997.
- [Vassilev et al., 2003] V.Vassilev, K.Genova, M.Vassileva and S.Narula. Classification-Based Method of Linear Multicriteria Optimization. International Journal on Information Theories and Applications, 10, 3, 266-270, 2003.
- [Vassileva et al., 2001] M.Vassileva, K.Genova and V.Vassilev. A Classification Based Interactive Algorithm of Multicriteria Linear Integer Programming. Cybernetics and Information Technologies, 1, 5 – 20, 2001.
- [Vassileva, 2004] M.Vassileva. A Learning-Oriented Method of Linear Mixed Integer Multicriteria Optimization. Cybernetics and Information Technologies, 4, 1, 13-25, 2004.
- [Vincke, 1992] P.Vincke. Multicriteria Decision-Aid. John Wiley & Sons, New York, 1992.

[Von Winterfeldt and Edwards, 1986] D.VonWinterfeldt and W.Edwards. Decision Analysis and Behavioral Research. Cambridge University Press, London, 1986.

[Wierzbicki, 1980] A.P.Wierzbicki. The Use of Reference Objectives in Multiobjective Optimization. In: Multiple Criteria Decision Making Theory and Applications. Lecture Notes in Economics and Mathematical Systems, 177. Ed. G.Fandel and T.Gal. Springer-Verlag, Berlin, 468-486, 1980.

Authors' Information

Vassil Vassilev, PhD – Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: vvassilev@iinf.bas.bg

Krasimira Genova, PhD – Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: kgenova@iinf.bas.bg

Mariyana Vassileva-Ivanova, PhD – Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: mvassileva@iinf.bas.bg

GENERALIZED SCALARIZING PROBLEMS *GENS* AND *GENSLEX* OF MULTICRITERIA OPTIMIZATION¹

Mariyana Vassileva

Abstract: Generalized scalarizing problems, called *GENS* and *GENSLEX*, for obtaining Pareto optimal solutions of multicriteria optimization problems are presented in the paper. The basic properties of these scalarizing problems are described. The existence of single-criterion problems with differentiable objective functions and constraints, which are equivalent to *GENS* and *GENSLEX* scalarizing problems, are pointed out.

Keywords: Multicriteria optimization, Interactive methods, Multicriteria decision support systems.

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

Introduction

Various real problems can be modelled as multicriteria optimization problems. In multicriteria optimization problems several criteria are simultaneously optimized in the feasible set of alternatives. In the general case, there does not exist one alternative, which optimizes all the criteria. There is a set of alternatives however, characterized by the following: each improvement in the value of one criterion leads to deterioration in the value of at least one other criterion. This set of alternatives is called a set of the Pareto optimal alternatives (solutions). Each alternative in this set could be a solution of the multicriteria optimization problem. In order to select one alternative, it is necessary to have additional information set by the so-called decision maker (DM).

¹ This paper is partially supported by the National Science Fund of Bulgarian Ministry of Education and Science under contract № I-1401\2004 "Interactive Algorithms and Software Systems Supporting Multicriteria Decision Making".

The information that the DM provides reflects his/her global preferences with respect to the quality of the most preferred alternative.

The general problem of multicriteria optimization (MO) can be represented in the following way:

$$\begin{aligned} & \max \{ f_k(x), k \in K \} \\ & \text{subject to: } x \in X \end{aligned}$$

where:

- $f_k(x)$, $k \in K = \{1, 2, \dots, p\}$ are different criteria (objective functions) of the type $f_k: R^n \rightarrow R$, which must be simultaneously maximized;
- $x = (x_1, \dots, x_j, \dots, x_n)$ is the vector of variables, belonging to the non-empty feasible set $X \subset R^n$;
- $Z = f(X) \subset R^p$ is the feasible set of the criteria values.

The scalarizing approach is one of the main approaches in solving MO problems. The basic representatives of the scalarizing approach ([Wierzbicki, 1980], [Sawaragi, Nakayama and Tanino, 1985], [Steuer, 1986], [Narula and Vassilev, 1994], [Buchanan, 1997], [Miettinen, 1999], [Vassileva, 2004], [Ehrgott and Wiecek, 2004]) are the interactive algorithms. The MO problem in these algorithms is treated as a decision-making problem and the emphasis is placed on the real participation of the DM in the process of its solution. Each interactive algorithm consists of two procedures in the general case – an optimization one and an evaluating one, which are cyclically repeated until the stopping conditions are satisfied. During the evaluating procedure the DM estimates the current Pareto optimal solution obtained, either approving it as the final (the most preferred) one, or setting his/her preferences in the search for a new solution. On the basis of these preferences a scalarizing problem is formed and solved in the optimization procedure and a new Pareto optimal solution is obtained with its help, which is presented to the DM for evaluation and choice. The main feature of each scalarizing problem is that every optimal solution is a Pareto optimal solution of the corresponding MO problem. The scalarizing problem is a single-criterion optimization problem, which allows the application of the theory and methods of single-criterion optimization. A number of scalarizing problems and a set of interactive algorithms developed on their basis have been proposed up to now. The different algorithms offer different possibilities to the DM in the control or in stopping the process of the final solution finding. On its hand, this searching process can be divided into two phases. In the first phase (the learning phase), the DM usually defines the region, in which he expects to find the most preferred solution, whereas in the second phase (the concluding phase) he is looking for this solution namely in this region.

The present paper describes generalized scalarizing problems, called *GENS* and *GENSLex*. They are extensions of the generalized scalarizing problem GENWS [Vassilev, 2004] and enables the obtaining of Pareto optimal solutions. Almost all scalarizing problems known up to now can be obtained from *GENS* and *GENSLex* problems, as well as new scalarizing problems with different properties can be generated from these problems.

Generalized Scalarizing Problems *GENS* and *GENSLex*

For easier description of the topic further on, the following definitions will be introduced:

Definition 1: The solution $x \in X$ is called a Pareto optimal solution of the multicriteria optimization problem, if there does not exist another solution $\bar{x} \in X$, satisfying the following conditions:

$$f_k(\bar{x}) \geq f_k(x), k \in K \text{ and } f_k(\bar{x}) > f_k(x) \text{ for at least one index } k \in K.$$

Definition 2: The vector $z = f(x) = (f_1(x), \dots, f_p(x))^T \in Z$ is called a Pareto optimal solution in the criterion space, if $x \in X$ is a Pareto optimal solution in the decision variable space.

Definition 3: The current preferred solution $z = (f_1, \dots, f_k, \dots, f_p)^T \in Z$ is a Pareto optimal solution in the criterion space, selected by the DM at the current iteration.

Definition 4: The most preferred solution is the current preferred solution, which satisfies the DM to the highest extent.

Definition 5: The criteria classification is called the implicit division of the criteria into classes, depending on the alterations in the criteria values at the current solution, which the DM wishes to obtain.

In order to obtain Pareto optimal solutions starting from the current preferred solution, GENS scalarizing problem is proposed. It has the following type:

Minimize

$$(1) \quad T(x) = \max_{k \in K^{\geq}} (F_k^1 - f_k(x))G_k^1 R_1 \max_{k \in K^{\leq}} (F_k^2 - f_k(x))G_k^2 R_2 \max_{k \in K^{<}} (F_k^3 - f_k(x))G_k^3 \\ + R_3 \max_{k \in K^{>}} (F_k^4 - f_k(x))G_k^4 + \sum_{k \in K^0} (F_k^5 - f_k(x))G_k^5 + \\ + \rho \left(\sum_{k \in K^{\geq}} (F_k^1 - f_k(x))G_k^1 + \sum_{k \in K^{\leq}} (F_k^2 - f_k(x))G_k^2 + \sum_{k \in K^{<}} (F_k^3 - f_k(x))G_k^3 + \right. \\ \left. + \sum_{k \in K^{>}} (F_k^4 - f_k(x))G_k^4 - \sum_{k \in K^{\neg} \cup K^{\succ}} f_k(x)G_k^6 \right),$$

subject to:

- (2) $f_k(x) \geq f_k, k \in K^{>} \cup K^=$
- (3) $f_k(x) \geq f_k - D_k, k \in K^{\leq}$
- (4) $f_k(x) \geq f_k - t_k^-, k \in K^{\succ}$
- (5) $f_k(x) \leq f_k + t_k^+, k \in K^{\succ}$
- (6) $x \in X$

where:

- K is the set of all the criteria;
- $G_k^1, G_k^2, G_k^3, G_k^4, G_k^5$ are scaling, normalizing or weighting positive coefficients, $k \in K$;
- $F_k^1, F_k^2, F_k^3, F_k^4, F_k^5$ are parameters, connected with aspiration, current or other levels of the criteria values, $k \in K$;
- R_1, R_2, R_3 are equal to the arithmetic "+" or to a separator " , " ;
- D_k is the value, by which the DM agrees the criterion with an index $k \in K^{\leq}$ to be deteriorated ($D_k > 0$);

- t_k^- and t_k^+ are the lower and upper bound of the feasible for the DM interval of alteration of the criterion with an index $k \in K^{\times}$ ($t_k^- > 0$; $t_k^+ > 0$);
- f_k is the value of the criterion with an index $k \in K$ in the current solution obtained;
- K^{\geq} is the set of criteria, the current values of which the DM wishes to be improved up to desired by him/her levels F_k^1 ;
- $K^>$ is the set of the criteria, the current values of which the DM wishes to be improved;
- K^{\leq} is the set of the criteria, for which the DM agrees their current values to be deteriorated up to set by him/her feasible levels F_k^2 , but not more than certain values D_k ($D_k > 0$);
- $K^<$ is the set of criteria, for which the DM agrees their current values to be deteriorated;
- $K^=$ is the set of criteria, for which the DM agrees their current values not to be deteriorated;
- K^{\times} is the set of the criteria, for which the DM agrees their values to alter in defined intervals;
- K^0 is the set of criteria, for which the DM does not set explicit preferences concerning the change of their values;
- ρ is a small positive number.

The constraints (2) - (6) define a subset of X , containing Pareto optimal solutions.

Theorem 1: *The optimal solution of GENS scalarizing problem is a Pareto optimal solution of the multicriteria optimization problem.*

Proof:

Let $K^{\geq} \neq \emptyset$ and/or $K^> \neq \emptyset$, or $K^0 = K$ and let $x^* \in X$ be an optimal solution of GENS scalarizing problem. Then the constraints (2) - (6) are satisfied for $x^* \in X$, together with the following condition:

$$(7) \quad T(x^*) \leq T(x), x \in X.$$

Let us assume that $x^* \in X$ is not a Pareto optimal solution of the multicriteria optimization problem. Then, another $x' \in X$ must exist, for which the constraints (2) - (6) are satisfied, as well as the conditions given below:

$$(8) \quad f_k(x') \geq f_k(x^*), k \in K \quad \text{and} \quad f_k(x') > f_k(x^*) \quad \text{for at least one index } k \in K.$$

Inequality (8) follows from the definition of a Pareto optimal solution.

Using constraint (8) and the definitions of R_1, R_2, R_3 , the objective function $T(x)$ of scalarizing problem GENS can be transformed, obtaining the following inequality:

$$(9) \quad T(x') = \max_{k \in K^{\geq}} (F_k^1 - f_k(x')) G_k^1 R_1 \max_{k \in K^{\leq}} (F_k^2 - f_k(x')) G_k^2 + \\ R_2 \max_{k \in K^<} (F_k^3 - f_k(x')) G_k^3 R_3 \max_{k \in K^>} (F_k^3 - f_k(x')) G_k^4 + \\ + \sum_{k \in K^0} (F_k^5 - f_k(x')) G_k^5 +$$

$$\begin{aligned}
& + \rho \left(\sum_{k \in K^{\geq}} (F_k^1 - f_k(x')) G_k^1 + \sum_{k \in K^{\leq}} (F_k^2 - f_k(x')) G_k^2 + \sum_{k \in K^{<}} (F_k^3 - f_k(x')) G_k^3 + \right. \\
& \quad \left. + \sum_{k \in K^{>}} (F_k^4 - f_k(x')) G_k^4 - \sum_{k \in K^{\circ} \cup K^{>\circ}} f_k(x') G_k^6 \right) = \\
& = \max_{k \in K^{\geq}} \left((F_k^1 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^1 R_1 \\
& \quad \max_{k \in K^{\leq}} \left((F_k^2 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^2 R_2 \\
& \quad \max_{k \in K^{<}} \left((F_k^3 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^3 R_3 \\
& \quad \max_{k \in K^{>}} \left((F_k^4 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^4 + \\
& \quad + \sum_{k \in K^0} \left((F_k^5 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^5 + \\
& + \rho \left(\sum_{k \in K^{\geq}} \left((F_k^1 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^1 + \right. \\
& \quad + \sum_{k \in K^{\leq}} \left((F_k^2 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^2 + \\
& \quad + \sum_{k \in K^{<}} \left((F_k^3 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^3 + \\
& \quad + \sum_{k \in K^{>}} \left((F_k^4 - f_k(x^*)) + (f_k(x^*) - f_k(x')) \right) G_k^4 - \\
& \quad \left. - \sum_{k \in K^{\circ} \cup K^{>\circ}} (f_k(x^*) + (f_k(x') - f_k(x^*))) G_k^6 \right) < \\
& < \max_{k \in K^{\geq}} (F_k^1 - f_k(x^*)) G_k^1 R_1 \max_{k \in K^{\leq}} (F_k^2 - f_k(x^*)) G_k^2 \\
& \quad R_2 \max_{k \in K^{<}} (F_k^3 - f_k(x^*)) G_k^3 R_3 \max_{k \in K^{>}} (F_k^4 - f_k(x^*)) G_k^4 + \\
& \quad + \sum_{k \in K^0} (F_k^5 - f_k(x^*)) G_k^5 + \\
& + \rho \left(\sum_{k \in K^{\geq}} (F_k^1 - f_k(x^*)) G_k^1 + \sum_{k \in K^{\leq}} (F_k^2 - f_k(x^*)) G_k^2 + \sum_{k \in K^{<}} (F_k^3 - f_k(x^*)) G_k^3 + \right. \\
& \quad \left. + \sum_{k \in K^{>}} (F_k^4 - f_k(x^*)) G_k^4 - \sum_{k \in K^{\circ} \cup K^{>\circ}} f_k(x^*) G_k^6 \right) = \\
& = T(x^*).
\end{aligned}$$

It follows from (9) that $T(x') < T(x^*)$, which contradicts to (7). Hence, $x^* \in X$ is a Pareto optimal solution of the multicriteria optimization problem.

The scalarizing problem *GENS* guarantees that Pareto optimal solutions are generated. The common drawback [Miettinen, 1999] is how to select the coefficient ρ . An alternative way is to use a lexicographic approach. The following *GENSLex* problem in two phases is a lexicographic variant of scalarizing problem *GENS*.

The first problem *GENSLex1* to be solved is the following:

Minimize

$$(10) \quad T_1(x) = \max_{k \in K^{\geq}} (F_k^1 - f_k(x))G_k^1 \quad R_1 \max_{k \in K^{\leq}} (F_k^2 - f_k(x))G_k^2 \quad R_2 \max_{k \in K^{<}} (F_k^3 - f_k(x))G_k^3 \\ R_3 \max_{k \in K^{>}} (F_k^4 - f_k(x))G_k^4 + \sum_{k \in K^0} (F_k^5 - f_k(x))G_k^5$$

subject to:

$$(11) \quad f_k(x) \geq f_k, k \in K^{>} \cup K^{=}$$

$$(12) \quad f_k(x) \geq f_k - D_k, k \in K^{\leq}$$

$$(13) \quad f_k(x) \geq f_k - t_k^-, k \in K^{><}$$

$$(14) \quad f_k(x) \leq f_k + t_k^+, k \in K^{><}$$

$$(15) \quad x \in X$$

Let us denote the optimal objective function value of (10) by T_1^* . The final solution is obtained by solving the following problem GENSLex2:

Minimize

$$(16) \quad T_2(x) = \sum_{k \in K^{\geq}} (F_k^1 - f_k(x))G_k^1 + \sum_{k \in K^{\leq}} (F_k^2 - f_k(x))G_k^2 + \sum_{k \in K^{<}} (F_k^3 - f_k(x))G_k^3 + \\ + \sum_{k \in K^{>}} (F_k^4 - f_k(x))G_k^4 - \sum_{k \in K^{=} \cup K^{><}} f_k(x)G_k^6$$

subject to:

$$(17) \quad T_1(x) = \max_{k \in K^{\geq}} (F_k^1 - f_k(x))G_k^1 \quad R_1 \max_{k \in K^{\leq}} (F_k^2 - f_k(x))G_k^2 \quad R_2 \max_{k \in K^{<}} (F_k^3 - f_k(x))G_k^3 \\ R_3 \max_{k \in K^{>}} (F_k^4 - f_k(x))G_k^4 + \sum_{k \in K^0} (F_k^5 - f_k(x))G_k^5 \leq T_1^*$$

and constraints (11) - (15).

Theorem 2: The optimal solution of GENSLex scalarizing problem is a Pareto optimal solution of the multicriteria optimization problem.

Proof:

Let $K^{\geq} \neq \emptyset$ and/or $K^{>} \neq \emptyset$, or $K^0 = K$ and let $x^* \in X$ be an optimal solution of GENLex scalarizing problem. Then the constraints (11) - (15) are satisfied for $x^* \in X$, together with the following conditions:

$$T_1(x^*) \leq T_1(x) \text{ and } T_2(x^*) \leq T_2(x), x \in X.$$

Let us assume that $x^* \in X$ is not a Pareto optimal solution of the multicriteria optimization problem. Then there must exist another $x' \in X$, for which the constraints (11) - (15) are satisfied, as well as the condition given below:

$$(18) \quad f_k(x') \geq f_k(x^*), k \in K$$

and $f_k(x') > f_k(x^*)$ for at least one index $k \in K$.

It is clear that independently of defined values of R_1, R_2, R_3 and from (18) and (10 - 17) follows that:

$$T_1(x') \leq T_1(x^*) \text{ and } T_2(x') < T_2(x^*)$$

or

$$T_1(x') < T_1(x^*) \text{ and } T_2(x') \leq T_2(x^*),$$

which contradicts with x^* being an optimal solution of *GENLex* scalarizing problem.

Scalarizing problem *GENS* is in the general case an optimization problem with a non-differentiable objective function. Every *GENS* scalarizing problem (defined values of R_1, R_2, R_3) can be reduced to an equivalent optimization problem with a differentiable objective function on the account of additional variables and constraints. The equivalency of each pair of optimization problems is in relation to the obtained values of the objective functions (criteria) and the main variables. Different types of equivalent problems are obtained at different values of R_1, R_2, R_3 .

Every equivalent problem can be presented as follows:

$$\min \left(\mu + \sum_{k \in K^0} y_k + \rho \sum_{k \in K \setminus K^0} y_k \right)$$

and satisfies two groups of constraints.

The first group of constraints is equal for all types of equivalent problems and has the following form:

$$(19) \quad \alpha \geq (F_k^1 - f_k(x))G_k^1, k \in K^{\geq}$$

$$(20) \quad \beta \geq (F_k^2 - f_k(x))G_k^2, k \in K^{\leq}$$

$$(21) \quad \gamma \geq (F_k^3 - f_k(x))G_k^3, k \in K^{<}$$

$$(22) \quad \Omega \geq (F_k^4 - f_k(x))G_k^4, k \in K^{>}$$

$$(23) \quad (F_k^1 - f_k(x))G_k^1 = y_k, k \in K^{\geq}$$

$$(24) \quad (F_k^2 - f_k(x))G_k^2 = y_k, k \in K^{\leq}$$

$$(25) \quad (F_k^3 - f_k(x))G_k^3 = y_k, k \in K^{<}$$

$$(26) \quad (F_k^4 - f_k(x))G_k^4 = y_k, k \in K^{>}$$

$$(27) \quad (F_k^5 - f_k(x))G_k^5 = y_k, k \in K^0$$

$$(28) \quad -f_k(x)G_k^6 = y_k, k \in K^= \cup K^{\times}$$

$$(29) \quad f_k(x) \geq f_k, k \in K^{>} \cup K^=$$

$$(30) \quad f_k(x) \geq f_k - D_k, k \in K^{\leq}$$

$$(31) \quad f_k(x) \geq f_k - t_k^-, k \in K^{\times}$$

$$(32) \quad f_k(x) \leq f_k + t_k^+, k \in K^{\times}$$

$$(33) \quad x \in X$$

$\alpha, \beta, \gamma, \Omega, y_k / k \in K$ - arbitrary.

The second group of constraints has different type and number of constraints depending on the values of R_1, R_2, R_3 . The constraints from the second group for one equivalent problem of scalarizing problem *GENS*, which is obtained when R_1 is equal to the separator “,”, R_2 and R_3 are equal to the arithmetic operation “+”, have the following form:

$$(34) \quad \mu \geq \alpha$$

$$(35) \quad \mu \geq \beta + \gamma + \Omega$$

μ - arbitrary.

The constraints from the second group in the other equivalent problems can be stated in a similar way.

Scalarizing problems *GENSLex1* and *GENSLex2* are in the general case optimization problems with a non-differentiable objective functions and constraints. Every scalarizing problem of both types *GENSLex1* and *GENSLex2* (defined values of R_1, R_2, R_3) can be reduced to an equivalent optimization problems with a differentiable objective functions and constraints on the account of additional variables and constraints.

Different types of equivalent problems of scalarizing problem *GENSLex1* are obtained at different values of R_1, R_2, R_3 . Each equivalent problem can be presented as follows:

$$(36) \quad \min \left(\mu + \sum_{k \in K^0} y_k \right),$$

satisfying two groups of constraints. The first group of constraints is equal for all types of equivalent problems and has the following form:

$$(37) \quad \alpha \geq (F_k^1 - f_k(x))G_k^1, k \in K^{\geq}$$

$$(38) \quad \beta \geq (F_k^2 - f_k(x))G_k^2, k \in K^{\leq}$$

$$(39) \quad \gamma \geq (F_k^3 - f_k(x))G_k^3, k \in K^{<}$$

$$(40) \quad \Omega \geq (F_k^4 - f_k(x))G_k^4, k \in K^{>}$$

$$(41) \quad (F_k^5 - f_k(x))G_k^5 = y_k, k \in K^0$$

$$(42) \quad f_k(x) \geq f_k, k \in K^{>} \cup K^{=}$$

$$(43) \quad f_k(x) \geq f_k - D_k, k \in K^{\leq}$$

$$(44) \quad f_k(x) \geq f_k - t_k^-, k \in K^{\times<}$$

$$(45) \quad f_k(x) \leq f_k + t_k^+, k \in K^{\times>}$$

$$(46) \quad x \in X$$

$\alpha, \beta, \gamma, \Omega, y_k / k \in K^0$ - arbitrary.

The second group of constraints has different type and number of constraints depending on the values of R_1, R_2, R_3 . The constraints from the second group for one equivalent problem of scalarizing problem

GENSLex1, which is obtained when R_1 is equal to the separator “,”, R_2 and R_3 are equal to the arithmetic operation “+”, have the following form:

$$(47) \quad \mu \geq \alpha$$

$$(48) \quad \mu \geq \beta + \gamma + \Omega$$

μ - arbitrary.

Different types of equivalent problems of scalarizing problem *GENSLex2* are obtained at different values of R_1, R_2, R_3 . Each equivalent problem can be presented as follows:

$$(49) \quad \min \left(\sum_{k \in K \setminus K^0} y_k \right)$$

and satisfies two groups of constraints.

The first group of constraints is equal for all types of equivalent problems and has the following form:

$$(50) \quad (F_k^1 - f_k(x))G_k^1 = y_k, k \in K^{\geq}$$

$$(51) \quad (F_k^2 - f_k(x))G_k^2 = y_k, k \in K^{\leq}$$

$$(52) \quad (F_k^3 - f_k(x))G_k^3 = y_k, k \in K^{<}$$

$$(53) \quad (F_k^4 - f_k(x))G_k^4 = y_k, k \in K^{>}$$

$$(54) \quad -f_k(x)G_k^6 = y_k, k \in K^= \cup K^{\gg}$$

$$(55) \quad f_k(x) \geq f_k, k \in K^{>} \cup K^=$$

$$(56) \quad f_k(x) \geq f_k - D_k, k \in K^{\leq}$$

$$(57) \quad f_k(x) \geq f_k - t_k^-, k \in K^{\gg}$$

$$(58) \quad f_k(x) \leq f_k + t_k^+, k \in K^{\gg}$$

$$(59) \quad x \in X$$

$y_k / k \in K \setminus K^0$ - arbitrary.

The second group of constraints has different type and number of constraints depending on the values of R_1, R_2, R_3 . The constraints from the second group for one equivalent problem of scalarizing problem *GENSLex2*, which is obtained when R_1 is equal to the separator “,”, R_2 and R_3 are equal to the arithmetic operation “+”, have the following form:

$$(60) \quad \alpha \geq (F_k^1 - f_k(x))G_k^1, k \in K^{\geq}$$

$$(61) \quad \beta \geq (F_k^2 - f_k(x))G_k^2, k \in K^{\leq}$$

$$(62) \quad \gamma \geq (F_k^3 - f_k(x))G_k^3, k \in K^{<}$$

$$(63) \quad \Omega \geq (F_k^4 - f_k(x))G_k^4, k \in K^{>}$$

$$(64) \quad (F_k^5 - f_k(x))G_k^5 = y_k, k \in K^0$$

$$(65) \quad \mu \geq \alpha$$

$$(66) \quad \mu \geq \beta + \gamma + \Omega$$

$$(67) \quad \left(\mu + \sum_{k \in K^0} y_k \right) \leq T_1^*$$

$\alpha, \beta, \gamma, \Omega, \mu, y_k / k \in K^0$ - arbitrary.

Conclusion

The interactive algorithms solving different types of multicriteria optimization problems use different scalarizing problems. The features of each scalarizing problem are defined by the possibilities offered to the decision maker to set his/her preferences, as well as by the quality of the Pareto optimal solutions obtained. Altering the parameters of the generalized scalarizing problems *GENS* and *GENSLex*, a great part of the already known scalarizing problems can be obtained and also new scalarizing problems can be generated. In connection with this, generalized interactive algorithms with alterable scalarization and parameterization can be designed, which expand to a great extent the possibilities of the decision-maker in describing his/her preferences.

Bibliography

- [Buchanan, 1997] J.T. Buchanan. A Naive Approach for Solving MCDM Problems: The GUESS Method. Journal of the Operational Research Society, 48, pp. 202 – 206, 1997.
- [Ehrgott and Wiecek, 2004] M.Ehrgott and M.Wiecek. Multiobjective Programming. In: Multiple Criteria Decision Analysis: State of the Art Surveys. Ed. J.Figueira, S.Greco and M.Ehrgott. Springer, London, 2004.
- [Miettinen, 1999] K.Miettinen. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston, 1999.
- [Narula and Vassilev, 1994] S.Narula and V.Vassilev. An Interactive Algorithm for Solving Multiple Objective Integer Linear Programming Problems. European Journal of Operational Research, 79, pp. 443–450, 1994.
- [Sawaragi, Nakayama and Tanino, 1985] Y.Sawaragi, H.Nakayama and T.Tanino. Theory of Multiobjective Optimization. Academic Press, Inc., Orlando, Florida, 1985.
- [Steuer, 1986] R.E.Steuer. Multiple Criteria Optimization: Theory, Computation and Applications. John Wiley & Sons, Inc., 1986.
- [Vassilev, 2004] V.Vassilev. A Generalized Scalarizing Problem of Multicriteria Optimization. Working papers of IIT-BAS, IIT/ WP-187B, 2004.
- [Vassileva, 2004] M.Vassileva. A Learning-Oriented Method of Linear Mixed Integer Multicriteria Optimization. Cybernetics and Information Technologies, vol. 4, No 1, pp. 13-25, 2004.
- [Wierzbicki, 1980] A.P.Wierzbicki. The Use of Reference Objectives in Multiobjective Optimization. In: Multiple Criteria Decision Making Theory and Applications, Lecture Notes in Economics and Mathematical Systems, vol. 177, pp. 468-486. Ed. G.Fandel and T.Gal. Springer-Verlag, Berlin, Heidelberg, 1980.

Author Information

Mariyana Vassileva – Ivanova, PhD – Research Associate, Institute of Information Technologies, BAS; Acad. G. Bonchev Str., bl. 29A, Sofia 1113, Bulgaria; e-mail: mvassileva@iinf.bas.bg.

SOFTWARE DEVELOPMENT FOR DISTRIBUTED SYSTEM OF RUSSIAN DATABASES FOR ELECTRONICS MATERIALS

Valery Kornyshko, Victor Dudarev

Abstract: Current state of Russian databases for substances and materials properties was considered. A brief review of integration methods of given information systems was prepared and a distributed databases integration approach based on metabase was proposed. Implementation details were mentioned on the posed database on electronics materials integration approach. An operating pilot version of given integrated information system implemented at IMET RAS was considered.

Keywords: distributed database integration, metabase, Web services, database on electronics materials.

ACM Classification Keywords: C.2.4 Distributed applications, Distributed databases; D.4.4 Network communication.

Introduction

Development and utilization of databases for substances and materials properties is a basis in providing information service for specialists in chemistry and materials science. Every research organization aimed at its own data center creation. Such data centers contain information closely related to research areas of a particular organization. Historically several data centers were formed for data storage and processing in every organization (scientific research institute or university). This can be explained not only by administrative reasons, but rather by significant differences in the problem domain. Existing situation creates great problems for accessing such data, because this information is dispersed over numerous data sources.

At present time, period of such an information fragmentation is coming to the end due to rapid IT-industry development. Present-day progress in science and technique stimulates concentration of diverse information on physicochemical substances properties. Modern polyfunctional materials development requires from us high standard of knowledge in different properties of substances. Efficient online information service (for materials science engineers and chemists providing full data from reliable sources) decreases baseless papers' duplication and ultimately it reduces cost and time required for modern materials development. Inaccessibility and frequently dispersion of information over different heterogeneous data sources makes great difficulties in decision-making process considering application of one or another material.

During development integrated information system presented in this paper the key task was to create an intelligent, simple in architecture and effective software infrastructure. This software infrastructure should integrate data on properties of substances and materials rationally and reasonably. Integration means are required that should be capable to provide not only unified access to operating data centers, but these integration means should allow us to create comprehensive data access infrastructure based on unified standards and also on uniform network interconnection principles.

Database Integration Approaches Overview

Principally there are two approaches to database integration.

The first one implies full merging of existing resources. That is the case when database complex is a single information system (megabase) for end users, operators and administrators. Database exploitation costs reduction and information duplication decrease can be mentioned among advantages of this very variant.

Every data center is a point of information concentration and online data analytical processing. In addition, technology of information accumulating and data processing is settled down in each organization. Moreover, great investments that were made in hardware and software do not allow solving the data dissociation problem by mechanical transportation to some centralized database of all data. Moreover Russian databases for electronics

materials have been developed in various organizations and thus they took advantage of different database management systems (DBMS). Taking into consideration differences in data quality, data expertise, data store types and many other troubles emerging when changing existing systems operating principles it should be stated that full and smooth integration is practically impossible for above mentioned resources.

The second integration approach main essence is that we are not going to integrate databases themselves, but we want to integrate their proprietary user interfaces only. From the one hand this approach allows us not to change every integrated database structure dramatically (and thus established database utilization and administration technology – data update and insert). From the other hand, this approach allows the end user to get access to the whole information picture on chemical substances stored in different databases. So called “virtual” database integration (or in other words heterogeneous information system creation) implies independence in evolution of separate subsystems and at the same time end user gets access to the whole information array on a particular chemical substance or material stored in databases of a virtually united system (fig. 1). And that fact solves the main integration goal truly.

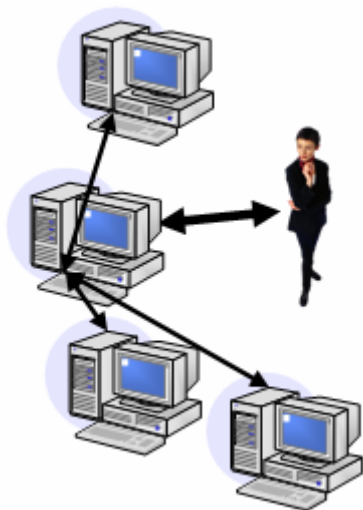


Fig. 1. Database integration at Web interface level

Taking into consideration current development conditions of Russian databases on physicochemical substances' properties the second integration approach – integration at interface level only – is more appropriate and quite perspective. It's worth mentioning that Web-interfaces have been developed for IMET RAS databases on physicochemical substance properties (“Crystal” database on acoustic-optical, electro-optical and non-linear optical substances and “Diagram” database on semiconductor systems phase diagrams). These Web-interfaces allow users to get remote access via Internet to data stored in these databases using any Web browser.

Searching for Relevant Information in Integrated System

When integrating databases at Web-interfaces level it's required to provide facilities for browsing information contained in other databases. This information should be relevant to the data on some chemical system currently being browsed by user. Let's consider this in the following example. User who browses information on Ga-As system from “Diagram” database should have an opportunity to get information for example on piezoelectric effect or non-linearoptical properties of GaAs substance contained in “Crystal” database.

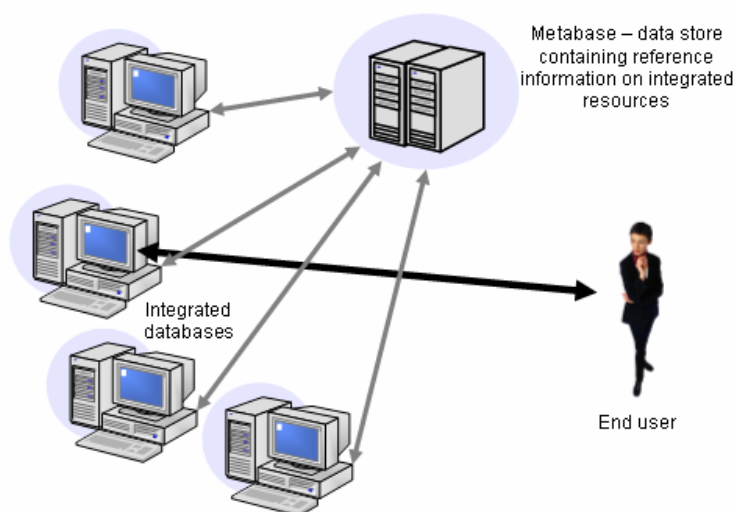


Fig. 2. Metabase concept

So it's obvious that when designing distributed information system, it's required to provide search for relevant information contained in other databases of distributed system. Thus, we hardly need to have some active data center that should know what information is contained in every integrated database. Obviously some data store should exist that somehow describes information contained in integrated database resources. In this manner, we come to the metabase concept – a special database that contains some reference information on integrated databases contents (fig. 2).

In our case, it is information on chemical systems and their properties. The amount of this metainformation should be enough to perform search for relevant information on systems and corresponding properties.

Let's try to formalize the problem in terms of set-theoretic approach. Hence, metabase should contain information on integrated databases (D set), information on chemical substances and systems (S set) and information on their properties (P set). To describe correlation between elements of D , S and P sets let's define ternary relation called W on set $U = D \times S \times P$.

Here U is a Cartesian product of D , S and P . Membership of a (d, s, p) triplet to the W relation, where $d \in D, s \in S, p \in P$, can be interpreted in the following way: "Information on property p of chemical system s is contained in integrated database d ". Having defined three basic sets it can be seen that search for information relevant to s system can be localized to determination of R relationship, that is a subset of Cartesian product $S \times S$ (or in other words, $R \subset S^2$). Thus, it can be stated about every pair $(s_1, s_2) \in R$ that chemical system s_2 is relevant to the system s_1 . So all we need to solve the task of searching for relevant information in integrated databases is to determine somehow the R relation. It is significant to note that R relation can be created or complemented by means of either of two variants. The first variant is via using predefined rules by a computer. The second one is that experts in chemistry and materials science can be engaged to solve this task.

The second variant is quite clear – experts can form relationship R manually following some multicriterion rules affected by their expert assessments. So let's consider possible variants of automatic R relation generation. One of such variants can be like this one based on the following rules:

1. For any chemical systems $s_1 \in S, s_2 \in S$ composed from chemical elements e_{ij} $s_1 = \{e_{11}, e_{12}, \dots, e_{1n}\}, s_2 = \{e_{21}, e_{22}, \dots, e_{2m}\}$ it is true, that if $s_1 \subseteq s_2$ (i.e. all chemical elements of system s_1 are contained in system s_2), then $(s_1, s_2) \in R$.
2. R relation is symmetric. In other words for any $s_1 \in S, s_2 \in S$, it is true, that if $(s_1, s_2) \in R$, then $(s_2, s_1) \in R$ as well.

These two rules allow us to determine a set of chemical systems relevant to the given one. It should be noticed that this automatic R relation generation variant is just one of the simplest and most obvious variants of such rules, and in fact more complex mechanisms can be used to get R relation. For example, browsing information on a particular property of a compound in one of integrated databases (in fact, it is information defined by (d_1, s_1, p_1) triplet), we consider (d_2, s_2, p_2) triplet to be relevant information. (d_2, s_2, p_2) triplet characterizes information on some other property of a system from another integrated database. In this case, we have got more complex relevance relation like this $R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2)$, where $d_1, d_2 \in D; s_1, s_2 \in S; p_1, p_2 \in P$. In fact we can even define a set of several R relations (R_1, R_2, \dots, R_n) by applying different rules. Thus we'll be able to perform search for relevant information based on wide variety of R interpretations.

Loading Information into Metabase

As it was stated above, Russian databases on materials science have been developed in different organizations on various platforms and that fact makes integration significantly more complicated. Metabase should store reference data on integrated resources contents. It's obvious that in this situation it's required to use open network interconnection principles and standards supported on multiple platforms. If we consider present technology stack then it's quite clear that currently Web services are connection links between different platforms and heterogeneous environments. Web services are based on common standards such as SOAP (Simple Object Access Protocol) and XML (eXtensible Markup Language). Nowadays these technologies are capable to provide reliable infrastructure for cross platform message exchange.

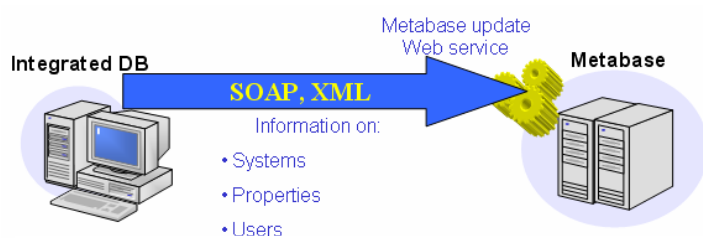


Fig. 3. Metabase update Web service

In that way reference information loading into metabase was implemented by means of metabase update Web service, so-called MUSEvice (fig. 3). Let's consider metadata updating mechanisms in detail. System that is to be integrated with others should generate XML document which contains information on updates in that very system.

The layout format of this XML document is

generally standardized for all integrated subsystems and it is strictly fixed by means of specially developed XML schema [1]. Thus, all subsystems being integrated should generate valid XML document that meets XML schema requirements to notify metabase of information changes that occurred in their state. After being generated, XML document is being sent to the metabase update Web service for processing and metabase update. Interaction with this MUSEvice is implemented by means of SOAP protocol according to the Web service WSDL (Web Services Description Language) description [2]. In that way client databases report about updates to the metabase and so actual information on integrated resources contents appears in the metabase.

It's important to note that security issues were among primary concerns while designing and implementing the resulting system. In that way symmetric encryption mechanism was implemented with the aid of DES (Digital Encryption Standard). It guarantees secure metadata exchange with metabase update Web service. Additionally an option for data archiving was included into the system. A kind of zip-achieving was implemented that allows us to package data transmitted to the metabase server. This feature allows dramatically decreasing data volumes being transmitted via public networks (taking into consideration a high level of compression for XML documents). Thus this feature lowers requirements to network bandwidth and it is very important and actual for Russia since high-speed Internet access is not available everywhere in the country. Compressing techniques enable us to decrease information volumes so that it becomes possible to use old-fashioned data modems on telephone wires to transmit data.

As it can be seen, metabase update Web service supports some rather complex additional data transformations (encryption and archiving) requiring some extra coding. So to simplify the interaction process with the Web service a special Web client has been created. It was implemented as a COM object and thus it can be easily accessible from any environment which supports Microsoft COM. Created Web client addresses issues connected with encryption and compression of information that is to be sent to the Web service. It controls all network interconnection aspects also. All this functionality just simplifies routine database attachment to the integrated system.

It should be mentioned that at present time only integrated database systems (as client systems to the metabase) could initiate data update process with metabase update Web service. This technique of course is not the only variant of interaction scheme. Thus it is planned to redesign metabase update mechanism to enable metabase to inquire integrated resources on demand and thus to query information updates occurred.

It should be mentioned that after every metabase update session incremental population crawl is started on the metabase to update or to reindex relevant chemical systems list regarding information changes. This allows metabase to maintain actual information on relevancy relations of chemical systems contained in integrated resources. Currently relevant system reindexing is performed by means of approach of two rules proposed in this paper earlier. If it is necessary, these rules can be easily modified. And the main advantage is that in this very case there is no need to redesign the whole system concept. All we have to do is just to write a new piece of software to provide a new method of searching for relevant systems and replace the old module with the new one.

Metabase Integration – How It Works

Let's consider the operating process of the integrated system from the end user's point of view. In a general sense, the integration of information resources of materials science is in consolidation of available Web applications serving users of different materials science databases. This consolidation is provided by means of

specialized software but user should not be aware of it if possible. The software should be transparent in this sense.

When designing the integrated system special attention was paid to security system development. It should be mentioned that every developed information system has its own proprietary security facilities protecting the system and giving permissions to access it. Security system of a particular information system is responsible for granting permissions to the registered users of a given system only. It's obvious that in the context of integrated security system authorized users should have permissions required to get access to the information in integrated resources within their privileges strictly.

For example, let's consider the possible user work session scenario. A user has been granted access to database on semiconductor system phase diagrams "Diagram" and currently he or she is browsing information on In-Sb system. Obviously the user should have an opportunity to get information on elastic constants of In-Sb system from "Crystal" database. But that user should not be granted privileges to observe information on chemical systems other than In-Sb since he or she is not a registered user of "Crystal" database. And vice versa, if the user is a registered user of "Crystal" database too then he or she will be granted full access to "Crystal" as integrated resource. From our point of view, the described approach is an appropriate one and so it is used to design the distributed security system of integrated databases. It should be mentioned that user credentials of every integrated resource are also transmitted to the metabase via MUserService as well. It is done to organize distributed security system operation in cooperation with corresponding security systems of integrated resources. It should be emphasized that open user passwords are not transmitted to the metabase, instead of open passwords, password MD5 hashes are transmitted in fact. This substitution (MD5 hash instead of open password) allows integrated information system to authenticate active user and at the same time this technique excludes possibility of using open user password to login to the integrated system database. In other words, there is no place for vulnerabilities here. So even if this data are stolen integrated resources can not be compromised.

Let's assume that in one of integrated system a user browses information on some particular chemical system. In other words, the user is in Web application of a particular information system (fig. 4). If it is necessary to get relevant information this Web application is capable to send a request to specially developed Web service [3] that serves users of integrated system. The request aim is to get information contained in integrated resources that is relevant to the currently browsed data. After the request the Web service sends a response to the Web application in a form of XML document. It describes what relevant information on chemical systems and properties is contained in integrated resources. As it was mentioned, earlier data in XML format are properly understood on all major platforms. That information can be output to user for example by means of a XSL-transformation in form of HTML document (XML + XSL = HTML) containing hyperlinks to special gateway. The user can follow from one Web application to another to browse relevant information via this gateway only.

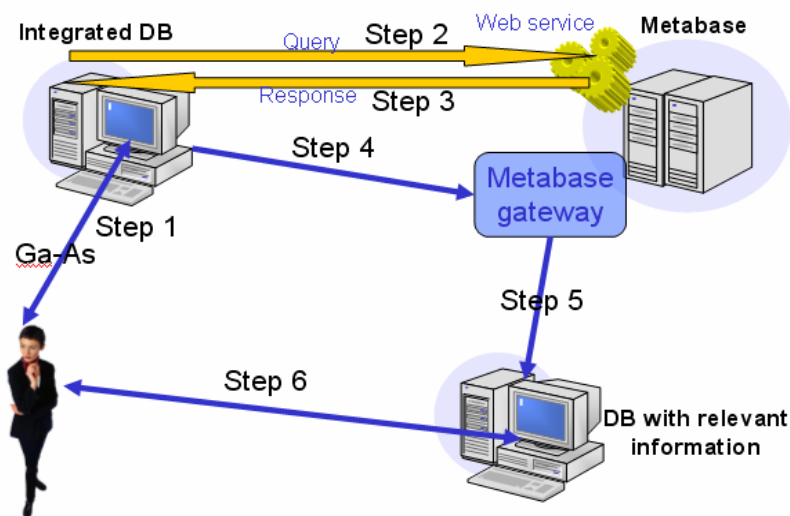


Fig. 4. Metabase integration – how it works.

Imagine that the user clicks on a hyperlink to start browsing information from some other integrated system. First of all, when the user has clicked the hyperlink, he is forwarded to the special gateway. Actually it is a specialized Web application that runs on the metabase Web server. The gateway main purpose is to perform security-dispatching function in distributed system. According to the task stated it is responsible for user authentication and it also checks whether the user has required privileges to address the information requested.

Let's imagine that authentication is successful and the user is eligible to address the data so the metabase security gateway performs redirection to a specialized entry point of desired Web application adding some additional information to create proper security context and a kind of digital signature. It should be stated that the entry point is a specialized page in target Web application that is to perform service functions for integrated system users. At this very page target Web application checks digital signature of the metabase security gateway and if everything is ok the page creates special security context for user with given access rights within target Web application. Finally, the user is automatically redirected to the page with the information required. In spite of redirection process apparent complexity, user transition from one Web application to another is absolutely transparent. Thus, end user can even not note that some complex processing has been done to perform redirection. So, it is an illusion created that having clicked on a hyperlink the user is simply transferred from one information system to another.

Conclusion

It's high time to draw a conclusion. The proposed database integration approach based on metabase was successfully applied at A.A. Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences (IMET RAS). "Crystal" and "Diagram" databases were the very first systems connected to the metabase integrated solution. This fact allows users of either information system to browse information from these databases. Now several words should be said about the project perspectives. First perspectives are connected with the resulting system extension due to addition of already developed Russian databases on materials science: IVTAN and MITHT databases. This integration will allow creating distributed database complex on electronics materials that has no analogs worldwide at present time. Besides numerical growth of integrated system there are plans for functional capabilities extension i.e. qualitative leap is planned. For example, it is projected to provide capability to perform complex distributed database queries that allow searching for substances that satisfy some defined complex criteria while information on criteria values is distributed over several databases. Consequently, to successfully perform such query it's required that metabase information system has an opportunity to query distributed databases impersonating acting user who initiates the initial complex query. After that the metabase should gather information from several sources, process it and output to the end user. Integration at that level undoubtedly will expand distributed information resources capabilities significantly.

Bibliography

- [1] XML-schema that standardized XML document format for metabase update Web service is available at <http://meta.imet-db.ru/MUService.xsd>
- [2] WSDL-contract that defines methods that can be utilized to interact with metabase update Web service is available at <http://meta.imet-db.ru/MUService/MUService.asmx?wsdl>
- [3] WSDL-contract that defines methods that can be utilized to interact with Web service that serves integrated resources is available at <http://meta.imet-db.ru/Service/Service.asmx?wsdl>

Authors' Information

Valery Kornyshko – MITHT, Head of IT department; 119571, pr. Vernadskogo, 86, Moscow, Russia; e-mail: inftech@mitht.ru

Victor Dudarev – MITHT, junior member of teaching staff of IT department; 119571, pr. Vernadskogo, 86, Moscow, Russia; e-mail: vic@osg.ru

THE INFORMATION-ANALYTICAL SYSTEM FOR DIAGNOSTICS OF AIRCRAFT NAVIGATION UNITS

Ilya Prokoshev, Vyacheslav Suminov

Abstract: The operation of technical processes requires increasingly advanced supervision and fault diagnostics to improve reliability and safety. This paper gives an introduction to the field of fault detection and diagnostics and has short methods classification. Growth of complexity and functional importance of inertial navigation systems leads to high losses at the equipment refusals. The paper is devoted to the INS diagnostics system development, allowing identifying the cause of malfunction. The practical realization of this system concerns a software package, performing a set of multidimensional information analysis. The project consists of three parts: subsystem for analyzing, subsystem for data collection and universal interface for open architecture realization. For a diagnostics improving in small analyzing samples new approaches based on pattern recognition algorithms voting and taking into account correlations between target and input parameters will be applied. The system now is at the development stage.

Keywords: technical diagnostics, fault detection, inertial navigation system, navigation, aircraft units, supervision, monitoring, fault diagnostics, diagnostic reasoning

ACM Classification Keywords: B.8.1 Reliability, Testing, and Fault-Tolerance; J.2 Computer Applications: Physical Sciences and Engineering: Aerospace

Introduction

Improvements in the reliability and safety of technical systems require advanced methods of supervision, including fault detection and fault diagnostics. Many modern systems are very complex and it is difficult to manually adjust control functions and settings when departures arise between a system and a system model. It is also difficult to respond manually to the onset of faults before they develop into huge failures. This group seeks to develop and apply effective methods to cope with these problems.

In the recent years, activities in the navigation field have been boosting. There are and will be more and more areas where navigation becomes an important part of a system solution. So far navigation systems have been of major importance for aircraft, missiles, ships, etc. In aircraft systems, the accurate navigation plays an important role. Typical tools today for navigation are inertial navigation systems (INS), which essentially means that acceleration and angular velocity measurements are integrated to a position.

The INS is based on the principle that a Schuler-tuned platform will remain aligned to the local vertical regardless of the movement of the vessel carrying it. Three mutually perpendicular sensitive gyros are gimbaled to create a stable platform on which is mounted a two-axis accelerometer.

Nowadays INS development is aimed to achieving navigation parameters in an unlimited range of mobile object orientation corners with the subsequent information digital output to the consumer.

Growth of complexity and functional importance of INS leads to high losses at refusals of the equipment. Development of INS diagnostic and prediction information systems is necessary for prevention of occurrence of refusals and the malfunctions leading to high breakdown of aircraft units and increase of expenses for its major overhaul.

In the most cases the records analysis and data processing of the flight information is performed by operator. That is a human based approach; it is aimed to compare the received data with the set ranges of control parameters to make a decision about system technical condition. The solution of diagnostic task is made practically in a visual form and not takes into account interrelation between parameters. Thus, in this case it's necessary to use the complex solution that considers the general structure of the output data.

Problem Statement

As one of the main research object, the integrated inertial navigation system INS-2000 developed by Ramensky Instrument Engineering Plant is considered. The development of The INS-2000 system provides definition and delivery of navigating parameters and is intended for new and modern helicopters and planes. The INS-2000 is made as a mono-block consisting of gyrostabilized platform on base of dynamically-tuned gyros, service electronics and computer interface units.

The technical acts analysis research of INS-2000 refusals has shown enough plenty of faults of a product at various production stages (adjustment, trial, refining and so forth).

The experience of inertial navigation systems development shows that the intrinsic error of these units defining their functional reliability is the random parametric drift called by dynamically-tuned gyros, interface electronic cards, control cards and couplers. The given task solution is impossible without more profound analysis of occurrence reasons and influence of design and technological parameters on values and stability of random drift.

According to stated the research of the factors' influential in involuntary drift of system and creation of the effective diagnostic technique permitting to estimate current technical condition of INS-2000 is the actual task.

The main work purpose is development of algorithms for the INS diagnostics, permitting to reveal reasons of refusals and faults on the data on the basis of structural adapting and identification of parameters of navigation model.

The offered technology of solution of the task includes the following stages:

- the structural adapting of the INS equations in view of the detected disorder and model defect in parametric type;
- retrospective estimation of the extended state INS error vector, originating because of defects;
- correlation processing of the received estimations of errors;
- solution of the algebraic equations on parameters, approximating correlation function and included in diagnostic model;
- INS state handle in view of the current state of meters, namely - retargeting of parameters of models of errors and restoring of working capacity of INS.

Given technology will allow solving the following problems:

- optimization malfunctions search strategy;
- separate system units technical condition estimation.

According to the purpose of work, it is possible to solve the following research problems:

- the statistical analysis of INS units parameters accuracy not meeting the quality specifications requirements
- refusals database development of INS interconnected units not past a trial stages;
- open architecture development for processing information from various data sources;
- development realizing automated information capturing for its subsequent processing.

The decision-making task in diagnostic problems starts with observation of behaviour recognized as a deviation from that which is expected or desirable and establishes some hypothesis about the cause of the malfunction. In recent years, two methodologies have been widely applied to approximate the nonlinear assignment rule from the set of observations to the hypothesis: Rule-based systems, characterized by linguistic, logical and cognitively oriented schemes, versus the paradigm of artificial neural networks, characterized by the numeric, associative and adaptive nature.

Fault detection is a key technology in automatic supervision of engineering systems, such as production facilities, machines, airplanes, and appliances. There are a great number of fault detection methods available, ranging from more traditional approaches, such as limit checking, to more advanced model-based methods.

Most model-based methods for fault detection and diagnostics rely on the idea of analytical redundancy that is the comparison of the actual behavior of a system to the behavior predicted on the basis of the mathematical system model. Typical model-based fault detection process consists of two steps: residual generation and residual assessing/classification. Residuals that are the difference between the measurements and the model

predictions are nominally zero, and become non-zero because of faults. Residual assessing is to make a detection decision for the monitored system through evaluating the residuals obtained. The decision making is actually a process of classifying the residuals into one of two classes: normal and fault. Technically, after obtaining residuals, the model-based fault detection becomes a pattern classification problem. Hence, different classification methods can be applied.

It is interesting to observe that almost all of the modern fault detection functions for both unmanned and piloted aircraft are designed by using model-based fault detection methods as described above. This probably contributes to the fact that the model-based fault detection methods have several advantages over the model-free methods, for example, the model-based methods have relatively higher performance and computational straightforwardness, have noise depression capability, and can provide more fault information that can facilitate the subsequent fault isolation and corrections.

Diagnostics Methods

Within the automatic control of technical systems, supervisory functions serve to indicate undesired or unpermitted process states, and to take appropriate actions in order to maintain the operation and to avoid damage or accidents. The following functions can be distinguished:

- (a) *monitoring*: measurable variables are checked with regard to tolerances, and alarms are generated for the operator.
- (b) *automatic protection*: in the case of a dangerous process state, the monitoring function automatically initiates an appropriate counteraction.
- (c) *supervision with fault diagnostics*: based on measured variables, features are calculated, symptoms are generated via change detection, a fault diagnostics is performed and decisions for counteractions are made.

The classical methods (a) and (b) are suitable for the overall supervision of the processes. To set the tolerances, compromises have to be made between the detection size of abnormal deviations and unnecessary alarms because of normal fluctuations of the variables. Most frequently, simple limit value checking is applied, which works especially well if the process operates approximately in a steady state. However, the situation becomes more involved if the process operating point changes rapidly. In the case of closed loops, changes in the process are covered by control actions and cannot be detected from the output signals, as long as the manipulated process inputs remain in the normal range. Therefore, feedback systems hinder the early detection of process faults.

The big advantage of the classical limit-value-based supervision methods is their simplicity and reliability. However, they are only able to react after a relatively large change of a feature, i.e. after either a large sudden fault or a long-lasting gradually increasing fault.

In addition, an in-depth fault diagnostics is usually not possible.

Therefore (c) *advanced methods of supervision and fault diagnostics* are needed, which satisfy the following requirements:

- Early detection of small faults with abrupt or incipient time behaviour,
- Diagnostics of faults in the actuator, process components or sensors.
- Detection of faults in closed loops.
- Supervision of processes in transient states.

The goal for the early detection and diagnostics is to have enough time for counteractions such as other operations, reconfiguration, maintenance or repair. The earlier detection can be achieved by collection more information, especially by using the relationship between the measurable quantities in the form of mathematical models. For fault diagnostics, the knowledge of cause-effect relations has to be used.

INS Monitoring System Developing

The full analysis of various methods has led to expediency of application of complex monitoring systems which use different by the physical nature research methods that, in turn, will allow excluding lacks of one method and use advantages of other methods to realize thus a principle of "redundancy" increasing reliability of INS systems.

Система сбора информации ИНС-2000

Дата: от 01.01.03 до 01.01.07 Искать только по первой дате Фильтр включён

Имя системы: 19-21 Вид проверки:

ФИО: В системе: 75 запись(ей)

Дата	Имя системы	Вид проверки	Название теста	Фамилия
13.01.2004 6:02:22	19-21	ГК К=0 Вх_контр	Погрешности системы после режима подготовки методом гироскопирования при Курс ист, = 0	Барков
13.01.2004 9:20:35	19-21	Период ШУЛЕРА	Проверка периода собственных колебаний гироскопирования	Кулагина
13.01.2004 13:56:13	19-21	Погрешности уг	Проверка погрешностей определения и выдачи углов крена, тангажа и ист, курса	Кулагина
28.12.2003 17:36:29	19-21	ПЭК. Тест-контр	Прохождение режима тест - контроль	Николаева
28.12.2003 15:21:27	19-21	период. Шулера	Проверка периода собственных колебаний гироскопирования	Николаева
29.12.2003 19:34:07	19-21	П,3,23 ВНЕШ,СВ	Проверка внешних связей	Киселевский
29.12.2003 19:47:29	19-21	П,3 Тест ко	Прохождение режима тест - контроль	Киселевский
29.12.2003 18:47:13	19-21	ГК Погрешность	Проверка погрешностей определения и выдачи углов крена, тангажа и ист, курса	Киселевский
20.01.2004 2:01:36	19-21	+60	Погрешности системы после режима подготовки методом гироскопирования при Курс ист, не равным 0	Лапшин
20.01.2004 4:28:39	19-21	+20 ну	Погрешности системы после режима подготовки методом гироскопирования при Курс ист, не равным 0	Лапшин
22.01.2004 18:52:13	19-21	+40	Устойчивость к воздействию пониженной температуры (-40°C)	ОТК+ПЗ
23.01.2004 4:50:27	19-21	+60	Погрешности системы после режима подготовки методом гироскопирования при Курс ист, не равным 0	ОТК+ПЗ

Редактировать Сохранить Удалить

Протокол испытаний	Начальные данные	Ошибки счисления	Максимальное отклонение
Время ИНС	Время выст.	Время нав.	Управляющее слово
00:00:00	00:00:00	00:00:00	УС=0024
00:00:01	00:00:01	00:00:01	+Preparation +GyroCo
00:00:31	00:00:31	00:00:31	CC=B000
00:00:32	00:00:32	00:00:32	CC=B900
00:00:49	00:00:49	00:00:49	CC=B500
00:01:21	00:01:21	00:01:21	CC=B300
00:02:00	00:02:00	00:02:00	CC=F300
00:09:57	00:09:57	00:09:57	CC=F500
00:10:37	00:10:37	00:10:37	CC=F300
00:14:54	00:14:54	00:14:54	CC=F308
00:15:01	00:15:01	00:15:01	УС=0000

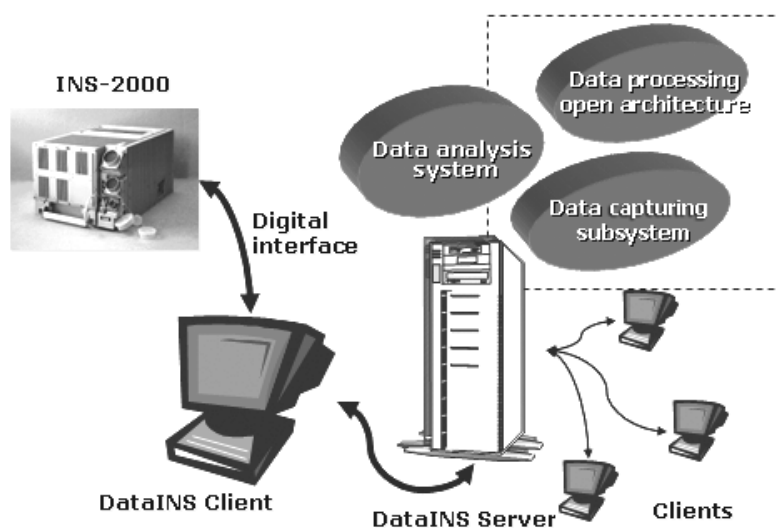
Pic. 1. The screenshot of DataINS system

The improvement of quality of diagnostics and prediction in conditions of small analyzed samples new approaches based on voting of algorithms of recognition and the account of correlations between target and output parameters are developing.

Nowadays, an intellectual system of information capturing and a subsystem of data analysing are developing.

Open architecture of a system allows using different data source based on Microsoft SQL Server, Microsoft Access, Oracle and others. That approach gives a universal model for data analyzing systems.

The general DataINS structure is shown on the following picture.



Pic. 2. DataINS system structure

Conclusion

An overview of the different approaches to fault diagnostics has been given. So far, none of the methods presented solves the remaining task of completeness. Thus, in practical application, principle of "redundancy" increasing reliability of INS systems is able to solve the defined problem. Complex application of quality monitoring and diagnostics methods for fault detection in units and systems is directed to increase the efficiency, validity check, prolongation of system resources working capacity.

For the first part of work the data capturing system is developed. At stage of development, there is an open architecture data processing system, allowing expanding a set of algorithms without system reconstruction.

Sharing a subsystem of capturing information with a subsystem of data analysis will allow eliminating in time the malfunctions both at a level of test of pre-production models, and in operation and perfection of serial samples.

Bibliography

- [Gertler, 1998] Janos Gertler, Fault Detection and Diagnosis in Engineering Systems, Marcel Dekker (May 1, 1998), 484 p.
- [Chiang,Russell,Braatz, 1991] Leo H. Chiang, Evan Russell, Richard D. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer-Verlag; 1st edition (February 15, 2001), 296 p.
- [Clark,1978] R.N. Clark. Instrument fault detection. *IEEE Trans. Aerospace Electron. Syst.*, Vol. 14, No. 3, pp. 456-465.
- [Frank,1990] P.M. Frank. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy. *Automatica*, Vol. 26, pp. 459-474.
- [Singh, 1987] M.G. Singh. Fault Detection and Reliability: Knowledge Based and Other Approaches (International Series on Systems and Control, V. 9) by Reliability, and Related knowl European Workshop on Fault Diagnostics, Pergamon Pr; 1st edition (December 1, 1987), 324 p.
- [Isermann,1984] R. Isermann. Process fault detection based on modelling and estimation methods - a survey. *Automatica*, Vol. 20, No. 4, pp. 387-404.

Author's Information

Ilya V. Prokoshev – Moscow State Aviation Technological University, 121552, Russia, Moscow, Orshanskaya, 3; post-graduate student, member of teaching staff; +7(926) 5239446; e-mail: mati@e-music.ru

Vyacheslav M. Suminov – Moscow State Aviation Technological University, 121552, Russia, Moscow, Orshanskaya, 3, Doctor of Science, Professor, "Instruments production engineering and aircraft control systems" department header; +7(095)9155441

USING ORG-MASTER FOR KNOWLEDGE BASED ORGANIZATIONAL CHANGE

Dmitry Kudryavtsev, Lev Grigoriev, Valentina Kislova, Alexey Zablotsky

Abstract: Enterprises in growing markets with transitional economy nowadays encounter extreme necessity to change their structures and improve business processes. In order to support knowledge processes within organizational change initiative enterprises can use business modeling tools. On one hand software vendors suggest many tools of this kind, but on the other hand growing markets with transitional economy determine quite special requirements for such tools. This article reveals these requirements, assesses existing business modeling tools using these requirements and describes ORG-Master as a tool specially created for support of process improvement initiatives in the growing markets with transitional economy.

Keywords: Business information modeling, business modeling, knowledge process, organizational change, business process improvement, growing markets, transitional economy.

ACM Classification Keywords: I.6.3 Simulation and Modeling: Applications

Introduction

ORG-Master is a business modeling software, which was initially created as a response to growing need for computer aid to consulting projects in the field of organizational change and business transformation. In spite of the diversity of products for business modeling ORG-Master has certain advantages that can be revealed in solving certain tasks in certain environment.

Certain tasks include such organizational change components as business process improvement, business restructuring, quality management implementation and holistic improvement of management system. In the current article, organizational development will be described by the example of business process improvement (BPI) initiative.

Certain environment includes growing markets with transitional economy (GMwTE) which determine specialties in organizational change initiatives. GMwTE include post-soviet countries (Russia, Ukraine, Belarus, Kazakhstan) and in the current article will be described by the example of Russia. In order to reveal these specialties Section 1 describes features of GMwTE from management point of view. Section 2 focuses on the flow of knowledge within BPI initiative and gives an ability to define requirements for business modeling tool at the GMwTE (section 3). Section 4 reveal imperfections of existing business modeling tools with respect to above-mentioned requirements and show the niche for ORG-Master. Section 5 explains the main concepts and consequent advantages of ORG-Master. Section 6 describes practical application of ORG-Master.

1. Business Process Improvement Initiatives in the Growing Markets with Transitional Economy

The most important features of GMwTE from management point of view are:

1. Extremely high pace of change in market conditions and business environment
2. Low level of managerial culture
3. Predominance of informal methods of management

Quick changes and competition growth make companies to change in the same pace and the main objectives in the organizational change is to fit company structure with business needs and to implement client-oriented business processes that allow to achieve company goals. This results in the necessity to launch restructuring or BPI initiatives.

The main prerequisite for BPI initiative is transparent management at every level of organization. In this context transparency implies holistic knowledge describing *What* functions and processes are realized in the company, *Who* performs the functions, *How* the functions are performed, *What for* are the functions performed. While low level of managerial culture results in absence of clear knowledge in this field. As a results BPI initiative in the GMwTE usually involve a wide range of preliminary stages directed towards understanding of company "big picture" in order to make conceptual changes and define the processes for improvement or re-engineering.

The third feature of GMwTE - predominance of informal methods of management results in small amount of documents and business rules. Such a situation has its roots either in skeptical attitude to archaic and out-of-date formal documents at post-soviet enterprises or in quick growth of small start-ups. In some situations, informal intuitive method of management brings fruits, but it is terminated by the scale of business and is one of the barriers in development of managerial culture. As a result, BPI initiative in the GMwTE has an important objective – to switch company from informal methods of management to formal procedures and business rules.

2. Knowledge Process in the Business Process Improvement Initiative

BPI or restructuring initiatives deal with business organization knowledge. Under *business organization knowledge* in the current article we will understand *knowledge domain* covering organizational goals, structure, processes, functions, rules, rights, authorities and relationships between this objects. Thus in order to raise effectiveness of BPI initiative project team should support knowledge process in the domain of business organization knowledge. As described in [Strohmaier, 03a] knowledge infrastructure¹ is determined by the nature

¹ Under Knowledge Infrastructure we imply all the means that enable effective knowledge management within organization ~ knowledge process support

of knowledge process, which in turn can be understood through analysis of business processes covered by improvement initiative (figure 1).

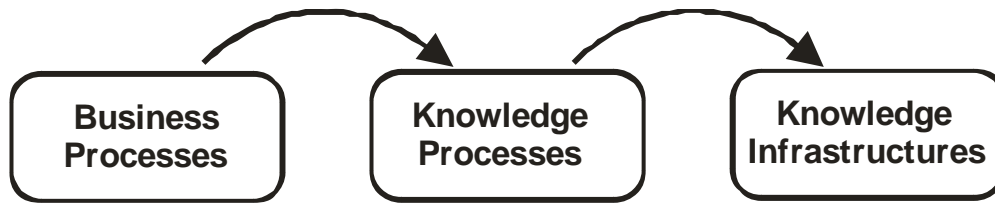


Figure 1: Business process and knowledge infrastructure relationship

The main constituents of improvement initiatives are *organizational change processes* - a subset of the whole system of business processes:

- business process analysis
- business process improvement
- organizational structure control
- performance management

Business analysts (either internal or external consultants) together with domain experts (head of departments and other managers) *generate* business organization knowledge, *store* and *transfer* it throughout the organization in these processes.

Application of business organization knowledge is distributed between all the other business processes - operating, management and support processes. Organizational roles of performers vary from workers to executives (top managers).

According to [Strohmaier, 03b] business organization knowledge can be visualized (figure 2).

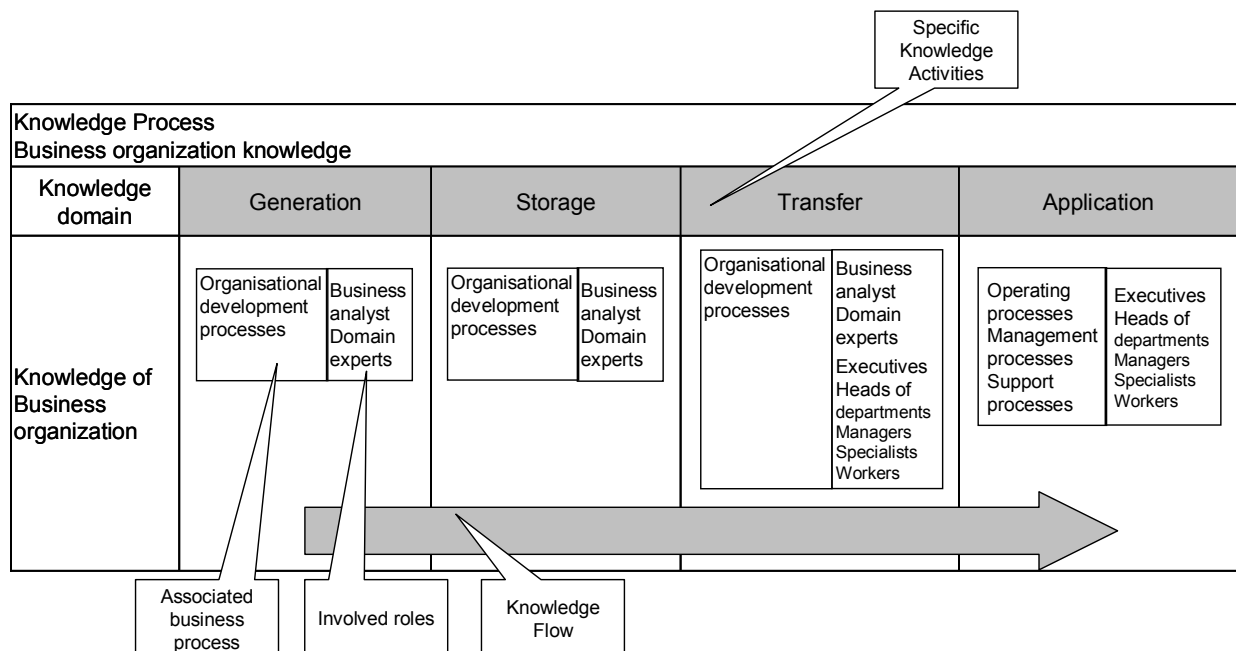


Figure 2: Knowledge process in the business improvement initiative

The most important and influential feature of this process consists in different organizational roles involved in it and especially in the knowledge transfer process. During transfer process business analysts deliver their knowledge through the mediation of domain experts to personnel from different domains and organizational

levels. As it was mentioned in [Section 1], one of the goals of BPI initiative in the GMwTE is to shift the priorities of management from informal methods to formal business rules. Thus the basis of knowledge transfer is formalized knowledge and the main factor of its successful internalization [see Nonaka, 03] by personnel is type of knowledge representation.

Type of knowledge representation depends on two specific knowledge processes generation on one hand and application on the other. While the way these processes are performed is determined by the involved organizational roles.

Business analysts have developed competencies in organizational management and system analysis. They have detailed understanding of business from different points of view and can operate with different objects and their relations (organizational units, functions, processes, goals etc).

Personnel from different domains and organizational levels have low competencies in organizational management see [Section 1]. They usually have only dim understanding of business from "organizational unit" point of view (answer for the question "who do what?").

As a result, business analysts primary use diagrams of different types and notations (IDEF0, UML, EPC) as a mean of knowledge representation. The other personnel use job descriptions, documents describing the functions of business units / departments and other regulating documents. In addition, these regulating documents for application usually prepared according to national, industrial or corporate standard. Namely these regulating documents solve one of the objectives of BPI initiative in the GMwTE - shift the priorities of management from informal methods to formal business rules see [Section 1].

Thus, the necessity to facilitate communication between people speaking "different languages" predetermines important requirement for knowledge infrastructure.

3. Requirements to Business Modeling Tool for GMwTE

For the current analysis we assume business modeling tool primary as a support system for knowledge generation and storage during BPI initiative.

Previous section described the necessity to have different types of knowledge representation during BPI initiative in the GMwTE. Assumption that the process of knowledge transfer do not change the type of knowledge representation imply the necessity to generate knowledge both in type of diagrams for analysis and regulating documents for application. This requirement for knowledge generation process determine the first requirement for business modeling tool:

1. Ability to represent knowledge in different types and formats

Section 1 highlights the necessity of preliminary stages within BPI initiative in the GMwTE. For example, companies should define goals, composition of functions, change organizational structure, reassign responsibilities for function realization, reveal a list of business processes. This tasks can be done both in series and in parallel and include several analysts concentrating either on different tasks or on different levels of detail. Such a nature of BPI initiative determine the next requirement:

2. Ability to work both with a complex model (e.g. business process model) and with separate parts of this model (relate functions with organization roles, roles with infrastructure etc) using different views of enterprise.

Fast dynamic of the enterprise development is especially relevant for GMwTE and require constant improvements in business processes thus a model once created should be constantly up-dated. Model is a system of constituent objects and their relationships, but both objects and their relations change constantly. This situation generate the third requirement:

3. Ability to reflect changes in objects and in their relationships throughout the whole model after changing any part of the model.

In order to reveal a tool, which satisfy all the requirements mentioned above an analysis of the tools existent in the Russian market was carried out.

4. Analysis of Existing Business Modeling Tools in the Russian Market

Although in some BPI initiatives knowledge is created and stored using typical office applications like MS Word or Excel or simple graphical packages like MS Visio this tools obviously do not satisfy requirements see [Section 3].

The main business modeling tools existent in the Russian market that are usually used for organizational development and BPI purposes are:

ARIS <http://www.ids-scheer.com/>

BPWin (AllFusion Modeling Suite) <http://ca.com/>

There are also some Russian products that contain either limited functionality or slight modifications of foregoing tools. Differences of these products are immaterial from point of view of chosen requirements and as a result, they appeared beyond the scope of our analysis.

There are also a broad range of CASE tools (e.g. Rational Rose) for corporate systems development. These tools include business process modeling, but their primary function is information architecture development and it determines their whole viewpoint for enterprise modeling. As a result they are nor convenient for organizational management and business process modeling, nor efficient. Thus, they appeared beyond the scope of our analysis.

Here is generalized result of the analysis:

Requirement 1: Ability to represent knowledge in different types and formats

ARIS: It includes a broad library of object types and corresponding diagrams, but it has a very complicated mechanism for generating regulating documents. It is hard to customize necessary templates and consequently requires unique and expensive specialists

BPWin: It allows generating IDEF0 diagrams, but it is also very hard to generate corresponding regulating documents in customary standards.

Requirement 2: Ability to work both with a complex model (e.g. business process model) and with separate parts of this model

ARIS: Satisfy. There are both a whole process model and separate constituent models.

BPWin: Dissatisfy. User works either with one object type (functions, roles) or with a whole model of business process (one type of composite diagram).

Requirement 3: Ability to reflect changes in objects and in their relationships throughout the whole model after changing any part of the model.

ARIS: Partially. Centralized library of modeling objects guarantee the reflection of changes in the particular object throughout the model (e.g. changing function name in one diagram cause changing this name in every diagram in the model), but changes in relationships between objects of different type do not appear automatically throughout all diagrams.

BPWin: Satisfy. All the objects stored in centralized library and are used in one type of diagram.

Thus, presented tools do not completely satisfy suggested requirements. Besides this tools are quite expensive and require extremely professional analysts to support business model.

There is a necessity for more effective business modeling tool for organizational development.

5. Main Concepts and Advantages of ORG-Master

Concepts and methodology

The main idea of ORG-Master consists in division of business modeling interface from model representation one. As a result, each interface and type of knowledge representation is optimized for the solution of own tasks.

This idea is contrary to an approach of ARIS and BPWin. In the foregoing product user input, editing and represent business model in the same knowledge representation type and format.

Division of interfaces in ORG-Master allows representing knowledge both in different types (diagrams in different notations, reports, tables) and from different point of views.

On the other hand business model editing interface has its own type of knowledge representation based on two instruments: classifier (ontological models, see [Gavrilova, 00]) and matrix (table).

Classifier – hierarchical tree of particular objects (e.g. organizational roles, functions, material resources, documents etc), that can have different attributes: type, meaning, comments etc. In the process of building classifier objects become structured into a hierarchy/ tree – they receives relationships of AKO (“A Kind Of” [Gavrilova, 00]) type (figure 3).

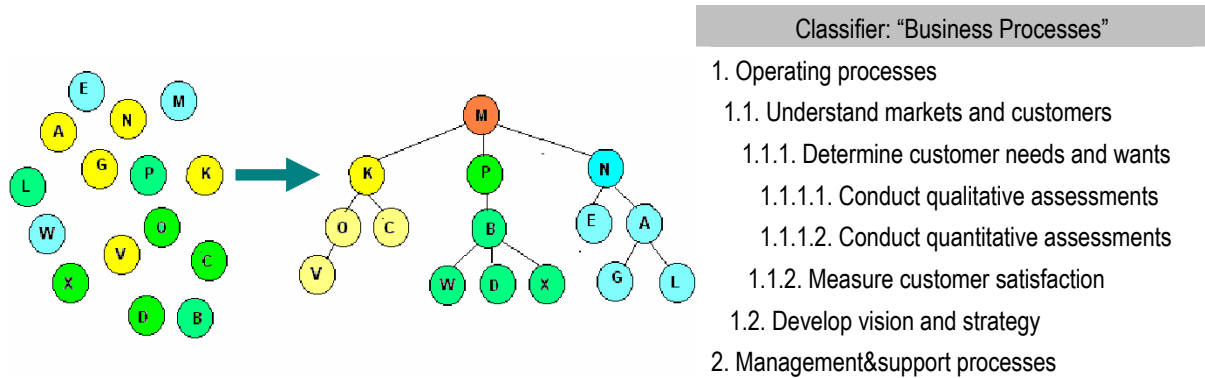


Figure 3 Structuring informational objects

Matrix (table) – models that define relationships between objects of different classifiers in any combination of the later (figure 4). Relationships can also have different attributes (directions, type, name, index, meaning).

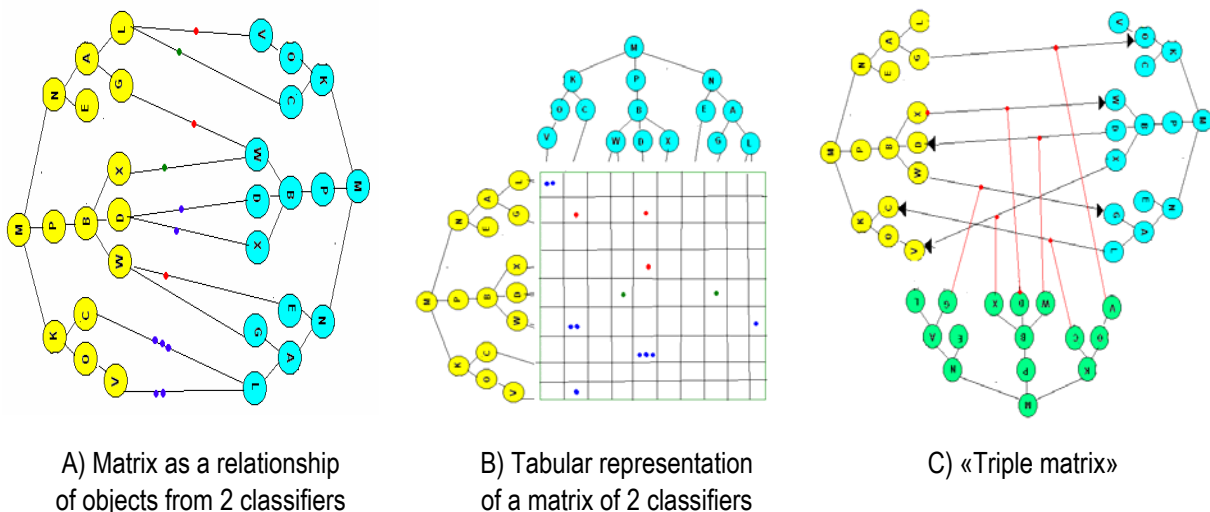


Figure 4: Conceptual framework of matrix (table)

As any material object of any complexity (e.g. building) can be described using definite number of 2-dimensional (flat) schemes (e.g. design drawings) so and several matrix allow to receive multidimensional description of complex business system and make it both holistic and visible (see figure 5).

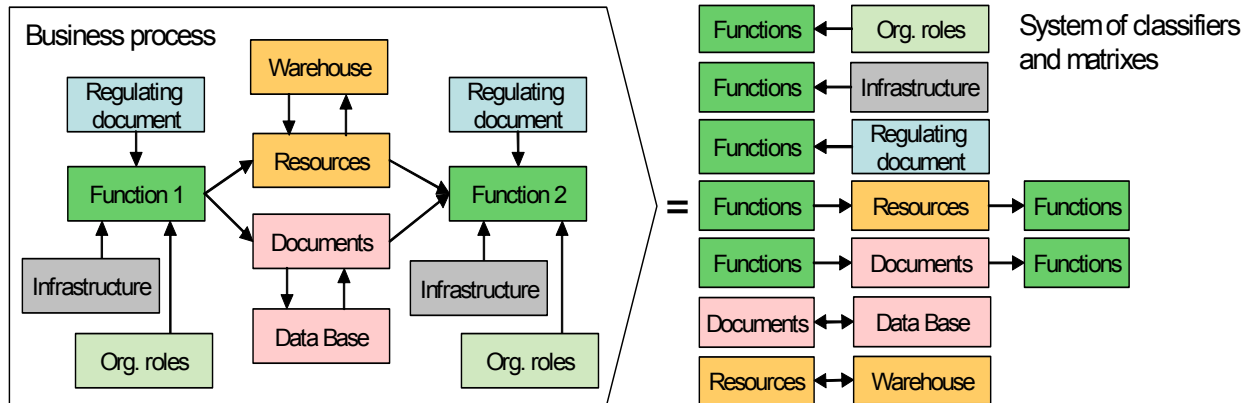


Figure 5: Business process model as a system of classifiers and matrixes

ORG-Master advantages for end user

Foregoing concepts of ORG-Master provide the following features of this tool:

- ability to generate multidimensional reports based on matrixes from business model with different level of detail, which allow to analyze company from many viewpoints for people at different levels of organizational hierarchy
- ability to fine-tune knowledge representation reports, that allow to generate regulating reporting on the basis of business model for particular needs and customary standards
- ability to generate visual diagrams of business processes that support business analysis
- ability to decompose the whole business model into constituent separate submodels, which allow to divide complex BPI initiative into manageable tasks and solve problems in separate domains with adequate (pared-down) tool
- all the objects and relationships from different submodels are integrated into centralized holistic model that allow to reflect changes in objects and in their relationships throughout the whole model after changing any part of the model

These features of ORG-Master satisfy requirements to business modeling tool for GMwTE and together with relatively low price and low training complexity characterize it as effective and efficient tool.

Among the relative disadvantages of this tool the most obvious is absence of quantitative analysis of business processes, but this feature is of low importance with respect to foregoing requirements.

6. Application of ORG-Master and Practical Results

ORG-Master has 6-year history of application in the organizational change and BPI initiatives. There is a broad range of ORG-Master clients located in Russia, Ukraine. Size of ORG-Master clients vary from small companies to large holding structures (up to 10000 people).

The results of typical ORG-Master application in BPI initiatives include:

- Business model which describes functions, organizational roles, goals and measures, function distribution among organizational roles and description of necessary business processes.
- Regulating documents based on business model (job descriptions, procedures etc)
- Diagrams of the necessary processes based on business model

Conclusions

Since organizational change and BPI initiatives become a life-style of every company, it is useful to support such an activity with adequate tools for business modeling. However, choice of the tool is determined by the objectives of tool application and business environment. Current paper revealed specialties of BPI initiatives in the GMwTE and analyzed existing business modeling tools from that perspective. Because of this analysis, ORG-Master can be considered as an effective and efficient for knowledge process support during organizational change initiatives in the GMwTE.

Acknowledgements

The authors of this paper are thankful to the advisor Dr. Prof. Tatiana Gavrilova (St. Petersburg State Polytechnic University) for her useful suggestions about content and destiny of this paper.

Bibliography

- [Strohmaier, 03a] M. Strohmaier Designing Business Process Oriented Knowledge Infrastructures Proceedings der GI Workshopwoche, Workshop der Fachgruppe Wissensmanagement, Karlsruhe (2003)
- [Strohmaier, 03b] M. Strohmaier A Business Process oriented Approach for the Identification and Support of organizational Knowledge Processes Proceedings der 4. Oldenburger Fachtagung Wissensmanagement, Oldenburg (2003)
- [BIG, 96] BIG&Expert "Seven notes of management", Moscow (1996)
- [Nonaka, 95] Nonaka I., Takeuchi H.: "The knowledge creating company"; Oxford University Press (1995)
- [Bukowitz, 99] Bukowitz W., Williams R.: "The knowledge management fieldbook"; Prentice hall, Pearson Education Limited (1999)
- [APQC, 96] APQC's International Benchmarking Clearinghouse Process Classification Framework www.apqc.org, (1996)
- [Gorelik, 01] Gorelik S., "Business-engineering and management of organizational change"; (2001) <http://www.big.spb.ru/publications/bigspb/metodology/>
- [Gavrilova, 00] Gavrilova T., Horoshevsky V. "Knowledge bases of intellectual systems"; Piter / Saint-Petersburg (2000)
- [Рубцов, 99] Рубцов С., Сравнительный анализ и выбор средств инструментальной поддержки организационного проектирования и реинжиниринга бизнес процессов <http://or-rsv.narod.ru/Articles/Aris-IDEF.htm>
- [Репин, 01] Репин В. Сравнительный анализ нотаций. <http://www.interface.ru/fset.asp?Url=/ca/an/danaris1.htm>
-

Authors' Information

Dmitry Kudryavtsev – Saint-Petersburg State Polytechnical University, Tkachy str., 24-24, Saint-Petersburg - 193029, Russia; e-mail: dk@big.spb.ru

Lev Grigoriev – BIG-Petersburg (consulting company), Sovetskaya str., 2, Saint-Petersburg - 191014, Russia; e-mail: spbbig@infopro.spb.su

Valentina Kislova – BIG-Petersburg (consulting company), Sovetskaya str., 2, Saint-Petersburg - 191014, Russia; e-mail: valya@big.spb.ru

Alexey Zablotsky – BIG-Petersburg (consulting company), Sovetskaya str., 2, Saint-Petersburg - 191014, Russia; e-mail: support@big.spb.ru

NEURAL NETWORK BASED APPROACH FOR DEVELOPING THE ENTERPRISE STRATEGY

Todorka Kovacheva, Daniela Toshkova

Abstract: Modern enterprises work in highly dynamic environment. Thus, the developing of company strategy is of crucial importance. It determines the surviving of the enterprise and its evolution. Adapting the desired management goal in accordance with the environment changes is a complex problem. In the present paper, an approach for solving this problem is suggested. It is based on predictive control philosophy. The enterprise is modelled as a cybernetic system and the future plant response is predicted by a neural network model. The predictions are passed to an optimization routine, which attempts to minimize the quadratic performance criterion.

Keywords: enterprise strategy, model predictive control, neural network, black-box modeling, business trends.

ACM Classification Keywords: 1.2.6 Artificial Intelligence: Neural nets; 1.6.3 Simulation and Modeling: Applications

Introduction

In the present paper, a Generalized Strategy Development (GSD) approach is suggested. Designing of the enterprise strategy is a very complicated process. It depends on many factors, which require a lot of variables to be taken into account. The relationships between them are complex and non-linear.

In the decision making process the managers need to know the environment characteristics in order to adapt the developed strategy. Therefore, the predictions of the environment changes are needed. They enable businesses make better strategic decisions and manage their activity more efficiently. It can also identify new opportunities for increased revenues and entering new markets. The prediction of the environment changes is a very difficult task. Price, advertising, goods seasonality, customers and competitors behaviour, global economic trends etc. are all factors that influence the overall performance of the enterprise.

Traditional forecasting methods such as regression and data reduction models are limited in their effectiveness as they make assumptions about the distribution of the underlying data, and often fail to recognize the inter-relatedness of variables. Now, a new forecasting tool is available – artificial neural networks (ANN). They are a form of artificial intelligence, which provide significant potential in economic applications by increasing the flexibility and effectiveness of the process of economic forecasting [Tal, Nazareth, 1995]. They are successfully used in various economic studies including investment, economic and financial forecast [Hsieh, 1993; Swales and Yoon, 1992; Hutchinson, Lo, and Poggio, 1994; Shaaf, 2000].

The enterprise strategy development requires not only predictions but also have to be optimized and adapted according to the environment changes. A suitable control design algorithm is needed. During the last years a number of methods for automatic control synthesis are applied for managing business processes. Many authors suggest the Model-Based Predictive Control (MBPC) algorithm to be used as a decision-making tool for handling complex integrated production planning problems [Tzafestas, Kapsiotis, Kyriannakis, 1997] and supply chain management [Braun et al., 2003]. MBPC is a very popular controller design method in the system engineering. It is a suitable technique for prediction of future behaviour of a plant.

Both, ANN and MBPC, are tools for solving complex problems under uncertainty by providing the ability to learn from the past experience and use information from various sources to control the enterprise performance. Generalized Strategy Development approach combines the advantages of artificial neural networks and Model-Based Predictive Control algorithm to increase the effectiveness of the enterprise management in the entire decision making process and development of all functional strategies (incl. production-, marketing-, financial-, sales-, innovation strategy etc.).

Business Trends and Management Theory

The contemporary business is accomplished in highly dynamic environment. The continuous changes in the internal and external environment of the enterprise force it to apply a number of adaptation mechanisms, which contribute to its surviving and competitive power. These adaptation mechanisms are based on the degree of information availability. This makes providing the information a necessary condition for adaptation process and the adaptation itself – the most important characteristic of each system. In this regard the developing and the implementing of tools, which enables the corporate adaptation according to the environment changes becomes a strategic need.

Globalization [Кирев, 2001; Голдштейн 2002, 2003; Ганчев, 2004; Стоилова, 2004; Стоянов, 2003; Краева, 2003], *virtualization* [Мейтус, 2004; Баксанский, 2000; Манюшис, Смольянинов, Тарасов, 2003; Вютрих, Филипп, 1999], *Internet* and the developing of the *Information Technologies* [Христова, 1997; Върбанов, 2000; Илиев, 2003; Седлак, 2001] have a deep impact on the economic and social life of the society. These global trends determine the transition from the traditional industrial society to the information age society. A *new economic based on knowledge* [Applegate et al., 1996] appears and as a result the traditional managerial hierarchy cease to exist and a horizontal relationships are formed. Enterprises of a new type appear, which accomplish their activity on the global market from their founding. They overcome the spatial and time boundaries. The common name for such structures is *"globally born"* [Андерссон, Виктор, 2004]. These enterprises have their own mechanisms for developing, which substantially differ from those of the traditional industrial enterprises. Thus, the small national companies become multinational very fast.

The adaptation to environment changes requires new knowledge for its elements, the relationships between them, and characteristics of their functioning. Thus the concept of *"Learning enterprise"* [Senge, 1990] comes into being. It is based on the continuous acquiring new knowledge regarding the environment, using it for innovation strategies and this process applies to the enterprise as a whole.

The global trends in business development mentioned above cannot be considered partially. There are mutual relationships and dependencies between them. The existing of certain trend is a prerequisite for appearing and developing of another one and vice versa. Therefore, they influence the contemporary enterprise activities by forming an integrated set of strategies.

Now let us consider the modern business trends in a management theory point of view. Many authors [Каменов, 1984; Камионский, 1998; Рубцов, 2001] state that an unified management theory does not exist. There are different managerial concepts. Some of them claim to be universal, other are a tool for solving particular problems, some are not developed enough, other are just catchwords, some contribute and expand each other, and other contradict each other [Айвазян, Балкинд, Баснина, 1998]. This causes difficulties for the development the enterprise strategy, which strongly depends on the environment changes.

The experience shows that there is time delay between the problems, which arise in the practice and the developing of methods for their solving, which constitute the theory. In this regard, the new structures mentioned above – *"globally born"* and *"learning enterprise"* are not considered in the general management theory. Therefore, there is a lack of methodologically developed and scientifically based management approaches. These enterprises do not respond to the traditional rules and concepts as they arise and perform in strongly uncertain and highly dynamic environment. They need new management, approaches, which have to correspond to their characteristics and meet their requirements. These enterprises can be presented as complex nonlinear cybernetic systems. Thus, the laws of system and control theory can be applied to their management.

Model-Based Predictive Control

Model-Based Predictive Control has established itself in industry as an important form of advanced control [Townsend, Irwin, 2001]. An overview of industrial applications of advanced control methods in general can be found in Takatsu et al. [Takatsu et al, 1998] and in Qin and Badgwell [1998].

The main advantage of MBPC algorithm is the simplicity of the basic scheme, forming a feedback, which combines with adaptation capabilities. This determines its successful applying in the practice of designing control systems.

MBPC is an efficient methodology to solve complex constrained multivariable control problems in the absence, as well as in the presence of uncertainties [Mayne et al., 2000]. It makes possible the uncertainty of the plant and disturbances to be taken into account and enables the on-line optimization and control synthesis.

In general, it is used to predict the future plant behaviour. According to this prediction in the chosen period (prediction horizon), the MBPC optimizes the manipulated variables to obtain an optimal future plant response. The input of chosen length (also known as control horizon) is sent into the plant and then the entire sequence is repeated again in the next period. An important advantage of MBPC is that it allows the inclusion of constraints on the inputs and outputs.

The prediction plant model is realized with neural network. It provides predictions of the future plant response over a specified time horizon. The predictions are passed to an optimization routine to determine the control signal that minimizes the following performance criterion over the specified time horizon:

$$J = \sum_{j=N_1}^{N_2} (y_r(t+j) - y_m(t+j))^2 + \rho \sum_{j=N_1}^{N_u} (u'(t+j-1) - u'(t+j-2))^2$$

subject to the constraints, which are imposed on the state and control variables. The constants N_1 , N_2 , N_u define the horizons over which the tracking error and control increments are evaluated. The u' variable is the tentative control signal, y_r is the desired response and y_m is the network model response. The ρ value is weight coefficient. Generalized Strategy Development approach will be introduced in Model-Based Predictive Control framework.

Generalized Strategy Development Approach

The purpose of the Generalized Strategy Development Approach is to transform the incomplete information about the environment and the processes inside the enterprise into complete strategy for its adaptation and evolution. From cybernetic point of view, this can be considered as a control system. The functional structure is given in Fig.1

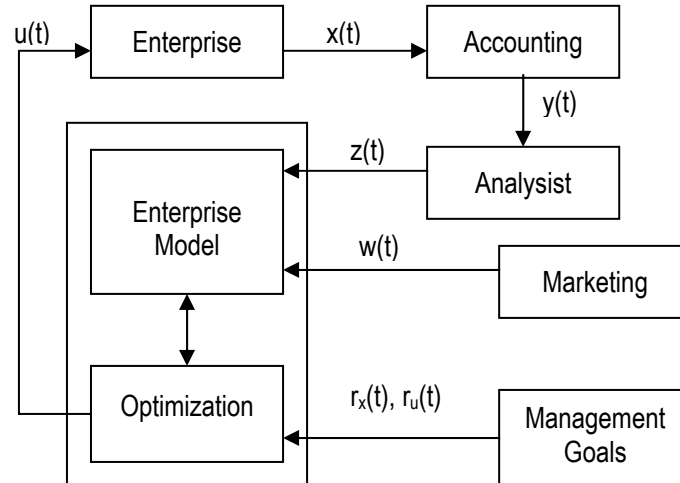


Fig.1 Global Strategy Development functional structure

Enterprise

The enterprise is a dynamic system with a high complexity. In order, the management and control $[u(t)]$ to be effective we need to know its physical structure, the relationships between the constituting elements, their dynamic behaviour and the characteristics of the environment. We have to compare the current state of the enterprise to the desired state. In case they coincide entirely the management goals are achieved. In order to register the difference we need to measure the state of the enterprise. This could be realized by performance measurement system.

Performance management is the prerequisite to performance improvement. For the enterprises to improve their performance, they must be able to measure how they are performing at present, and how they are performing after any changes. So, the companies will have the possibility to monitor if a chosen strategic direction is appropriate.

Traditional performance management systems are frequently based on cost and management accounting. There are five main difficulties with traditional management accounting techniques for performance measurement [Maskell, 1991]:

1. Management accounting reports are not relevant to strategy development;
2. Some of the data which are used for decision-making process can be distorted by cost accounting;
3. Traditional accounting reports are inflexible and are usually received too late to be of value;
4. The information about the pay-back on capital projects comes late;
5. To be of value, management accounting systems must be based on different methods and assumptions than on the financial accounts.

As traditional performance measurement systems are based on management accounting, they are primarily concerned with cost. But in today's manufacturing environment, cost based measures are no longer the only basis for decision making in enterprises. The new performance measurement systems should have some additional characteristics [Maskell, 1991].

Accounting

In Fig.1 the Accounting is the traditional performance measurement system. Therefore, the current state of the enterprise $[y(t)]$ is represented by the measured one from the accounting and measurement error. The reports, which are formed by the accounting, need to be interpreted in order to be useful for the management. This task is performed by analyst.

Analyst

The accounting information is now manipulated for giving proper estimate for the current enterprise state $[z(t)]$. The manipulations include: recapitulation, generalization, estimation, recalculation etc. in order to analyze the entire enterprise activity. The results are used by managers to make decisions about the future behaviour of the enterprise.

The analysis is performed on the basis of incomplete information about the environment changes. Another error is formed. The obtained information is passed to the prediction model of the enterprise in order to minimize the tracking error.

Enterprise model

The model is used to determine the direction in which changes in the manipulated variables will improve performance. The plant operating conditions are then changed by a small amount in this direction, and a new, updated model is evaluated. The enterprise is a complex, dynamic and non-linear plant. Also different disturbances affect the its performance. Because of that, there is a lack of knowledge on the function or construction of the system.

The process output can be predicted by using a model of the process to be controlled. Any model that describes the relationship between the input and the output of the process can be used and a disturbance or noise model can be added to the process model [Duwaish, Naeem, 2001]. We can build a model using the observations of the enterprise activities.

Therefore the enterprise can be viewed as a black-box [Sjoberg et al., 1995] which aims to describe the relationships between input/output data. The non-linearities and the disturbances are taken into consideration.

During the past few years, several authors [Narendra and Parthasarathy, 1990; Nerrand et al. 1994] have suggested neural networks for nonlinear dynamical black-box modeling. To date, most of the work in neural black-box modeling has been performed making the assumption that the process to be modeled can be described accurately by neural models, and using the corresponding input-output neural predictors [Rivals, Personnaz, 1996]. Therefore, artificial neural networks are used as effective black-box function approximators with learning and adaptation capabilities.

Marketing

We could receive information $[w(t)]$ about the environment changes from the Marketing Information System (MIS) which is used in the enterprise. It is passed to the enterprise model by taking into account the forming of a new error. This error is due to impossibility of MIS to register all trends in global economy and social life of the society.

Optimization and Management Goals

Using the enterprise model, we predict the future plant response and taking into consideration the management goals $[rx(t), ru(t)]$ we optimize it and develop a new management strategy. This is an iterative process, which provides the continuous enterprise adaptation to the environment changes.

Conclusion

Strategy development is a complex task in the continuously changing environment. The enterprise management must combine internal and external information in order to survive and evaluate. Therefore, the company needs an efficient control and strategy development and evaluation system to work in rapidly changing business conditions.

The Generalized Strategy Development approach suggested here is very suitable for this problem, namely for optimization and adaptation of the strategy development process. Thus, the effectiveness of management is increased.

Bibliography

- [Айвазян, Балкинд, Баснина, 1998] Айвазян, С. А., Балкинд, О. Я., Баснина, Т. Д., и др., Стратегии бизнеса: Аналитический справочник. / Под ред. Г. Б. Клейнера. – М.: КОНСЭКО, 1998
- [Андерссон, Виктор, 2004] Андерссон, С., Виктор, И., Инновационная интернационализация в новых фирмах, Международный журнал "Проблемы теории и практики управления", 1/2004
- [Баксанский, 2000] Баксанский, О.Е., Виртуальная реальность и виртуализация реальности, //Концепция виртуальных миров и научное познание, СПб.: РХГИ, 2000 (<http://sociology.extrim.ru>)
- [Върбанов, 2000] Върбанов, Р., Електронният бизнес като отражение на Интернет революцията в деловата сфера, Бизнес управление 3/2000, стр.83-95, СА "Д.А.Ценов", Свищов
- [Вютрих, Филипп, 1999] Вютрих, Х., Филипп, А., Виртуализация как возможный путь развития управления, Международный журнал "Проблемы теории и практики управления", 5/1999
- [Ганчев, 2004] Ганчев, П., Глобализацията и българските национални интереси, Диалог, СА "Д.А.Ценов" – Свищов, 2/2004, стр.32-47
- [Гольдштейн, 2003] Гольдштейн, Г. Я., Основы менеджмента, ТРТУ, Таганрог, 2003
- [Гольдштейн, 2002] Гольдштейн, Г. Я., Стратегический инновационный менеджмент: тенденции, технологии, практика, ТРТУ, Таганрог, 2002
- [Илиев, 2003] Илиев, П., Виртуална организация на електронния бизнес, Сборник доклади от научна конференция "Иновации и трансформации на организирани пазари в България", ИУ-Варна, 2003, стр.147-153
- [Каменов, 1984] Каменов, С., Эффективност на управлението, Издателство "Г.Бакалов", Варна, 1984
- [Камионский, 1998] Камионский, С. А., Менеджмент в российском банке: опыт системного анализа и управления/ Общая ред. и предисловие Д. М. Гвишиани. М.: Деловая библиотека "Омскпромстройбанка", 1998
- [Кирев, 2001] Кирев, Л., Относно факторите за глобализация на научноизследователската дейност от транснационалните корпорации, Бизнес управление, 4/2001, стр.49-64, СА "Д.А.Ценов" – Свищов
- [Краева, 2003] Краева, В., Глобални аспекти на електронната търговия, Сборник доклади от научна конференция "Иновации и трансформации на организирани пазари в България", ИУ-Варна, 2003, стр.192-197
- [Манюшис, Смольянинов, Тарасов, 2003] Манюшис, А., Смольянинов, В., Тарасов, В., Виртуальное предприятие как эффективная форма организации внешнеэкономической деятельности компании, Международный журнал "Проблемы теории и практики управления", 4/2003
- [Мейтус, 2004] Мейтус, В., Виртуализация производства, Международный журнал "Проблемы теории и практики управления", 1/2004
- [Рубцов, 2001] Рубцов, С., Целевое управление корпорациями, Москва, 2001
- [Седлак, 2001] Седлак, Я., Мировая экономика: возможность неожиданных потрясений, Международный журнал "Проблемы теории и практики управления", 5/2001
- [Стоилова, 2004] Стоилова, М., Глобализацията – познати плюсове и минуси, Диалог, СА "Д.А.Ценов" – Свищов, 1/2004, стр.63-68

- [Стоянов, 2003] Стоянов, В., Пазар, трансформация, глобализация, нов световен ред, Галик, София, 2003
- [Христова, 1997] Христова, С., Материали от дискусия, Предизвикателствата на информационните технологии на прага на 21-и век, Национална научна конференция АИ'97, Автоматика и информатика 5/6 – 1997, САИ, София, стр.10-11
- [Abdallah, Al-Thamier, 2004] Abdallah, J., Al-Thamier, A., The Economic-environmental Neural Network Model for Electrical Power Dispatching, *Journal of Applied Sciences* 4 (3): 340-343, 2004
- [Applegate et al, 1996] Applegate, L. M., McFarlan, F. W., McKenney, J. L., *Corporate information systems management: text and cases*, 4th ed., IRWIN, USA, 1996
- [Braun et al., 2003] Braun, M., Rivera, D., Carlyle, W., Kempf, K., A Model Predictive Control Framework For Robust Management Of Multi-Product, Multi-Echelon Demand Networks, *Annual Reviews in Control*, 2003, vol.27, no. 2, pp.229-245(17)
- [Duwaish, Naeem, 2001] Duwaish, H., Naeem, W., Nonlinear model predictive control of Hammerstein and Wiener Models using genetic algorithms, In *Proceedings of the 2001 IEEE International Conference on Control Applications (CCA'01)*, 5-7 September, Mexico City, Mexico, pp.465-469, IEEE
- [Hsieh, 1993] Hsieh, C., Some potential applications of artificial neural system in financial management. *Journal of System Management*, April, 1993, 12-15
- [Hutchinson, Lo, and Poggio, 1994] Hutchinson, J., Lo, A., and Poggio, T., A nonparametric approach to the pricing and hedging of derivative securities via learning networks. *Journal of Finance*, 49, 851-99, 1994
- [Maskell, 1991] Maskell, B. H., *Performance Measurement for World Class Manufacturing*, Productivity Press, Cambridge, Massachusetts, 1991
- [Mayne et al., 2000] Mayne, D. Q., J. B. Rawlings, C. V. Rao and P. O. M. Scokaert, Constrained model predictive control: Stability and optimality, *Automatica*, 36, 2000, 789-814
- [Narendra and Parthasarathy, 1990] Narendra, K. S., Parthasarathy, K., Identification and control of dynamical systems using neural networks, *IEEE Trans. on Neural Networks* 1 (1990), pp. 4-27
- [Nerrand et al. 1994] Nerrand, O., Roussel-Ragot, P., Urbani, D., Personnaz, L., Dreyfus, G., Training recurrent neural networks: why and how? An Illustration in Process Modeling, *IEEE Trans. on Neural Networks* 5 (1994), pp. 178-184
- [Qin, Badgwell, 1998] Qin, S. Joe and Badgwell, T.A., An Overview of Non-linear Model Predictive Control Applications, *Proc. Workshop on NMPC*, Ascona, Switzerland, 1998
- [Rivals, Personnaz, 1996] Rivals, I., Personnaz, L., Black-box modeling with state-space neural networks. In: "Neural Adaptive Control Technology", R. Zbikowski and K. J. Hunt eds., World Scientific (1996), pp. 237-264
- [Senge, 1990] Senge, P. M., *The Fifth Discipline. The Art & Practice of The Learning Organization*. – Currency Doubleday, 1990
- [Sjoberg et al., 1995] Sjoberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.Y., Hjalmarsson, H., Juditsky, A., Nonlinear black-box modeling in system identification: an unified overview. *Automatica*, 12. 1691-1724, 1995
- [Shaaf, 2000] Shaaf, M., Predicting recession using the yield curve: an artificial intelligence and econometric comparison. *Eastern Economic Journal*, 26(2), Spring, 2000
- [Swales, Yoon, 1992] Swales, G. S., Yoon, Y., Applying artificial neural networks to investment analysis. *Financial Analyst Journal*, September-October, 70-80, 1992
- [Takatsu et al., 1998] Takatsu, Haruo, Itoh, Toshiaki and Araki, Mituhiko; Future needs for the control theory in industries – report and topics of the control technology survey in Japanese industry, *J.Proc.Cont.*, Vol.8, no 5-6, 369-374, 1998
- [Tal, Nazareth, 1995] Tal, B., Nazareth, L., *Artificial Intelligence and Economic Forecasting*, Canadian Business Economics, Volume 3, Number 3, Spring 1995
- [Townsend, Irwin, 2001] Townsend, S., Irwin, G., Non-linear model based predictive control using multiple local models. In: B. Kouvaritakis, & M. Cannon (Eds.), *Non-linear predictive control: Theory and practice*, IEE Control Engineering Series, IEE, London, 61(11), 47-56
- [Tzafestas, Kapsiotis, Kyriannakis, 1997] Tzafestas, S., G. Kapsiotis and E. Kyriannakis (1997). Model-based predictive control for generalized production planning problems. *Computers in Industry* 34(2), 201-210

Authors' Information

Todorka Kovacheva – Economical University of Varna, Kniaz Boris Str, e-mail: todorka_kovacheva@yahoo.com
Daniela Toshkova – Technical University of Varna, 1, Studentska Str., Varna, 9010, e-mail: daniela_toshkova@abv.bg

GENERALIZATION BY COMPUTATION THROUGH MEMORY

Petro Gopych

Abstract: Usually, generalization is considered as a function of learning from a set of examples. In present work on the basis of recent neural network assembly memory model (NNAMM), a biologically plausible 'grandmother' model for vision, where each separate memory unit itself can generalize, has been proposed. For such a generalization by computation through memory, analytical formulae and numerical procedure are found to calculate exactly the perfectly learned memory unit's generalization ability. The model's memory has complex hierarchical structure, can be learned from one example by a one-step process, and may be considered as a semi-representational one. A simple binary neural network for bell-shaped tuning is described.

Keywords: generalization, 'grandmother' model for vision, neural network assembly memory model, one-step learning, learning from one example, neuron receptive field, bell-shaped tuning, semi-representation.

ACM Classification Keywords: Memory structures (B.3), associative memories; reliability, testing, and fault-tolerance (B.8.1); learning (I.2.6), connectionism and neural nets; vision and scene understanding (I.2.10), representations, data structures, and transforms; image representation (I.4.10), hierarchical.

1. Introduction

We know from our everyday experience that even under difficult observation conditions, the recognition of complex visual objects occurs in practice immediately, in an on-line regime. The ability to recognize visual objects regardless of the side of view, their illumination, occlusion, or particular distortion is called generalization ability; up to present its brain mechanisms remain unclear [1].

In real life, any two successive images, although they correspond to the same particular object, cannot coincide literally, point-by-point. As a result the amount of all possible images of all possible objects to be recognized is extremely large and, consequently, they the all cannot be stored in human memory even of very large but limited capacity. To overcome this difficult problem, it is supposed that it is enough to remember labels of only some typical images (examples) and to learn the common memory/generalization system to predict to a huge amount of unknown images, not storing in memory. Such a statement of the problem implies that for a given object each its particular image can continuously be transformed, possibly not too sharp, into any other its image through an infinite continuous series of its intermediate images.

The classic learning theory [2] gives a formal definition of generalization and rules to ensure it. For the generalization purposes, the learning theory provides the best possible functional relationship between an input image, x , and its label, y , by learning from a set of n examples, x_i, y_i . This problem is similar to the problem of fitting a continuous smooth function of some arguments to measurement data, x_i, y_i , or, in other words, the ability of estimating correctly the values of this function in points where data are not available (i.e. it is assumed implicitly that sets of unknown images and their labels are continuous).

Within this approach, for a given training set $(x_i, y_i; i = 1, 2, \dots, n)$, the empirical risk minimization (ERM) learning algorithm can find the estimated interpolating function f which minimizes empirical error — the quantity defining through a loss function the quality of fitting f to the training set of examples. To provide the good generalization, f should also guarantee the minimization of predictive error — the quantity defining through the same loss function the quality of fitting f to new samples — in such a way that the difference between empirical and predictive errors is zero in probability as $n \rightarrow \infty$. It may be possible if f , chosen from a given functional hypothesis space, is simple enough. In general case (for finite sets of examples and complex hypothesis spaces) by using the classic ERM learning, the solution needed it is not always possible to find [2,3]. For this reason a new paradigm of learning

was proposed which provides 'conditions for generalization in terms of precise stability property of the learning process: when training set is perturbed by deleting one example, the learned hypothesis does not change much' [3]. This stability criterion means that if after deleting any i th training sample (example) from any large training set of samples (examples) almost always the learned interpolating function f changes in small, then it generalizes well. Formally it is demanded a cross-validation leave-one-out stability with stability of empirical and expected errors: for any i , for sets of training samples S and S' (S' is the set S with the deleted item i), supremums of differences between corresponding loss functions, corresponding empirical errors, and corresponding expected errors equal zero in probability as $n \rightarrow \infty$. Such stability ensures that good (predictive) generalization functions may be found by not only the ERM process but also other learning algorithms [3] and, consequently, this method of generalization is suitable (see ref. 3 and references therein) for solving those practical problems where classic ERM learning [2] does not work. But for both cases (minimization of empirical error within a given hypothesis space or stabilization of the learning process), the important challenge of the finiteness of training sets remains unsolved because all above results are valid only asymptotically ($n \rightarrow \infty$), i.e. for a rather large amount of training examples.

The approach based on learning from a set of examples is not the only possible. Indeed, it is naturally to assume that in human visual system the real world is actually represented as a set/series of 'frames,' discrete and only perceived continuously (as in a movie). If it is, then the amount of information needed to be maintained reduces crucially and for this reason memory system, serving vision and dealing with a finite set of discrete images, may computationally become simpler. This work follows such an alternative approach.

2. Generalization by Interpolating among Examples

Within the classic learning theory [2], generalization by interpolating among examples supports a popular neural network (NN) architecture that combines the activity of some hidden broadly tuned 'units' (local NN circuits), each of which is learned to respond to one of the training examples optimally and to a variety of other images at sub-maximal firing rates. This idea is consistent with the fact that bell-shaped tuning is common among neurons in visual cortex and that in infero-temporal cortex, IT, there exist neurons tuned to different complex objects or their parts.

Mathematically, using the method of regularization, the learning from examples may be formulated as measurement data approximation by a smooth function, $f(x) = \sum w_i k(x, x_i)$, which minimizes the empirical error (error of training); here $f(x)$ is a weighted sum (weights w_i) of basis functions, $k(x, x_i)$, depending on a new (unknown) image, x . For example, function $k(x, x_i)$ may be a radial Gaussian centered on x_i , representing the i th neuron's receptive field, and responding optimally to (memorizing) x_i (that is so called radial basic function approach, RBF). The width of $k(x, x_i)$ defines also the unit's selectivity as a memory device: for broadly tuned k , its selectivity is poor but a linear combination of such functions provides a good generalization ability; for sharply tuned k (e.g., a delta function or very narrow Gaussian), its selectivity is perfect but such a $k(x, x_i)$ cannot be used for generalization. In $f(x)$, functions $k(x, x_i)$ may be learned from their inputs, x_i , in a passive regime (without the feedback) while weights, w_i , depend also on outputs, y_i , and demand more complicate iterative learning from examples, x_i , y_i . That is, the learning process splits into two parts: learning the basis functions (memory units and, simultaneously, neuron receptive fields) and learning the weights of the whole network (learning to generalize using already learned memory units). The algorithm described can implement a feedforward NN with one hidden layer containing as many units as training examples; parameters w_i are interpreted as synaptic weights between corresponding units and the output, $f(x)$ [1].

In this case [1] the ability to generalize is traditionally [2,3] grounded on the use of many training examples and is paid by the poor selectivity of all memory units (a large value of the regularization parameter), the assumption of low biological plausibility.

3. 'Grandmother' Model for Vision

In the classic 'grandmother' theory for vision, an image recognition happens when the combination of all its features precisely coincides with such a combination associated with particular grandmother neuron, i.e. in this case between the input image and different memory records a direct literally comparison is needed. The lack of generalization is the basic problem of such a model. To solve it, the model was essentially extended: it is supposed that 'generalization emerges from linear combinations of neurons tuned to an optimal stimulus' [1] (see also Section 2). We propose another extension solving the generalization problem under assumption that each memory unit itself can generalize.

As Figure 1 demonstrates, in our model all sensory-specific stages of input visual data processing coincide completely with those that Poggio & Bizzi [1] discussed and, consequently, in this part both models are biologically equally plausible. In particular, in the model proposed AIT neurons (open circles), tuned to respond to complex visual images, are also used although in present work the architecture and operation of local NN circuits, employed for tuning, are quite different (see Section 5). But the main distinction between our model (Figure 1) and Poggio & Bizzi model (Figure 2 in [1]) consists in the structure of their sensory-independent parts: in Figure 1, it is grounded on the neural network assembly memory model, NNAMM, discussed in Section 4 [4].

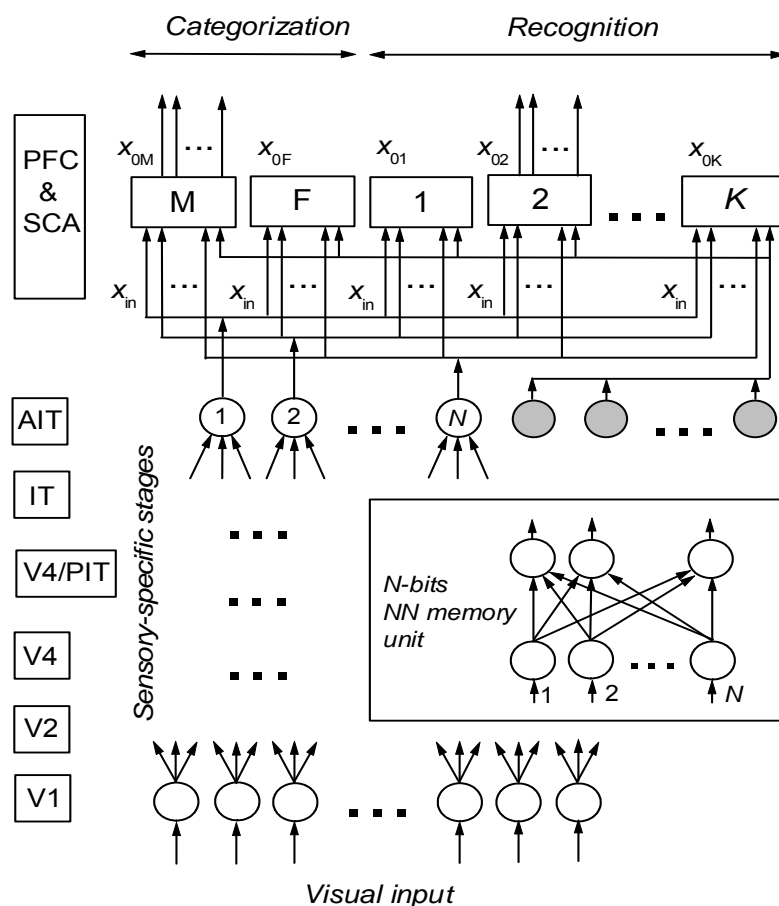


Figure 1. An oversimplified scheme of a 'grandmother' model for vision based on the NNAMM [4]. At the bottom, in V1, cells have small receptive fields and respond preferably to oriented bars; along the ventral visual stream they increase gradually their receptive fields and complexity of their preferable images and at the top, in AIT, neurons respond optimally already to rather complex objects. AIT neurons 1, ..., N (open circles) could code the image of current interest, e.g. a face, as a binary (± 1) feature vector x_{in} ; other similar neurons (filled circles) can

code (respond optimally to) other complex objects. Boxes M and F correspond to assembly memory units, AMUs (Figure 2), storing reference codes (representations) of the 'ideal' (reference) male, x_{0M} , and female, x_{0F} , faces; boxes $1, \dots, K$ denote AMUs storing the codes (representations) x_{01}, \dots, x_{0K} which represent known (previously encountered) faces $1, \dots, K$, regardless of their categorization. The case is presented where a current face feature code x_{in} , extracted from the current visual input, is recognized as the face number 2 and categorized as a male face (x_{in} initiates the correct retrieval of memory traces x_{0M} and x_{02} designated as output arrows from boxes M and 2, respectively). In the insertion, a feedforward NN, related to particular AMU_{*i*} and storing the code (representation) x_{0i} , is shown (see also box 2 in Figure 2; AIT neurons $1, \dots, N$ may be equivalent to exit-layer neurons of such an NN). If x_{in} does not correspond to one of the codes (representations) x_{01}, \dots, x_{0K} but is recognized as x_{0M} or x_{0F} then it can be remembered in the $(K + 1)$ th empty AMU, AMU _{$K+1$} , which is not shown. V1, primary visual cortex; V2 and V4, extrastriate visual areas; IT, infero-temporal cortex; AIT, anterior IT; PIT, posterior IT; PFC, prefrontal cortex; SCA, subcortical areas (e.g., as it is shown in Section 4.2, hippocampus).

We suppose that visual memory is constructed as a set of the NNAMM's assembly memory units, AMUs (Figure 2 in Section 4.2), interconnected between each other and storing only one memory trace per one AMU. Memory traces are N -dimensional binary (± 1) vectors represented particular images (e.g., known faces, x_{01}, \dots, x_{0K}) or categories of such images (e.g., male, x_{0M} , and female, x_{0F} , faces). Tuned AIT neurons $1, \dots, N$ (open circles) convey the code x_{in} , extracted from the current visual input at sensory-specific stages of data processing and representing the current face, to all AMUs devoted to vision. Similar codes of other images, available in the current visual input, are also extracted and other tuned neurons (filled circles) convey them to all AMUs devoted to vision. But by means of a spatio-temporal synchrony mechanism and anatomically in part, the AMUs shown select only the code of their interest, x_{in} ; other similar codes may be the codes of interest for other AMUs, which are not shown.

Even if the analyzed visual scene is stable, the current (at the moment t_0) visual input may slightly be changed, for example, due to a saccadic eye movement. In such a case, at the next moment, t_1 , the hierarchy of tuned local NN circuits, constituting the pathways of sensory-specific stages of initial visual signal processing (see Figure 1 and Section 5), produces, most probably, binary feature vector $x_{in}(t_1)$ which is equal to previous one, $x_{in}(t_0)$. As $x_{in}(t_0) = x_{in}(t_1)$, at sensory-independent but memory-specific stage of data processing, $x_{in}(t_1)$ initiates the recall/retrieval of memory patterns, the same as $x_{in}(t_0)$ initiates, e.g., x_{0M} and x_{02} (see Figure 1). That is, in numerous slightly (even continually) changed visual inputs, it takes place the recall/recognition of the same image of interest (e.g., a face) of the same category (e.g., male faces) whose binary representations in visual memory are vectors x_{0M} and x_{02} , respectively.

If in two successive visual scenes the difference between images of interest is not very small and not very large simultaneously then visual pathways mentioned at moments t_0 and t_1 may produce binary feature vectors $x_{in}(t_0)$ and $x_{in}(t_1)$ which are different but related to the same finite set of them characterized by the same value of the damage degree, d , or intensity of cue, q (see Section 4.1). In such a case, in spite of the fact that $x_{in}(t_0) \neq x_{in}(t_1)$, at sensory-independent but memory-specific stage of visual data processing, $x_{in}(t_1)$ initiates the recall/retrieval of memory patterns x_{0M} and x_{02} , the same as $x_{in}(t_0)$ initiates, with the same probabilities defined by Equation 5 or 6. That is, in numerous visual inputs containing rather changed or damaged images of interest, it also takes place their equally successful categorization and recall/recognition.

If the difference between images of interest in successive visual scenes is large then sensory-specific visual pathways may produce feature vectors $x_{in}(t_0)$ and $x_{in}(t_1)$ which are related to different sets of them characterized by different values of d (or q). In such a case, $x_{in}(t_1)$ also initiates successful recall/retrieval of patterns x_{0M} and x_{02} , the same as $x_{in}(t_0)$ initiates, but already with other probabilities. That is, even in numerous visual inputs containing essentially changed or damaged images of interest, their successful categorization and recall/recognition takes also place.

Consequently, the model for vision proposed provides successful categorization and recall/recognition of numerous changed, in particular essentially changed, versions of the same visual image employing its single

binary representation, x_0 , stored in visual memory. In other words, it implements the idea of generalization in its conventional form (Section 1) but in a new way — by generalization through a single NNAMM memory unit, AMU, storing only *one* binary representation, x_0 , of all possible versions of the image of interest, which may differ from each other in small as well as in large.

As one can see, a learned AMU itself ensures generalization (recall/generalization) of only its binary inputs, x_{in} , (Section 4.1) with the probability may be calculated exactly (Equation 5 or 6). To find the probability of generalization (recall/recognition) of any initial half-tone visual image completely, the probability of producing these binary feature vectors, x_{in} , is also required. For solving the latter problem, we should specify beforehand the architecture of sensory-specific visual pathways (Figure 1) as a hierarchy of tuned local NN circuits, extracting step-by-step from the initial image its more and more general features/properties (see also Section 5). When this hierarchical architecture (specific, in general, for each category of images of interest) will completely be constructed, its performance may be found as performance of a device built in a known manner from building blocks with known properties. Hence, the content of particular visual memory is *jointly* defined by the content of corresponding AMU (a rather short binary vector x_0) and complete hierarchical architecture of tuned local NN circuits, which perform a sensory-specific visual data processing and extract from complex initial input the feature vector x_{in} that, in turn, initiates the retrieval of x_0 . Very early (in the retina) stages of this many-stage process play a special role because here the binarization of initial half-tone images is carried out and the quality of binarization exerts an essential influence on the quality of the final representation of images in the entire visual system. As it was empirically demonstrated [5], the binarization required may be performed optimally, without the loss of information essential for the following binary data processing according to optimal binary algorithms described in Section 4.1.

Within the model proposed, representation of an image in visual system may be considered as a complex dynamic process consisting of three successive stages: i) binarization (in the retina) of an initial half-tone image; ii) allocating essential features of the image binarized and production of its rather short binary representation, x_0 (in a hierarchical architecture of local functionally similar tuned NN circuits which constitute visual data processing pathways); iii) storing x_0 (in visual memory) and its multi-purpose use for planning and maintaining different possible mental and behavioral operations. Owing to this three-level structure of data processing and due to the data graduate compression, the code (representation) x_0 stored in visual memory can along not specify completely its corresponding visual (perceptual) input and the same x_0 , but in memory devoted to another modality, could in general code (represent) a quite different object or idea, for example, the odour of a perfume (if x_0 is stored in olfactory memory). For the same reason, each visual (perceptual) memory (each AMU) should intimately be related to corresponding final stages of their sensory-specific pathways, strictly anatomically defined. Consequently, according to the model, in the brain should exist areas preferably devoted and responded to different specific memories and to specific categories of these memories. This theoretical prediction is completely consistent with the available anatomical findings demonstrating that is actually the fact. For example, the fusiform face area (a part of fusiform gyrus located in brain temporal lobe) is devoted to face perception in humans [6,7]. Brain damages to or near to the fusiform face area lead to specific mental disorder — prosopagnosia, an inability to perceive faces while all other mental properties remain intact [8]. Some persons suffered of prosopagnosia retain, in spite of that, the ability of face categorization (e.g., they differ males from females or olds from youngs) and can correctly identify faces of familiar persons unconsciously (e.g., their galvanic skin response rises when they hear the correct name). These neuropsychology findings are also consistent with the model proposed which predicts, in particular, that brain areas devoted to face recognition and face categorization should anatomically be segregated in part, that face perception is a many-stage process in a hierarchical brain structure (visual pathways in Figure 1) with anatomically segregated levels (areas) and damages to higher levels (areas) of visual pathways do not hinder the normal functioning of their lower levels (areas).

As it has been pointed out, an initial half-tone visual input can be binarized optimally, without the loss of information essential for the further processing of obtained binary data [5]. Perfectly learned local NN units (Section 5) and AMUs (Section 4.2) processing this data also operate over their binary inputs optimally (in the sense of pattern recognition quality, Section 4.1). Consequently, if the chain a binarization device (retina)—feature-extractive hierarchical architecture of tuned local NN units (sensory-specific visual pathways)—AMUs (visual memory, as in Figure 1) is constructed (hard-wired) in an optimal manner (that is the problem of animal evolution or an engineer who builds a machine, data processing system or device) then its operation performance may also be optimal. In sum: within the model proposed, for each specific category of images, the entire visual data processing system/algorithm may surely be optimal but the architecture, needed to implement this theoretical possibility as an algorithm or device, is not specified so far completely.

For the construction of optimal data processing architecture providing generalization through memory of visual images of different categories, only its building blocks (learned AMUs and tuned local NN units) having optimal operation performance are now available. But that is enough to conclude that such a future system/algorithm cannot solve the inverse problem: reconstruction of the initial visual input when its binary representation (x_0 stored in an AMU), properties of the retina, tuned local NN units, AMU and their connections are known. The reason is in the irreversibility of these all components constituting jointly the hierarchical architecture required. For example, a learned two-layer NN, the heart of all AMUs and tuned local NN units, is served by a finite set of its input binary vectors x_{in} and has only the single output, x_0 , providing the solution and specified strictly by the additional learned 'grandmother' neuron (Section 4.1). From these follows directly the convergence of learned AMUs and tuned local NN circuits (by definition, all their inputs, x_{in} , lead to the single solution, x_0 , stored in corresponding NN and its learned 'grandmother' neuron) and, simultaneously, their irreversibility (by definition, their the given solution, x_0 , cannot exactly specify that one of many particular NN inputs, x_{in} , which has earlier initiated the recall/retrieval of x_0). It is clear that the reversible processing system (computer algorithm) required to solve an inverse problem cannot be built from irreversible components.

There exists two opposite viewpoints of the nature of human memory. On the one hand, traditionally (e.g., [1]), it is supposed that objects, actions, etc are stored in memory as their representations, i.e., as coded messages may be used, if necessary, as instructions governing the learned mental or physical behavior. On the other hand, it was introduced the so called nonrepresentational memory, an ability of dynamic system 'to repeat or suppress a mental or physical act' or an 'ordered sequence of brain activities ... that, in time, leads to a particular neuron output.' 'In this view, a memory is dynamically *generated* from the activity of selected subsets of circuits' [9]. Within our BSDT/NNAMM approach, a memory is defined as consisting of two closely related parts: the representational code x_0 , stored in an AMU related to particular visual (perceptual) memory, and the stream of neuron activity, dynamically generated in visual pathways and directed from the retina to the AMU mentioned. That is, our memory model has some properties of representational as well as nonrepresentational memories and may be qualified as a *semi-representational* one.

In contrast to Section 2, the NNAMM's memory unit itself provides perfect memory trace selectivity as well as generalization through memory. Because each AMU contains a 'grandmother' neuron (for details see Sections 4 and 5), we can consider the model for vision introduced as a 'grandmother' one.

4. NNAMM as a Memory Model Used

P.M.Gopych has proposed a ternary/binary data coding and demonstrated [10] that corresponding NN decoding algorithm (inspired by J.J.Hopfield [11]) is simultaneously the retrieval mechanism for an NN memory. As NNs used for data decoding and memory storing/retrieval are the same (see insertion in Figure 1), they have also common data-decoding/memory-retrieval performance (Section 4.3). Later this data coding/decoding approach was developed into the binary signal detection theory (BSDT) [12] and neural network assembly memory model (NNAMM) [4] closely interrelated in their roots and providing the best quality performance. The price paid for the

NNAMM optimality is the fact that it places each memory trace in its own AMU (an estimation of human memory capacity, though it is possibly too optimistic — 10^{8432} bits [13], supports this assumption).

4.1 Formal Background

Let us denote a vector with components x^i ($i = 1, \dots, N$), whose magnitudes are ± 1 , as x . It can carry N bits of information and its dimension N is the size of a local receptive field for the NN/convolutional feature discrimination algorithm [5] or the size of an NN memory unit discussed below. If x represents information stored or that should be stored in the NN then we term it reference vector x_0 . If the signs of all components of x are randomly chosen with uniform probability, $1/2$, then that is random vector x_r or binary noise. We define also a damaged reference vector $x(d)$ with components

$$x_i(d) = \begin{cases} x_0^i, & \text{if } u_i = 0, \\ x_r^i, & \text{if } u_i = 1 \end{cases} \quad d = \sum u_i / N, \quad i = 1, \dots, N, \quad (1)$$

where marks u_i take magnitudes 0 or 1 and may randomly be chosen with uniform probability, $1/2$; d is a damage degree of x_0 . If the number of marks $u_i = 1$ is m then the fraction of noise components in $x(d)$ is $d = m/N$; $0 \leq d \leq 1$, $x(0) = x_0$ and $x(1) = x_r$. The fraction of intact components of x_0 in $x(d)$, $q = 1 - d$, is *intensity of cue* or *cue index*; $0 \leq q \leq 1$, $q + d = 1$, d and q are proper fractions. For a given $d = m/N$, the number of different vectors $x(d)$ is $2^m C_m^N$, $C_m^N = N! / (N - m)! m!$; for d ranged $0 \leq d \leq 1$, complete finite set of all vectors $x(d)$ consists of $\sum 2^m C_m^N = 3^N$ elements ($m = 0, 1, \dots, N$).

For decoding the data coded as described, we use a two-layer NN with N McCulloch-Pitts model neurons in its entrance and exit layers; these neurons are linked as in the insertion of Figure 1, 'all-entrance-layer-neurons-to-all-exit-layer-neurons.'

For a learned NN, its synapse matrix elements, w_{ij} , are

$$w_{ij} = \xi x_0^i x_0^j \quad (2)$$

where $\xi > 0$ is a parameter (below $\xi = 1$); x_0^i and x_0^j are the i th and the j th components of x_0 , respectively. Hence, the matrix w is defined by vector x_0 and Equation 2 unambiguously. We refer to w as the perfectly learned NN and it is of crucial importance that it remembers only *one* pattern x_0 (the available possibility of storing other memories in the same NN is intentionally disregarded). It is also assumed that the NN's input vector x_{in} is decoded (reference or state vector x_0 is extracted) successfully if the learned NN transforms an x_{in} into the output vector $x_{out} = x_0$ (an additional 'grandmother' neuron checks this fact; see also Sections 3, 4.2, 5, and ref. 14).

The transformation algorithm is the following. For the j th exit-layer neuron, its input signal, h_j , is

$$h_j = \sum w_{ij} v_i, \quad i = 1, \dots, N \quad (3)$$

where v_i is an output signal of the i th entrance-layer neuron. The j th exit-layer neuron's output, x_{out}^j , is calculated by a rectangular response function with the neuron's triggering threshold $\theta \geq 0$ (for the case $\theta < 0$, see ref. 14):

$$x_{out}^j = \begin{cases} +1, & \text{if } h_j > \theta \\ -1, & \text{if } h_j \leq \theta \end{cases} \quad (4)$$

where for $h_j = \theta$ the value $v_j = -1$ is arbitrary assigned.

Since entrance-layer neurons of the NN used play only the role of input fan-outs, which convey their inputs to all exit-layer neurons, in Equation 3 $v_i = x_{in}^i$. Of this fact and Equations 3 and 4 for the j th exit layer neuron we have: $h_j = \sum w_{ij} x_{in}^i = x_0^j \sum x_0^i x_{in}^i = x_0^j Q$ where $Q = \sum x_0^i x_{in}^i$ is a convolution of x_0 and x_{in} . The substitution of $h_j = x_0^j Q$ into

Equation 4 gives that $x_{out} = x_0$ and an input vector x_{in} is decoded (reference vector x_0 is extracted) successfully if $Q > \theta$. Since for each x_{in} exists such a vector $x(d)$ that $x_{in} = x(d)$, inequality $Q > \theta$ can also be written as a function of $d = m/N$: $Q(d) = \sum x'_0 x_i(d) > \theta$ ($i = 1, 2, \dots, N$) where θ is the threshold of Q and, simultaneously, the neuron's triggering threshold. Hence, for perfectly learned intact NNs, NN and convolutional decoding algorithms are functionally equivalent.

Since $Q > \theta$ and $D = (N - Q)/2$, where D is Hamming distance between x_0 and specific $x(d)$, the inequality $D < (N - \theta)/2$ is also valid and NN, convolutional, and Hamming distance decoding algorithms mentioned are equivalent. As Hamming decoding algorithm is the best (optimal) in the sense of statistical pattern recognition quality (i.e., there is no other algorithm outperforming it), NN and convolutional algorithms are also optimal (the best). Moreover, similar decoding algorithms based on locally damaged NNs may also be optimal [4,15] (see example in Table 1 of Section 6).

4.2 AMU's Architecture

We saw that a two-layer NN (as in the insertion of Figure 1) can be used for optimal one-trace memory storing/retrieval. But for such an NN, one separate randomly chosen vector $x_{in} = x(d)$ can initiate successful retrieval only randomly. Thus, to implement the model's possibilities completely, the retrieval should be initiated by a series of different vectors x_{in} and it happens when one of the next x_{in} leads suddenly to the emergence of the output $x_{out} = x_0$. For this reason the minimal architecture, needed to provide optimal memory trace retrieval from the learned NN (box 2), should be as in Figure 2. Because the retrieval is initiated by vectors $x(d)$, which constitute complete finite set of binary representations of those images (or 'frames') that were mentioned in the last paragraph of Section 1, such an architecture provides also the optimal generalization by computation through memory. The internal loop, 1-2-3-4-1, ensures the generation of different (e.g., random) vectors $x_{in} = x(d)$ with a given value of d while the external loop, 1-2-3-4-5-6-1, maintains the internal one.

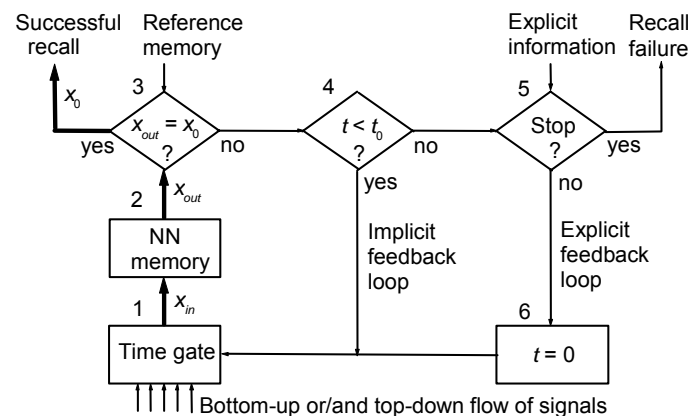


Figure 2. The flow chart (the architecture) of an assembly memory unit, AMU, and its short-distance environment adopted from [4]. The structure of the NN memory unit (box 2) specifies the insertion in Figure 1. Pathways and connections are shown in thick and thin arrows, respectively.

Within the NNAMM, the whole memory is a very large set of interconnected AMUs of rather small capacity ($N \sim 100$ or less), organized hierarchically. An AMU (Figure 2) consists of boxes 1, 2 and 6, diamonds 3, 4 and 5; their internal and external pathways and connections are designed to propagate synchronized groups of signals [vectors $x(d)$] and asynchronous control information, respectively. AMUs implement directly the BSDT for solving the problem of optimal generalization and memory storing/retrieval.

Box 1 (a kind of N -channel time gate) transforms initial ternary ($0, \pm 1$) sparsely coded very-high-dimensional vectors into binary (± 1) densely coded and rather low-dimensional ones. Here from the flood of asynchronous

input spikes, a synchronized pattern of signals in the form of N -dimensional feature vector $x_{in} = x(d)$ is extracted by a dynamical spatiotemporal synchrony mechanism. Box 2 is an NN learned according to Equation 2 (or Equation 7 from Section 4.4) where each input, x_{in} , is transformed into its corresponding output, x_{out} . Diamond 3 (a kind of comparator or familiarity/novelty detector) performs the comparison of x_{out} , just now emerged, with the reference vector (trace) x_0 from *reference memory* (RM, see below). If $x_{out} = x_0$, then the retrieval is successful and it is finished. In the opposite case, if current time of retrieval, t , is less than its given maximal value, t_0 , (this fact is checked in diamond 4) then the loop 1-2-3-4-1 is activated, retrieval starts again from box 1, and so forth. If t_0 , a parameter of time dependent neurons, was found as insufficient to retrieve x_0 then diamond 5 examines whether an external reason exists to continue retrieval. If it is, then the loop 1-2-3-4-5-6-1 is activated, the count of time begins anew (box 6), and internal cycle 1-2-3-4-1 is repeated again with a given frequency f , or time period $1/f$, while $t < t_0$.

The trace x_0 is held simultaneously in a particular NN memory (box 2) and in its auxiliary reference memory (RM) that may be interpreted as a *tag* of corresponding NN memory or as a *card* in a long-term memory catalog. An RM performs two interconnected functions: verification of current memory retrieval results (diamond 3 serves as a comparator) and validation of the fact that a particular memory record actually exists in the long-term memory store (diamond 3 serves as a familiarity/novelty detector). Thus, specific RM is a part of memory about memory or '*metamemory*,' in other words. In contrast to the NN memory, which is a kind of computer register and is conventionally associated with a real biological network, particular RM is a kind of *slot* devoted to the comparison of a current vector x_{out} with the reference pattern x_0 and may be associated with a coincidence integrate-and-fire '*grandmother*' neuron (cf. Sections 3, 4.1, 5, and ref. 14).

All elements of the internal feedback (reentry) loop, 1-2-3-4-1, run routinely in an automatic regime and for this reason they may be interpreted as related to an *implicit* (unconscious) memory. Consequently, under the NNAMM, all operations at synaptic and NN memory levels are unconscious. External feedback (reentry) loop, 1-2-3-4-5-6-1, is activated in an unpredictable manner because it relies on external (environmental and, consequently, unpredictable) information and in this way provides unlimited diversity of possible memory retrieval modes. For this reason, an AMU can be viewed as a particular *explicit* (conscious) memory unit. An external information used in diamond 5 can be thought of as an explicit or conscious one.

Recent evidences demonstrate that learning induces molecular changes in neocortex and hippocampus; this finding, along with based on it physiological theory assuming that any long-term memory record is stored in parallel in the neocortex and hippocampus [16], supports the NNAMM's idea of storing simultaneously each memory trace in an NN (a counterpart to a neocortex network) and in a '*grandmother*' neuron (probably, a cell in hippocampal structures). This point of view is also consistent with the content of ref. 17,18 where a hippocampal comparator or familiarity/novelty detector is considered. For some other arguments in favor of the NNAMM's biological plausibility see ref. 4.

4.3 AMU's Basic Performance

The best data-decoding/memory-retrieval algorithms considered have common quality performance function, $P(d,\theta)$, the probability of correct decoding/retrieval or generalization, conditioned under the presence or absence of x_0 in the data analyzed, as a function of d and θ ($d = 1 - q$, all notations are as in Section 4.1).

The finiteness of the set of vectors $x(d)$ makes possible to find $P(d,\theta)$ by multiple computations [10]:

$$P(d,\theta) = n(d,\theta)/n(d) \quad (5)$$

where $n(d)$ is a given number of different inputs with a given value of d , $x_{in} = x(d)$; $n(d,\theta)$ is the number of those $x(d)$ which are leading (under condition that for their decoding the NN algorithm with triggering threshold θ is applied) to the NN's response $x_{out} = x_0$. For small N , $P(d,\theta)$ can be calculated exactly because the number of items in the complete set of $x(d)$, $n(d) = 2^m C_m^N$, is small and they the all can be taken into account. For large N ,

$P(d, \theta)$ can be estimated by multiple computations approximately but, using a sufficiently large set of randomly chosen inputs $x(d)$, with any given accuracy.

For intact perfectly learned NNs, convolutional (Hamming) version of the BSDT/NNAMM formalism allows to derive an expression for $P(d, \theta)$ analytically [15]:

$$P(d, \theta) = \sum_{k=0}^{k_{\max}} C_k^m / 2^m, \quad k_{\max} = \begin{cases} (N - \theta - 1)/2, & \text{if } N \text{ is odd} \\ (N - \theta)/2 - 1, & \text{if } N \text{ is even.} \end{cases} \quad (6)$$

Here if $k_{\max} \leq m$ then $k_{\max} = m$ else $k_{\max} = k_{\max_0}$, C_k^m denotes a binomial coefficient.

Since θ (triggering threshold) and F (false-alarm probability), d (damage degree) and q (intensity of cue) are related (see details in ref. 12), functions $P(d, \theta)$ can, for example, be written as ROCs [receiver operating characteristic curves, $P_q(F)$], or BMPs [basic memory performance curves, $P_F(q)$] [4].

4.4 AMU's Learning

Equation 2 defines perfect one-step learning from one example because in this case for the NN considered its input, x_0 , and its output, x_0 (the label, 'teacher,' or 'supervisor'), are exactly known. But often unsupervised learning is also needed.

Let us use the traditional delta learning rule in the form

$$w_{ij}^{(n+1)} = w_{ij}^{(n)} + \eta v_j^{(n)} h_i^{(n)} \quad (7)$$

where n and $\eta > 0$ are an iteration number and a learning parameter, respectively; $v_j = x_{in}^j$; $h_i = \sum w_{ik} v_k$, $k = 1, \dots, N$. Here the training set consists of only one sample, $x_{in} = x_0$, and, consequently, Equation 7 describes learning from one example (such an iteration process does not feedback the NN's output x_{out} to the NN's input; the current value of w_{ij} is estimated using its previous value, the values of η and components of x_0).

If η is small ($\eta < 1$) then the learning rate achieved is low and asymptotic values of w_{ij} are not reached. This case has no essential practical significance. If η is large ($\eta > 100$) then the iteration process leads to a fast, one-trial, without the 'catastrophic forgetting' learning because already the first iteration gives the result which is close to the asymptote and next iterations do not lead to the essential advance.

Let us consider the NN with $N = 40$, continuous w_{ij} , v_j , h_i , x_{in}^i , x_{out}^i and all initial values of w_{ij} chosen randomly with uniform probability from the range $[-1, 1]$. If the initial learning pattern is $x_{in} = x_0$ then after each next iteration an NN with the next version of its weight matrix w_{ij} provides the emergence of the next version of x_{out} (the next approximation of x_0). For example, for $\eta = 400$ already the first iteration gives the approximation's quality estimation $\sum |x_{out}^i - x_0^i| < 10^{-30}$ ($i = 1, \dots, N$) which is more than enough for practical purposes.

Simultaneously with the NN itself, its specific reference memory (RM) should also be learned (for example, by direct recording the components of x_0 into RM).

5 Neuron RFs and NNs for Tuning

Figure 3 illustrates the process of visual data processing using the NN described in Section 4.1 and AMU described in Section 4.2. A binarization algorithm (e.g., [5]) transforms vector y , a half-tone image, into spinlike vector x_{in} without loss of information important for the following feature discrimination procedure (e.g., if $y_i > bd_i$ then $x_{in}^i = 1$ else $x_{in}^i = -1$). It is supposed that binarization of components of y or h means spike generation; h may be interpreted as a simplified 1D profile of a 'grandmother' neuron's receptive field (RF) which results in the process of internal weighted network computations (Equation 3, see also ref.14). Profiles of such RFs can be as it is typical for on-cells (panels a, c, d) or for off-cells (panel b) and, as panels a and b demonstrate, noise $x_{in} = x_r$ can initiate the reverse of the RF polarity (these predictions are consistent with current physiological results [19]). At a given level of data processing, the set of outputs of 'grandmothers' of different NNs (the top row in Figure 3)

reduces the redundancy of initial data and can constitute an x_{in} for NNs at the next level of data processing hierarchy; in particular, in AIT, an x_{in} could already represent a face (Section 3).

From the above consideration and example (Figure 3) follows that within the theory for vision proposed for image recognition at any level of data processing hierarchy, the NNs of the same structure are used [Equations 2-4, Figures 1 (the insertion) and 3]. These NNs perform a given normalization of a current input (Equation 3) and thresholding the result (Equation 4). The main distinction between them consists in the fact that they (and their 'grandmothers') are learned ('tuned') to recognize different binary patterns x_0 which, depending on the context, can code simple elements/features of a visual scene (e.g., oriented bars in V1) as well its rather complex objects or their parts (e.g., human faces in AIT). Tuning the NNs considered to recognize optimally (to respond preferably to) x_0 is extremely simple because that is a one-step learning from one example, according to Equation 2 or 7. As NNs can generalize, they can recognize patterns $x(d)$, which are x_0 damaged by noise, and the more the d the smaller $P(d,\theta)$ is [d is a damage degree of x_0 , $0 \leq d \leq 1$, $x(0) = x_0$; $P(d,\theta)$ is the probability of recognizing x_0 in $x(d)$; how at $\theta = 0$ $P(d,\theta)$ depends on d , one can see from examples in Figure 3 of ref. 4]. Thanks to this property of functions $P(d,\theta)$, the NN's tuning is 'bell-shaped' (at $\theta = 0$, Figure 3 of ref. 4 displays examples of some possible bell-shaped profiles of tuning). Hence, the NN introduced may be interpreted as an universal circuit underlying bell-shaped tuning in different visual brain areas (here, it is important to note that because within our 'grandmother' theory for vision each NN discussed is connected to its 'grandmother' neuron, the terms 'tuning the NN' and 'tuning the neuron' are synonymous).

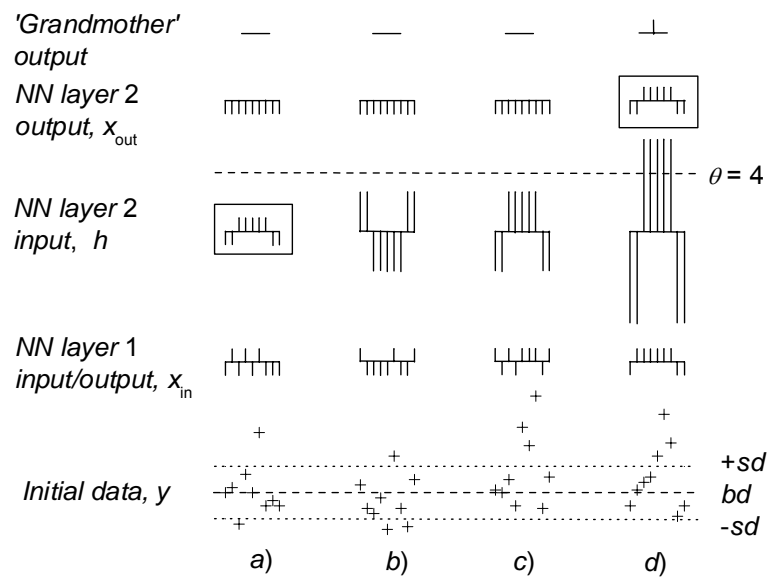


Figure 3. Computer simulated samples of initial visual data, y (e.g., an electric output of light-sensitive retina cells), and their processing results, x_{in} , h , x_{out} , in four N -channel data processing windows (crosses are values of y in each channel). In panels a and b , y is a fixed background, bd , damaged by Poisson-like noise, $bd = 100$; in panels c and d , y is a Gaussian peak on the background, bd , their sum is damaged by Poisson-like noise, the peak's amplitude is $a = 20$, its full width at half maximum is $fwhm = 5$. Vectors y , x_0 (boxed), x_{in} , h (1D profile of an RF), and x_{out} are N -dimensional ones, $N = 9$; positive and negative components of x_0 , x_{in} , h , x_{out} correspond to upward and downward bars, respectively; intact NN and its 'grandmother' hold $x_0 = (-1, -1, 1, 1, 1, 1, 1, -1, -1)$, a kernel for the convolutional decoding/retrieval algorithm (the neuron's triggering threshold is $\theta = 4$); peak is identified in panel d ; $sd = bd^{1/2}$, standard deviation of bd .

It is clear that the employment of NNs discussed for solving the problem of generalization may be considered as a kind of alternative to radial basic function approach (RBF), mentioned in Section 2.

6 Generalization by Computation through Memory Performance

Since $P(q, \theta)$ defines (Equation 5) the fraction of vectors $x_{in} \neq x_0$ leading, along with x_0 , to successful retrieval of the trace x_0 from the learned NN (the insertion in Figure 1 and box 2 in Figure 2), the probability of memory retrieval, $P(q, \theta)$, and generalization ability by computation through memory, $g(q, \theta)$, are numerically equal, $g(q, \theta) = P(q, \theta)$.

Table 1

Generalization ability, $g(q, \theta) = n(q, \theta)/n(q)$, for an AMU storing the trace $x_0 = (-1, -1, 1, 1, 1, 1, 1, -1, -1)^1$.

q	Intact NN, $g(q, 6), \%^2$	Damaged NN, $g(q, 0), \%^3$	q	Intact NN, $g(q, 6), \%$	Damaged NN, $g(q, 0), \%$
1	2	3	4	5	6
0/9	10/512 = 1.953	10/512 = 1.953	5/9	5/16 = 31.250	630/2016 = 31.250
1/9	9/256 = 3.516	81/2304 = 3.516	6/9	4/8 = 50.000	336/672 = 50.000
2/9	8/128 = 6.250	288/4608 = 6.250	7/9	3/4 = 75.000	108/144 = 75.000
3/9	7/64 = 10.938	588/5376 = 10.938	8/9	2/2 = 100.000	18/18 = 100.000
4/9	6/32 = 18.750	756/4032 = 18.750	9/9	1/1 = 100.000	1/1 = 100.000

¹ q , intensity of cue ($q = 1 - d = 1 - m/N$, $0 \leq m \leq N$, $N = 9$; $q = 0$, free recall; $0 < q < 1$, cued recall; $q = 1$, recognition); θ , the neuron's triggering threshold; for definitions of $n(q, \theta)$ and $n(q)$, see Section 4.3.

² Values of $g(q, 6)$ were calculated by Equations 5 and 6, results are equal.

³ Values of $g(q, 0)$ were calculated by Equation 5; 30 disrupted interneuron connections (entrance-layer neuron, exit-layer neuron) are as follows: (2,1), (4,1), (5,1), (6,1), (8,1), (3,2), (5,2), (7,2), (1,3), (4,3), (5,3), (2,4), (4,4), (2,5), (3,5), (7,5), (9,5), (3,6), (7,6), (8,6), (9,6), (1,7), (2,7), (4,7), (8,7), (1,8), (5,8), (3,9), (6,9), (7,9); this set of disrupted connections was chosen to illustrate the fact that similar to intact NNs, damaged NNs can also provide the best decoding/retrieval/generalization performance (in columns 2 and 3, 5 and 6, generalization abilities coincide completely).

In Table 1, generalization abilities for two AMUs, containing an intact and a damaged NN, are compared. In columns 2, 3, 5, and 6, values of $g(q, \theta)$ provide optimal (the best in the sense of pattern recall/recognition quality) generalization abilities; $g(0, 6) = g(0, 0) \sim 1\%$ was, for example, chosen as that is typical for professionals [5].

Usually, generalization is considered as a function of the relative size $\alpha = n/N$ of the training set of n examples and the learning strategy. For very large networks ($N \rightarrow \infty$) and $\alpha \gg 1$, the error of generalization decreases as $\sim \alpha^{-1}$ [20]; for small networks (for learning from few examples), the problem of generalization remains unsolved in theory [1]. The approach proposed in this work gives a solution of this problem because it makes possible learning even from one example (Section 4.4).

7 Conclusion

The first solution of the problem of generalization through memory has been proposed and illustrated by an original 'grandmother' theory for vision, here introduced using the recent neural network assembly memory model, NNAMM [4]. For the NNAMM's intact NN memory unit, analytical formulae and a numerical procedure are found to calculate exactly optimal values of generalization as a function of the cue index, q , and the neuron's triggering threshold, θ ; for two specific NNs their generalization abilities are numerically calculated (it is important that in all calculations simple binary/digital mathematics is only used). It has been demonstrated that the approach proposed provides generalization for the case of learning even from one example and that binary NNs discussed can also be interpreted as universal circuits underlying bell-shaped tuning of neurons in different visual brain areas.

Acknowledgments

I am grateful to the Health InterNetwork Access to Research Initiative (HINARI) for free on-line access to current full-text research journals, to an anonymous reviewer for useful comments, and to my family and my friends for their help and support.

Bibliography

1. T.Poggio and E.Bizzi. Generalization in vision and motor control. *Nature*, 2004, 431(7010), 768-774.
2. V.N.Vapnik. *Statistical learning theory*. Wiley, New York, 1998.
3. T.Poggio, R.Rifkin, S.Mukherjee and P.Niyogi. General conditions for predictivity learning. *Nature*, 2004, 428(6981), 419-422.
4. P.M.Gopych. A neural network assembly memory model based on an optimal binary signal detection theory. *Problemy Programirovaniya (Programming Problems, Kyiv, Ukraine)*, 2004, no. 2-3, 473-479; see also <http://arXiv.org/abs/cs.AI/0309036>.
5. P.M.Gopych. Identification of peaks in line spectra using the algorithm imitating the neural network operation. *Instruments and Experimental Techniques*, 1998, 41(3), 341-346.
6. N.Kanwisher, J.McDermott and M.M.Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neurosciences*, 1997, 17(11), 4302-4311.
7. E.Tong, K.Nakayama, M.Moskovitch, M.Weinrib and N.Kanwisher. Response properties of human fusiform face area. *Cognitive Neuropsychology*, 2000, 17(1), 257-280.
8. Y.Wada and T.Yamamoto. Selective impairment of face recognition due to a haematoma restricted to the right fusiform and lateral occipital region. *Journal Neurology, Neurosurgery and Psychiatry*, 2001, 71(2), 254-257.
9. G.Edelman and G.Tononi. *A universe of consciousness: How matter becomes imagination*. Basic Books, New York, 2000.
10. P.M.Gopych. Determination of memory performance. *JINR Rapid Communications*, 1999, 4[96]-99, 61-68 (in Russian).
11. J.J.Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings National Academy of Sciences USA*, 1982, 79(8), 2554-2558.
12. P.M.Gopych. Sensitivity and bias within the binary signal detection theory, BSDT. *Int. Journal Information Theories & Applications*, 2004, 11(4), 318-328.
13. Yingxu Wang, D.Liu and Ying Wang. Discovering the capacity of human memory. *Brain and Mind*, 2003, 4(2), 189-198.
14. P.M.Gopych. Neural network computations with negative triggering thresholds. In *ICANN 2005, Lecture Notes in Computer Sciences 3696*, W. Duch et al. editors, pages 223-228. Springer-Verlag, Berlin-Heidelberg, 2005.
15. P.M.Gopych. ROC curves within the framework of neural network assembly memory model: Some analytic results. *Int. Journal Information Theories & Applications*, 2003, 10(2), 189-197.
16. P.K.Dash, A.E.Hebert and J.D.Runyan. A unified theory for systems and cellular memory consolidation. *Brain Research Reviews*, 2004, 45(1), 30-37.
17. J.Gray. The content of consciousness: A neuropsychological conjecture. *Behavioral Brain Sciences*, 1995, 18(4), 659-722.
18. D.C.Dennett. Overworking the hippocampus. *Behavioral Brain Sciences*, 1995, 18(4), 677-678.
19. G.C.DeAngelis, I.Ohzawa and R.D.Freeman. Receptive-field dynamics in the central visual pathways, *Trends in Neurosciences*, 1995, 18(10), 451-458.
20. M.Opper. Statistical mechanics of generalization. In *The Handbook of Brain Theory and Neural Networks*, Michael A. Arbib editor, pages 922-925. The MIT Press, Cambridge, Massachusetts, 1995.

Author's Information

Petro Mykhaylovych Gopych – V.N.Karazin Kharkiv National University; Svoboda Sq., 4, Kharkiv, 61077, Ukraine; e-mail: pmg@kharkov.com.

About ...

FP6-IST INFRAWEBBS EUROPEAN RESEARCH PROJECT

INFRAWEBBS
europe



<http://www.fh-bochum.de/infrawebs/>

Web services define new paradigm for the Web in which a network of computer programs becomes the consumers of information. The platform and language independent interfaces of such services allow the easy integration of heterogeneous systems. Current Web service technologies describe the syntactical aspects of a Web service providing only a set of rigid services that cannot be adapted to a changing environment without human intervention. Realization of the full potential of the Web services requires further technological advances in the areas of service interoperation, service discovery and service composition. A possible solution to these problems can be provided by application of Semantic Web technologies to converting Web services to Semantic Web Services (SWS).

Semantic Web Services are "self-contained, self-describing, semantically marked-up software resources that can be published, discovered, composed and executed across the Web in a task driven semi-automatic way". Semantic Web services can constitute a solution to the integration problem enabling dynamic, scalable and reusable cooperation between different systems and organizations. These great potential benefits have led to the establishment of an important research area, both in industry and academia, to realize Semantic Web services.

There are two major initiatives aiming at developing world-wide standard for the semantic description of Web services. The first one is OWL-S, a collaborative effort by BBN Technologies, Carnegie Mellon University, Nokia, Stanford University, SRI International and Yale University. OWL-S is intended to enable automation of web service discovery, invocation, composition, interoperation and execution monitoring by providing appropriate semantic descriptions of services. The second one is Web Service Modelling Ontology (WSMO), a European initiative intending to create an ontology for describing various aspects related to Semantic Web Services and to solve the integration problem. WSMO consortium includes more than 50 academic and industrial partners. The next, more technology-oriented step in the process of development of semantic Web services is proposed in the ongoing FP6-IST INFRAWEBBS European research project.

The primary project's objective is to develop an ICT framework, which enables software and service providers to generate and establish open and extensible development platforms for creating and maintaining Semantic Web Services supporting specific applications based on WSMO framework.

INFRAWEBBS is an EU FPG STREP project, involving leading authorities on Semantic Web Technologies to develop an application-oriented software toolset for creating, maintaining and executing open and extensible development platforms for Semantic Web services. INFRAWEBBS is an Intelligent Framework for Generating Open (Adaptable) Development Platforms for Web - Service Enabled Applications Using Semantic Web Technologies, Distributed Decisions Support Units and Multi - Agent - Systems. It is a Specific Target Research Project of the European Commission 6th Framework Programme - Priority 2 "Information Society Technologies"; Proposal Number: 511723

Project Focus

The main INFRAWEB project focus and objective is the development of an application-oriented software toolset for creating, maintaining and executing WSMO-based Semantic Web Services (SWS) within their whole life cycle.

This next generation of tools and systems will enable software and service providers to build open and extensible development platforms for web service applications. These services will run on open standards and specifications, such as BPEL4WS, WSMO, WSMX, WSML, SPARQL, RDF, etc. In particular, they will be compliant with WSMO (Web Services Modelling Ontology), a W3C initiative in Semantic Web services.

The systems generated will consist of loosely-coupled and linked INFRAWEB units, with each unit providing tools and adaptable system components. Developers will be able to use these components to analyse, design and maintain WSMO-based Semantic Web services across the whole lifecycle.

These Semantic Web services offer a new dimension in collaborative work and service production, service provision and service maintenance in run-time environments.

In the first step, the INFRAWEB units are being used to establish an open development platform for SMEs and industrial vendors.

Project Activities

The essential project activities are to build up software modules within an integrated framework - IIF - the Integrated Infrawebs Framework. Several functionalities are provided for the usage in design-time as well as in run-time (by service providers, service designers or service brokers). For the design-time phase they are:

- The SWS Designer, which is aimed at designing a WSMO-based Semantic Web service from an existing non-semantic Web service.
- The Organizational Memory, a Web Service implementation of a case-based memory (learning from the past), which stores and categorises non-logical representations of WSMO objects as well as additional non-semantic data (like graphical models and templates of SWS).
- The Semantic Information Router as a metadata based content management and aggregation platform (endowed with a SPARQL query interface), used by other components to query for annotated and categorized service descriptions.
- The SWS Composer for creating a Semantic Web service through composition of existing WSMO-based SWSs. It uses a case base memory for retrieving service composition templates quasi-similar to the orchestration interface of the service to be composed.

The run-time modules (for service consumers, service providers, service brokers) are given by:

- The Distributed Repository for effective storing and retrieving all semantic elements of the WSMO Framework: Goals, Ontologies, SWS and Mediators (written in WSML), whereas each repository consists of two parts: a local repository (storing of all WSMO objects created in the Semantic Web service Unit, and a local registry for advertising the SWSs).
- The Service Access Middleware provides a retrieval and execution interface for advertised SW services. The user mandates a user interface agent for fulfilling the service demand and the agent provides recommendations based on the user's query. The matchmaking between user request and service capability are similarity and logic based.
- The SWS Executor module processes Semantic Web service WSMO descriptions using choreography and orchestration engines for executing specific SWS related rules.
- The QoS (Quality of Service) Broker provides functionalities for monitoring the SW service execution process by feeding back extracted metric data.
- The Security and Privacy enabler realised as an artificial "immune defence system" allowing the INFRAWEB framework to function properly under changing conditions

INFRAWEBs represents a novel approach to problem solving in the creation of SWS applications. It involves a tight integration of similarity-based (non-semantic) and logic-based (semantic) reasoning.

Impact and Exploitation

In the quest of competitive edge, companies in Europe are pressed to gain innovation, become faster and more flexible (i.e highly dynamic and adaptable), but also offer a wider range of stable and reliable services along with a personalized interaction with customers, clients and partners.

Undoubtedly, ICT supporting modern, dynamic, reconfigurable, and scalable technologies like the INFRAWEBs approach, do play an important role in tackling these challenges and implementing advanced semantically based knowledge domains.

INFRAWEBs is a flexible, interoperable and reconfigurable framework, enabling organisations to build up partnerships faster and in a more effectively way with respect to the service generation, execution and distribution process. By allowing peers to change their role - to be client, broker and service provider within one environment - INFRAWEBs ensures a highly dynamic and efficient service production process and workflow, and one which spans the whole service lifecycle.

About papers presented in IJ ITA

The first paper in the proposed set from the ongoing project works "INFRAWEBs Semantic Web Service Development on the Base of Knowledge Management Layer" (Nern et al) describes the knowledge management layer for developing of Semantic Web Service that is embedded in an application oriented realization framework.

An important part of INFRAWEBs is a Semantic Web Unit (SWU) – a collaboration platform and interoperable middleware for ontology-based handling and maintaining of SWS. INFRAWEBs Designer is sub-module of SWU responsible for creating Semantic Web Services.

According to WSMO, functional and behavioral descriptions of a SWS may be represented by means of complex logical expressions (axioms). The paper "INFRAWEBs Axiom Editor – a Graphical Ontology-Driven Tool for Creating Complex Logical Expressions" (Agre et al) describes a specialized user-friendly tool for constructing and editing such axioms – INFRAWEBs Axiom Editor that is a part of INFRAWEBs Designer.

"A Survey on The Integration of Enterprise Applications as a Service" (Hristina Daskalova, Vladislava Grigorova) discusses the integration process of Web services using business logic in multi-lateral integration of business applications.

INFRAWEBs project considers usage of semantics for the complete lifecycle of Semantic Web processes, which represent complex interactions between Semantic Web Services. In "Semantic Description of Web Services and Possibilities of BPEL4WS" (Vladislava Grigorova) methods of using of BPEL4WS as a component of web services technology for the purposes of Semantic Web Services semi-automatic integration are suggested.

In "INFRAWEBs BPEL-Based Editor for Creating the Semantic Web Services Description" (Tatiana Atanasova) the conceptual architecture for BPEL-based INFRAWEBs editor is proposed that is intended to construct a part of WSMO descriptions of the Semantic Web Services

In "Adjusting WSMO API Reference Implementation to Support More Reliable Entity Persistence" (Ivo Marinchev) the WSMO technology concerning issues are discussed.

INFRAWEB S SEMANTIC WEB SERVICE DEVELOPMENT ON THE BASE OF KNOWLEDGE MANAGEMENT LAYER

Joachim Nern, Gennady Agre, Tatiana Atanasova, Zlatina Marinova,
Andras Micsik, Laszlo Kovacs, Janne Saarela, Timo Westkaemper

Abstract: The paper gives an overview about the ongoing FP6-IST INFRAWEB S project and describes the main layers and software components embedded in an application oriented realisation framework. An important part of INFRAWEB S is a Semantic Web Unit (SWU) – a collaboration platform and interoperable middleware for ontology-based handling and maintaining of SWS. The framework provides knowledge about a specific domain and relies on ontologies to structure and exchange this knowledge to semantic service development modules. INFRAWEB S Designer and Composer are sub-modules of SWU responsible for creating Semantic Web Services using Case-Based Reasoning approach. The Service Access Middleware (SAM) is responsible for building up the communication channels between users and various other modules. It serves as a generic middleware for deployment of Semantic Web Services. This software toolset provides a development framework for creating and maintaining the full-life-cycle of Semantic Web Services with specific application support.

Keywords: Semantic Web Services, Fuzzy Set, Ontologies, Case-Based Reasoning

ACM Classification Keywords: H.3.4 Systems and Software: Information networks

Introduction

The Infraweb s open platform and framework is divided into 3 layers:

- a knowledge management layer
- a service development layer
- a service deployment layer

The project's current relation to the state of the art in services engineering and semantic web service area is characterised by a principally innovative approach: based on a bottom-up approach. The project tries to comprehensively design and establish a software tool set as well as a modelling framework covering the full life cycle of semantic web services. The innovative feature is given in the degree of comprehending this life cycle by embedding the cycle components in closed loop structures.

It starts from the "bottom" (knowledge management layer) by providing contemporary as well as future-oriented knowledge management tools, semantic information routing, and enrichment facilities for knowledge objects, which represents the conventional service generation and handling process. This semantic based knowledge management layer is grounded to existing well defined standards like WSDL, SOAP, UDDI and acts as well structured connection to the existing web service related world.

The semantically enriched knowledge artefacts are "shifted to" and "accessed by" a service development layer (Semantic Web Service Unit -SWU) via the novel SPARQL RDF query language. This layer provides tools for creating and composing of Semantic Web Services embedded in a semantic based interoperable middleware, consisting of Semantic Web Service Designer & Composer, Distributed Semantic Web Service Registries, and discovery modules. The INFRAWEB S specific Semantic Web Services are WSMO [WSMO] compliant to parallel European research efforts.

As the top of the overall structure the service deployment layer provides tools for the execution and monitoring of Semantic Web Services. Extracting execution and monitoring information (Quality of Service Brokering) and

feeding back this information to the underlying layers guarantees a stable bottom-up to top-down cycle, which inherently optimises itself.

Generated in this way, the open platform (Fig. 1) consists of coupled and linked INFRAWEBs units, whereby each unit provides tools and system components to analyse, design and maintain WEB-Services realised as Semantic-Web-Services within the whole life cycle.

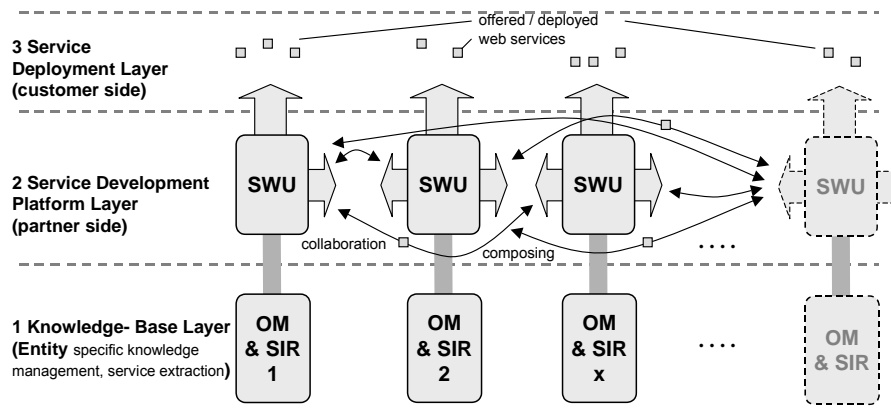


Fig. 1: Open and extensible development platform for the design, deployment and maintenance of Semantic Web Services as a net of coupled INFRAWEBs units.

As illustrated in Fig.1 the overall design is structured in three main layers:

- 1) a knowledge management layer for handling service related knowledge artefacts realised as an organisational memory coupled to semantic information routing components (OM&SIR),
- 2) a service development layer for creating and maintaining Semantic Web Services embedded in a semantic based interoperable middleware, consisting of Semantic Web Service Designer & Composer, Distributed Semantic Web Service Registries, and an agent based discovery module (Semantic Web Service Unit -SWU)
- 3) a service deployment layer for the execution and monitoring of Semantic Web Services exploiting closed loop feedback information (Quality of Service brokering) provided for distributed decision support issues.

The software tool-set building components map the specific modules of the INFRAWEBs framework. With regard to the SWS technology the position of the project is to leave the conceptual level towards practical and reasonable applicable software tools and components. In the rest of the paper the knowledge management and service development layers are considered in detail.

Knowledge Management Layer

Concerning the knowledge management layer, the base module for the organizational memory (OM) is specified and designed as a Fuzzy Concept Matching OM (FCM-OM). It acts as a repository for semi-structured knowledge artefacts (knowledge content objects).

The OM management module is endowed with tools for knowledge acquisition and knowledge representation [Nern, 2005a]. This module is responsible for the collection, organisation, refinement, and distribution of knowledge objects handled and managed by the service providers. The current specification and realisation of the INFRAWEBs OM is based on the novel results of research activities in the area of Fuzzy Set theory:

Considering the ambiguity, imprecision of concepts (electronic knowledge objects, knowledge content objects of an entity, including the objects handled and received in Internet related environments) a useful approach is the adaptation and application of FCM (Fuzzy Conceptual Matching) methods [Zadeh, 2002]. Within this approach a

“concept” is defined (and represented) by a sequence (a set) of weighted keywords. Ambiguity in these concepts is defined by a set of imprecise concepts. Each imprecise concept is defined as a set of fuzzy concepts (using methods of implicit semantics), which is related to “a set of imprecise terms representing the context” [Zadeh, 2002]. Involving and considering also formal semantics, these imprecise terms (words) are “translated” into precise terms (words) formalised as an ontology.

Within the INFRAWEBBS project two streams are focused in realising this system component:

- one component of the OM (FCM-OM) is designed following the rules of implicit & soft semantics (using statistical based AI methods like clustering and classification)
- a second component (O-OM) reflects the Knowledge handling based on methods related to formal semantics (Ontologies) – hard semantics [Zadeh, 2002].

The FCM-OM is coupled to the SIR – a Semantic Information Router [Westkaemper, 2005], that is a further module within the first platform layer.

Semantic and non-semantic components of INFRAWEBBS are interconnected by SIR, which is responsible for:

- Locating all resources needed for problem solving either in the local SWU or outside.
- Creation of non-semantic content (knowledge objects) by means of semantic content stored in the Distributed SWS Repository.
- Creating an effective system of indexes allowing fast communication between semantic and non-semantic modules of SWU.
- INFRAWEBBS proposes a bottom-up approach to the problem of converting regular Web services to semantic ones. Based on some initial description of a Web service to be converted, SIR finds the appropriated WSDL files corresponding to the query and extracts metadata information about the service (in Dublin Core metadata standard) [Dublin Core] from the UDDI (if the service is registered in it). Such metadata is further used by the Service Modeller (a sub-component of the INFRAWEBBS CBR Service Designer tool) for filling in non-functional properties of the constructed semantic Web service according to the WSMO Framework.
- The interface between the Semantic Information Router (SIR) and the OM is based on the SOAP protocol. The interface is used to provide WSDL based service descriptions to the OM component and to query for content item references that are related to given service description references.
- The query interface of the SIR component is based on the SPARQL which is a RDF metadata query language designed by the W3C DAWG working group with protocol bindings defined for JDBC, HTTP and other protocol stacks. Technically the SPARQL interface is a metadata query language with syntax close to SQL which offers querying metadata as table and graph result sets. The protocol stack most suitable for J2EE environments is a JDBC type 4 (Java Database Connectivity) compliant drivers. The implementation of a JDBC based protocol binding for SPARQL is named SPARQL4j (<http://sourceforge.net/projects/sparql4j>).
- Using this RDF query language the connection to the service development layer is performed. Registration of non-semantic atomic services is accomplished via a web-based GUI interface. This interface is mainly used to input WSDL and BPEL4WS based service descriptions and enter additional related non-functional properties. Additionally a UDDI registry interface is provided for the SIR component. This will be established via SOAP as a UDDI Subscription Listener.

Service Development Layer - Semantic Web Unit (SWU)

The main module within the second – the service development – layer is the Semantic Web Unit (SWU) [Atanasova, 2005], [Nern, 2005b], [Agre, 2005]. SWU provides knowledge about a specific domain and relies on ontologies to structure and exchange this knowledge.

The following challenges for developing of SWU have to be taken into account:

- Converting Web Services from available descriptions and domain knowledge (organizational memory) to the semantic ones;
- Composition of Web Services, combining and orchestrating them in order to deliver added-value services;

- Dependencies that arise when a service integrates with external services and becomes dependent on them;
- Integration Processes of several business partners situated on different locations that have to be integrated with each other.

SWU is embedded in INFRAWEBs Environment, responsible for communicating with different users, agents and other SWUs.

SWS Unit ensures designing SWS from the domain knowledge. All knowledge objects from the organizational phenomena influent on the constructed semantic web service and consist of WSDL, BPEL4WS and UDDI files as web service descriptions together with WSML and metadata as ontologies and non-functional properties carriers.

One promising solution to SWS design is to define a library of reusable aspects that would allow the service developer to dynamically instantiate and configure all the needed aspects to deal with different SWS parts. These reusable aspects can be seen as generic templates that can be customized and integrated "on demand" to accommodate to service requirements. This consideration leads naturally to using Case-Based Reasoning approach for service development.

Within the SWU the Designer and Composer modules are responsible for decision supported creation of Semantic Web Services using the Case-Based Reasoning approach. The Designer is a tool for semiautomatic conversion of non-semantic Web services to Semantic Web Services, whereas the Composer enables the semiautomatic creation of new Semantic Web Services via the composition of existing Semantic Web Services.

The architecture of both modules is based on such general principles as:

- Specialization: Each tool is carefully designed based on analysis of specificity of the task it is intended to be used for. It leads to minimization of efforts the service provider should apply for creating a semantic web service. Such minimization is achieved via fully utilization of all available information resources about the service as well as CBR-based mechanism for improving the behaviour of a tool through accumulating and using experience of the service provider to work with this tool.
- User-friendliness: it is assumed that the users of our tools will be semantic Web service providers as well as customers of such services. In both cases the users will not be specialists in first-order logic that is why we implement a self-explained graphical way for constricting and editing of all elements of a semantic web service.
- Intensive use of ontologies: ontologies are the core concept of the Semantic Web technology; however, we consider that creating ontologies for different application domains requires very intensive cooperation of highly qualified domain knowledge engineers and logicians. Both categories of the users do not belong to the range of potential customers of our tool. That is why we assume that our customer will be mainly a user of already created ontologies rather than a creator of new ontologies. However, we foresee that in some cases the service providers have to be able to create some specialized versions of (general) existing ontologies. Means for creating such (restricted) ontologies are also included in our tools.
- Semantic consistency: operation with each tool is organized via ontology-based system-driven interaction with the service creator, which prevents him/her from possible errors and allows being concentrated on the relevant part of knowledge to be acquired. Application of context-sensitive syntactical and completeness checks at each step of the semantic service creation prevents the user from constructing semantically inconsistent and incomplete models.

Designer With the Case-based Designer SWS a service provider creates semantic descriptions of the services on the base of set of ontologies, preferences (QoS) and business logic of services using service design templates (DST).

The Designer consists of several sub-modules responsible for WSMO compliant creation of main elements of INFRAWEBs specific Semantic Web Service.

SWS-Designer has to add the semantic meaning to Web Services about: Data, Functioning, Execution, Discovery, and Selection. This can be done by the following modules: Capabilities editor, Interface editor, Grounding editor via using of DST with appropriate validation and indexing. Creating, storing, and retrieving of similar DST are organized using Case-based Reasoning (CBR) approach. A retrieved template can be further

used or adapted by the user for designing the desired functional model of a new semantic Web service and/or to be stored in case-based memory for later re-use.

As a graphical user-friendly tool the Capability Editor facilitates construction and editing of complex WSML-based logical expressions used for representing service capabilities. The BPEL4WS-based editor serves for creating WSMO-based service choreography and orchestration as SWS interface. The Grounding Editor provides facilities for semiautomatic creating of WSMO-based grounding on the base of WSDL descriptions.

A first prototype of an Axiom Editor - an ontology-driven user-friendly tool for graphical creating complex WSML logical expression - was developed as a part of INFRAWEB SWS Designer. This module is the main part of the INFRAWEB SWS Designer that is responsible for creating the capability description of a Semantic Web Service according to WSMO framework.

A basic feature of the Designer is the use of Design Service Templates (DST), representing graphical models of capability and functionality (or their parts) of Semantic Web Services, which have been designed by the user in the past.

Composer With the Case-Based Semantic Web Service Composer a service provider constructs SWS semi-automatically in design-time by composing descriptions of existing SWS and using domain knowledge.

The SWS Composer has to resolve two problems during composition: planning of the process (service scenarios) and orchestration of services. It uses the previous service compositions that form the general tasks. Such compositions are represented by Service Composition Templates (SCT).

The SWS Composer provides:

- Similarity-based retrieval of an appropriate semantic service template based on the description of capability of the desired service and description of its functionality
- Semi-automatic adaptation of service functional model based on the results of discovery of sub-services matching the template proxies
- Advertising the created service and its generalization and storing as new template for later re-use.

The Case Base of SCT consists of references to complex service scenarios constructed by SWS Composer in the past and associations between problem solutions (particular description of request for servicing made in the past) and founded solutions.

The Composer presents an interactive approach for composition of WSMO compliant Semantic Web Services. The SCT represents graphical models of the service composition as a control and a data flow between several semantic sub-services given by incomplete description of their capabilities. On the (Service) provider side the Composer enables the creation of a new composed "static" Semantic Web Service by discovering appropriate Semantic Services matching the required capabilities. Selecting of such services is implemented as an interactive system-driven semiautomatic process.

Eclipse RCP (Rich Client Platform) is used for developing the basic platform components; plug-in infrastructure and graphical user interface components, whereas Eclipse GEF (Graphical Environment Framework) is the basis for implementing the graphical editors. Access to WSMO-based repositories (ontologies, Semantic Web Services, etc.) is realized via the WSMO API.

WSMO API - WSMO4J For ensuring compatibility and interoperability within and between the INFRAWEB SWS framework modules the WSMO API (WSMO4J) is applied. The WSMO4J is an open-source project (distributed under a LGPL licence) with two parts: a) WSMO API - application programming interfaces for WSMO, which allow basic manipulation of WSMO descriptions, e.g. creation, exploration, storage, retrieval, parsing, and serialization and b) WSMO4J - a reference implementation of the WSMO API, including a WSML parser.

One of the major advantages of using the WSMO API in INFRAWEB SWS is to assure the compatibility and interoperability between the SWS Designer and Composer modules, and the repository component. The distributed SWS registry uses WSMO4J in the process of transforming WSMO element descriptions into RDF triples stored into an RDF triple repository for efficient query and management. Using WSMO4J enables easy integration and interoperability within the framework as well as with the WSMO Studio, thus some of the components can be realized as extensions (plug-ins) for the WSMO Studio.

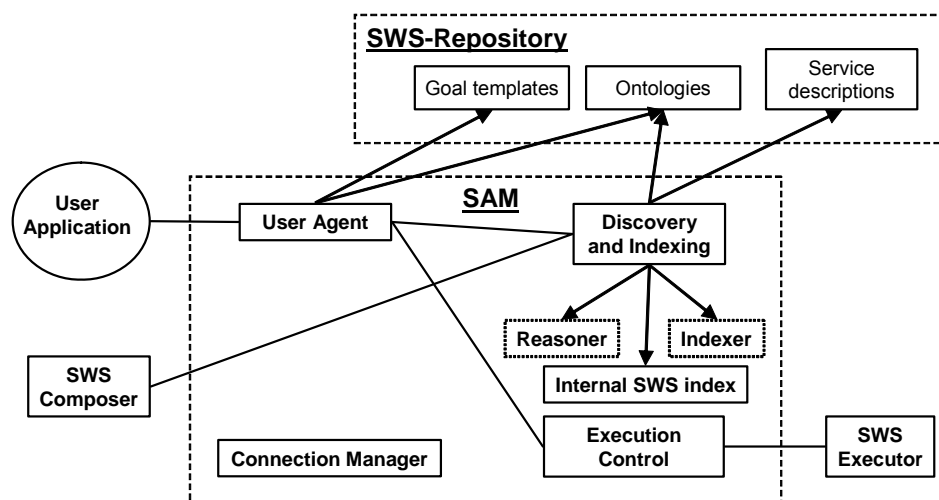


Fig. 2: Internal Architecture of Middleware Layer and its Dependencies with Other Modules

SWS Discovery - SAM The SAM (Service Access Middleware) [Kovacs, 2005] unit is a basic layer of the INFRAWEBS software environment (Fig. 2), guiding the user applications through the steps of semantic web service discovery, selection and execution. The Connection Manager is responsible for building up the communication channels between users and various other modules. The User Application circle denotes the application acting on behalf of the user: for example a GUI interface or an intelligent agent. The User circle and the SWS Composer need to discover and select existing web services with specified capabilities. The Discovery component supports this task, while the Execution Control component oversees the execution of the selected web service in cooperation with the SWS Executor module.

The discovery process is planned as a hybrid approach combining text processing and reasoning. Therefore, this process needs other external information and tools as well: domain ontologies contain the background information about service domains and organizational memories gathered from the INFRAWEBS framework and transformed into Semantic Web compatible format.

The SAM unit within INFRAWEBS acts as a middleware layer for user applications, connecting these applications to the functionality available in the form of Semantic Web Services. An important decision is whether to hide the semantic approach and accompanying logic-based framework of SWS from the user applications or not. A present investigation is to clarify: What level to choose for communication? On the level of plain web services, simple XML data structures are exchanged, which are easy to generate and process, but lack the possibilities of the semantic approach. On the level of semantic web services, facts and rules are exchanged in the form of logical expressions. As the latter case puts extra requirements for the user application, it is decided to find a middle course between the two solutions. Communication is kept on the semantic level, but its form is either hidden or aided with special functionalities and automatic translations.

Apart to SWS discovery the simplest way of discovery is based on keyword matching within the descriptive metadata of web services (similar to the UDDI approach). The most complex method is to logically prove that the web service is able to fulfil the given goal. Currently, response time of discovery process is a significant trade-off for these approaches.

The keyword-based method is improved if the goals and capabilities are used for keyword generation instead of metadata. Goals and capabilities are described using ontology terms, so more homogeneous and precise keywords can be extracted this way. The logic-based methods differ in aspects of goals and capabilities considered. It is simpler and faster to match only by post-conditions, but then matching services might not be executable because of missing or unacceptable information (input).

The project approach is to split the discovery into two phases. In the first phase, the keywords of capabilities are used to filter the possible web services. Then, logical matching is applied only for the filtered services.

The list of matching web services is returned to the user application enriched with descriptive and qualifying metadata.

A Web service is the access point to the real service: its quality determines the quality of access, but might be independent of the quality of service (e.g. automatic translation or flight reservation). Therefore, an iterative selection process (similarly to iterative query refinement in information retrieval) guides the selection of the best service in INFRAWEBs. This is based on a simple two-phase workflow agreement (or business logic) between the web services and the middleware layer.

The Service Access Middleware (SAM) provides support for the usual steps of goal construction, discovery, selection and execution of Semantic Web Services. This support is achieved through a neutral interface, which hides the complexities of Semantic Web Services, therefore applications can be easily adapted to it, and also it can be equivalently used in variants of Semantic Web Services, such as WSMO and OWL-S. SAM also features an iterative selection refinement process for finding not only the suitable web services, but also the best service offers for users' goals.

SWS deployment layer The SWS deployment layer consist of SWS executor that is split up into three main components, namely, the Communication Manager, Choreography Engine and the Invoker [Polleres, 2005] and QoS broker. At present it is defined that the executor should mainly interact with the distributed registry and the SAM components.

Conclusion

The primary objective of the INFRAWEBs project is to develop an ICT framework consisting of several specific software tools, which enables software and service providers to generate and establish open and extensible development platforms for Semantic Web Service based applications [Nern, 2004]. This software tool set facilitates the establishment of virtual development platforms as well as interoperable middleware designed for a semantic and ontology-based handling of Semantic Web Services oriented on given conception WSMO specifications.

One of the goals of the INFRAWEBs project is the development of a SWS full-life-cycle software toolset for creating and maintaining Semantic Web Services with specific application support. An important part of INFRAWEBs is a Semantic Web Unit (SWU) – a collaboration platform and interoperable middleware for ontology-based handling and maintaining of SWS. The SWU provides knowledge about a specific domain and relies on ontologies to structure and exchange this knowledge.

INFRAWEBs Designer and Composer are sub-modules of SWU responsible for creating Semantic Web Services using Case-Based Reasoning approach to fulfil decision support demands.

The architecture of both modules is based on such general principles as:

- service-oriented architecture with bottom-up approach for semi-automatic constructing of semantic web services;
- system driven syntactic consistent and completeness checking;
- past experience utilizing.

The SAM unit within INFRAWEBs acts as a middleware layer for user applications, connecting these applications to the functionality available in the form of Semantic Web Services. These software toolsets are developing for creating and maintaining of full-life-cycle Semantic Web Services with specific application support.

Bibliography

- [Nern, 2004] H Joachim Nern, G. Agre, T. Atanasova, J. Saarela. System Framework for Generating Open Development Platforms for Web-Service Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems - INFRAWEBs II. In: WSEAS TRANS. on INFORMATION SCIENCE and APPLICATIONS, 1, Vol. 1, 286-291, 2004.
- [WSMO] Web Services Modelling Ontology, SDK WSMO working group, <http://www.wsmo.org>
- [Nern, 2005a] H Joachim Nern, A Dziech, E Tacheva, Fuzzy Concept Sets (FCS) applied to semantic organizational memories within the Semantic Web Service designing and composing cycle, In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005.
- [Zadeh, 2002] Lotfi A. Zadeh. Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. In: Journal of Statistical Planning and Inference 105, 233-264, 2002.
- [Westkaemper, 2005] T Westkaemper, J Saarela, H Joachim Nern, Semantic Information routing as a pre-process for Semantic Web Service generation - SIR & OM, In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005.
- [Dublin Core] The Dublin Core Metadata Initiative, <http://dublincore.org/>
- [Atanasova, 2005] Tatiana Atanasova, Gennady Agre, H Joachim Nern, "INFRAWEBs Semantic Web Unit for design and composition of Semantic Web Services INFRAWEBs approach", In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005.
- [Nern, 2005b] H Joachim Nern, A Dziech, T Atanasova, Applying Clustering and Classification Methods to distributed Decision Making in Semantic Web Services Maintaining and Designing Cycles, EUROMEDIA 2005, In: Proc. Workshop for "Semantic Web Applications", IRIT, Université Paul Sabatier, Toulouse, France, April 11-13, 2005.
- [Agre, 2005] Gennady Agre, Tatiana Atanasova, H Joachim Nern, "Case Based Designer and Composer", In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005.
- [Kovacs, 2005] L Kovacs, A Micsik, "The SUA-Architecture within the Semantic Web Service Discovery and selection process", In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005.
- [Polleres, 2005] A Polleres, J Scicluna, "Semantic Web Execution for WSMO based choreographies", In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005
-

Authors' Information

Joachim Nern – Scientific coordinator of INFRAWEBs project; big7.net GmbH & Aspasia Knowledge Systems Germany, e-mail: nern@aspasia-systems.de

Gennady Agre – Institute of Information Technologies, Acad. G. Bonchev 29-A, 1113 Sofia, Bulgaria, e-mail: agre@iinf.bas.bg

Tatiana Atanasova – Institute of Information Technologies, Acad. G. Bonchev 2, 1113 Sofia, Bulgaria, e-mail: atanasova@iinf.bas.bg

Andras Micsik – MTA SZTAKI, H-1111 Budapest XI. Lagymanyosi u. 11, Hungary, e-mail: micsik@sztaki.hu

Laszlo Kovacs – MTA SZTAKI, H-1111 Budapest XI. Lagymanyosi u. 11, Hungary, e-mail: Laszlo.kovacs@sztaki.hu

Zlatina Marinova – SAI, Ontotext Lab, Sirma Group Corp., 135 Tsarigradsko Shosse Blvd., 1784 Sofia, Bulgaria, e-mail: zlaty@sirma.bg

Janne Saarela – Profium Ltd, Lars Sonckin Kaari 12, 02600 Espoo, Finland, e-mail: janne.saarela@profium.com

Timo Westkaemper – Profium Ltd, Lars Sonckin Kaari 12, 02600 Espoo, Finland, e-mail: timo.westkamper@profium.com

INFRAWEB S AXIOM EDITOR – A GRAPHICAL ONTOLOGY-DRIVEN TOOL FOR CREATING COMPLEX LOGICAL EXPRESSIONS

Gennady Agre, Petar Kormushev, Ivan Dilov

Abstract: The current INFRAWEB S European research project aims at developing ICT framework enabling software and service providers to generate and establish open and extensible development platforms for Web Service applications. One of the concrete project objectives is developing a full-life-cycle software toolset for creating and maintaining Semantic Web Services (SWSs) supporting specific applications based on Web Service Modelling Ontology (WSMO) framework. According to WSMO, functional and behavioural descriptions of a SWS may be represented by means of complex logical expressions (axioms). The paper describes a specialized user-friendly tool for constructing and editing such axioms – INFRAWEB S Axiom Editor. After discussing the main design principles of the Editor, its functional architecture is briefly presented. The tool is implemented in Eclipse Graphical Environment Framework and Eclipse Rich Client Platform.

Keywords: Web services, Semantic Web Services, Web Service Modelling Ontology framework.

ACM Classification Keywords: H.5.2 User Interfaces: Graphical user interfaces (GUI)

Introduction

Current Web service technologies describe the syntactical aspects of a Web service providing only a set of rigid services that cannot be adapted to a changing environment without human intervention. Realization of the full potential of the Web services requires further technological advances in the areas of service interoperability, service discovery, service composition and orchestration. A possible solution to these problems is likely to be provided by application of Semantic Web technologies.

Semantic Web Services (SWSs) are self-contained, self-describing, semantically marked-up software resources that can be published, discovered, composed and executed across the Web in a task driven semi-automatic way. There are two major initiatives aiming at developing world-wide standard for the semantic description of Web services – the American OWL-S [OWL-S 2004] and the European WSMO [Roman et al. 2005]. The INFRAWEB S European research project is based on the WSMO framework for service modelling and proposes a next, more technology-oriented step in the process of semantic Web service development [Nern et al. 2004]. One of the concrete project objectives is developing a full-life-cycle software toolset for creating and maintaining SWSs supporting specific applications based on Web Service Modelling Ontology (WSMO) framework.

A main part of WSMO-based SWS is service capability – a declarative description of Web service functionality. A formal syntax and semantics for such a description is provided by Web Service Modelling Language (WSML), which is based on different logical formalisms, namely, Description Logics, First-Order Logic and Logic Programming [de Bruijn et al. 2005]. The conceptual syntax for WSML has a frame-like style. The information about a class and its attributes, a relation and its parameters and an instance and its attribute values is specified in one large syntactic construct, instead of being divided into a number of atomic chunks. It is possible to spread the information about a particular class, relation, instance or axiom over several construct. WSML allows using of variables that may occur in place of concepts, attributes, instances, relation arguments or attribute values. A variable may not, however, replace a WSML keyword. Furthermore, variables may only be used inside logical expressions. A WSML description of a Web service capability is represented as a set of complex logical expressions called axioms. Machines can easily handle these axioms. However, it is very difficult for humans to create and comprehend complex logical expressions. Therefore, the construction of axioms needs to be supported by some easy-to-use graphical tools. It should allow a non-specialist to create highly complex axioms in WSML language through simple graphical interaction.

This paper describes a specialized tool called INFRAWEB S Axiom Editor, which is aimed at constructing and editing WSMO-based SWS capabilities. The structure of the paper is as follows – the next section discusses the

basic design principle of the Editor. Then the models for representing and constructing the axioms are presented. Next two sections are devoted for describing the main functionality of the Editor and its graphical user interface. In conclusion some implementation details and future trends are discussed.

Basic Design Principle of INFRAWEBs Capability Editor

INFRAWEBs Axiom Editor is a specialized user-friendly tool for constructing and editing complex WSMML logical expressions based on available set of WSMML ontologies. It is a core part of a more complex tool – INFRAWEBs Semantic Service Designer, which is aimed at converting existing Web services to WSMO-based semantic Web Services [Agre et al. 2005].

The main design principles of the Axiom Editor are:

1. *Specialization*: the tool is intended to be used mainly for constructing logical expressions representing capabilities of WSMO-based semantic Web services rather than axioms on ontologies. Our analysis has shown that the logical structure of such expressions is rather simple and in most cases does not require using of such complex WSMML logical operators (connectives) as *Implies*, *ImpliedBy* or *Equivalent*.
2. *User-friendliness*: it is assumed that the users of our tool will be semantic Web service providers as well as customers of such services. In both cases the users will not be specialists in first-order logic, so using of some (even rather advanced) text editor for constructing logical expressions seems for us an inappropriate solution. That is why we propose a graphical way for constructing and editing the axioms abstracting away as much as possible from a concrete syntax of logical language used for implementing them.
3. *Intensive use of ontologies*: it is well known that the core concept of the Semantic Web is ontologies – “formal, explicit specification of a shared conceptualization” [Gruber 1993]. In our opinion, creating such formal and consensual specifications for different application domains requires very intensive cooperation of highly qualified domain knowledge engineers and logicians. Both categories of the users do not belong to the range of potential customers of our tool; for such users are more appropriate such general ontology editors like *Protégé2000* [Protégé 2005] or *Ontology Management Suit* which is currently under development in the frame of WSMO project. So we assume that our customer will be mainly a *user* of already created ontologies rather than creator of new ontologies. However, we foresee that in some cases the service providers need to be able to create some *specialized versions* of (general) existing ontologies containing specific instances or subconcepts of general ontology concepts. Means for creating such (restricted) ontologies are going to be included in our Editor.
4. *Semantic consistency*: our analysis has shown that the main difficulties of the process of constructing complex logical expressions are associated with use of correct names of concepts, attributes, relations and parameters as well as their types rather than with expressing logic itself. That is why the process of constructing logical expression in INFRAWEBs Axiom Editor is *ontology-driven*, which means that in each step of this process the user may select only such elements of existing ontologies that are consistent with already constructed part of the axiom. From this point of view the created axiom is always semantically consistent with ontologies used for its construction.

Representation of Axioms

According to the formulated above requirements the Axiom Editor should allow automatic generating correct WSMML logical expressions from some graphical representation (model) of such expressions. As a graphical model of WSMML axiom we have selected a direct acyclic graph (DAG). Such a graph can contain four kinds of nodes:

- A single node called *Root*, which may have only outgoing arcs. This node corresponds to WSMML statement *defineBy*. Graphically the root node is represented as a circle named “Start”.
- Intermediate nodes called *variables*. Such nodes have one or more incoming arcs and can have several outgoing arcs. Each variable has a unique name and a frame-like structure consisting of slots represented by pairs attribute – attribute value (WSMML variable). Such a variable corresponds to a notion of compound molecule in WSMML [de Bruijn 2005] consisting of an *a*-molecule of type Var_i memberOf Γ and conjunction of

b-molecules of type $Var_i[p_1 \text{ hasValue } Var_{j_1}]$ and $Var_i[p_k \text{ hasValue } Var_{k_l}]$ respectively, where $Var_i, Var_{j_1}, Var_{k_l}$ are WSML variables and Γ is a concept from a given WSML ontology. Graphically each variable is represented as a rectangle with a header containing variable name and type (i.e. the name of concept, which has been used for crating the variable), and a row of named slots.

- Intermediate nodes called *relations*. Such a node corresponds to a WSML statment $r(Par_1, \dots, Par_n)$, where r is a relation from a given ontology, and Par_1, \dots, Par_n are WSML variables – relation parameters. Graphically each relation node is represented as a rectangle with a header containing relation name and a row of relation parameters.
- Intermediate nodes called *operators* that correspond to WSML logical operators *AND*, *OR* and *NOT*. Each node can have only one incoming arcs and can have one (for *NOT*) or several (two or more – for *AND* and *OR*) outgoing arcs. Graphically each operator is represented as an oval, containing the name of the corresponding operation.
- Terminal nodes (leaves) that can not have any outgoing arcs. Such terminal nodes are called *instances*. Each instance corresponds to the WSML statement $Var \text{ hasValue } Instance$, where Var is a WSML variable and $Instance$ is an instance of a concept from a given ontology. Graphically an instance is represented by a rectangle with header containing the name of concept, an instance of which the Instance is, and the concrete name of the instance.

Directed arcs of a graph are called *connections*. A connection outgoing from a variable or relation has the meaning of refining the variable (or relation parameter) value and corresponds to WSML logical operator *AND*. A connection outgoing from an operator has the meaning of a pointer to the operator operand.

The proposed model allows to consider the process of axiom creation as a formal process of DAG expanding (and editing) and to formulate formal rules for checking syntactic and semantic (in relation to given ontologies) correctness of constructed axioms.

An advantage of the proposed model is ability to separate logical *AND* (represented as the model *AND* operator) used by the axiom creator for describing logical conjunction at a high level of abstraction from a “hidden”, “technical” *AND* (represented by the model connection) used for specifying more concrete values of variable attributes. As a result, the explicit logic conjunction may be used in the model only as a part of a path starting from the axiom root and ending in an intermediate variable node or in a terminal node. This has a very important consequence for the semantic service discovery process. First, if a represented in such a way axiom is interpreted as a user goal (i.e. a request for desired service functionality), the proposed mechanism gives a very simple method for splitting the goal to sub-goals. And second, if such an axiom is interpreted for example as a service post-condition, the proposed mechanism allows easily determining if the service offers a single functionality of a set of different functionalities.

An Informal Model of the Axiom Construction Process

A process of axiom creation may be considered as a repetitive process consisting of combination of three main logical steps – definition, refinement (or specialization) and logical development (or elaboration). The *definition* step is used for defining some general concepts needed for describing the meaning of axioms. The *refinement* step is used for more concrete specification of desired properties of such concepts. Such a step may be seen as specialization of too general concepts introduced earlier. The *logical development* step consists of elaborating logical structure of the axioms, which may be achieved by combination of general concepts by means of logical operators *AND*, *OR* and *NOT*.

Syntactic and semantic checks applied during the all phases of axiom creation process are based on the following properties:

- Subsumption relation between different elements of ontologies: such a relation determines compatibility between axiom variables;
- Acyclic property of the selected model (DAG) for representing an axiom;
- Uniqueness of the names of variables used for constructing an axiom (if contrary is not explicitly specified);
- Arity of logical operators used for constructing an axiom.

Definition Step

During the definition step the nature of a main variable defining the axiom is specified. Such a step is equivalent to creating a WSML statement *?Concept memberOf Concept*, which means that the WSML variable *?Concept* copying the structure of the *Concept* from a given WSML ontology is created. Attributes of the concept, which are "inherited" by the axiom model variable, are named *variable attributes*. By default the values of such attributes are set to free WSML variables with type defined by the definition of such attributes in the corresponding ontology.

It should be mentioned that in the definition step every concept, instance or relation from an arbitrary WSML ontology may be used as a template for creating the corresponding axiom variable.

Refinement Step

The refinement step is a recursive procedure of refining values of some attributes (relation parameters) defined in previous step(s). In terms of our model each cycle in such a step means an expansion of an existing non-terminal node – variable (or relation). More precisely that means a selection of an attribute from a list of available attributes of an existing axiom variable, and binding its value (which in this moment is a free WSML variable) to another (new or existing) node of the axiom model. The main problem is to ensure semantic correctness of the resulted (extended) logical expression. Such correctness is achieved by applying explicit rules determining permitted expansion of a given node.

An attribute value¹ of an axiom variable may be refined by binding it to:

- A. A new variable produced from the ontology concept specified by *ofType* or *impliesType* WSML statement for the corresponding attribute (default binding);
- B. A new variable produced from a subconcept of the ontology concept specified by *ofType* or *impliesType* WSML statement for the corresponding attribute;
- C. A new terminal node – instance produced from an instance of the corresponding concept or of its subconcepts;
- D. A relation which parameters are compatible with the type of the selected attribute;
- E. An existing axiom variable, which are compatible with the type of the selected attribute and which does not lead to creation of cycles in the model.
- F. A shared variable with compatible type.
- G. A complex logical expression composed from all mentioned above items by logical operators OR and NOT.

Logical Development Step

This step of the axiom construction process consists in adding logical operations (*AND*, *OR* and *NOT*) to the current logical expression. Such operators may be added to connect two independently constructed logical expressions or be inserted directly into already constructed expressions. In both cases it leads to creating more complex logical expressions.

A logical operator can be inserted only into a connection that has been already created as a part of the axiom model. Such an insertion "splits" the connection on two parts, which are linked by newly inserted logical operation. Since operators *AND* and *OR* should have at least two operands, the addition of such logical operators requires creating the second operand, which can be either a new or an existing axiom element. The operation is controlled by context-dependent semantics and syntactic checks so different logical operators can be inserted only in some allowed places in the axiom. Such checks analyze the whole context of the axiom, which in some cases leads for necessity to verify the path from the edited element till the starting axiom element – the axiom *Root*.

It should be underlined that during this step the user is constructing the axiom by logical combination of main axiom objects defined in the previous steps. In other words, the logical operators are used not for refining or clarifying the meaning of some parameters of already defined objects, but for complicating the axiom by specifying the logical connections between some axiom parts which are independent in their meaning.

¹ The same rules are applicable to every unbound relation parameter.

Functional Architecture

The functional architecture of the Axiom Editor provides a complete set of functions (operations) needed for graphical constructing WSML logical expressions. The top-level functional components of the Editor are:

- *Ontology Store* – a set of operations for maintaining ontologies used for creating and editing axioms.
- *Axiom Model Generator* – a set of operations for graphical constructing and editing an axiom.
- *Axiom Text Generator* – the module providing automatic generation of the WSML text corresponding to the current graphical model of an axiom.
- *Axiom Persistence* – the module providing saving and retrieving axioms as well as all information needed for axiom creation.

Ontology Store

The Ontology Store is an *in-memory* set of ontologies providing the semantic elements for constructing axioms. These elements are concepts, attributes, instances, relations and parameters¹. The Ontology Store is global to all axioms opened in the Editor.

In order to be used in the Axiom Editor ontologies should be defined in the WSML language. To start creating a new axiom at least one ontology is needed. The Axiom Editor reads ontologies from *.wsml files. The parsing of these files is done by a standard WSML parser which is a part of the WSMO4J API [WSMO4J 2005].

A *tree structure* is used for graphical representation of ontologies. Since *.wsml files are flat (they have no hierarchical structure), additional information is obtained from the WSMO4J API to construct a tree from the lists of concepts, relations etc. The API provides information about concept and relation inheritance by a special *SuperConcepts* property that every ontology element possesses. It should be noted that this property is a set, which means that one element can have more than one parent in the hierarchy. In tree-structured visualization every child element appears as many times in the tree as there are concepts in its *SuperConcepts* property. A visualized ontology may be browsed and all properties associated with each ontology element are shown in a special window.

Ontologies may be loaded manually by the user from the file system or loaded automatically on-demand. Ontologies describe inheritance between concepts. A concept usually has one or more super-concepts. Super-concepts may be defined in other ontologies. For example the concept "Person", defined in the ontology "Sociology", may have the concept "Human", defined in the ontology "Biology" as its super-concept. In such a case, the "Sociology" ontology declares "Biology" as an *imported ontology*. The "load imported ontology" operation can be applied to such concepts displayed in the Ontology View which are defined in imported ontologies. Since an imported ontology is declared with its identifier, the URI is used to locate that ontology and load it to the Ontology Store. The concept is automatically located in the new tree, the concept's attributes become available so variables of that type can be now created.

A concept inherits all its super-concepts' attributes. If a super-concept is defined in an imported ontology, which is not currently loaded to the Ontology Store, then the super-concept's attributes are unavailable. The mechanism for on-demand loading of imported ontologies provides automatic updating concepts' attributes inherited from super-concepts belonging to such ontologies.

Axiom Model Generator

As it has been already mentioned, the main concern of the Axiom Editor is to *guarantee the semantic consistence* of the constructed logical expressions since the users of this tool are assumed to be non-specialists in the first-order logic. Such a consistence is achieved by a semantically-aware construction process, in each step of which the user is allowed to perform only such operations that are consistent with the already constructed part of the axiom.

¹ Functions are not supported in the current implementation of the Capability Editor. Such elements of a WSML ontology as non-functional properties and ontology axioms are shown in the Editor but currently are not used in the process of axiom constructing.

Two modes for axiom construction are available:

- *Standard mode* involves only extending an existing part of the axiom by selecting semantically compatible elements from context-sensitive menus. This method is construction-driven and is suitable for novice users.
- *Advanced mode* allows adding isolated elements to the modelling area, which can be later combined in various semantically correct ways. This allows advanced users to be more efficient.

The axiom construction process begins by selecting a concept from Ontology Store. This concept is used to create the first variable in the axiom model. The variable's type is equal to the selected concept. Automatically, just after adding the first variable to the model, it is connected to the Axiom root element "Start".

From this moment on, the construction process continues by performing semantically-correct operations on different elements in the axiom model which can be: variables, variable attributes, instances, connections, operators, relations and relation parameters. A summary of the most important semantically-correct operations in Axiom Model Generator are shown in Table 1.

Operations for creating elements of Axiom Model	
<i>Create a variable</i>	Creates a new variable in the graphical axiom modelling area (window). The type of the variable is selected by the user from Ontology Store. The name of the variable is automatically generated from the name of the selected concept guaranteeing the uniqueness of variable names across the axiom.
<i>Create an operator</i>	Creates a new logical operator of a specified type in the modelling area. The operator's type is selected from the menu – it can be <i>OR</i> , <i>AND</i> or <i>NOT</i> .
<i>Create an instance</i>	Adds an instance to the graphical modelling area. The user is given the opportunity to select the instance from Ontology Store.
<i>Create a connection (advanced mode)</i>	Creates a new connection between two elements placed on the modelling area. The user selects a source and a target element for the new connection. The selection is restricted only to semantically-compatible source and target elements.
<i>Create a relation</i>	Adds a relation to the modelling area. The user is given the opportunity to select an arbitrary relation from Ontology Store.
Operations on Variables	
<i>Rename a variable</i>	The user can change the automatically generated variable name as long as the uniqueness of names is not violated. The Axiom Editor takes care of changing the variable's name from the old one to the new at all its occurrences in the model.
<i>Involve a variable in a relation (Advanced mode)</i>	A variable that has been already placed at the axiom modelling area may be further involved in a relation also presented at this area. More exactly, such an operation creates a connection linking the variable with a parameter of the relation. Operation is possible only when the variable and the selected relation parameter have compatible types.
<i>Delete variable</i>	Deletion of a variable leads to deletion of all incoming and outgoing connections of the selected variable in the model, thus keeping the axiom consistent.
Operations on attributes of a variable	
<i>Refine an attribute by a variable</i>	Creates a new variable at the modelling area and links the selected attribute value to it with a connection. The meaning of the operation is that the value of the attribute is equal to this new variable. The name of the new variable is automatically set equal to the name of the selected attribute value being refined.
<i>Refine an attribute by an instance</i>	Adds to the current axiom a new instance selected from an ontology and links it to the attribute value to be refined by a connection. The user is given the opportunity to select such an instance from a special dialog window containing a subset of instances from the Ontology Store. More exactly, in order to preserve the semantic consistence of the axiom, the selection is limited only to those instances, whose concepts are equal to or are sub-concepts of the concept specified as the type of the chosen attribute.

<i>Refine an attribute by involving into a relation</i>	A value of an attribute of a variable from the current axiom may be further refined by specifying that it is involved in a relation defined either in the Ontology Store or already placed at the modelling area. Selecting the attribute to be refined restricts a set of relations that may be applied to the value of such an attribute – that are all relations, which have parameters with types compatible with the type of that attribute.
---	---

Operations on relations and relation parameters	
<i>Refine a relation parameter</i>	A set of available operations on relation parameters is practically the same as the operations working on values of attribute variables (see "Operations on attribute of a variable").
<i>Delete relation</i>	Deletion of a relation leads to deletion of all its incoming and outgoing connections in the model, thus keeping the axiom consistent.
Operations on operators	
<i>Change operator type</i>	Is used for changing the type of an operator selected from the modelling area.
<i>Delete operator</i>	Potentially leads to creating some orphaned axiom model elements. In order to preserve the semantic consistence of the axiom, such "orphaned elements" are not included in the axiom text generation.
<i>Add operand</i>	Adds a new operand to a selected operator placed at the modelling area and links them by a connection. The new operand can be either an existing model element from the modelling area (variable, relation, instance, etc.) or a new element that can be created by means of already described operations, which the user may select from right-click sub-menu.
Operations on instances	
<i>Edit an instance of WSML built-in data types</i>	Can be performed on such instances of the axiom model which have a WSML built-in data type or a subtype of such type. The value of these instances is entered by the user and can be edited later.
<i>Delete an instance</i>	Leads to deletion of the instance along with all connections incoming to it from the model, thus keeping the axiom consistent.
Operations on Connections	
<i>Insert an alternative</i>	The main meaning of a connection in the axiom model is that the target element of the connection is used as a refinement of its source element. It is natural to allow the user to define an alternative (or several alternatives) for such a refinement. In order to insure that such an operation will be meaningful, it is necessary to restrict its application domain.
<i>Insert an AND operator</i>	Aims at allowing the user to specify explicitly logical conjunction of two axiom model elements and is also used during the "Logical development" phase of the axiom model construction process operator as its second operand.
<i>Insert a NOT operator</i>	Inserts a NOT operator in the middle of any connection.
<i>Reconnect a source/target element (Advanced mode)</i>	Moves the starting/ending point of the connection to another element in the Axiom Model. In order to preserve the semantic consistence of the axiom, the operation can be performed only if the new source/target element is semantically-compatible with the type of the edited connection.

Table 1. The most important semantically-correct operations in the Axiom Model Generator

Axiom Text Generator

The Axiom Text Generator dynamically generates text representation of the graphical axiom model in the human-readable WSM-L-Core syntax. That allows to observe and control (for experience users) the result of each operation accomplished on the axiom model. It should be mentioned that only elements of the modelling area having connections with the root elements of the axiom model (*Start* element) are considered as parts of the current axiom and, hence, are mapped to its WSM-L text representation.

Axiom Persistence

Creating a semantic Web service is a rather complex process, which may need a lot of time, so it is necessary to have a module for storing all intermediate results and supplemented data structures facilitating such a process. The Axiom Persistence is such a module that is used for storing and retrieving axioms created by the Axiom Editor. Since an axiom has no meaning without the ontologies used for its creation, loading an axiom leads to automatic loading of all ontologies associated with it.

Axioms are persisted in binary files which can only be opened by the Axiom Editor. Besides the semantic content, all elements store their graphical coordinates so that the graphical model of an axiom can be fully restored. During the loading process different validations are made. If any of them fails, an error message is displayed and the axiom file is not loaded. For implementation of these operations Java serialization is extensively used.

Currently the Axiom Editor uses a predefined directory called the Ontology File Store in the file system to store *.wsm-l files containing ontologies. Every ontology has a unique identifier, which is a URI written in the *.wsm-l file defining the ontology. When an ontology, whose identifier is known, must be loaded, the Axiom Editor searches the Ontology File Store for that identifier and loads the respective ontology to the Ontology Store.

Graphical User Interface

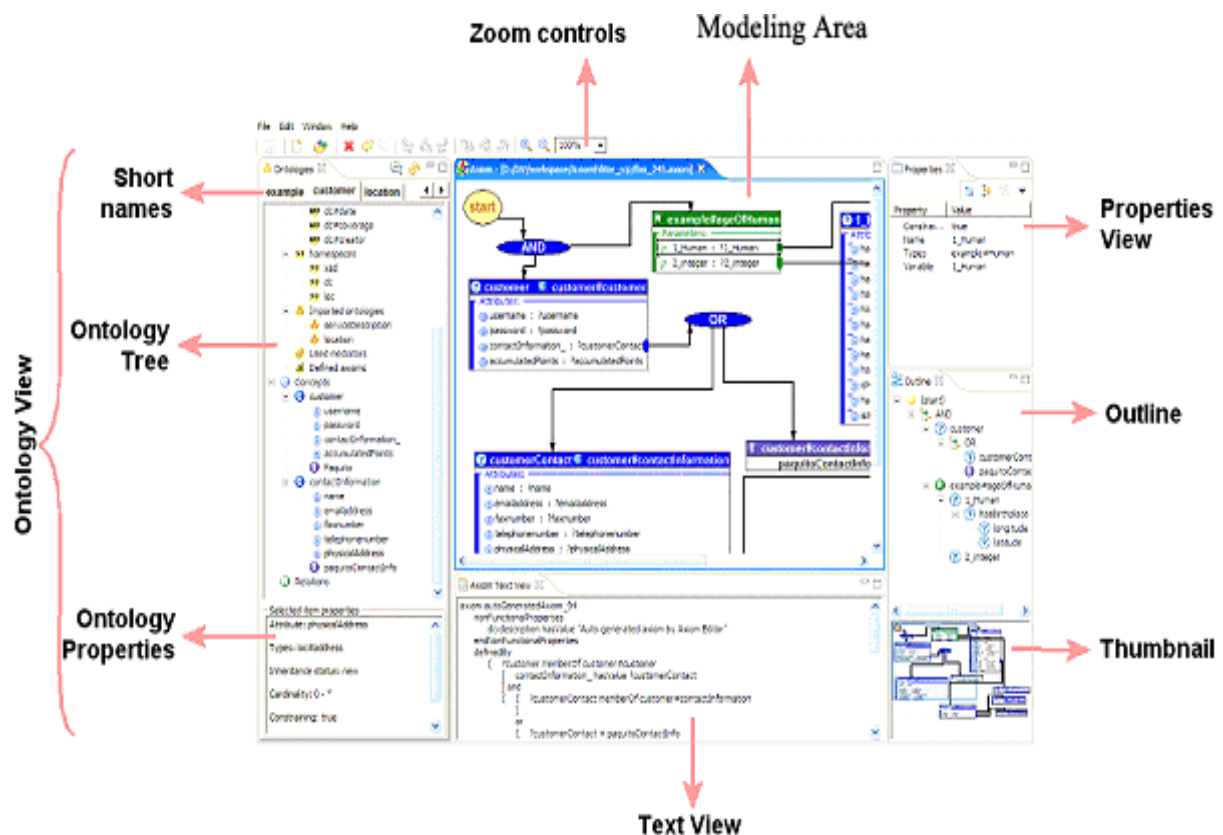


Figure 1. An overview of the Axiom Editor workspace

The Axiom Editor runs as an Eclipse plug-in. Eclipse [Des Rivieres and Wiegand 2004] is a free, integrated development environment (IDE) which can host different third-party applications, providing a unified visual outlook and better integration between them.

The Axiom Editor is bundled as a standalone application on top of the Rich Client Platform (RCP). The RCP is a compact Eclipse core which can also host plug-ins. It provides a startup executable which runs a lightweight version of the IDE and automatically loads the appropriate plug-in (in this case – the Axiom Editor).

An overview of the Axiom Editor workspace is shown on Figure 1. The screen is divided in several major areas: Ontology View, Modelling Area, Properties View, Outline View, Thumbnail + zoom controls and Text View.

You can find a more detailed description of the workspace areas in Table 2.

















Workspace area	Description
Ontology View	Contains all loaded WSMO ontologies (Ontology Store). At the top of the view there is a list of tabs used to switch between different ontologies.
Ontology Tree	The centre part of the ontology view containing all ontology elements, structured in a hierarchy. The nodes represent:  concepts,  attributes,  instances,  relations,  parameters,  relation-instances;  non-functional properties,  namespaces;  imported ontologies,  used mediators;  defined ontology axioms.
Ontology Properties	The bottom part of the Ontology View section. It contains details in plain text about the selected element in the Ontology Tree such as the non-functional properties of a concept, the definition of an ontology axiom etc.
Modelling Area	Contains the graphical representation of the axiom model. The model is displayed as a directed acyclic graph, reflecting the tree structure of the logical expression. This is where the user creates axiom elements out of ontology elements and adds dependencies between them through the use of semantically consistent operations. Axiom elements are:  Variables,  Instances,  Relations,  Logical Operators,  The start element. Dependencies between these elements are introduced through the use of connections displayed as directed arrows.
Properties View	Displays the properties of the selected element in the Modelling Area. Different kinds of elements have different sets of properties – some of them read-only, others - editable.
Outline View	Displays a classical tree representation of the logical expression. The branches of the tree can be expanded or collapsed to help the viewer better perceive the high-level structure of the expression. It also allows for easier navigation among the elements. If an element is selected in the Outline View, it becomes also selected in the Modelling Area and its properties are displayed in the Properties View.
Thumbnail	A mini-map of the whole modelling area. On large models it helps the user to not lose the whole picture, makes navigation easier and always highlights the part of the model being displayed in the Modelling Area.
Zoom Controls	Provide a way of getting a larger part of the model into view by selecting zoom-factor less than 100%. If the user selects a zoom-factor above 100% details can be clearly seen and elements can be more precisely aligned in the Modelling Area.
Text View	Contains the WSMML representation of the axiom. It is automatically refreshed whenever something changes in the graphical axiom model to reflect the current state of the expression. It is useful for advanced users who want to always know the exact impact of their actions on the capabilities of the web-service they are designing.

Table 2. Description of Axiom Editor Workspace areas

Conclusion

The Axiom Editor is implemented in J2SDK 1.4.2 runtime environment and uses basic platform components, plug-in infrastructure, graphical user interface components (menus, buttons, tree views, event handling) from Eclipse RCP (Rich Client Platform). For development of visual designers the Eclipse GEF (Graphical Environment Framework) is used. Access to WSMO-based ontologies is accomplished via *WSMO4J* (*WSMO API*).

Main directions or future development of the Editor are as follows:

- Transformation of the Axiom Editor to an integrated Service Capability Editor by extending it with some customized modules of WSMO Studio [[WSMO Studio 2005] and integrating with the Axiom Case-base Memory.
 - Extending application domain of the Axiom Editor by expanding the range of logical operations used (e.g. including *implies*, *impliedBy*, *:-* and *!* operators). As a result the Editor could be used not only for creating the SWS capabilities but for constructing axioms in WSML ontologies as well.
-

Acknowledgement

This work is carried out under EU Project INFRAWEBBS - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

Bibliography

- [Agre et al. 2005] G. Agre, T. Atanasova, J. Nern. Migrating from Web Services to Semantic Web Services: INFRAWEBBS Approach. In: *Proceeding of Eleventh Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Applications, EUROMEDIS'2005*, April 11-13, 2005, Toulouse, France, 221-225.
- [de Bruijn et al. 2005] J. de Bruijn, H. Lausen, R. Krummenacher, A. Polleres, L. Predoiu, M. Kifer and D. Fensel. *D16.1v0.2 The Web Service Modeling Language WSML, WSML Final Draft 20 March 2005*, available at: <http://www.wsmo.org/TR/d16/d16.1/v0.2/20050320/>.
- [Des Rivieres and Wiegand 2004] J. Des Rivieres and J. Wiegand. Eclipse: A platform for integrating development tools. *IBM Systems Journal*, 43(2), 2004.
- [Enderton, 2002] H. B. Enderton. *A Mathematical Introduction to Logic (2nd edition)*. Academic Press, 2002.
- [Gruber 1993] T. Gruber: A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5:199-220, 1993.
- [Kifer et al. 1995] M. Kifer, G. Lausen, and J. Wu: Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42:741-843, July 1995.
- [Nern et al. 2004] H.-Joachim Nern, G. Agre, T. Atanasova, J. Saarela. System Framework for Generating Open Development Platforms for Web-Service Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems - INFRAWEBBS II. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS, ISSN 1790-0832, Issue 1, Volume 1*, July 2004, 286-291.
- [OWL-S 2004] *The OWL Services Coalition: OWL-S: Semantic Markup for Web Services, version 1.0*; available at <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>
- [Protégé 2005] <http://protege.stanford.edu/index.html>
- [Roman et al. 2005] D. Roman, H. Lausen, U. Keller, J. de Bruijn, Ch. Bussler, J. Domingue, D. Fensel, M. Kifer, J. Kopecky, R. Lara, E.I Oren, A.I Polleres, M. Stollberg. *D2v1.1. Web Service Modeling Ontology (WSMO) - WSMO Final Draft 10 February 2005*, <http://www.wsmo.org/TR/d2/v1.1/>
- [WSMO4J 2005] <http://wsmo4j.sourceforge.net/>.
- [WSMO Studio 2005] <http://www.wsmostudio.org>.
-

Authors' Information

Gennady Agre – Institute of Information Technologies – Bulgarian Academy of Sciences, Acad. G. Bonchev St., block 29A, Sofia 1113, Bulgaria; e-mail: agre@iinf.bas.bg

Petar Kormushev – Sofia University St. Kliment Ohridski; e-mail: pkormushev@gmail.com

Ivan Dilov – Sofia University St. Kliment Ohridski; e-mail: idilov@gmail.com

A SURVEY ON THE INTEGRATION OF ENTERPRISE APPLICATIONS AS A SERVICE

Hristina Daskalova, Vladislava Grigorova

Abstract: The present paper discusses the process of multi-lateral integration of the business applications, which requires the construction of a common infra-structure, acquires the format of a service and leads to release of the individual construction of a private infra-structure by every participant in the process.

Keywords: Web services, BPEL, Enterprise Applications Integration.

ACM Classification Keywords: H.3.4 Systems and Software: Information networks

Introduction

The concept of Web-services application has appeared after the failure to find another successful mechanism for interaction in the enormous variety of information systems and due to business requirements as well. The modern commercial enterprises use book-keeping, financial, production, store and other information systems. The large enterprises possess multi-functional information systems and also providers, customers and partners with their specific information systems, which necessitate interaction. That is why the area of Enterprise Applications Integration (EAI) attracts the research attempts and the Web-services are expected to prove as the most efficient instrument for Web-services solution.

Decomposition of the Business Processes

The efficient organization of the information systems interaction is preceded by evaluation of the necessary business processes realizing the business functions of a given organization. For this purpose the functional blocks of the business processes are decomposed in order to obtain a loop of business processes – from the loop of the business processes up to obtaining the separate business process, and from the independent business processes up to their comprising functions. The business function gives a certain measurable result, it is the essence, defining the process and hence this function may be identified with the service. It could be considered as a resource realizing the business function and having the following features:

- it enables a repeated use,
- it is determined by one or more technologically independent interfaces,
- it is weakly connected with other similar resources and may be requested by protocols providing the possibility for resources interaction.

In this way, from a functional view point, the business applications are decomposed to a set of interacting services. This set of interacting services, subordinate to certain common rules, interfaces and mechanisms of interaction, is the basis of the Service-Oriented Architecture (SOA).

The representative level and the level of data bases in such architecture are traditional and use the level of portals, different user's interfaces and relational data bases. The level of business logic realizes the systems and the working flows within the frames of the business processes. For this purpose, objects of different types have been introduced [Fagin, 2005]: entity objects, which derive the functions from the information at the level of business logic, realize the selection and modification of the data without details about their physical storing; activity objects which support the transactions, ensure the interactive operation of the users with the system, the access being done by an API; service objects accomplishing the service-oriented architectural principles, providing functionality and data as services for other applications. The objects from the level of the business logic guarantee the integration of the business components from a high level. The activity objects realize the customer's access to the system functional modules by standard interfaces of different types. The service objects

supply a new type of communication for the business modules on XML basis and provide the interaction of these components with the base communication services.

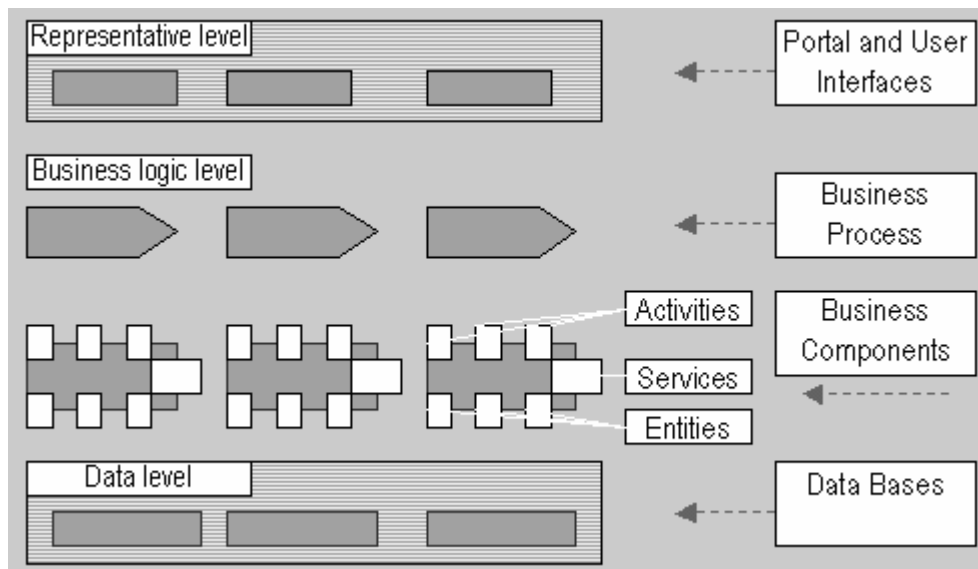


Fig. 1

Web services and Business Process Execution Language for Web services

The Web services are based on documents exchange, on appropriately designed interface interactions and on continuous expansion of the Simple Object Access Protocol (SOAP) [Curbera, 2002] [SOAP, 2003]. The Message Exchange Patterns (MEP) of SOAP protocol suppose single-directional messages transfer, as well as a "question-answer" mode, but it does not ensure sufficient semantics for control of the interaction at an infrastructure level. Some organizations like OASIS, W3C and WS-I work for improvement of the applications controlled by the events. This is aided by the opportunity to expand SOAP protocol which is assured by its composition properties. The SOAP messages can be accompanied by appropriate electronic instructions controlling the messages acquisition and processing (for example a channel and a specific quality of the message transfer may be selected, etc). Thanks to this mechanism some specifications such as WS-Addressing, WS-Eventing, WS-Notification, WS-ReliableMessaging and WS-Security [WS-Security, 2003] together with WSDL 2.0 [WSDL, 2004] [Curbera, 2002] guarantee the realization of the functionality of modern tools at the intermediate layer, oriented towards messages.

The application of the Business Process Execution Language (BPEL) [BPEL, 2003] is a natural way to realize the service-oriented options when executing the business processes. It gives the possibility to represent an abstract model of the business process with the help of information flows connecting the system and the algorithm corresponding to this model. The algorithm may be presented as an executable Web-service with an interface depicted by WSDL language. The request for the service operation, realized with the help of BPEL causes a sequence of steps which are executed by BPEL container in compliance with the formal description of the process logic. Such steps are the synchronous and asynchronous calls of other services; in this way the process subject interacts with the external systems, combining them in one flow according to the rules of the routine and of messages conversion. BPEL enlarges the area of processes models application from the analysis up to the realization phase.

At the same time the use of such executable models implies some new questions, like:

- which aspects can be described by BPEL, and which – by WSDL and what is the difference between the process models and the traditional program models;

- in what way such elements like non-functional requirements and characteristics of the services quality could be accounted by the models;
- how to evaluate the quality of the executable process models and what tools to apply for this purpose;
- what is the role of the business flows, how the processing modeling could be synchronized with the modeling of the use cases at the design phase, etc.

The classic demonstration of BPEL possibilities can be seen in tourist services, for example in reservation of air plane tickets, accommodation booking and their payment. Certain problems appear when setting the optimizing constraints, such as total cost of the trip. This constraint depends on the price of the flight tickets and the price of the hotel rooms, which usually vary. Besides, the tickets may be sold and the hotel rooms – already booked. Only after the determination of the optimal dates and accomplishment of successful reservations, the process is paid by a credit card; otherwise there is no reservation and an error is signaled.

Local memory, branches, cycles and exclusions processing enable the automation of the process with the help of BPEL. It is also important to provide such infra-structure that guarantees integration of new services (new air lines, new hotel chains), modification of the parameters of the already used, removal of some services from the process consideration, not altering its structure. The provider of tourist services should not react independently to every alteration accomplished by the contra-agent side.

The Universal Description Discovery and Integration standard (UDDI) [UDDI, 2003] is suitable for support of the partners' self service. It is a universal method for description, discovery and integration of the Web-services in B2B systems for e-commerce.

UDDI business register is a data base for common use, in which the interested organizations register themselves (by corresponding operators' nodes) and enter information concerning their business. On the basis of UDDI standard, the description of complex business processes is possible after decomposition to their components. At that the information exchange is increased due to the easier perception of standardized information. UDDI does not provide the service, but creates the technical possibility to search for the necessary services until the technologies of the desired partners are defined, to look for compatibility interfaces, to provide standardized formats for program search of business and services.

UDDI is a manual of the available Web-services, including the types of business, names, post addresses, persons for contacts, phone and fax numbers, email addresses, URL of the Web-services offered, meta-data describing the interfaces towards existing Web-services and others. This speeds up the search for appropriate partners in Internet; assures optimal way for interaction and possibility to organize shared access to the information by a global business register.

Using this approach, the tourist operator, who keeps business relations with the air lines and hotel chains, gives them the right to autonomously publish their services and to control them in his partners' register-catalogue. The client of these services is the reservation process. When operating the requests, the process is connected to the catalogue in order to find the accessible services, after that it operates in conformance with the business logic. Using the options of the catalogue, that classifies the services and their providers, the process may be improved, avoiding consideration of services not connected with the request considered. Thus the number of flights and hotels combinations to be processed, could be diminished.

The catalogue provides also one and the same interface for the services. It must contain a description of the meta-data associated with the services, including the specification of their interfaces described by WSDL. Every booking service may be referred to a given group. During the reservation process only the services corresponding to this type, will be discovered and considered. Different autonomous business applications are combined in this example of tourist services. Inside the companies the services are also integrated in order to realize the functions needed for external interactions. The direct interaction with the partners and customers is prepared by the integration of the internal business processes of the enterprises.

Conclusion

Some integration service providers already possess a built-in logic for events control in their own infra-structure, accessible through Internet. The optimal solution of the problem for multi-lateral integration will require the construction of a common infra-structure. The users themselves could control the logic of interaction with the partners using BPEL realization. One of the best fulfilled realizations of BPEL is known as the product Oracle BPEL Process Manager. It has got a strategic position in the service-oriented investigations of the company and is presented within the composition of Oracle Application Server 10g (tools for processes execution) and JDeveloper (tools for processes design). In this way BPEL is the linking element among the variety of Oracle Applications products [Oracle, 2004].

In its essence the integration process of an infinite number of terminals acquires the format of a service, supplied on a request, and this releases the participants in the process from the necessity to develop their own infra-structure. A still more intensive advance of the companies specializing in the area of integration technologies is expected in the future.

Acknowledgement

This work is carried out under EU Project INFRAWEBs - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

Bibliography

- [BPEL, 2003] Business Process Execution Language for Web Services Version 1.1
<http://www-106.ibm.com/developerworks/library/ws-bpel/>
- [BPML, 2002] BPML working draft March 25, 2002. <http://www.bpml.org/>, <http://xml.coverpages.org/bpml.html>
- [Curbera, 2002] F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana. Unraveling the web services web: An introduction to SOAP, WSDL, UDDI. IEEE Internet Computing, 6(2):86-93, March-April 2002.
- [Fagin, 2005] D. Fagin. Architectures and tools for applications integration. Open systems. # 02/2005. (In Russian).
- [Oracle, 2004] Oracle BPEL Process Manager. <http://www.oracle.com/technology/products/ias/bpel/index.html>
- [Shapiro, 2005] R. Shapiro. A Comparison of XPDL, BPML and BPEL4WS. <http://xml.coverpages.org/Shapiro-XPDL.pdf>
- [SOAP, 2003] SOAP Version 1.2 Part 1: Messaging Framework. <http://www.w3.org/TR/2003/REC-soap12-part1-20030624/>
- [UDDI, 2003] UDDI Version 3.0. <http://uddi.org/pubs/uddi-v3.00-published-20020719.htm>
- [WSDL, 2004] Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language.
<http://www.w3.org/TR/2004/WD-wsdl20-20040326/> Author information
- [WSFL, 2003] Web Services Flow Language. <http://www-3.ibm.com/software/solutions/Webservices/pdf/WSFL.pdf>
<http://xml.coverpages.org/wsfl.html>, <http://www.ebpm.org/wsfl.htm>
- [WS-Security, 2003] Web Services Security (WS-Security).
<http://www-106.ibm.com/developerworks/Webservices/library/ws-secure/>
- [WS-Transaction, 2003] Web Services Transaction (WS-Transaction).
<http://www-106.ibm.com/developerworks/Webservices/library/ws-transpec/>

Authors' Information

Hristina Daskalova – Institute of Information Technologies; BAS, Acad. G. Bontchev St., block 2, Sofia-1113, Bulgaria; e-mail: daskalovahg@abv.bg

Vladislava Grigorova – Institute of Information Technologies; BAS, Acad. G. Bontchev St., block 2, Sofia-1113, Bulgaria; e-mail: v.grigorova@abv.bg

SEMANTIC DESCRIPTION OF WEB SERVICES AND POSSIBILITIES OF BPEL4WS

Vladislava Grigorova

Abstract: The using of the upsurge of semantics web technologies gives a possibility for an increasing of the flexibility, extensibility and consistency of the existent industrial standards for modeling of web services. In the paper the types of semantic description of web services and the degree of their realization in BPEL4WS (Business Process Execution Language for Web Services) respectively on the abstract and executable level are treated. The methods for using of BPEL4WS for the purposes of semantic web services in the direction of their semi-automatic integration are suggested.

Keywords: Semantics, Web Services, BPEL4WS.

ACM Classification Keywords: H.3.4 Systems and Software: Information networks

Introduction

Technologies for creation, composition, description, publication, discovery, implementation and execution of web services are developing strenuously now for the fulfillment of exchanges and interactions of data and functions for dynamic integration of distributed computer systems in heterogeneous networks. Web services are software platform-independent, self-contained, modular business process applications. The main components of web services technology are: Web Service Description Language (WSDL) for interface description which details indicate how the web services to be called; Simple Object Access Protocol (SOAP) for messages exchange, that is XML-based protocol for communication between the web services and client applications; UDDI Universal Description, Discovery and Integration (UDDI) for registration, publication and location of web services and their characteristics.

The protocols and languages for business processes modelling are developed for the aims of the enterprise application integration, for creation and management of logic of connection between service and application during execution of business function. It was found that these basic components of web services are not enough for dealing with modern requirements – dynamic composition, flexible discovery, good initialization and selection of appropriate service.

In parallel with the elaboration of web services standards the investigations connected with application of semantic technologies in the web space are carrying out. There already exist semantic repositories of data, ontologies, rules of ontologies, and engines for data interpretation – taxonomies modeling, sets of tools and applications that are necessary base for the semantic web.

The use of the potential of the semantic web together with industrial standards for web services would help to resolve many present problems in business processes integration.

Business Process Execution Language for Web Services

A set of articles exists which compare and evaluate current languages for web services modeling and composing and show why Business Process Execution Language for Web Services (BPEL4WS or BPEL) [Andrews, 2003] is established as an industrial standard. It is made in co-authorship between IBM, Microsoft, BEA, SAP и Siebel Systems, combines IBM's Web Services Flow Language and Microsoft's XLANG specifications, superseding both these specifications. BPEL supports the constructions for presenting of complex models of web services compositions. The business process includes self-contained set of activities that are predetermined according given cases. There are two ways for specification of distributed business process – a global model specifies orchestration of the whole set of web services, on the other hand, each web service specifies the interaction between partners.

BPEL extends the possibilities of WSDL in direction of the integration of complex connections between web services on the base of the principles for modeling of business logic and the corresponding business processes [Farahbod, 2004]a.

The definition of business process in BPEL expresses business logic and describes process model elements in terms of business messages exchange and contains process partners to which web services are connected; variables, which define the state of the process; activities (basic and structural), which define the behaviour of business process. The basic activities define tasks, which are accomplished in business process. The structural activities define the control flow.

The figure 1 is a good illustration of the connections between the process model elements and the artifacts of BPEL4WS file and the relevant WSDL files [Beck, 2005].

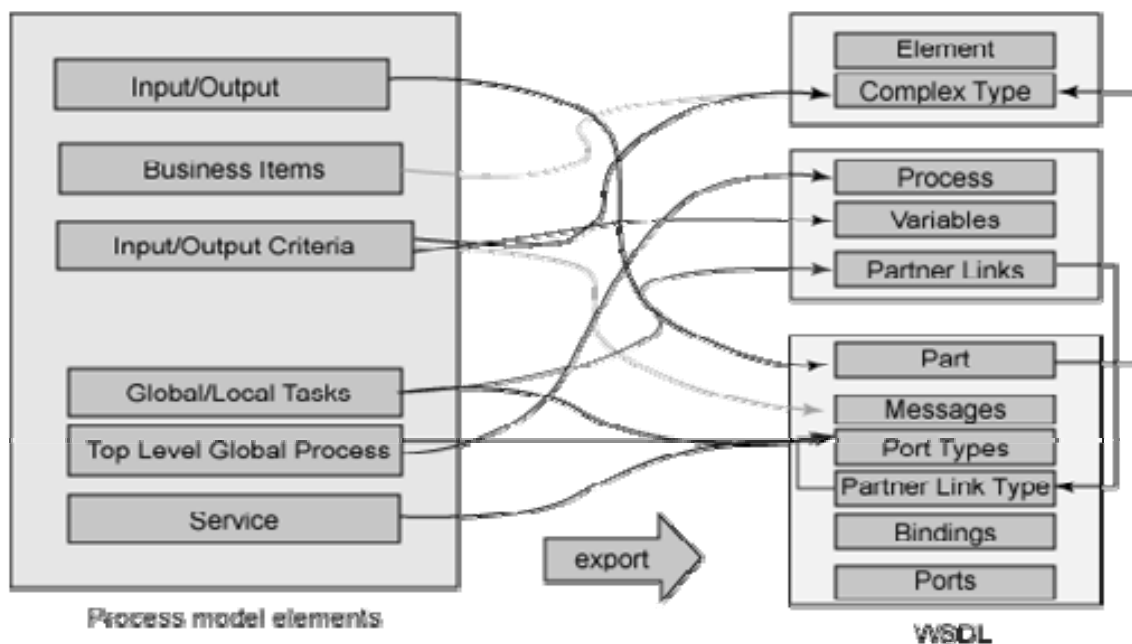


Figure 1

BPEL4WS is based on XML. A business process, described in BPEL4WS is executable by BP4WSJ engine, realized by IBM. It allows an automatic web services execution and is a precondition for automatic web services discovery [McIlraith, 2003]a.

Web services from partners that are not using BPEL can be included in the process. Mechanisms of dealing with errors and exceptions, events, compensations in long running transactions, dynamic process origination, and dynamic messages connection in different processes are supported.

The basic concepts of BPEL4WS are declared in two aspects: abstract and executable processes. The executable process reflects on the actual behavior of the process participants with all their specifications. The abstract process or the business protocol for interactions has a descriptive role with more than one aim – it can define openly the behavior of several or all services, as the operational details are skipped or opaque, or it can define a model, representative of the best domain specific practice.

Semantics

Semantic approaches give possibilities for unambiguous interpretation of the web services content, for description of their properties and potentialities in machine-understandable form. The higher layer on the base of traditional web is constructed, in which the information is given through its precise defined meaning.

The decisions that can be made on the base of web services by using their related semantics are illustrated by [IBM © Corporation, 2004] in the figure 2.

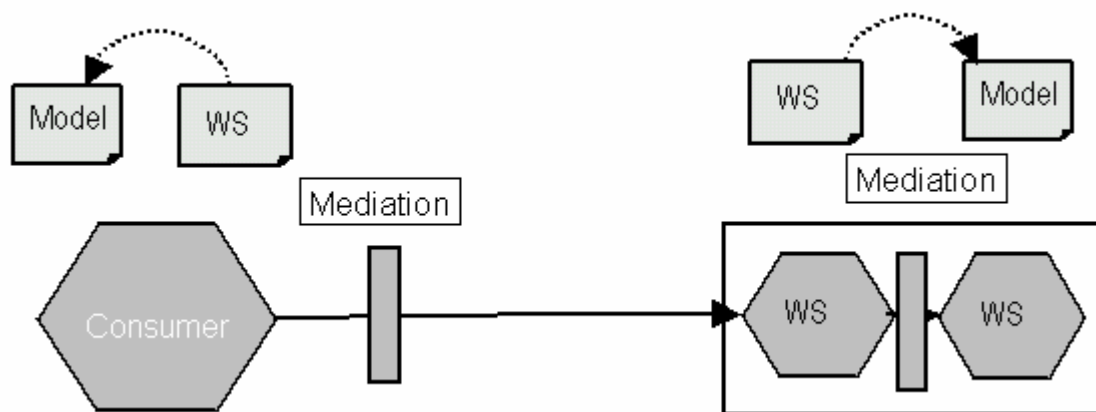


Figure 2

The semantics can be divided into different types on the base of different indications. For the purposes of this work, it is transformed as:

- functional, that expresses business logic and corresponds to web services capabilities;
- data semantics, semantics of inputs / outputs of Web Services;
- QoS semantics, that is connected with requirements and reflects on the results;
- Execution semantics that is connected to execution and dynamic service invocation.

[Patil, 2005] [Sivashanmugam, 2005] [Verma, 2005].

The functional semantics is connected with inputs/outputs as with functioning of web service, it is a collection of them. The functional ontologies are semantics of operation. The appropriate services are discovered on the base of their functional semantics.

The data semantics is the description of operation data. The ontology on their meaning is constructed.

QoS semantics is used during selection of the services. Different quality aspects (time, price, security, etc.) can be important for different composition of services in different range.

The execution semantics is precondition for realizing the dynamic discovery, binding and monitoring of process execution.

All of these kinds of semantics have to be recognized, marked and taken into consideration for establishing of the semantic web services.

Associating of Semantic Meanings with BPEL4WS Constructions

In [Mandell, 2003] it is mentioned the restrictions of XML schema, which is not able to be appropriate tool for expressing of the semantics of the exchanged data as to deal with problems of semantic interoperability, that reflects on the possibilities of BPEL [Mandell, 2003] [McIlraith, 2003]b.

XML is oriented to the document structure and does not propose data interpretation. BPEL is not a declarative language, which makes the task of its enriching with semantic technologies difficult.

However, it is necessary to exert efforts in this direction because there is no better industrial standard today than BPEL4WS, which envelops all life cycle of web services and is the essence of all similar languages. From the other side the semantics is proofed its necessity.

The semantic web services design from existing web services has to go through extracting and marking of all four kinds of semantics.

The data semantics can be described with using of corresponding WSDL files that belong to the BPEL-files, because WSDL is connected to the description of web services. On the base of input/output data, the ontology for the business process can be built.

By tracing the operation names in the corresponding constructions together with the names of partners and inputs/outputs, the collection of functionalities is formed that is connected to functional semantics. From other

side the possibility to represent the business process as directed graph helps for formal description of operational behaviour and extraction of abstract functional semantics [Farahbod, 2004]b.

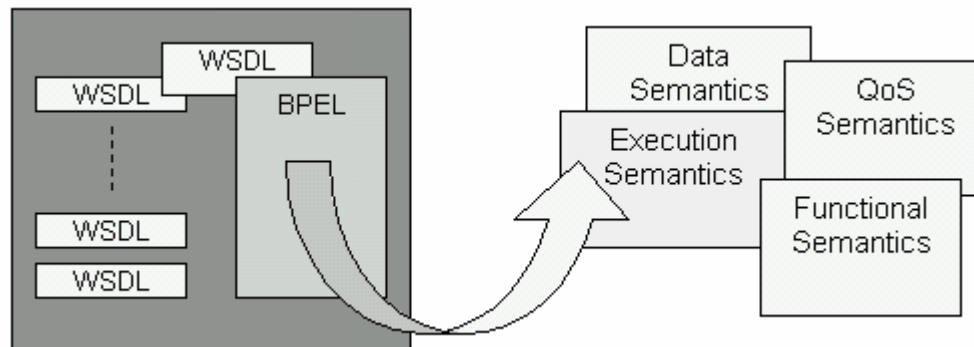


Figure 3

QoS may be referred to WS-Security, WS-Reliability, WS-Transactions, WS-Coordination, WS-Context and eventually can be extracted from them. In complex processes, they can be presented as compositional models, based on workflows models, and in such a way, it is possible to aggregate some numerical metrics of QoS [Cardoso, 2004] and corresponding semantic meanings. Another approach is in the extending of non-functional properties with QoS from the providers, receiving the certificate of quality and registering in UDDI with the corresponding description. [Al-Ali, 2003].

The execution semantics coordinates with the way of realizing in the process of the mechanisms for supporting of exceptions and error handling as with correct design of accordance between inputs and their outputs. Every activity is associated with the state of execution (completed, stopped, interrupted, unsuccessful, skipped, etc) which have to be defined for the process be executed and the semantics of alternative ways of execution can be extracted.

Deriving of different kinds of semantics not differs principally for abstract and executable BPEL process. But the results may be different because for the abstract process it is allowed to hide its behaviour, to use different levels of restrictions and transparency. Using templates can contribute to defining business processes from the application domain and to receiving the semantic knowledge more comprehensively.

Conclusion

Extant web services technologies for modeling, publishing, discovery and services connecting guarantee syntactic compatibility. That means that in the standard web services the discovery and the composition are syntactic based, and semantic compatibility does not exist. The using of the semantic description would help the dynamic composition of web services for concrete request. The dynamic composition of web services is necessary when the client request cannot be realized directly from extant web services.

For these purposes and in the same time to respond qualitative to the given request the existing web services should be described by their meanings, which would help the better comprehensions between computers and people in their work together.

The solution is in recognizing and marking-up of the web content with well-defined semantics.

Acknowledgements

This work is carried out under EU Project INFRAWEBs - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

Bibliography

- [Al-Ali, 2003] R. Al-Ali, O. Rana, D. Walker, S. Jha, and S. Sohail. G-QoS: Grid service discovery using QoS properties. *Computing and Informatics Journal, Special Issue on Grid Computing*, 21(5), 2003.
- [Andrews, 2003] T. Andrews et al. Business Process Execution Language for Web Services, Version 1.1. <ftp://www6.software.ibm.com/software/developer/library/ws-bpel.pdf>, 2003.
- [Arroyo, 2004] S. Arroyo, R. Lara, J. M.Gomez, D. Bereka, Y. Ding, D.Fensel. Semantic Aspects of Web Services. In: *Practical Handbook of Internet Computing*. Munindar P. (Ed.) Chapman Hall and CRC Press, Baton Rouge, 2004.
- [Beck, 2005] K. Beck, J. Joseph, G. Goldszmidt. Learn business process modeling basics for the analyst, <http://www-128.ibm.com/developerworks/library/ws-bpm4analyst/> (1 of 10)4/13/2005.
- [Cardoso, 2004] J. Cardoso, A. Sheth, J. Miller, J. Arnold and K. Kochut. Quality of Service for Workflows and Web Service Processes. *Journal of Web Semantics*, 1 (2004) pp. 281-308, Elsevier.
- [Farahbod, 2004]a R. Farahbod, U. Glaesser and M. Vajihollahi. Specification and Validation of the Business Process Execution Language for Web Services. SFU-CMPT-TR-2003-06 (revised). Feb 2004.
- [Farahbod, 2004]b R. Farahbod, U. Glaesser, and M. Vajihollahi. Abstract operational semantics of the business process execution language for web services. Technical Report TR 2004-03, "School of Computing Science, Simon Fraser University", Apr. 2004.
- [Gomez, 2005] J. M. Gomez, A. Haller and Ch.Bussler. A Conversationoriented language for B2B integration based on Semantic Web Services. WWW 2005. Chiba, Japan. May 10--14, 2005.
- [IBM® Corporation, 2004] IBM® Corporation. Web Services Semantics: A View of Semantics in Services Oriented Architecture. <http://awwebx04.alphaworks.ibm.com/ettk/psme/WebServicesSemanticsWhitePaper.htm>, December, 2004
- [Jaeger, 2004] M. C. Jaeger, Gr. Rojec-Goldmann and G. Muehl. QoS Aggregation for Web Service Composition using Workflow Patterns. In *Proceedings of the 8th International Enterprise Distributed Object Computing Conference (EDOC 2004)*, pages 149–159, Monterey, California, USA, September 2004. IEEE CS Press.
- [Mandell, 2003] D. J. Mandell and S. A. McIlraith. Adapting BPEL4WS for the Semantic Web: The Bottom-Up Approach to Web Service Interoperation. In *Proc. of the International Semantic Web Conference (ISWC)*, pages 227–241, 2003.
- [McIlraith, 2001]a S. McIlraith, T. Son, and H. Zeng. Semantic Web Services. *IEEE Intelligent Systems*, 16(2):46 – 53, 2001.
- [McIlraith, 2003]b S. McIlraith and D. Martin. Bringing Semantics to Web Services. *IEEE Intelligent Systems*, 18(1) 90-93. 2003.
- [Patil, 2005] A. Patil, S. Oundhakar, A. Sheth, K. Verma, METEOR-S Web service Annotation Framework, *Proceeding of the World Wide Web Conference*, July 2004 (*Proceeding of the World Wide Web Conference*, July 2004) WWW 2004, May 17–22, 2004, New York, New York, USA.
- [Sivashanmugam, 2005] Sivashanmugam K., Verma K., Sheth A., Miller J. Adding Semantics to Web Services Standards. *Proceedings of the 1st International Conference on Web Services (ICWS'03)*, Las Vegas, Nevada (June 2003) pp.395-401.
- [Verma, 2005] K. Verma, K. Sivashanmugam, A. Sheth, A. Patil, S. Oundhakar, and J. Miller. METEOR-S WSDI: A Scalable P2P Infrastructure of Registries for Semantic Publication and Discovery of Web Services. *Journal of Information Technology and Management*, 6(1):17–39, 2005.
- [Zaremba, 2005] M. Zaremba and Ch. Bussler. Towards Dynamic Execution Semantics in Semantic Web Services. WWW 2005. Chiba, Japan. May 10–14, 2005.

Author's Information

Vladislava Grigorova – Institute of Information Technologies; BAS, Acad.G.Bontchev St., bl.2, Sofia-1113, Bulgaria; e-mail: v.grigorova@abv.bg

INFRAWEBBS BPEL-BASED EDITOR FOR CREATING THE SEMANTIC WEB SERVICES DESCRIPTION

Tatiana Atanasova

Abstract: INFRAWEBBS project [INFRAWEBBS] considers usage of semantics for the complete lifecycle of Semantic Web processes, which represent complex interactions between Semantic Web Services. One of the main initiatives in the Semantic Web is WSMO framework, aiming at describing the various aspects related to Semantic Web Services in order to enable the automation of Web Service discovery, composition, interoperability and invocation. In the paper the conceptual architecture for BPEL-based INFRAWEBBS editor is proposed that is intended to construct a part of WSMO descriptions of the Semantic Web Services. The semantic description of Web Services has to cover Data, Functional, Execution and QoS semantics. The representation of Functional semantics can be achieved by adding the service functionality to the process description. The architecture relies on a functional (operational) semantics of the Business Process Execution Language for Web Services (BPEL4WS) and uses abstract state machine (ASM) paradigm. This allows describing the dynamic properties of the process descriptions in terms of partially ordered transition rules and transforming them to WSMO framework.

Keywords: Semantic Web Services, BPEL, ASM, WSMO

ACM Classification Keywords: H.3.4 Systems and Software: Information networks

A little semantics goes a long way.

James Hendler

Introduction

The power of Web services can be realized only when appropriate services are discovered based on the functional requirements given by functional semantics. Including the functional semantics to the web services descriptions is a step forward to the Semantic Web.

The Semantic Web is an extension of the current World Wide Web. It builds on the current World Wide Web constructs and topology, but adds further capabilities by defining machine-processable data and relationship standards along with richer semantic associations [KM-GOV, 2005]. Semantic Web Services are integrated solution for realizing the vision of the next generation of the Web; they define semantically driven technologies for automation of the Web Service usage process.

The semantics of a Web Service is the shared expectation about the behavior of the service [W3C, 2004] and give more expressive descriptions about the service, for example, the intent of the service or kind of data being deal with.

Semantics for a service can be:

- Implied
- Expressed in human-understandable form
- Expressed in machine-readable form

As a step in the direction to the semantic services creation, accession and reusing within the Semantic Web the current INFRAWEBBS research project [Nern et al, 2004] may be considered. The project uses semantics during the complete lifecycle of Semantic Web processes. One of the main components of INFRAWEBBS project is SWU – Semantic Web Services Unit [Atanasova et al, 2005]. SWU is intended to converting web services from available descriptions and domain knowledge to the semantic ones and to Compose Web Services, through combining and orchestrating them in order to deliver added-value services.

The appropriate solution for realizing these aims may be found using semantic approach to describing and processing of Web Services. One of the initiatives in Semantic Web is WSMO – Web Service Modelling Ontology (www.wsmo.org). The INFRAWEB project relies on the WSMO framework.

Current standards for service description (for example, WSDL, UDDI, BPEL) have no semantically marked up content. In [Cardoso, 2005] it is pointed that *“During Web service development data, functional and QoS semantics of the service needs to be specified... it is very important to use semantics at this stage.”*

Several approaches have already been suggested for adding semantics to Web services. Semantics can either be added to currently existing syntactic Web service standards like UDDI or WSDL or services can be described using some ontology based description language like DAML-S [Patil et al, 2004]. Only in [Mandell, 2003] it is made an assumption for using BPEL as a base for semantic annotation by proposing to “take a bottom-up approach to integrating Semantic Web technology into Web Services”. But they mainly focus on introducing a semantic discovery service and facilitate semantic translations.

In this paper an approach is proposed for migrating from BPEL to WSMO description by semi-automatic mapping BPEL constructions to WSMO descriptions of service interface using Abstract State Machine paradigm.

BPEL

The Business Process Execution Language for Web Services (BPEL) at present is the most well-known language to specify and execute business processes, using Web Services as its technological basis. BPEL is built on top of WSDL and indirectly on SOAP and introduces a stateful interaction model that allows exchanging sequences of messages between business partners (i.e. Web Services).

The definition of business process in BPEL expresses business logic and describes process model elements in terms of business messages exchange and consists of:

- process partners to which web services are connected;
- variables, which define the state of the process;
- activities (basic and structural), which define the behaviour of the business process. The basic activities define tasks, which are accomplished in business process. The structural activities define the control flow.

BPEL is a block-structured programming language; it allows recursive blocks, but restricts definitions and declarations to the top level. BPEL document contains partner links, variables (message containers), and activities (process programs). Partners are external Web services. Activities are the actions or tasks performed within the business process and are the basic components of a process definition. Structured activities prescribe the order in which a collection of activities take place. Sequential control between activities is provided by *sequence*, *switch*, and *while*. Concurrency and synchronization between activities is provided by *flow*. Nondeterministic choice based on external events is provided by *pick*.

Variables are process instance-relevant data, providing their definitions in terms of WSDL message types, XML Schema simple types, or XML Schema elements. Variables allow processes to maintain state data and process history based on messages exchanged.

BPEL defines a mechanism for catching and handling faults. Fault handlers catch faults when they are thrown by other actions. There is also a compensation handler to enable compensatory activities in the event of actions that cannot be explicitly undone. BPEL does not support nested process definition.

It is possible to use abstract and executable BPEL processes. Abstract process is a business interaction protocols; executable process models actual behaviour of a participant in a business interaction and it may be compiled into invocable services. The structure of BPEL-based process is as follows:

```
<process name = "...">
  <partnerLinks>
    ...
  </partnerLinks>
  <variables>
    ...
  </variables>
```

```

    <flow>
      <links>
        ...
      </links>
      <!-- activities -->
        ...
    </flow>
  </process>

```

The possibility to represent the BPEL-based business process as directed graph helps for formal description of operational behaviour and extraction of abstract functional semantics [Farahbod, 2004].

ASM and BPEL

In [Farahbod, 2004] an Abstract Syntax Tree is proposed to capture the complete structure of a BPEL process. This representation will be used for formalization of BPEL file by ASM.

First of all, let's recall a definition of functional (operational) semantics: *Operational semantics* consists in describing the steps of the computation of a program by giving formal rules to derive judgments of the form $\{p, a\} \rightarrow r$, to be read as "the program p , when applied to the input a , terminates and produces the output r ".

Operational semantics is a way to give meaning to computer programs in a mathematically rigorous way [Wikipedia]. The operational semantics for a programming language describes how a valid program is interpreted as sequences of computational steps. These sequences are the meaning of the program.

For the formal definitions of functional semantics of programming languages Abstract State Machines (ASM) are now widely used.

ASM's definition is given in [Börger, 1999]:

- An ASM M is a finite set of rules for guarded multiple function updates;
- Applying one step of M (a set of rules) to a state (algebra) A produces as next state another algebra A' of the same signature obtained as follows:
 - First evaluate in A using the standard interpretation of classical logic all the guards of all the rules of M
 - Compute in A for each of the rules of M whose guard evaluates to true all the arguments and all the values appearing in the updates of this rule
 - Replace simultaneously for each rule and for all the locations in question the previous A -function value by the newly computed value
 - The algebra A' thus obtained differs from A by the new values for those functions at those arguments where the values are updated by a rule of M which could fire in A

A distributed ASM – DASM [Börger, 1997] is given by a set of agents and a program function Mod which assigns to each agent a module ("sequential" program) consisting of a finite number of so called transition rules of the form "If $Cond$ then $Updates$ ", where $Cond$ is any expression (of first order logic) and $Updates$ is a finite set of function updates, i.e. of updates $f(t_1; : : : t_n) := t$. The states of ASMs are arbitrary structures, i.e. domains with predicates and functions defined on them.

DASM generalize runs from sequences of moves of a basic ASM to partial orders of moves of multiple agents, each executing a basic ASM.

So ASM is usually a state represented by algebra and transition rule. The algebra consists of superuniverse divided into universes and a number of dynamic functions. The interpretation of the transition rule generally causes an update of the state by modifying the dynamic function or by inserting a new element into the universe.

The DASM for formalizing of BPEL may be defined. One such attempt is made in [Farahbod, 2004] as:

A DASM M has a finite set AGENT of autonomously operating *agents*.

- The set of agents changes dynamically over runs of M

- The behavior of an agent a in a given state S of M is defined by its program $programs(a)$
- To introduce a new agent a in state S , a valid program has to be assigned to $programs(a)$; to terminate a , $programs(a)$ is reset to the distinguished value $undef$
- In any state S reachable from an initial state of M , the set of agents is well defined as $AGENTS \equiv \{x \in S : programs(x) \neq undef\}$.

Asynchronous (distributed) ASMs generalize sequences of transitions of basic ASM to partial orders of transitions of multiple agents, each executing a basic ASM [Börger, 1997].

Modeling service contract as an abstract machine needs to define its dynamics as functions and its static as state variables. The idea is that the execution of BPEL produces a computation which is a sequence of states. Each next state in the sequence is the result of applying in parallel the updates of exactly one *rule* to the current state. The rule that is chosen to be executed is a rule whose guard is true in the current state. If there is more than one rule whose guard evaluates to true in the current state, one of them is chosen non-deterministically. If no guard is true, then the computation halts.

The structure of BPEL Abstract Machine consists of different types of DASM agents representing process instances and Activity agents - created dynamically by process agents for executing BPEL structured activities.

Each BPEL pattern {<on message> operation; logic of transition; final operation} corresponds to *transition*.

The full BPEL abstract machine is a DASM that has components, each of them is again a DASM. The [Farahbod, 2004] formalization of the key semantic aspects of BPEL in terms of a hierarchically defined BPEL Abstract Machine can be further developed for the aims of mapping such formalization to semantic services description using WSMO framework.

WSMO

The Web Services Modeling Ontology framework - WSMO [WSMO Specification, 2005] considers the semantic description of Web Services as including Capability (functional) and Interfaces (usage).

WSMO Web Service Interfaces are concerned with service consumption and interaction. Service Interface Description uses Ontologies as data model. This means that all data elements interchanged are ontology instances and service interface represents evolving ontology. Sub-concepts of Service Interface are Choreography and Orchestration. In WSMO service interface description represents the dynamics of information interchange during service consumption and interaction and supports ontologies as the underlying data model.

Service Interface Description Model in WSMO framework consists of:

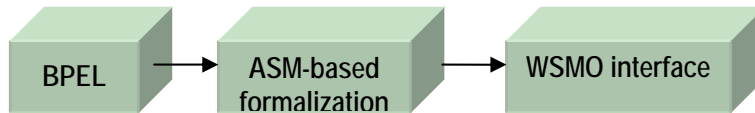
- Vocabulary Ω , that is ontology schema(s) used in service interface description and usage for information interchange: *in, out, shared, controlled*.
- States $\omega(\Omega)$ that are a stable status in the information space, defined by attribute values of ontology instances;
- Guarded Transition $GT(\omega)$ representing state transition with general structure: *if (condition) then (action)* but different for Choreography and Orchestration.

The conceptual model of WSMO orchestrations and WSMO Choreography [Roman et al., 2005] may be described with ASM-based formal semantics.

So ASM is a conceptual base for describing of the formal semantics of non-semantic web services descriptions and the semantic ones. But there are two different layers of abstraction in BPEL-based ASM formalization and WSMO description of SWS interface. To construct a bridge between these representations it is proposed to provide Case-based mapping from BPEL to WSMO through ASM paradigm. As WSMO is still developing the proposed approach is the conceptual one. The proposed tool for semi-automatically mapping of BPEL functional semantic to WSMO description is a part of INFRAWEBs Case-Based Designer intended to semiautomatic conversion of non-semantic Web services to Semantic Web Services.

BPEL-based Editor

The approach consists in constructing a bridge between two different levels of abstractions in the two ASM-based formalizations (Fig. 1). We have to find relation between BPEL and WSMO service model description. To develop the BPEL-based editor first it is necessary to formalize the semantics in imported BPEL-file.



Because of no formal definition of the semantics is provided by the BPEL so we need to explore the formal meaning of activities in BPEL using ASM paradigm. The semantics of BPEL describe how the language provides interactions. Modelling these semantics requires mapping between activities. An approach to mapping of BPEL to ASM is given in [Farahbod et al, 2004], but we need to enrich it by annotating it with ontologies.

The BPEL is a directed graph. N transitions from state S_i are mapped. A link connects one source activity to one target activity. An activity can have multiple incoming and outgoing links that can be guarded. On the basis on the graph the mapping transitions, mapping states and connecting state skeletons can be made. Each activity is mapped to different kinds of agents in DASM formalization.

To make a matching to WSMO service interface let's consider:

Data-flow aspect - the concrete syntax of activity is reviewed; then a semantic mapping for the abstract syntax representation of this fragment of the activity is defined. This needs to include the mapping for basic control flow.

Control-flow aspect - The abstract definition of a web service (via WSDL as a port type with an operation) can be mapped to an activity since it comprises a self-contained functionality. The `<partnerLinkType>` definitions describe relationships between different activities without detailing the type of dependency (task, resource, goal).

Data Handling - variables hold message data and state information.

The proposed Functionality of the editor is as follows:

- load BPEL file,
- parse BPEL file,
- generate an ASM for each of the component processes, the component processes may be viewed either graphically or textually,
- load ontologies,
- matching/annotation,
- results of matching/annotation,
- export WSMO service interface (formal behaviour model).

The editor is been realizing as a plug-in to Eclipse (Fig. 2).

The following panes are proposed:

- *BPEL XML Tree View*
- *BPEL Process Activities List*
- *WSDL Service Description List*
- *ASM formalization*
- *Mapper/annotator*

At present we have modelled and formalized the web service compositions constructed in the BPEL standard specification. Further work is required with respect to transactional modelling. We are also working in decomposition based upon a resource model of how BPEL processes are composed to distributed requirements.

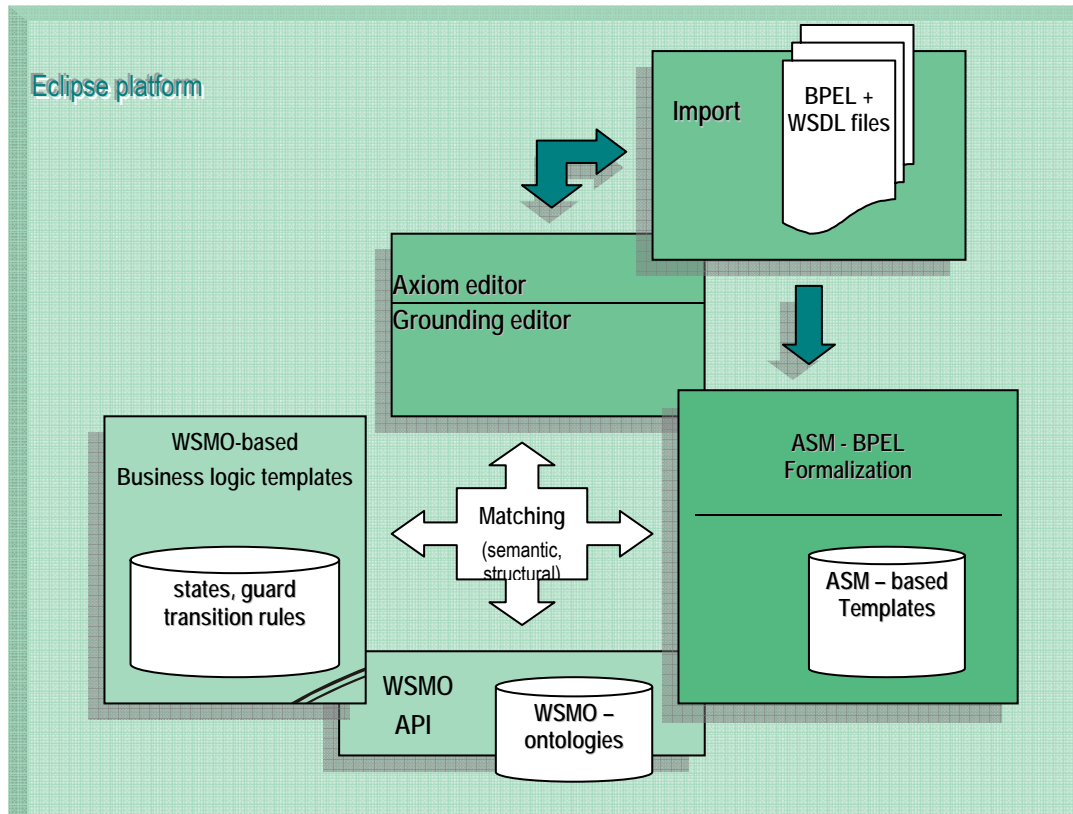


Fig.2

Conclusion

BPEL provides the model and XML-based grammar that define the interactions between a process and its partners using Web Services interfaces. BPEL also defines the states and logic of coordination between these interactions and systematic ways of dealing with exceptional conditions. But BPEL itself cannot explicitly describe the content and meaning of a Web service/process. Formalizing BPEL by ASM aims to get strict meaning of the functional semantics of the BPEL-based Web service and to migrate to WSMO description of semantic Web service interface. By using ASM the semantics of BPEL may be defined by the work of an abstract interpreter whose behavior is expressed in terms of transition rules. The states of the interpreter are represented by a number of dynamic functions. The semantics of the BPEL components is describing by a corresponding transition rule which indicates in what way a dynamic function should be updated in the process of execution of its component. Mapping to WSMO service interface is proposed by matching/annotation a state with ontology, and guarded transition by transition rules that express changes of states from activities in BPEL-file with ontologies.

Acknowledgement

This work is carried out under EU Project INFRAWEEBS - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

Bibliography

- [INFRAWEB, 2004] <http://www.fn-bochum.de/infraweb/>
- [W3C, 2004] Web Services Architecture, W3C Working Group Note 11, <http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/> February 2004.
- [Cardoso, 2005] J. Cardoso and A. Sheth, Introduction to Semantic Web Services and Web Process Composition, in Semantic Web Process: powering next generation of processes with Semantics and Web Services, Lecture Notes in Computer Science, Springer, 2005
- [Patil et al, 2004] A. Patil, S. Oundhakar, A. Sheth, K. Verma, METEOR-S Web Service Annotation Framework, WWW 2004, May 17–22, 2004, New York, New York, USA.
- [Sivashanmugam et al. 2003] K. Sivashanmugam, J. A. Miller, A. P. Sheth, K. Verma, Framework for Semantic Web Process Composition, – Technical Report 03-008, LSDIS Lab, Dept of Computer Science, UGA. June 2003.
- [Nern et al, 2004] H Joachim Nern, G. Agre, T. Atanasova, J. Saarela. System Framework for Generating Open Development Platforms for Web-Service Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems - INFRAWEB II. In: WSEAS TRANS. on INFORMATION SCIENCE and APPLICATIONS, 1, Vol. 1, 286-291, 2004.
- [Atanasova et al, 2005] Tatiana Atanasova, Gennady Agre, H Joachim Nern, INFRAWEB Semantic Web Unit for design and composition of Semantic Web Services INFRAWEB approach, In: Proc. 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005, IRIT, Université Paul Sabatier, Toulouse, France, April 2005.
- [Farahbod et al, 2004] Roozbeh Farahbod, Uwe Glässer and Mona Vajihollahi, Specification and Validation of the Business Process Execution Language for Web Services, Software Engineering Lab School of Computing Science, Simon Fraser University, Canada, Technical Report SFU-CMPT-TR-2003-06, Feb 2004
- [Börger, 1999] Egon Börger, High Level System Design and Analysis using Abstract State Machines. Current Trends in Applied Formal Methods (FM-Trends 98). Springer LNCS 1641, 1999.
- [Börger, 1997] Egon Börger, Integrating ASMs into the Software Development Life Cycle, Journal of Universal Computer Science, vol. 3, no., 603-665, Springer Pub. Co., 1997
- [Gurevich, 1995] E. Börger (ed.): Yuri Gurevich: Evolving Algebras 1993: Lipari Guide, Specification and Validation Methods, Oxford University Press, 1995, 9-36.
- [Mandell, 2003] Mandell, D.J., McIlraith, S.A.: Adapting BPEL4WS for the semantic web: The bottom-up approach to web service interoperation. In: Proc. of the 2nd Int. Semantic Web Conf. (ISWC), 2003.
- [WSMO Specification, 2005] Roman, D.; Lausen, H.; Keller, U. (eds.): Web Service Modeling Ontology, WSMO Working Draft D2, final version 1.2, 13 April 2005.
- [WSMO Choreography and Orchestration] Roman, D.; Scicluna, J., Feier, C. (eds.): Ontology-based Choreography and Orchestration of WSMO Services, WSMO Working Draft D14, 01 March 2005.
- [Berners-Lee et al. 2001] Tim Berners-Lee, James Hendler, and Ora Lassila, "The Semantic Web". Scientific American, 284(5):34-43, 2001.
- [Zamulin, 2003] A. Zamulin, A state-based semantics of Pascal-like language, IIS, Novosibirsk, 2003
- [KM-GOV, 2005] Semantic Interoperability Community of Practice –Introducing Semantic Technologies and the Vision of the Semantic Web, White Paper Series Module, KM-GOV, 2005
- [Patil and Wagh, 2002] N. Patil & S. Wagh, Automating Business Process & SOA Testing using Optimyz WebServiceTester™, 2002.

Author's Information

Tatiana Atanasova – Institute of Information Technologies, Acad. G. Bonchev 2, 1113 Sofia, Bulgaria, e-mail: atanasova@iinf.bas.bg

ADJUSTING WSMO API REFERENCE IMPLEMENTATION TO SUPPORT MORE RELIABLE ENTITY PERSISTENCE¹

Ivo Marinchev

Abstract: In the presented paper we scrutinize the persistence facilities provided by the WSMO API reference implementation. It is shown that its current file data-store persistence is not very reliable by design. Our ultimate goal is to explore the possibilities of extending the current persistence implementation (as an easy short-run solution) and implementing a different persistent package from scratch (possible long-run solution) that is more reliable and useful. In order to avoid "reinventing the wheel", we decided to use relational database management system to store WSMO API internal object model. It is shown later that the first task can be easily achieved although in not very elegant way, but we think that the later one requires some changes in the WSMO API to smooth out some inconsistencies in the WSMO API specification in respect to other widely used Java technologies and frameworks.

Keywords: Semantic Web Services, Web Service Modelling Ontology (WSMO), WSMO API, WSMO4J.

ACM Classification Keywords: H.3.2 Information Storage: File organization; I.2.4 Knowledge Representation Formalisms and Methods: Representation languages

Introduction

Web services are defining a new paradigm for the Web in which a network of computer programs becomes the consumer of information. However, Web service technologies only describe the syntactical aspects of a Web service and, therefore, only provide a set of rigid services that cannot be adapted to a changing environment without human intervention. Realization of the full potential of the Web services and associated service oriented architecture requires further technological advances in the areas of service interoperation, service discovery, service composition and orchestration. A possible solution to all these problems is likely to be provided by converting Web services to *Semantic Web* services. *Semantic Web* services are "self-contained, self-describing, semantically marked-up software resources that can be published, discovered, composed and executed across the Web in a task driven semi-automatic way" [Arroyo et al, 2004].

There are two major initiatives aiming at developing world-wide standard for the semantic description of Web services. The first one is OWL-S [OWL-S, 2004], a collaborative effort by BBN Technologies, Carnegie Mellon University, Nokia, Stanford University, SRI International and Yale University. The second one is Web Service Modelling Ontology (WSMO) [Roman et al, 2004], a European initiative intending to create an ontology for describing various aspects related to Semantic Web Services and to solve the integration problem.

As part of the later initiative the WSMO API specification and reference implementation [WSMO4J] has been developed. WSMO4J is an API and a reference implementation for building Semantic Web Services applications compliant with the Web Service Modeling Ontology. WSMO4J is compliant with the WSMO v1.0 specification [WSMO v1.0] (20 Sep 2004). At the time of this writing the WSMO API reference implementation is version 0.3 (alpha), that means that it is far from completed product and is subject to changes without prior notice. Nevertheless the API is incomplete as part of our work on the INFRAWEBs project [Nern et al, 2004] we have to utilize the WSMO4J package as it is the only available working implementation of the WSMO specifications.

¹ The research has been partially supported by INFRAWEBs - IST FP62003/IST/2.3.2.3 Research Project No. 511723 and "Technologies of the Information Society for Knowledge Processing and Management" - IIT-BAS Research Project No. 010061.

Current State of the Art

At the time of this writing the reference implementation of the WSMO API [WSMO4J] can export its internal data model to a set of binary files organized in a bundle of directories that correspond to the major entity (entities that are identifiable according to WSMO API terms) types. Every identifiable entity is saved as a separate file that contains the serialized entity identifier, then the Java object that represents the entity and at the end several lists of identifiers corresponding to the different groups of entities that are subordinates of the current entity. The order in which these lists are serialized to the output file (stream) is implementation specific and is implemented by several internal classes named entity-type processors. Every entity type has separated processor class that serializes/deserializes the corresponding objects to/from their persistent state. The subordinate entities are stored in the same way in separate files and so on. Fig. 1 shows the directories created by the file data-store. In practice, this simple solution appears to be very unreliable and a relatively small problem may incur enormous data-losses. The reader also has to keep in mind that this storage mechanism is intended to be used for processing ontologies. And all of the ontologies that are applicable to real-world problem tend to be extremely large.

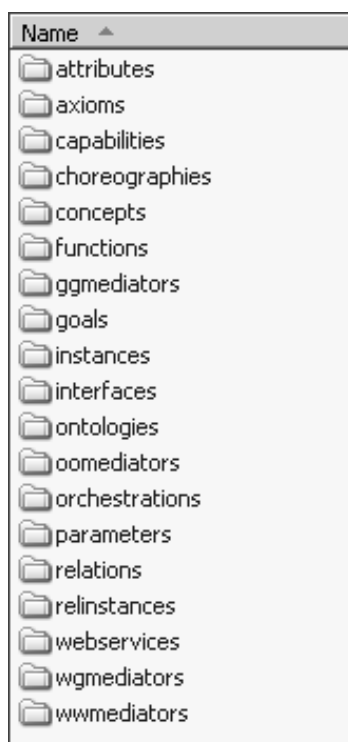


Fig. 1: Directories created by the "FileDataStore" class of the reference implementation.

This ad hoc solution is simple and does not require any third party programs/libraries but it has several major drawbacks that prevent it to be used in a production system. Some of them are implementation problems and can be easily circumvented but others require rather sophisticated solutions to general purpose problems. Below we enumerate some of the problems:

1. If a certain entity is loaded, all its dependent entities are loaded as well no matter whether they are needed or not.
2. If any object has to be changed all its subordinate objects are overwritten again no matter if they are changed or not. Moreover because the implementation of the ObjectOutputStream in Java serializes objects by reachability (when an object is serialized all objects that are referenced by it are serialized as well) a certain object is serialized as many times as the number of objects that refers to it in *direct and indirect way!* Such persistence scheme brings enormous excessive overhead in the serialization of large object graphs and the worst is that the overhead increases exponentially with the size of the object graph. A possible solution to the later problem is all objects to implement Externalizable interface in order to control their serialization process but even it will not remove all unneeded read/write operations.
3. In the current implementation, the file names correspond to the entity identifiers. These identifiers may become rather long. This is especially important as many identifiers are created from URLs and at the same time, many of the entities have the identifiers that extend the identifier of its parent entity (for example axioms that are part of the ontology). The implementation makes the file names additionally longer by encoding some special characters that may be prohibited by certain file systems (for example / is encoded with .fslash., : with .colon., * with .star.). At the same time most of the file systems do not allow file names with more than 255 characters.
4. The store operations are not atomic. Thus, there is no guarantee that the data will remain consistent and the original object graph can be recreated from its serialized state. For example if an exception occurs when the data is being saved, the operation is terminated and the on-disk structures remain in an unpredictable and

undeterminable state – the user neither can fix them, nor can turn them to their state before the last operation occur (roll back the last operation).

5. The store implementation is not thread-safe due to the usage of fields to transfer data between methods. But even that one can instantiate several different data-store objects they can not work on the same store simultaneously because of the lack of any locking or synchronizations.

In short, if we use database terminology, the current implementation is very far from being ACID¹ compliant. It is obvious that any of the above issues can be solved but the solutions are usually very sophisticated, and one ultimately will implement complete transactional database storage engine in order to solve all of them.

Using Relational Database as a Data Store

The first improvement that we implemented was to move the persistent data to the relational database by just replacing the file data-store directories with the database tables. The tables consist of two columns: the first one for the entity identifier, and the second column of type BLOB (Binary Large Object) that stores the serialized Java objects. This extension was relatively easy to be implemented. We changed several private methods that deal with the file names and entity types (getFileNameFor, getEntityType, etc.) to work with the database tables and records, and then we changed all of methods that serialize and deserialize data to store/load it to the corresponding BLOB fields instead of using file output and input streams. In order to be able to use the transaction facilities provided by all modern relational database management systems (RDBMS) we use a single database connection that is initiated at the beginning of the store process and get committed at its end (or rolled back in case of exceptional circumstances).

Even with these simple modifications, we get several important advantages:

1. We get atomic changes – all of the changes are written or all are discarded at once.
2. No data is lost in case the storing gets terminated - not only by checked exception but even if the whole process is terminated by the unchecked one.
3. The store may be physically located on the remote system.
4. Several different client processes can use the store simultaneously.
5. The store may use the back-up, replication, and clustering facilities provided by the underlying RDBMS.
6. The store uses the data caching provided by the database.
7. The database can be changed at will if one does not use proprietary database extensions.

Avoiding Identity-Lists Serialization

The next logical consequent step is to start removing object serializations. As it was discussed earlier when a certain entity is persisted the file data-store in the reference implementation serializes first the entity identifier, then the entity object and at the end several lists of the identifiers of the entities that depend on the current one in the entity hierarchy. This last step is entity type specific and is implemented by a specialized entity processors for different type of entities that take care of saving, loading the lists (Vectors in Java terms) of identifiers.

Using the information from these entity processor classes, we created a separate table columns that hold the lists of identities of a given type. For example, for the “capability processor”, we created columns for Assumptions list, Pre-Conditions list, Post-Conditions lists, and Effects list. Thus, the content of the database tables gets more human readable and it is easier to debug potential problems, but we have to emphasize that the database is still not even in the first normal form (1NF) - it requires all table columns to be atomic.

Utilizing Object-Relational Frameworks

The newly created columns in fact represent the relationships between the entities represented by the corresponding table rows and their dependent entities. That is why we can remove these columns and replace them with foreign key columns in the dependent tables for one-to-many relationships and with relationship tables

¹ ACID – Atomicity (states that database modifications must follow an “all or nothing” rule), Consistency (states that only valid data will be written to the database), Isolation (requires that multiple transactions occurring at the same time not impact each other’s execution), Durability (ensures that any transaction committed to the database will not be lost).

for the many-to-many relationships. Dealing with one-to-many relationships with hand-written code is boring but not a complicated task. But the many-to-many relationships can become really problematic to be manipulated as they use additional (relationship) tables and the WSMO object model even have many-to-many reflexive relationships (for example the one between concepts and sub-concepts of the ontology). These facts imply that the programming code needed to deal with all these "housekeeping" activities will be much more than the code that implements the actual business logic.

At this point one can realize that the required changes to the original reference implementation become rather complex and in fact we start "reinventing the wheel" that is already created by others. So, the wise approach to this problem is to use some object-relational mapping frameworks to do the work for us. There are a lot of such frameworks available (for example Java Data Objects [JDO] implementations, Hibernate [Hibernate], Oracle TopLink [TopLink], and others) and many of them are open-source and free even for commercial use. The common feature of all these frameworks is that they use XML configuration files to specify how the objects, fields and relationships (object model) are mapped to the corresponding database tables and columns (relational model). The basic idea behind these mapping files is to keep the object model and the relational model loosely coupled so that the two models can be changed independently. Utilizing such framework provides other useful features:

1. Automatic generation of database queries;
2. Loading/saving/updating the complete object graph with a single method call;
3. Lazy-loading (or on-demand loading)¹;
4. Tracking the user changes and updating just the changed fields;
5. Support for many different RDBMS;
6. Object caching - even distributed caching is possible;
7. Other specific features.

For our purpose the lazy-loading is extremely useful because if one wants to load and change a certain entity it does not need to load and save the complete sub-graph that originates from this entity.

Unfortunately, it appears that several significant issues arise in any attempt of integration between object-relational mapping framework and the current version of the WSMO API and its reference implementation. These issues are discussed in the next section. At the end of it, we represent one possible solution of the problem and why we think the proposed changes are appropriate.

Problems with the Current Version of WSMO API and its Reference Implementation

The most serious problem concerning the applicability of the OR mapping framework with the WSMO API (and its reference implementation) is that the object-relational frameworks work with JavaBeans classes/objects. We do not know why WSMO API was specified and implemented in its current form, but the fact is that all of the entity classes in it deviate from the JavaBeans specification [JavaBeans]. Specifically they lack the properly named accessor and mutator methods for the non-primitive types. At the same time, all methods for accessing non-primitive types are named as listXXX. We do not know why such naming scheme has been selected but we think that it is even not very intuitive. The worst is that at the same time listXXX methods return value is of type `java.util.Set`. In fact, the following issues appear:

1. The names of the property accessor methods are misleading for the user and deviate from other well-known framework and the JavaBean specification.
2. The semantics of the Set and List data types are significantly different as the list is an ordered collection of elements. More over unlike sets, lists typically allow duplicate elements. More formally, lists typically allow pairs of elements e_1 and e_2 such that $e_1.equals(e_2)$, and they typically allow multiple null

¹ The framework loads the expensive (in memory footprint and construction time) object fields and referenced objects just before they are accessed by the client program. This feature is usually very flexible and can be configured in the mapping files on a field level.

elements if they allow null elements at all. So the words list and set is not very appropriate to be used interchangeably.

3. Using "un-typed" return types in the listXXX methods in the otherwise very strongly typed specification is somewhat strange decision.

It is true that we can overcome the first problem by sub-classing all needed entity classes of the reference implementation and turn them to regular JavaBeans by adding the missing accessor and mutator methods and then create mappings for the newly introduced classes. But we do not want any consequent version of the reference implementation to break our "extension", or to require conversions of the database schema. So, this solution does not seem appropriate in the long-run.

At the end we will express our inner conviction that the persistence package has to be as loosely coupled as possible to the rest of the implementation, and to be written as much as possible against the specification not against the implementation as it is now. In the confirmation of the later we propose the WSMO API specification to be changed in the following way:

1. Add the missing get/set methods to the entity interfaces to turn the implementation classes in correct JavaBeans.
2. Introduce type-safe sets for any entity type that is needed and return them in the corresponding property accessor methods instead of java.util.Set.

Conclusion

As a conclusion we want to point out that although it is in its early stage of development and the fact that it is or still may be immature in some of its parts, the WSMO4J is sound enough to be used as a development tool in the research projects, and facilitates researchers in the early adoption of the WSMO related technologies. We hope that the WSMO API working group will take into account our remarks and suggestions and even they are rejected they will contribute in some way in the future improvements of the specification and/or implementation.

Bibliography

- [Arroyo et al, 2004] Arroyo, S., Lara, R., Gomez, J. M., Bereka, D., Ding, Y., Fensel, D. Semantic Aspects of Web Services. In: Practical Handbook of Internet Computing. Munindar P. (Editor). Chapman Hall and CRC Press, Baton Rouge, 2004
- [JDO] <http://java.sun.com/products/jdo/>
- [Hibernate] <http://www.hibernate.org>
- [JavaBeans] <http://java.sun.com/products/javabeans/>
- [Nern et al. 2004] H.-Joachim Nern, G. Agre, T. Atanansova, J. Saarela. System Framework for Generating Open Development Platforms for Web-Service Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems - INFRAWEBs II. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS*, ISSN 1790-0832, Issue 1, Volume 1, July 2004, 286-291.
- [OWL-S, 2004] The OWL Services Coalition: OWL-S: Semantic Markup for Web Services, *version 1.0*; available at <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>
- [Roman et al, 2004] D. Roman, U. Keller, H. Lausen (eds.): Web Service Modeling Ontology (WSMO), version 0.1; available at: <http://nextwebgeneration.com/projects/wsmo/2004/d4/d4.1/v01/index.html>
- [TopLink] <http://www.oracle.com/technology/products/ias/toplink/>
- [WSMO v1.0] <http://www.wsmo.org/2004/d2/v1.0/20040920/>
- [WSMO4J] <http://wsmo4j.sourceforge.net>

Author's Information

Ivo Marinchev – Institute of Information Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 29A, Sofia-1113, Bulgaria; e-mail: ivo@iinf.bas.bg

TABLE OF CONTENTS OF VOLUME 13, NUMBER 2

A Multicriteria Decision Support System <i>MultiDecision-1</i>	103
<i>Vassil Vassilev, Krasimira Genova, Mariyana Vassileva</i>	
Generalized Scalarizing Problems <i>GENS</i> and <i>GENSLex</i> of Multicriteria Optimization	111
<i>Mariyana Vassileva</i>	
Software Development for Distributed System of Russian Databases for Electronics Materials.....	121
<i>Valery Kornyshko, Victor Dudarev</i>	
The Information-analytical System for Diagnostics of Aircraft Navigation Units	127
<i>Ilya Prokoshev, Vyacheslav Suminov</i>	
Using ORG-Master for Knowledge Based Organizational Change	131
<i>Dmitry Kudryavtsev, Lev Grigoriev, Valentina Kislova, Alexey Zablotsky</i>	
Neural Network Based Approach for Developing the Enterprise Strategy	139
<i>Todorka Kovacheva, Daniela Toshkova</i>	
Generalization by Computation Through Memory.....	145
<i>Petro Gopych</i>	
FP6-ist INFRAWEBs European Research Project.....	158
INFRAWEBs Semantic Web Service Development on the Base of Knowledge Management Layer	161
<i>Joachim Nern, Gennady Agre, Tatiana Atanasova, Zlatina Marinova, András Micsik, László Kovács, Janne Saarela, Timo Westkaemper</i>	
INFRAWEBs Axiom Editor – A Graphical Ontology-Driven Tool for Creating Complex Logical Expressions	169
<i>Gennady Agre, Petar Kormushev, Ivan Dilov</i>	
A Survey on the Integration of Enterprise Applications as a Service.....	179
<i>Hristina Daskalova, Vladislava Grigorova</i>	
Semantic Description of Web Services and Possibilities of BPEL4WS.....	183
<i>Vladislava Grigorova</i>	
INFRAWEBs BPEL-Based Editor for Creating the Semantic Web Services Description.....	188
<i>Tatiana Atanasova</i>	
Adjusting WSMO API Reference Implementation to Support More Reliable Entity Persistence	195
<i>Ivo Marinchev</i>	
Table of Contents of Volume 13, Number 2.....	200