Doktorandské dny ’09

Ústav informatiky AV ČR, v. v. i.

Jizerka

21. – 23. září 2009

vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze

DRAFT
14. srpna 2009 - 14:32

Ústav informatiky AV ČR, v. v. i., Pod Vodárenskou věží 2, 182 07 Praha 8

DRAFT
14. srpna 2009 - 14:32

Všechna práva vyhrazena. Tato publikace ani žádná její část nesmí být reprodukována nebo šířena v žádné formě, elektronické nebo mechanické, včetně fotokopií, bez písemného souhlasu vydavatele.

© Ústav informatiky AV ČR, v. v. i., 2009
© MATFYZPRESS, vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze 2009

ISBN *– not yet –*

Doktorandské dny Ústavu informatiky AV ČR, v. v. i., se konají již počtrnácté, nepřetržitě od roku 1996. Tento seminář poskytuje doktorandum, podílejícím se na odborných aktivitách Ústavu informatiky, možnost prezentovat výsledky jejich odborného studia. Současně poskytuje prostor pro oponentní připomínky k přednášené tematice a použité metodologii práce ze strany přítomné odborné komunity.

Z jiného úhlu pohledu, toto setkání doktorandu podává pruřezovou informaci o odborném rozsahu pedagogických aktivit, které jsou realizovány na pracovištích či za spoluúčasti Ústavu informatiky.

Jednotlivé příspěvky sborníku jsou uspořádány podle jmen autoru. Uspořádání podle tematického zaměření nepovažujeme za účelné, vzhledem k rozmanitosti jednotlivých témat.

Vedení Ústavu informatiky jakožto organizátor doktorandských dnu věří, že toto setkání mladých doktorandu, jejich školitelu a ostatní odborné veřejnosti povede ke zkvalitnění celého procesu doktorandského studia zajišťovaného v součinnosti s Ústavem informatiky a v neposlední řadě k navázání a vyhledání nových odborných kontaktu.

1. září 2009

DRAFT
14. srpna 2009, 14:32

Obsah

Branislav Bošanský:
Medical Processes Agent-Based Critiquing System **5**

Jan Dědek:
Fuzzy Classification of Web Reports with Linguistic Text Mining **12**

Jakub Dvořák:
Porovnání optimalizačních metod pro změkčování rozhodovacího stromu **15**

Tomáš Dzetkulič:
Verification of Hybrid Systems Using Slices of Parallel Hyperplanes **21**

Alan Eckhardt:
How to Learn Fuzzy User Preferences with Variable Objectives **22**

Alena Gregová:
Modulárne ontológie **24**

Lukáš Hošek:
Gradient Learning of Spiking Neural Networks **29**

Karel Chvalovský:
Syntactic Approach to Fuzzy Modal Logics in MTL **35**

František Jahoda:
Signature Provenance obtained from the Ontology Provenance **44**

Kateřina Jurková:
Cost Functions for Graph Repartitionings Motivated by Factorization **47**

Robert Kessl:
Parallel Mining of Frequent Itemsets **53**

DRAFT
14. srpna 2009 – 14:32

Tomáš Kulhánek:
Virtual Distributed Environment for Exchange of Medical Images **62**

Miroslav Nagy:
Clinical Contents Harmonization of EHRs and its Relation to Semantic Interoperability **65**

Radim Nedbal:
Preference Handling in Relational Query Languages **75**

Vendula Papíková:
Databáze biomedicínských informačních zdrojů **83**

Milan Petrík:
Properties of Fuzzy Logical Operations **89**

Petra Přečková:
Mezinárodní klasifikace nemocí a její využití v Minimálním datovém modelu pro kardiologii **97**

Martin Řimnáč:
Experimenty s RDF úložištěm dat a reputacemi zdrojů **102**

Stanislav Slušný:
Pose Estimation Algorithms Based on Particle Filters **103**

Petra Šeflová:
Metody modularizace rozsáhlých ontologií **108**

David Štefka:
Assessing Classification Confidence Measures in Dynamic Classifier Systems **113**

Pavel Tyl:
COMP – Comparison of Matched Ontologies in Protégé **125**

Karel Zvára:
Information Extraction from Medical Texts **126**

Miroslav Zvolský:
Základní parametry dokumentů doporučených postupů českých lékařských společností publikovaných prostřednictvím Internetu **128**

DRAFT
14. srpna 2009 - - 14:32

DRAFT
14. srpna 2009 -- 14:32

Medical Processes Agent-Based Critiquing System

Post-Graduate Student:
MGR. BRANISLAV BOŠANSKÝ

Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2

182 07 Prague 8, CZ

bosansky@euromise.cz

Supervisor:
DOC. ING. LENKA LHOTSKÁ, CSC.

Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2

166 27 Prague 6, CZ

lhotska@labe.felk.cvut.cz

Field of Study:
Biomedical Informatics

This research was partially supported by the project of the Institute of Computer Science of Academy of Sciences AV0Z10300504, the project of the Ministry of Education of the Czech Republic No. 1M06014 and by the research program No. MSM 6840770012 "Transdisciplinary Research in Biomedical Engineering II" of the CTU in Prague.

Abstract

Processes and process modelling have proven themselves as a useful technique for capturing the work practice in business. We focus on their usage in the domain of healthcare and define two main types of processes in medicine – medical guidelines and organizational processes. Based on these types we present the architecture of a multi-agent system that is able to work with them and describe application of this multi-agent system as a critiquing decision support system for healthcare specialists.

1. Introduction

Developement of a system that will support the decision making of physicians and healthcare specialists is a long-term goal for researchers in artificial intelligence. Recently, there has been given an emphasis to monitoring systems that control and evaluate current situation (e.g. patient data, therapy, etc.) and alerts the medical staff in case of inconsistencies or possible danger. In order to recognize the occurrence of these situations such systems need to operate with appropriate knowledge. In the healthcare domain they can profit from medical guidelines which are sets of directions or principles that assist the physician [1] and are considered to be a good approach to standardize and improve health care [2]. When formalized, i.e. captured in a computer-interpretable form, they are being used in various decision support systems (e.g. in HeCaSe2 [3]).

The medical guidelines, however, can be seen as a specific way of process modelling. In our research we want to develop a system that would be able to work with knowledge captured in form of general processes – i.e. as with formalized medical guidelines, but also with organizational processes which are specific in each healthcare facility (e.g. activities necessary for transferring a patient from one department to a different one). Both of these types of processes were usually considered separately which resulted in different languages and different approaches (e.g. using Event-Driven Process Chains (EPC) to model organizational processes and GLIF for medical guidelines). In this paper we present the architecture of a multi-agent system that (1) is able to work with these general processes in healthcare domain, (2) can simulate them in given environment opening that way a possibility for future planning or process reengineering, and finally (3) can act as a critiquing and monitoring system that controls their adherence and can alert the medical staff.

The paper is organized as follows: in Section 2 we define the problem and theoretical foundations together with related approaches. Section 3 is focused on description of the architecture of the multi-agent system and behaviour of single agents. We describe the usage of the multi-agent system as a process-critiquing system in Section 4, following by an illustrative example and implementation issues in Section 5. We conclude and discuss the future work in Section 6.

2. Processes in Medicine and Related Work

The work practice (i.e. duties of employees, organizational procedures, specification of the order of activities, or necessary resources for each activity) can be captured using process modelling technique – i.e. as a sequence of actions, states, decision points, or steps splitting or joining the sequence. There are

DRAFT
PhD Conference 2009, 1.4.2009

various levels of processes in medical domain and with respect to terminology in [4] we can differentiate the *organizational processes* and the *medical treatment processes*.

2.1. Organizational Processes

The organizational processes in the healthcare domain are closely related to processes in other business areas, where the work practice has been captured for a long time using business process modelling languages. There are several studies [4, 5, 6] that analyze the problems of applying process modelling or usage of workflow management systems in medical care. They all agree that the implementation of this approach can improve current problems with organization, reduce the time of hospitalization and finally reduce the costs. However, they also point out, that till now, usage of processes is rather low and insufficient. The main reasons are more complex processes than in other fields of industry, or problems with interoperability resulting from inconsistencies of databases and used ontology or protocols. Finally, the captured work practice in healthcare is often very variable and closely depends on specific treatment of the patient. All these factors complicate successfull usage of classical business process management, or workflow management systems. Therefore, while working with organizational processes, we also need to take medical treatment of patients into consideration as well.

2.2. Medical Guidelines

Standardization of medical treatment processes is being done for a long time now known as medical guidelines. They contain recommended actions, directions, and principles for specific diseases, and they are all approved by appropriate expert committees helping that way physicians with clinical decisions. Several crucial positive factors have been identified when using guidelines [1]:

- they improve the quality of decisions as healthcare professionals can consult complicated situations in unknown areas and minimize the risk for a patient (e.g. to forget an examination that is important for this patient according to her/his condition)
- they are based on evidence-based medicine and help to reuse and disseminate the knowledge
- they help to standardize provided health care

However, the standard method of work with the guidelines (such as consulting, or using in practice) is solely based on a textual form. This, on one hand, helps the healthcare professionals to capture the knowledge in a straightforward way. On the other hand, such approach brings several complications. It is hard for physicians to do a quick consultation with the guideline during the examination of a patient, or to keep up with the relevant changes in new versions of the document.

Therefore a wide part of research in biomedical informatics is related to the formalization of medical guidelines into an electronic form. There are several workgroups and several languages (PRO*forma*, GLIF, Asbru, etc.) that captures the knowledge of a textual medical guideline into an electronic and structured form. All of them focus on specific parts – e.g. logic background in Asbru, or automatic execution and patient data retrievement in GLIF. They are all based on a process-oriented approach and specify the guideline as a sequence of actions, states, decision, or synchronization points. Research in decision support systems that work with formalized medical guidelines focuses mostly on acquisition, verification, or automatic execution of guidelines [1].

2.3. Related Work

The area of medical guidelines' execution is closest to our problem. There are several systems that can connect the guideline with the patient's health record, retrieve and store appropriate data and guide the physician by executing next steps and waiting for appropriate data to be entered. Within these systems, only a few ones profit from principles of multi-agent systems: Arezzo™[7], HeCaSe2 [3], or the work presented in [8].

Our approach differs from existing systems in several ways: firstly, as the guidelines as such are transformed into agents, which allows simultaneous work with a set of guidelines, not only with the selected one as in existing work. Secondly, our system is based on more general concept, therefore beside monitoring the proceeding of the guideline, it can also be used for simulation or general computing purposes. Finally, thanks to the distribution of knowledge, agents can focus on the specific activities.

3. Process-Based Multi-Agent System

In this section we present the architecture and the functioning of the multi-agent system (MAS) that realize the critiquing system. The architecture is based on the one presented in [9] later enhanced in [10]. As discussed in Section 2, the architecture is more general and it can be used on simulating other process-based systems as well.

DRAFT 4. apr 2009 14:32

The architecture and different types of agents are depicted in Figure 1. Let us now describe these agents and their purpose more in detail.

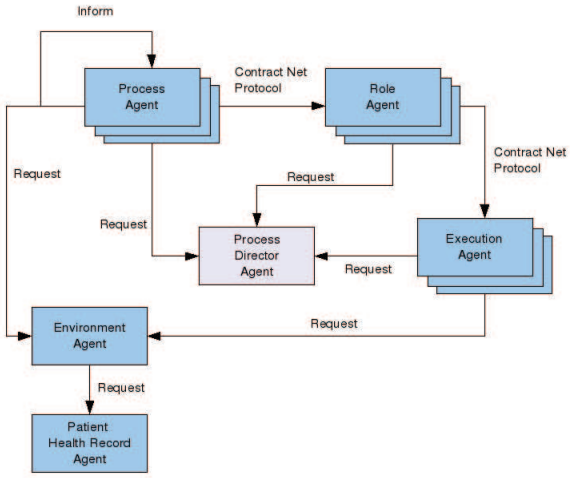


Figure 1: The architecture of a multi-agent system that is able to work (e.g. simulate, critique) with processes.

3.1. Environment Agent

Every agent-based simulation is situated in some environment which is represented by the Environment Agent in this architecture. With respect to the level of detail that we want model using this system the environment could represent the virtual world (e.g. a department of a hospital, etc.) with existing objects (e.g. RTG or EEG machines, wheel-chairs, beds, etc.).

3.2. Execution Agents

Execution Agents (EA) are representing concrete physicians, nurses, patients, or other employees of the facility that are involved in the processes. These agents are based on a reactive architecture in the form of hierarchical rules, which can be automatically generated based on possible activities that the specific EA can participate in. Each EA has several pre-defined rules, that for basic behaviour in the environment (i.e. responding to messages sent by other agents, sending appropriate messages to the Environment Agent during the execution of the activity, etc.). Then, for each activity that the agent (hence the represented person) can participate in, one additional rule is generated. These rules can be activated (when the condition for the process execution are met, and the EA can perform this action) or deactivated (execution of this process is no longer possible) by a message sent by appropriate Role Agent (see below). Finally, the Execution Agent autonomously chooses which of the activated processes it will execute based on the priority in which the rules are ordered.

3.3. Role Agents

Role Agents (RA) represent the roles in the environment (i.e. general roles for patient, nurse, physician, etc.). RA receives the proposal from a Process Agent (see below) and finds appropriate Execution Agent(s) (EA). The reason of using special agents for roles is in a typical usage of hierarchical structure of roles at workplace (e.g. a secretary, a nurse, or a doctor are all also employees, etc.). Therefore, when a RA receives a proposal from a Process Agent, it starts to find the appropriate EA between agents that posses this role (using contract-net protocol (CNP)), but also roles, that are more general in the hierarchy.

If multiple EAs are able to execute given activity and only one is needed, RA will choose the most suitable of them according to its internal rules, which are always domain or role dependent (e.g. in a simulation that occurs in some virtual world, the EA that is closest to the place of execution can be notified, or in another case the EA that is currently idle).

3.4. Process Agent

For every step in the process notation (i.e. activity, event, decision point, etc.) there is one Process Agent (PA) created in the system. The PA is responsible for a proper execution of the activity. Firstly it controls whether the initial conditions for the process are met: if the predecessing PA has successfully finished its execution, if all input objects have the needed values (using simple request protocol to Environment Agent), and if there exist appropriate agents that will execute this action (using CNP to those RAs that are connected with this activity). When all mandatory conditions hold, the PA starts the execution of the process (e.g. the simulation, calculation or a decision process, etc.) and after successfull finish, the PA is responsible for notifying the Environment Agent about the results of the activity (using simple request protocol) and the next succeeding Process Agent about the successfull finish (using simple inform protocol). Our approach takes into account the possibility of temporal suspension of the activity and reflecting the partial results in the environment, replacing the EA with another, coordination of several EAs participating on a single activity, or optional input objects.

Note, that each step of the process has its Process Agent – i.e. not only active steps (steps that represent activities as such) by also so called passive steps (usually related to the events (in EPC) or patient state (in GLIF)), flow-splitting (i.e. decision points), and flow-joining elements have appropriate Process Agent as well.

4. Critiquing System

We have described in detail the architecture of the multi-agent system that is able to work (e.g. simulate them) with the processes. One of possible application of such multi-agent system can be in critiquing – monitoring the correct execution of processes such as formalized medical guidelines or organizational processes in healthcare facilities.

We accentuate the Process Agents (PAs) and description of their behaviour, while other agents behave exactly in the way described in previous section. The main idea is that each PA is responsible for one step in the guideline, it monitors data fields in patient's health record related to the given step, and tries to estimate the outcome of the step simulating that way future developement of diagnostics or therapy. Whenever appropriate input data changes PAs update predicted output values and simulate the process further. Therefore, whenever the output data fields are changed by the physician in the way which PA has not expected an alert occurs.

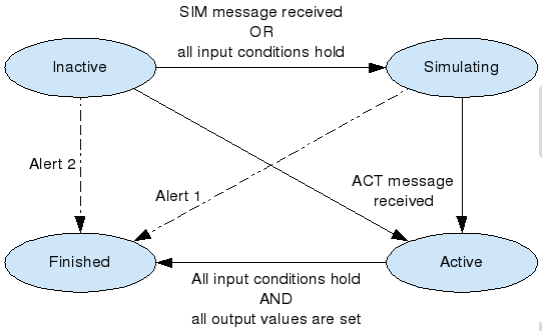


Figure 2: The states of Process Agents during critiquing. The solid arrow indicates valid transition, the dashed arrows indicate possible inconsistencies.

Let us now describe the critiquing more in detail. We distinguish four basic states of a PA (see Figure 2) – *inactive*, *simulating*, *active*, and *finished*. At the beginning, each PA is in the *inactive* state. PA in this state behaves the same way as in simulation before the execution of the activity – it periodically checks the objects in input condition whether they hold. In the critiquing phase therefore is PA periodically checking the associated fields in patient's data model (such as blood pressure, height, etc.) together with the message from predecessing PA (whether it has finished the activity or not).

The agent can get to the *simulating* state when at least one of two following conditions holds: (1) the agent receives the SIM message (i.e. predecessing agent has finished the simulation of the process), or (2) all input conditions for the process execution are met, and the agent has not received ACT message from its predecessor. When the PA is in *simulating* state, it checks again all of its input conditions and in case that some of them are not evaluable (i.e. data in the patient's data record are missing), they are estimated using k-means technique with respect to other patients' data. Such an estimation is necessary for proper running the correspondent action (e.g. setting the concrete diagnose, measuring the blood pressure, etc.) that would yield the simulation output of the process that can be temporarily stored in the simulation environment (but not the patient's data record) and other PAs can work with them. After finishing the simulation of the process, the PA sends a SIM message to the appropriate successor meaning the simulation of its activity has finished.

The agent gets to the *active* state when it receives the ACT message from its predecessor. In this state the PA behaves very similarly to *simulating* state with one difference: in case that all input conditions are met and the output value has been updated in the patient's data record (by the doctor), the agent moves to *finished* state and sends the ACT message to the appropriate successor.

The alert for the doctor occurs in the case when the output values of the process are updated but the agent is not in the *active* state. This can happen because of (1) the step was executed before its predecessors were successfully finished, or (2) the step was not expected to be executed. We can recognize these cases based on the current state the PA is in, when the output values are updated. For the first case the PA would be in *simulating* state, for the second case the PA would be in *inactive* state.

5. Experiments and Implementation

In this section we present an illustrative example, which is the basis for our preliminary experiments of presented process agent-based critiquing system. We demonstrate a possible application using a simplified version of the guideline for a hypertension treatment following by the biref description of the implementation details.

5.1. Guideline Critiquing

In Figure 3 we depicted a very simplified version of a hypertension guideline for demonstrating exemplary situations that can arise during the critiquing of medical processes. Note, that the guideline is simplified for explanatory reasons and in the system full medical processes representing the real diagnosis and

therapy processes (corresponding to formalized medical guidelines used in practice) would be used. Moreover, the description for two decision steps are shortened: (*) under the term "patient with high pressure" we understand a patient with blood pressure value at least 180/110 (values for systolic pressure/diastolic pressure), or at least one blood pressure value of at least 140/90 from three different sessions; (**) there are several possible complications for hypertension therapy such as SCORE value [11] over 5%, patient diagnosed with diabetes mellitus, and many others.

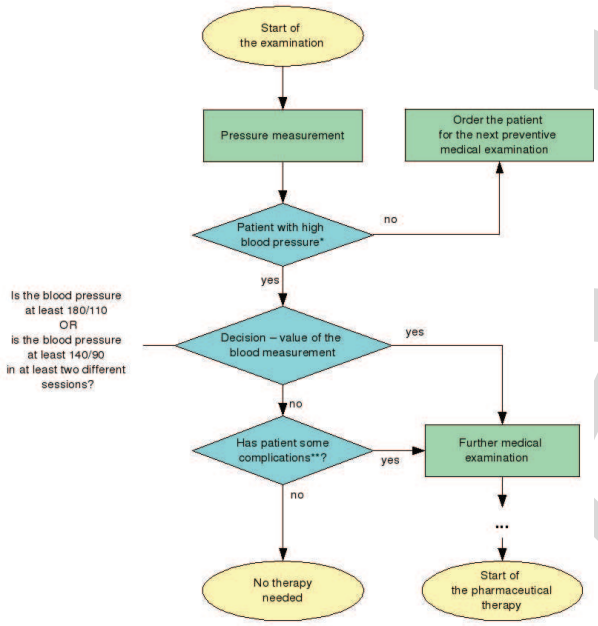


Figure 3: Simplified guideline for hypertension in GLIF

When a patient comes to a preventive examination (or he/she is examined during a longer stay in a hospital) his/her blood pressure is measured and then several decision steps (with possibly further necessary examination) is performed in order to decide whether to begin a pharmaceutical therapy or not. Let us now consider a patient that has a high value of blood pressure (over 180/110). After setting these values into patient's health record, Process Agents (PAs) in the right branch of the guideline would change their state to *simulating* as it is expected that this patient would be treated pharmaceutically[1]. However, in case the physician enters the data for a pharmaceutical treatment without performing further necessary examination, the PA associated with the "Start of pharmaceutical therapy" state would alert the system, as it would change its state in an unexpected way (from the *simulating* state into the *finished* state). In the other case, if the physician enters the data indicating no pharmaceutical treatment at all, the PA associated with the "No therapy needed" state would alert as it would unexpectedly change its state from *inactive* to *finished*.

The second type of alert can be more useful when a set of multiple medical processes is considered concurrently. Let us assume there also is a process describing a diagnosis and a therapy for diabetes mellitus present in the critiquing system. Now let us have a patient that has only one value of blood pressure over 140/90 and other values are from the interval 130-139/85-89. For such a patient no pharmaceutical therapy is needed in case he/she does not have any complications stated above. However, the patient could have results from previous laboratory examinations in his/her data record and in the process related to the diabetes mellitus diagnosis could diagnose this patient with a second type of diabetes mellitus. This diagnosis, as it is being only estimated by PAs in *simulating* state, is set in the environment using only Environment Agent, not storing this prediction in the patient's health record. Therefore the PA related to "Has patient some complications" would send the SIM message to the right branch of the guideline (hence the PA related to 'Start of pharmaceutical therapy" would be in *simulating* state) and the physician can be alerted when he/she indicates that there is no therapy needed.

5.2. Implementation

We implement described multi-agent system using the JADE framework[2], with the JADEX [12] as the extending reasoning engine for the agents. The implementation follows the architecture presented in previous section and depicted in Figure 1. Thanks to using JADE, the communication between agents is designed with respect to FIPA communication standards and as such can be extended with appropriate ontologies and communication standards in healthcare (e.g. designing the communication between the Environment Agent and Patient Health Record Agent with respect to HL7 version 3 standard).

During the implementation we decided not to follow the principles of offline transformation of the process knowledge into the rules for agents as described in [10]. In the approach presented in this paper each agent, that participates in the execution (i.e. Process Agents, Role Agents, and Execution Agents), requests the necessary information (e.g. predecessors of the Process Agent, necessary inputs, etc.) from the Process Director

[1]Note, that if further medical examination is needed, but has not been done yet, the PA connented to "Further medcial examination" would esimate the appropriate output values based on existing data from other patients and passes forward the SIM message.

[2]http://jade.tilab.com/

Agent (PDA). PDA reads the formalized processes in a relevant formalism (medical guidelines, organizational processes) and answers agents to their requests. This approach is equivalent to the offline transformation (by means of usage of processes), but more adaptive in case a change in processes occurs.

6. Discussion and Conclusion

In this paper we have presented the novel way of using the multi-agent system (MAS) as a technological framework for medical processes critiquing decision support system. The approach has several crucial advantages that differentiate it from existing approaches. Firstly, it uses the architecture of the MAS that can work with organizational processes and medical guidelines together. This creates a possibility to develop appropriate monitoring system that is able to control the work practice in a healthcare center jointly on several levels – the procedures for examination reservation or transportation of a patient on one hand, but also the treatment of specific diseases on the other one.

Secondly, it offers several possible ways how to alert healthcare personnel. In the Section 4 we described only the basic one regarding to correct sequence of the performed actions (i.e. whether executed action was executed before its predecessors or the action was not expected to be executed at all). However, thanks to the distributed nature of the system, it can be further improved and specific Process Agents can be enhanced with machine learning techniques that would also alert the doctor about the quality of the entered data.

Finally, such a system can also be used as a simulation tool for processes analysis during organizational process reengineering in healthcare environment, as it also can work with the appropriate medical knowledge, that is necessary to gaining proper results.

In future work, we intend to practically test the presented architecture as a critiquing system in a hospital department, to practically evaluate the approach and identify further improvements. Our critiquing system would focus on hypertension together with related diseases (such as diabetes mellitus and dyslipidemia).

References

[1] D. Isern and A. Moreno, "Computer-based execution of clinical guidelines: A review," *International Journal of Medical Informatics*, vol. 77, no. 12, pp. 787 – 808, 2008.

[2] R. Lenz, R. Blaser, M. Beyer, O. Heger, C. Biber, M. Baumlein, and M. Schnabel, "It support for clinical pathways–lessons learned," *International Journal of Medical Informatics*, vol. 76, no. Supplement 3, pp. S397 – S402, 2007. Ubiquity: Technologies for Better Health in Aging Societies - MIE 2006.

[3] D. Isern, D. Sánchez, and A. Moreno, "Hecase2: A multi-agent ontology-driven guideline enactment engine," in *CEEMAS '07: Proceedings of the 5th international Central and Eastern European conference on Multi-Agent Systems and Applications V*, (Berlin, Heidelberg), pp. 322–324, Springer-Verlag, 2007.

[4] R. Lenz and M. Reichert, "IT support for healthcare processes - premises, challenges, perspectives," *Data Knowl. Eng.*, vol. 61, no. 1, pp. 39–58, 2007.

[5] X. Song, B. Hwong, G. Matos, A. Rudorfer, C. Nelson, M. Han, and A. Girenkov, "Understanding requirements for computer-aided healthcare workflows: experiences and challenges," in *ICSE '06: Proceedings of the 28th international conference on Software engineering*, (New York, NY, USA), pp. 930–934, ACM, 2006.

[6] A. Kumar, B. Smith, M. Pisanelli, A. Gangemi, and M. Stefanelli, "Clinical guidelines as plans: An ontological theory," *Methods of Information in Medicine*, vol. 2, 2006.

[7] J. Fox, A. Alabassi, V. Patkar, T. Rose, and E. Black, "An ontological approach to modelling tasks and goals," *Computers in Biology and Medicine*, vol. 36, no. 7-8, pp. 837 – 856, 2006. Special Issue on Medical Ontologies.

[8] T. Alsinet, C. Ansótegui, R. Béjar, C. Fernández, and F. Manya, "Automated monitoring of medical protocols: a secure and distributed architecture," *Artificial Intelligence in Medicine*, vol. 27, no. 3, pp. 367 – 392, 2003. Software Agents in Health Care.

[9] B. Bosansky and C. Brom, "Agent-based simulation of business processes in a virtual world," in *HAIS '08: Proceedings of the 3rd international workshop on Hybrid Artificial Intelligence Systems*, pp. 86–94, Springer-Verlag Berlin, Heidelberg, 2008.

[10] B. Bosansky and L. Lhotska, "Agent-based simulation of processes in medicine," in *Procceding of PhD. Conference*, pp. 19–27, Institute of Computer Science/MatfyzPress, 2008.

[11] R. Conroy, K. Pyorala, A. Fitzgerald, S. Sans, A. Menotti, G. De Backer, D. De Bacquer, P. Ducimetiere, P. Jousilahti, U. Keil, I. Njolstad, R. Oganov, T. Thomsen, H. Tunstall-Pedoe, A. Tverdal, H. Wedel, P. Whincup, L. Wilhelmsen,

DRAFT
k: srpna 2009 - 14:52

and I. Graham, "Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project.," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, 2003.

[12] B. Lars, P. Alexander, and L. Winfried, "Jadex: A bdi-agent system combining middleware and reasoning," 2005.

DRAFT
14. srpna 2009 - 14:32

Fuzzy Classification of Web Reports with Linguistic Text Mining

Post-Graduate Student:
MGR. JAN DĚDEK
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25
118 00 Prague 1, CZ
jan.dedek@mff.cuni.cz

Supervisor:
PROF. RNDR. PETER VOJTÁŠ, DRSC.
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25
118 00 Prague 1, CZ
peter.vojtas@mff.cuni.cz

Field of Study:
Software Engineering

This work was partially supported by Czech projects: IS-1ET100300517, GACR-201/09/H057 and GAUK 31009.

Abstract

In this paper we present a fuzzy system which provides a fuzzy classification of textual web reports. Our approach is based on usage of third party linguistic analyzers, our previous work on web information extraction and fuzzy inductive logic programming. Main contributions are formal models and prototype implementation of the system and evaluation experiments.

The abstract was originally published in the paper [1]. Due to the copyright issues, only the abstract is presented here, extended with some additional information that is not included in the original paper.

1. Introduction

In this contribution we would like to present our latest work [1] and extend it with some additional information about the issues that are closely related to the original paper. As the original paper has only four pages, we present more details and references here.

The original paper deals with a structured data that could be extracted from web reports. The original paper is closely concentrated on the use of the structured data for a fuzzy classification of the reports. The original paper refers to our previous works where our method for extraction of a structured data form web reports is presented and gives very little details about it. In this contribution we present:

- more details about our extraction method (see in Section 2),
- a richer discussion of the related work (in Section 3) and
- the current state of our development and our plans for the future work (in the last section).

1.1. Motivation

Big amount of information on the web increases the need of automated precessing. Especially textual information are hard for machine processing and understanding. Crisp methods have their limitations. In this paper we present a fuzzy system which provides a fuzzy classification of textual web reports.

Messages of accident reports on the web (Fig. 1) are our motivating examples. We would like to have a tool which is able to classify such message with degree of being it a serious accident.

Our solution is based first on information extraction (see emphasized information to be extracted in Fig. 1) and second on processing this information to get fuzzy classification rules. The description of the the fuzzy classification is presented in [1], here we will present only some additional information.

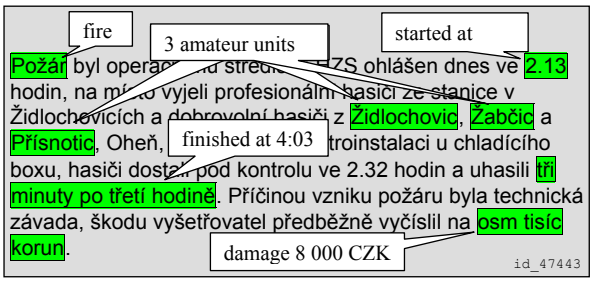


Figure 1: Example of analyzed web report.

DRAFT 14. Srpna 2009 - 14:52

2. Our Information Extraction Method

2.1. Linguistic Analysis

In this section we will briefly describe the linguistic tools that we have used to produce linguistic annotations of texts. These tools are being developed in the Institute of Formal and Applied Linguistics in Prague, Czech Republic. They are publicly available – they have been published on a CDROM under the title PDT 2.0 [2] (first five tools) and in [3] (Tectogrammatical analysis). These tools are used as a processing chain and at the end of the chain they produce tectogrammatical [4] dependency trees.

Tool 1. Segmentation and tokenization consists of tokenization (dividing the input text into words and punctuation) and segmentation (dividing a sequences of tokens into sentences).

Tool 2. Morphological analysis assigns all possible lemmas and morphological tags to particular word forms (word occurrences) in the text.

Tool 3. Morphological tagging consists in selecting a single pair lemma-tag from all possible alternatives assigned by the morphological analyzer.

Tool 4. Collins' parser – Czech adaptation. Unlike the usual approaches to the description of English syntax, the Czech syntactic descriptions are dependency-based, which means, that every edge of a syntactic tree captures the relation of dependency between a governor and its dependent node. Collins' parser gives the most probable parse of a given input sentence.

Tool 5. Analytical function assignment assigns a description (analytical function – in linguistic sense) to every edge in the syntactic (dependency) tree.

Tool 6. Tectogrammatical analysis produces linguistic annotation at the tectogrammatical level, sometimes called "layer of deep syntax". Such a tree can be seen on the Fig. 2. Annotation of a sentence at this layer is closer to meaning of the sentence than its syntactic annotation and thus information captured at the tectogrammatical layer is crucial for machine understanding of a natural language [3].

2.2. Web Information Extraction

Having web resource content analyzed by above linguistic tools, we have data stored in the form of tectogramatical trees. To achieve our objectives we have to extract information from this representation. Here we refer to our previous work [5, 6, 7]. A long path of tools starting with web crawling and resulting with the extracted structured information is developed in our previous work. In Fig. 2 we can see nodes of a tree where a piece of information about damage (8000 CZK) is located. We have used Inductive logic Programming to learn rules which are able to detect such nodes. The extraction process requires a human assistance when annotating a training data.

Note that our method is general and is not limited to Czech and can be used with any structured linguistic representation.

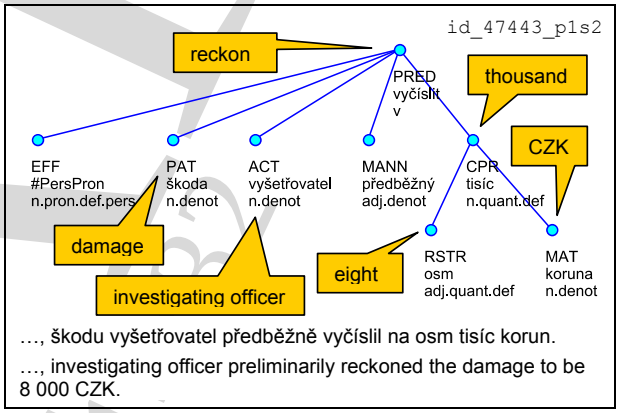


Figure 2: Example of a linguistic tree of one of analyzed sentences.

3. Related Work

There is a plenty of systems dealing with text mining and text classification, let us mention at least some. In [8] authors use ontology modeling to enhance text identification. In [9] authors use preprocessed data from National Automotive Sampling System and test various soft computing methods to modeling severity of injuries (some hybrid methods showed best performance). Methods of Information Retrieval (IR) are very numerous, with extraction mainly based on key word search and similarities. Connecting IR and text mining techniques with web information retrieval can be found in Chapter Opinion mining in the book of Bing Liu [10].

4. Conclusion and Future Work

Currently we are working on the integration of our method with further linguistic tools and we work on a graphical user interface so the whole system could be distributed as a software package and used by arbitrary users.

DRAFT 14. srpna 2009 - 13:32

We have made first experiments with the TectoMT system [11], which can replace the older tools from the PDT2.0 CD-ROM mentioned above and currently used in our system. TectoMT can bring us many befits like named entity recognition, better morphology and parsing (made by the McDonalds' parser [12]), but the biggest advantage is that we can use the same linguistic formalism (tectogrammatical trees) for English (and probably for other languages in the future).

On the other hand our approach is not limited to the tectogrammatical trees and we have made first experiments with *Stanford typed dependencies* [13] as an alternative linguistic formalism.

We will probably use The GATE architecture [14] as the platform for integration of our method with other systems and we can use it also as the graphical user interface. The GATE features will also bring a very modular fashion to the final system.

References

[1] J. Dědek and P. Vojtáš, "Fuzzy classification of web reports with linguistic text mining," in *Web Intelligence/IAT Workshops, Soft approaches to information access on the Web*, (Milan, Italy), Accepted for publication, 2009.

[2] J. Hajič, E. Hajičová, J. Hlaváčová, V. Klimeš, J. Mírovský, P. Pajas, J. Štěpánek, B. Vidová-Hladká, and Z. Žabokrtský, "Prague dependency treebank 2.0 cd–rom." Linguistic Data Consortium LDC2006T01, Philadelphia 2006, 2006.

[3] V. Klimeš, "Transformation-based tectogrammatical analysis of czech," in *Proc. 9th International Conference, TSD 2006*, no. 4188 in Lecture Notes In Computer Science, pp. 135–142, Springer-Verlag Berlin Heidelberg, 2006.

[4] M. Mikulová, A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá, and Z. Žabokrtský, "Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual," Tech. Rep. 30, ÚFAL MFF UK, Prague, Czech Rep., 2006.

[5] J. Dědek and P. Vojtáš, "Linguistic extraction for semantic annotation," in *2nd International Symposium on Intelligent Distributed Computing* (C. Badica, G. Mangioni, V. Carchiolo, and D. Burdescu, eds.), vol. 162 of *Studies in Computational Intelligence*, (Catania, Italy), pp. 85–94, Springer-Verlag, 2008.

[6] J. Dědek and P. Vojtáš, "Computing aggregations from linguistic web resources: a case study in czech republic sector/traffic accidents," in *Second International Conference on Advanced Engineering Computing and Applications in Sciences* (C. Dini, ed.), pp. 7–12, IEEE Computer Society, 2008.

[7] J. Dědek, A. Eckhardt, and P. Vojtáš, "Experiments with czech linguistic data and ILP," in *ILP 2008 (Late Breaking Papers)* (F. Železný and N. Lavrač, eds.), (Prague, Czech Republic), pp. 20–25, Action M, 2008.

[8] M. Reformat, R. R. Yager, and Z. Li, "Ontology enhanced concept hierarchies for text identification," *Journal Semantic Web Information Systems*, vol. 4, no. 3, pp. 16–43, 2008.

[9] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using machine learning paradigms," *Informatica*, vol. 29, pp. 89–98, 2005.

[10] B. Liu, *Web Data Mining*. Springer-Verlag, 2007.

[11] Z. Žabokrtský, J. Ptáček, and P. Pajas, "TectoMT: Highly modular MT system with tectogrammatics used as transfer layer," in *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, (Columbus, OH, USA), pp. 167–170, Association for Computational Linguistics, 2008.

[12] R. McDonald, F. Pereira, K. Ribarov, and J. Hajic, "Non-projective dependency parsing using spanning tree algorithms," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (Vancouver, British Columbia, Canada), pp. 523–530, Association for Computational Linguistics, October 2005.

[13] M. C. de Marneffe and C. D. Manning, "The Stanford typed dependencies representation," in *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, (Manchester, UK), pp. 1–8, Coling 2008 Organizing Committee, August 2008.

[14] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "Gate: A framework and graphical development environment for robust nlp tools and applications," in *Proceedings of the 40th Annual Meeting of the ACL*, 2002.

14. srpna 2009 - 12:32

Porovnání optimalizačních metod pro změkčování rozhodovacího stromu

doktorand:
MGR. JAKUB DVOŘÁK
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
dvorak@cs.cas.cz

školitel:
RNDR. PETR SAVICKÝ, CSC.
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
savicky@cs.cas.cz

obor studia:
Teoretická informatika

Tento výzkum byl podporován institucionálním výzkumným záměrem AV0Z10300504 a také projektem T100300517 programu „Informační společnost" AV ČR.

Abstrakt

Zkoumáme metody změkčování rozhodovacích stromů, které vycházejí z hotového rozhodovacího stromu získaného metodou CART a při zachování jeho struktury hledají změkčení tak, že optimalizují kvalitu klasifikátoru na trénovací množině. Představené metody používají dvě různé míry kvality klasifikátoru. Jednou z nich je součet jistým způsobem transformované chyby na jednotlivých vzorech trénovací množiny, druhou je plocha pod ROC křivkou (AUC). K hledání co nejlepšího změkčení je použita randomizovaná strategie iterované optimalizace, která v každém cyklu modifikuje pouze několik parametrů. V rámci této strategie využíváme jako optimalizační metody simulované žíhání a simplexový algoritmus pro minimalizaci — Nelder-Mead. Jako ukončující kritérium pro iterační proces změkčování za účelem porovnání metod používáme dosažení limitu reálného času výpočtu. V experimentech s daty „Magic Telescope" při porovnání podle AUC se ukazuje jako nejlepší optimalizace AUC pomocí metody Nelder-Mead.

1. Úvod

Změkčování rozhodovacích stromů je cesta ke zlepšení kvality predikce klasifikátoru na prostoru vzorů s reálnými atributy. Jestliže výstupem klasifikátoru je reálné číslo, změkčené stromy umožňují, aby tento výstup byl spojitou funkcí atributů. V nezměkčeném rozhodovacím stromu jsou ve vnitřních uzlech podmínky, které na základě předloženého vzoru rozhodují, zda v prohledávání pokračovat v levém nebo pravém potomkovi daného uzlu. Když prohledávání dosáhne listu, je z něj zjištěn výsledek klasifikace. Změkčování je založeno na tom, že rozhodovací podmínky ve vnitřních uzlech („splitech") jsou nahrazeny pravidlem výpočtu poměru, v jakém jsou zkombinovány výsledky levého a pravého podstromu.

Zde budeme vycházet z nezměkčeného klasického rozhodovacího stromu pro klasifikaci získaného metodou CART [2], jehož rozhodovací podmínky budeme následně změkčovat. Pro takové změkčování jakožto postprocessing budeme s použitím trénovací množiny optimalizovat kvalitu klasifikátoru. Zaměřujeme se na klasifikaci do dvou tříd nazývaných „signal" a „background", výstupem klasifikátoru je pro každý předložený vzor reálné číslo v intervalu $[0, 1]$, které je odhadem pravděpodobnosti, že vzor patří do třídy „signal".

Metoda změkčování, která byla v našich předchozích výzkumech nejúspěšnější [3], hledala co nejlepší vektor parametrů změkčení tak, že opakovaně vybírala několik parametrů, a na této podmnožině parametrů používala simulované žíhání k optimalizaci cílové funkce, ostatní parametry zůstávaly zatím zafixovány. Cílová funkce v této metodě byla počítána tak, že pro vzory z trénovací množiny byla vypočtena klasifikace stromem změkčeným s danými parametry, pro každý vzor byla absolutní hodnota rozdílu získané klasifikace od správné hodnoty (tj. 0 nebo 1) transformována exponenciální funkcí a výsledky sečteny přes celou trénovací množinu. Tato metoda však konvergovala velmi pomalu, získání dostatečně kvalitního změkčeného stromu trvalo několik hodin.

Následujícím cílem bylo nalezení metody, jež umožní získání alespoň stejně dobrého změkčeného stromu v rozumném čase. Zde budeme porovnávat změkčující metody, pro něž jsme zvolili jako ukončující kritérium cyklů optimalizačních iterací vyčerpání časového limitu.

DRAFT Prague, November 2009

Získané klasifikátory budeme porovnávat podle plochy pod ROC křivkou (Area Under Curve, AUC) [4] naměřené na testovacích datech. AUC je standardní míra kvality klasifikátoru. Hodnota AUC leží v intervalu $[0, 1]$, čím je vyšší, tím je klasifikátor lepší. AUC pro náhodný klasifikátor je $1/2$. Interpretace je následující: Vybereme-li náhodně rovnoměrně jeden pozitivní případ a jeden negativní případ, potom AUC je pravděpodobnost, že klasifikátor pro vybraný pozitivní případ vydá vyšší výstupní hodnotu, než pro vybraný negativní případ.

Jedna ze zde zkoumaných metod ještě aplikuje myšlenku výše zmíněné cílové funkce. Další metody jako cílovou funkci používají odhad AUC na základě trénovacích dat. Tuto funkci maximalizují v jednom případě opět iterovaným simulovaným žíháním a v druhém případě je v rámci iterační strategie použita metoda „Nelder-Mead”.

2. Změkčování rozhodovacího stromu

Mějme (nezměkčený) rozhodovací strom, pro klasifikaci vzorů s m reálnými atributy. Označme s počet vnitřních uzlů (splitů), tedy včetně listů má strom $2s + 1$ uzlů. Uzly stromu označujme $v_j, \quad j = 1, \ldots, 2s + 1$, předpokládejme, že splity mají indexy $1, \ldots, s$ a listy $s + 1, \ldots, 2s + 1$. Každému splitu v_j je přiřazen vektor $\mathbf{w}_j = (w_{j,1}, \ldots, w_{j,m})$ a reálné číslo c_j, každému listu v_j je přiřazeno reálné číslo r_j. Metoda CART používá jako hodnotu r_j relativní četnost signal případů v listu v_j, takže

$$0 \leq r_j \leq 1, \quad r = s + 1, \ldots, 2s + 1$$

Klasifikace vstupního vzoru $\mathbf{x} = (x_1, \ldots, x_m)$ probíhá tak, že od kořene stromu jakožto výchozího aktuálního uzlu se provádí následující proces: Když v_j je vnitřní uzel, provede se test

$$\mathbf{w}_j \mathbf{x} \leq c_j \tag{1}$$

a pokud je nerovnost (1) splněna, aktuálním uzlem se stává levý potomek uzlu v_j, jinak pravý potomek. Je-li aktuální uzel v_j list, potom výstupem klasifikátoru je hodnota r_j.

Tímto způsobem získáme jako výstup klasifikátoru reálné číslo (jedno z čísel r_j přiřazených některému listu). Pokud je hodnota výstupu klasifikátoru větší než zvolený práh, zařazujeme předložený vzor do třídy „signal”, jinak do třídy „background”.

V experimentech používáme stromy, kde vektory $\mathbf{w}_j$ obsahují právě jednu jedničku a jinak samé nuly, tzn. výraz $\mathbf{w}_j \mathbf{x}$ má hodnotu x_i, kde i je takový index, že $w_{j,i} = 1$, čili vektor $\mathbf{w}_j$ vyjadřuje výběr jednoho atributu předloženého vzoru. To opět odpovídá klasické metodě CART, i když existuje i modifikace, která ve splitech pro porovnávání používá obecné netriviální lineární kombinace atributů.

Rozhodovací strom rozděluje prostor vzorů na oblasti, v našem případě s právě jednou jedničkou ve vektoru $\mathbf{w}_j$ se jedná o hyperkvádry se stěnami kolmými na osy souřadné soustavy určené atributy. V každém hyperkvádru je výstup klasifikátoru pro všechny body stejný. Změkčování splitů ve stromu se projeví tím, že skokové přechody výstupu klasifikátoru na hranicích hyperkvádrů se změní na spojité.

Při změkčování každému splitu $v_j, \quad j = 1, \ldots, s$, přiřadíme parametry změkčení

$$a_j, b_j \geq 0 \tag{2}$$

Potom definujeme změkčující funci (viz obrázek 1) příslušnou uzlu v_j, parametrizovanou čísly a_j, b_j, c_j a vektorem $\mathbf{w}_j$:

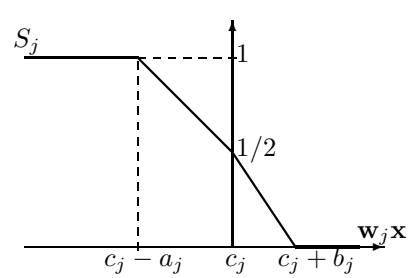
$$S_j(\mathbf{x}) = \sigma_{a_j, b_j}(\mathbf{w}_j \mathbf{x} - c_j)$$

kde $\sigma_{a,b}$ lineárně interpoluje body uvedené v tabulce:

t	$-\infty$	$-a$	0	b	∞
$\sigma_{a,b}(t)$	1	1	1/2	0	0

Pro případ, že pro nějaké $j \leq s$ je $a_j = 0$, nebo $b_j = 0$ je potřeba dodefinovat

$$\begin{aligned} \sigma_{0,b}(0) &= 1 && \text{pro každé } b \geq 0 \\ \sigma_{a,0}(0) &= 1/2 && \text{když } a > 0 \end{aligned}$$



Obrázek 1: Změkčující funkce

Pro vstupní vzor $\mathbf{x}$ a uzel v_j změkčeného stromu definujeme výstup uzlu $v_j(\mathbf{x})$ následující rekurzí: Je-li v_j list, je jeho výstupem r_j. Pro vnitřní uzel v_j označme v_j^L resp. v_j^R jeho levého resp. pravého potomka. Potom

$$v_j(\mathbf{x}) = S_j(\mathbf{x}) v_j^L(\mathbf{x}) + (1 - S_j(\mathbf{x})) v_j^R(\mathbf{x})$$

DRAFT 4. srpna 2009 - 14:32

Výstupem klasifikátoru je výstup kořene stromu. I u změkčeného stromu pro finální klasifikaci použijeme porovnání výstupu klasifikátoru se zvoleným prahem.

Jestliže $a_j = 0$ a $b_j = 0$ pro všechna $j \leq s$, potom výstup změkčeného stromu je pro každý vstupní vzor roven výstupu původního nezměkčeného stromu.

Úlohou změkčování daného stromu je hledání parametrů a_j, b_j pro všechny vnitřní uzly $j = 1, \ldots, s$. Vektor všech parametrů a_j, b_j budeme také označovat $\mathbf{p}$. Změkčený strom s parametry změkčení $\mathbf{p}$ budeme označovat $T^{(\mathbf{p})}$ přičemž $T^{(\mathbf{p})}(\mathbf{x})$ znamená výstup tohoto klasifikátoru pro vzor $\mathbf{x}$.

3. Iterování optimalizace na podmnožinách parametrů

V této sekci popíšeme strategii hledání parametrů minimalizujících cílovou funkci, která používá opakovaně optimalizační metodu, vždy pouze na podmnožině parametrů a ostatní parametry zůstávají konstantní. Tedy jednotlivé optimalizační běhy řeší úlohu nižší dimenze.

K tomuto účelu zavedeme následující značení: Nechť $Q \subseteq \{1, \ldots, 2s\}$ a $\mathbf{z} \in \mathrm{R}^{2s}$. R^Q bude označovat množinu vektorů $\{x_i\}_{i \in Q}$, tzn. vektorů z $\mathrm{R}^{|Q|}$, jejichž složky jsou indexovány prvky Q místo čísel $1, \ldots, |Q|$. Máme-li cílovou funkci změkčování $f(\mathbf{p})$ definovanou na R^{2s}, potom pro $\mathbf{z} = (z_1, \ldots, z_{2s})$ nechť $f[Q, \mathbf{z}] : \mathrm{R}^Q \to \mathrm{R}$ je funkce definovaná na každém $\mathbf{x} \in \mathrm{R}^Q$ tak, že $f[Q, \mathbf{z}](\mathbf{x}) = f(\mathbf{y})$, kde

$$y_i = \begin{cases} x_i & \text{když } i \in Q \\ z_i & \text{jinak} \end{cases}$$

Iterační strategie opakovaně aplikuje vybranou optimalizační metodu, cílovou funkcí je $f[Q, \mathbf{p}_0](\mathbf{p}')$, která má $|Q|$ argumentů, kde $\mathbf{p}_0$ je výsledek předchozího volání, nebo v případě první iterace iniciální hodnota změkčování. Restrikce $\mathbf{p}_0$ na vybranou množinu indexů Q je iniciální hodnota pro $\mathbf{p}'$ v aktuální iteraci.

Výběr podmnožiny Q parametrů v každém cyklu je randomizovaný a založený na struktuře stromu. Nejprve je náhodně zvolen jeden z parametrů změkčení jako kořenový parametr. Jestliže kořenový parametr je a_j pro nějaké j, potom v_j^X označíme v_j^L, tedy levého potomka uzlu v_j. Jestliže kořenový parametr je b_j pro nějaké j, potom v_j^X rozumíme v_j^R — pravého potomka uzlu v_j.

Kořenový parametr volíme tak, aby v_j ani v_j^X nebylo listem stromu, mezi takovými parametry je náhodný výběr kořenového parametru rovnoměrně rozdělen. Do výsledné podmnožiny parametrů pro optimalizaci Q zahrneme kořenový parametr, oba parametry příslušné uzlu v_j^X a všechny parametry příslušné přímým potomkům uzlu v_j^X. Množina Q tak může mít 7 prvků, ale protože jeden nebo oba z přímých potomků uzlu v_j^X mohou být listy, může mít tato podmnožina také 5 nebo 3 prvky.

4. Optimalizační metody

Porovnávané metody změkčování jsou založeny na dvou standardních optimalizačních metodách — simulovaném žíhání a simplexovém algoritmu pro minimalizaci „Nelder-Mead" [5]. Obě tyto metody hledají optimum iterativně pouze na základě funkčních hodnot cílové funkce, tedy bez potřeby výpočtu diferenciálu, takže jsou použitelné i na optimalizace nespojitých funkcí.

Protože použité implementace těchto optimalizačních metod neumožňovaly omezení definičního oboru cílové funkce, ale přitom na parametry změkčení stromu jsou kladeny podmínky (2), je cílová funkce dodefinována tak, aby pro vstupní hodnoty porušující (2), generovala vysokou hodnotu (viz níže) a tím uměle nutila optimalizační metody se této oblasti vyvarovat.

V rámci strategie iterované optimalizace popsané v předešlé sekci jsou metody použity následujícím způsobem: Simulované žíhání — v každé iteraci se provádí 100 kroků metody, nový kandidátský bod se generuje na základě Gauss-Markovova kernelu, iniciální teplota je nastavena na hodnotu 10 a update teploty se provádí po každém kroku. Nelder-Mead — v každé iteraci je provedeno 30 kroků metody.

Pro určení škály a iniciální hodnoty pro optimalizační metody používáme vzdálenost splitu od hranice hyperkvádru. Tyto hodnoty definujeme takto: Nejprve celý prostor vzorů ve směrech všech atributů omezíme nejzazšími trénovacími vzory, tak získáme základní hyperkvádr. Když v uzlu v_j podmínka (1) rozděluje hyperkvádr vyšší úrovně, který je v testované proměnné x_i omezen hodnotami $z_{j,1}, z_{j,2}$, kde $z_{j,1} < c_j < z_{j,2}$, potom za vzdálenost od hranice hyperkvádru příslušnou pro parametr a_j resp. b_j považujeme

$$h_j^a = c_j - z_{j,1}; \quad h_j^b = z_{j,2} - c_j$$

DRAFT 14 srpna 2009 - 14:32

metoda	A	B	C	D
iniciální hodnota	$a_j^0 = 0; \quad b_j^0 = 0$	$a_j^0 = 1/4\, h_j^a; \quad b_j^0 = 1/4\, h_j^b$		
ukončující kritérium	50 iterací po sobě bez zlepšení	vyčerpaný časový limit 5 minut		
cílová funkce	φ_{A}	φ_{B}	AUC	
optimalizační metoda	iterované simulované žíhání			iter. Nelder-Mead
škála	$(h_j^a, h_j^b),\ j = 1, \ldots, s$			$(1/16\, h_j^a, 1/16\, h_j^b)$ $j = 1, \ldots, s$

Tabulka 1: Přehled metod změkčování

Základní charakteristiky zkoumaných metod ukazuje tabulka 1. Písmeno A označuje metodu z [3]. Tato metoda používá cílovou funkci definovanou pro legální parametry změkčení $\mathbf{p}$ (viz (2)) jako

$$\varphi_{\mathrm{A}}(\mathbf{p}) = \sum_{i=1}^{n} \exp\left(4\left(\left|T^{(\mathbf{p})}(\mathbf{x}_i) - y_i\right| - 1\right)\right)$$

kde $\mathbf{x}_i,\ i = 1, \ldots, n$ jsou prvky testovací množiny a y_i jsou jim příslušné klasifikace, tedy hodnoty 0 nebo 1 pro background resp. signal případy. Pro nelegální parametry je $\varphi_{\mathrm{A}}(\mathbf{p}) = n + 1$.

Pod písmenem B uvádíme metodu, která je založena na stejném základním principu, jako metoda A, ale obsahuje několik zlepšení nalezených od publikování [3], zejména nenulový inicializační vektor parametrů, normalizaci funkční hodnoty a výstupní hodnotu při nelegálních parametrech, jež roste se vzdáleností hodnoty každého nelegálního parametru od legálních hodnot. Pro tuto metodu již používáme jako ukončující kritérium časový limit. To nám umožní relevantnější porovnání myšlenky metody A s ostatními metodami.

Cílová funkce pro metodu B při legálních parametrech je

$$\varphi_{\mathrm{B}}(\mathbf{p}) = \frac{1}{n}\varphi_{\mathrm{A}}(\mathbf{p})$$

Pro nelegální parametry, tzn. pokud některá ze složek vektoru $\mathbf{p}$ je záporná, definujeme

$$\mu(\mathbf{p}) = 1 + \sum_{i=1}^{2s} \max(0, -p_i)$$

$$\varphi_{\mathrm{B}}(\mathbf{p}) = \mu(\mathbf{p})$$

Pro další metody je v tabulce 1 uvedena jako cílová funkce „AUC”, což přesněji znamená, že se minimalizuje funkce, která pro legální parametry $\mathbf{p}$ má hodnotu záporně vzaté AUC pro $T^{(\mathbf{p})}$ naměřené na trénovací množině, pro nelegální parametry má hodnotu $\mu(\mathbf{p})$.

5. Implementace

Protože jsme použili v ukončujícím kritériu optimalizace reálný čas, je důležitá implementace experimentů. Základním frameworkem byl systém R [6], který zahrnuje interpret jazyka a rozšiřitelný systém balíčků, díky němuž mohou být jednotlivé metody naprogramovány např. v jazyce C a integrovány pomocí zkompilované sdílené knihovny.

V jazyce R byla naprogramována nejvyšší úroveň sestavení experimentů s využitím následujících komponent:

- Klasifikace množiny vzorů změkčeným stromem, byla implementována v jazyce C.
- Výpočet AUC — byl v jazyce C.
- Metoda simulovaného žíhání byla v jazyce C, jednalo se o mírně upravenou implementaci, jež je součástí systému R.
- Metoda Nelder-Mead byla integrovaná implementace z knihovny GNU Scientific Library[1] pomocí R package „gsl”.
- Strategie iterované optimalizace na podmnožinách množiny parametrů byla implementována v jazyce R.

Výpočty běžely na procesoru Intel® Xeon™ CPU 2.80GHz, v systému se 4GB operační paměti.

[1] http://www.gnu.org/software/gsl/

DRAFT 14. srpna 2009 – 4:52

6. Experimenty

V experimentech byla použita data „Magic Telescope"[2], která jsou zkoumána také v [3]. Problematikou klasifikace těchto dat se více zabývá [1]. Data mají 10 reálných atributů, obsahují přibližně 65 % signal případů.

Trénovací množina obsahující 12680 vzorů, byla rozdělena na dvě části v poměru velikostí 2:1, první část byla použita pro růst stromu a druhá část jako validační množina pro prořezávání. Stromy byly vytvořeny metodou CART, nastavením různých hodnot parametrů prořezávání byla získána sekvence stromů různých velikostí. V této sekvenci byl na počátku největší strom a každý další vzniknul prořezáním předchozího, tedy byl jeho podstromem. Pro změkčování byly použity z celé sekvence pouze ty stromy, které nebyly větší, než strom, krerý měl na validační množině nejmenší chybu. Jako data sloužící k výpočtu cílové funkce při změkčování byla potom použita celá trénovací množina.

Z důvodu velké časové náročnosti metody A byl vygenerován pouze malý počet stromů změkčených touto metodou, změkčovány byly stromy z počátku sekvence prořezávání, čili největší, tedy nejpřesnější stromy (podrobnosti viz [3]). Pro porovnání jsme pro každý strom ze sekvence spočetli tolik změkčení každou z metod B, C, D, kolik bylo k dispozici změkčení metodou A.

Na základě testovací množiny obsahující 6340 vzorů byla vypočtena hodnota AUC pro každý takto získaný klasifikátor. Pro každou z metod jsme vypočetli průměrnou a maximální hodnotu AUC ze všech změkčených stromů.

Celý popsaný postup byl opakován desetkrát s tím, že pro každý experiment byla dostupná data nově rozdělena na trénovací a testovací množinu. Díky odlišným trénovacím množinám byly v různých experimentech odlišné primární stromy, které byly základem pro prořezávání. Tabulka 2 ukazuje počet vnitřních uzlů největšího stromu ze sekvence použitého pro změkčování a jeho hodnotu AUC naměřenou na testovacích datech.

Průměrné hodnoty AUC stromů změkčených jednotlivými metodami porovnává tabulka 3, maximální hodnoty tabulka 4.

Metody B a C mají výsledky obecně horší, než metoda A. Výsledky metod A a D porovnává tabulka 5. Průměrné hodnoty AUC metody D jsou pouze ve dvou případech z 10 nepatrně horší, než u metody A, maximální hodnoty dokonce jen v jednom případě z 10.

[2] http://wwwmagic.mppmu.mpg.de

	počet splitů	AUC		počet splitů	AUC
1	45	0.887254	6	69	0.886673
2	49	0.882268	7	38	0.881902
3	69	0.886131	8	52	0.880057
4	64	0.892350	9	75	0.893006
5	43	0.894513	10	56	0.885681

Tabulka 2: Vlastnosti nezměkčených stromů.

	A	B	C	D
1	0.909050	0.903945	0.904057	0.912239
2	0.907109	0.896344	0.898889	0.908804
3	0.914037	0.902580	0.906368	0.914319
4	0.913617	0.903198	0.903999	0.915339
5	0.913058	0.905683	0.907888	0.917001
6	0.909587	0.897866	0.899282	0.909323
7	0.908522	0.901193	0.904355	0.909901
8	0.907109	0.897537	0.899769	0.908703
9	0.916947	0.906992	0.909030	0.917126
10	0.913255	0.903054	0.904603	0.912520

Tabulka 3: Průměrné hodnoty AUC.

	A	B	C	D
1	0.913832	0.907214	0.907320	0.913885
2	0.909889	0.898939	0.904546	0.910490
3	0.917025	0.905709	0.910231	0.917092
4	0.918478	0.908102	0.908378	0.918705
5	0.916306	0.909574	0.911750	0.919407
6	0.913164	0.903933	0.903855	0.911248
7	0.910700	0.905010	0.911652	0.914576
8	0.909786	0.901399	0.903703	0.911067
9	0.919530	0.911230	0.915396	0.920273
10	0.915340	0.907896	0.910725	0.916069

Tabulka 4: Maximální hodnoty AUC.

	$\oslash$ D / $\oslash$ A	max D / max A
1	1.0035087	1.0000580
2	1.0018689	1.0006610
3	1.0003088	1.0000739
4	1.0018849	1.0002472
5	1.0043183	1.0033848
6	0.9997104	0.9979017
7	1.0015178	1.0042566
8	1.0017566	1.0014079
9	1.0001951	1.0008084
10	0.9991960	1.0007963

Tabulka 5: Poměry hodnot AUC metody D a A.

DATA 14.srpna 2009 14:32

7. Závěr

Porovnali jsme 4 metody pro změkčování rozhodovacího stromu založené na optimalizaci kvality klasifikátoru na trénovací množině. Cílem bylo dosáhnout v rozumném čase alespoň srovnatelných výsledků získaného klasifikátoru, jaké dávala metoda založená na iterovaném simulovaném žíhání, která používala jako cílovou funkci φ_{A}, tedy součet exponenciální funkcí transformovaných vzdáleností výstupu klasifikátoru od správné klasifikace.

Představili jsme metodu podobnou — také založenou na iterovaném simulovaném žíhání a uvedené cílové funkci, ale se zlepšeními v oblasti inicializace, řešení ilegálních hodnot a normalizace funkční hodnoty. Tato zlepšení nevedla k tomu, že by metoda v daném pětiminutovém časovém limitu dosahovala dostatečně kvalitních změkčení.

V experimentech se ukázalo, že lepší cílovou funkcí je plocha pod ROC křivkou (AUC). Pro tuto cílovou funkci jsme použili jako optimalizační strategie opět iterované simulované žíhání a také iterovaný simplexový algoritmus (Nelder-Mead). Poslední z uvedených metod dosáhla na datové množině „Magic Telescope" při výpočtu omezeném v čase 5ti minut výsledků srovnatelných s těmi, které původní metoda počítala několik hodin.

Literatura

[1] R.K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, and A. Vaicilius, "Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope." *Nucl. Instr. Meth.*, A 516, pp. 511–528, 2004.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Belmont CA: Wadsworth, 1993.

[3] J. Dvořák and P. Savický, "Softening Splits in Decision Trees Using Simulated Annealing", *Adaptive and Natural Computing Algorithms*, LNCS vol. 4431/2007, pp. 721–729, 2007.

[4] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, pp. 861—874, 2006.

[5] J.A. Nelder and R. Mead, "A simplex algorithm for function minimization.", *Computer Journal* vol. 7, pp. 308–313, 1965.

[6] R Development Core Team (2008), "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

DRAFT 14. srpna 2009 - 14:57

Verification of Hybrid Systems Using Slices of Parallel Hyperplanes

Post-Graduate Student:
TOMÁŠ DZETKULIČ
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
dzetkulic@cs.cas.cz

Supervisor:
STEFAN RATSCHAN
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
ratschan@cs.cas.cz

Field of Study:
Verification of Hybrid Systems

My work has been supported by GAČR grants 201/08/J020 and 201/09/H057.

Abstract

A hybrid system is a dynamic system that exhibits both continuous and discrete behavior. With hybrid systems we can model traffic protocols, networking and locking protocols, microcontrollers and many other systems where a discrete system interacts with some continuous environment. Usually in such applications there are some unsafe states, that is, states that will be dangerous for the system or its user. Safety verification algorithms are algorithms that automatically check that a given hybrid system never reaches an unsafe state.

In our work we improve the method for verification of hybrid systems by constraint propagation based abstraction refinement [1]. That algorithm allows the verification of a very general class of hybrid systems (e.g., with non-linear ordinary differential equations), but does not exploit the structure of special cases. Our proposed improvement [2] still allows very general inputs, but exploits the fact that parts of the input might represent linear time evolution (so called clocks). In the algorithm, we compute slices of parallel hyperplanes separating reachable from unreachable parts of the state space for a given abstraction of the input system. We demonstrate the usefulness of such slices within an abstraction refinement algorithm based on hyper-rectangles.

References

[1] S. Ratschan and Z. She, "Safety Verification of Hybrid Systems by Constraint Propagation Based Abstraction Refinement", *ACM TECS*, vol. 6, 2007.

[2] T. Dzetkulič and S. Ratschan, "How to Capture Hybrid Systems Evolution Into Slices of Parallel Hyperplanes", to appear in the proceedings of *ADHS'09*: 3rd IFAC Conference on Analysis and Design of Hybrid Systems.

DRAFT
14. srpna 2009, 14:33

How to Learn Fuzzy User Preferences with Variable Objectives

Post-Graduate Student:
RNDR. ALAN ECKHARDT
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25
118 00 Prague 1, CZ
eckhardt@ksi.mff.cuni.cz

Supervisor:
PROF. RNDR. PETER VOJTÁŠ, DRSC.
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25
118 00 Prague 1, CZ
vojtas@ksi.mff.cuni.cz

Field of Study:
Software Engineering

This work was supported by Czech projects MSM 0021620838, 1ET 100300517 and GACR 201/09/H057.

Abstract

This paper studies a possibility to learn a complex user preference model, based on CP-nets, from user ratings. This work is motivated by the need of user modelling in decision making support, for example in e-commerce. We extend our user model based on fuzzy logic to capture variation of preference objectives. The proposed method 2CP-regression is described and tested. 2CP-regression uses CP-nets idea behind and can be considered as learning of a simple CP-net from user ratings.

The abstract was taken from [1]. Due to the copyright issues, only the abstract is presented here.

We add a brief overview of contributions of the paper, extended with work done after publishing the paper.

1. Ceteris paribus

One of the main contributions of the paper was to address the issue of ceteris paribus phenomenon in preferences [2]. Ceteris paribus means "all else being equal" and is applied when two attribute values are compared. Let us adopt the example from the paper - a user buying a notebook. When we want to compare e.g. two sizes of harddisk, we say that 250GB is preferred to 80GB ceteris paribus. This can be translated to a sentence:"Imagine two notebooks x and y, where x has the size of the harddisk 250GB and y has 80GB. All other attribute values are the same. Then x is always preferred to y."

The opposite consequence of ceteris paribus is that there can be a relation between two attributes A_1 and A_2, so that the ceteris paribus can not be applied for them. In Figure 1 is the example of the price being dependent on the value of the producer that was given in the paper. For producers HP, IBM, Lenovo, Toshiba and Sony the ideal price is 2200$ and for Fujitsu-Siemens, Acer, Asus and MSI the ideal price is 750$.

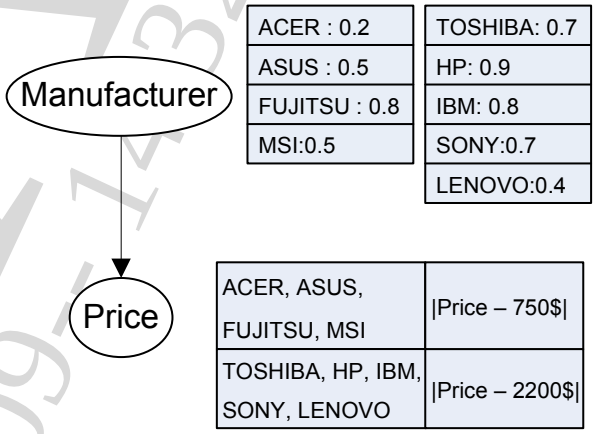


Figure 1: Example of a CP-net representing data about notebooks.

The relation between attribute preferences can be learnt, which was the task of the paper. We want to find the relation between the producer and the price of a notebook, having a small number of notebooks rated by the user. The rating can be represented by stars or school marks, but often it can be transformed to the set {1,2,3,4,5}. A general user preference model that would be able to predict the rating of all objects is constructed on the basis of these ratings.

Our user model was described in the paper, but the main focus was on the learning of the relation between attribute preference between a numerical attribute such as the price and a nominal one, such as the producer. We wanted to extend the approach to nominal attributes,

DRAFT
A. Jirka 2009 - 14:32

which we have already done. 2CP regression is the method proposed in the paper. For a numerical attribute A_1, it tries the values of other nominal attributes (e.g. A_2) and tries to find, if there is a relation between the values of A_2 and the preference of the values of A_1. In our example, instead of trying to make the regression of the price over all training set, we do it for a set of notebooks of a particular producer, which is significantly smaller. We are able to distinguish the two ideal prices (750$ and 2200$) of notebooks in this way at the cost of reducing the training set size.

1.1. Results

We present also a sample of results of the experiments from the paper.

Experiments was done on a set of 200 notebooks. The rating of notebooks was calculated by a set of functions - every attribute had an ideal value, the aggregation of preferences of attributes was done using a weighted average function. Price was transformed according to the example described in the text - for producers HP, IBM, Lenovo, Toshiba and Sony the ideal price was set to 200$ and for the rest to 750$.

We tested our method Statistical with preprocessing using linear regression and also using 2CP regression. For comparison, support vector machines and multilayer perceptron was also tested. Method Mean always returns the average rating from the training set, so it can be considered as the most simple method. Deeper description of the methods are in [1].

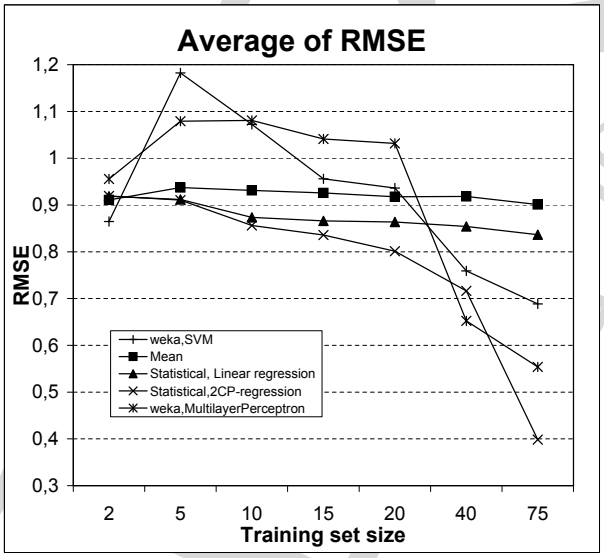


Figure 2: RMSE.

The results for various sizes of the training set are presented in Figure 2. The error measure in the figure is RMSE - root mean squared error. We can see that 2CP regression performs the best in the average. This was confirmed also by other error measures, which are described in [1].

2. Future work

In the future, we plan to find a measure of relation between two attributes. The relation is always used in our current approach, no matter if it is only a chance or statistical variation for a particular attribute value. If we could quantify the amount of relation between the two attributes, this added information may be used to improve the process of learning.

When 2CP regression is applied, the size of the training set decreases proportionally to the size of the domain of the influencing attribute (the producer). This affects the reliability of the learning - the smaller is the set, the more role the noise present in the data plays. One of the possible solutions for this is the clustering of attribute values - in our example, there were only two sets of producers, but the 2CP regression learns for each producer alone. The possibility to find a similar results of learning and cluster them together may radically improve the overall reliability and robustness of the algorithm. Initial experiments with clustering have not turned out very well, but we still think that this can be a good way.

Acknowledgment

The work on this paper was supported by Czech project 1ET 100300517.

References

[1] A. Eckhardt and P. Vojtáš, "How to learn fuzzy user preferences with variable objectives," in *To appear in proceedings of International Fuzzy Systems Association World Congress, 2009 (IFSA 2009)*, 2009.

[2] C. Boutilier, R.I. Brafman, H.H. Hoos, and D. Poole, "Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements," *Journal of Artificial Intelligence Research*, vol. 21, p. 2004, 2004.

Modulárne ontológie

doktorand:
ING. ALENA GREGOVÁ
Fakulta mechatroniky, informatiky a mezioborových studií
Technická univerzita v Liberci
Hálkova 6
461 17 Liberec, Česká republika
alena.gregova@tul.cz

školitel:
ING. JÚLIUS ŠTULLER, CSC.
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
stuller@cs.cas.cz

obor studia:
Technická kybernetika

Abstract

Problémy s veľkými monolitickými ontológiami z hľadiska rozšíriteľnosti, znovupoužitia, dostupnosti a podpory viedli k narastajúcemu záujmu o modularizáciu ontológií.
Modularizácia, ako taká, uľahčuje jednoduchšie doplňovanie nových poznatkov do existujúcich znalostí. Okrem toho môže zaistiť pre človeka aj ich väčšiu zrozumiteľnosť.
Hlavný cieľ modularizácie sa týka problematiky, ako môžu byť moduly navrhnuté, charakterizované a riadené. Využíva sa pritom deskripčná logika (DL), grafické komponenty a konceptuálne modelovanie.

1. Úvod

Nedeliteľnou súčasťou Sémantického Webu sú *ontológie*. Problému, ako *opakovane používať* ontológie sa venuje *modularizácia*. V súčasnosti existujú dve hlavné úrovne znovupoužitia ontológií, a to v rámci:

1. `ontologického jazyka OWL,` ktorý ponúka možnosť importovať OWL ontológie príkazom `<owl:imports>` [1]
2. `ontologických editorov,` napríklad Protegé, PATO, SWOOP, KMi, Prompt.

Prostredníctvom OWL jazyka je možné spojiť niekoľko OWL ontológií do jednej väčšej. Avšak takéto *syntaktické riešenie* nemusí byť v obecnom prípade úplne dostačujúce a v obecnosti *neumožňuje efektívne opakované použitie* určitých častí ontológie (tzv. *modulov*). Dôsledok tohto nedostatku môže spôsobiť neočakávanú nezlúčiteľnosť alebo nedostatočnú výkonnosť [2].

V prípade, ak sa jedná o naozaj veľké ontológie, editory ako Protegé a iné sú schopné spracovávať iba učitú časť z pôvodnej ontológie, čo by mohlo viesť k strate znalosti. Tento nedostatok je jedným z dôvodov skúmania modularizácie ontológií (MO).

Cieľom MO je analyzovať špecifické podmnožiny pôvodnej ontológie, ktoré nazývame **modulmi** [2]. To však nie je jediným cieľom modularizácie. Zaoberá sa aj *rozšíriteľnosťou z pohľadu získavania, vývoja a údržby znalosti*. Medzi ďalšie patrí *pochopiteľnosť* a *personalizácia* [4].

Jedno z intuitívnych chápaní modulu je *podmnožina celku*, kde celok predstavuje samostatnú ontológiu.

Modularizácia môže byť chápaná dvojakým spôsobom:

1. **Dekompozícia:** predstavuje *proces rozkladu veľkej ontológie na malé moduly*, kde začiatočným bodom je ontológia ako celok a finálnym nové moduly [4].
2. **Kompozícia:** predstavuje opačný proces k dekompozícii, to znamená, že *(menšie) ontologické moduly sú skladané do väčšej ontológie*. Štartovacím bodom je sada modulov, ktoré predstavujú *budúce zoskupenie*, a výstupným nová ontológia [4].

Pre pochopenie podstaty modularizácie je potrebné objasniť nasledujúce body:

1. Modul predstavuje zoskupenie sady konceptov, relácií, axióm a inštancií. Zásadnou otázkou je, *ako presne definovať takúto množinu*.
2. Hlavné použitie modulu je ako *komponenta prispievajúca k vytvoreniu novej ontológie*. Otázkou je, *ako takáto kompozícia môže byť riešená*.

DRAFT 14 srpna 2009 - 12:52

3. *Ako môže byť modul spojený s ďalším modulom, aké mapovanie môže byť definované medzi modulmi, ako môžu byť použité* [4].

2. Ciele modularizácie

Pre presné pochopenie čo modularizácia znamená, aké má výhody a nevýhody v spojitosti s ontológiami, je potrebné definovať jej základné ciele, ktorými sú:

1. **Rozšíriteľnosť**
   - *Rozšíriteľnosť pre* **vyhľadávanie znalosti:** Tu je kritériom modularizácie určenie a vymedzenie priestoru pre vyhľadávanie znalostí, čo si vyžaduje potrebné vedomosti o skúmanom priestore.
   - *Rozšíriteľnosť pre* **vývoj a údržbu:** Tu je kritériom modularizácie *aktualizácia ontológií*. Tento prístup si vyžaduje pochopenie stability informácií o ontológiach.
2. **Zrorumiteľnosť:** Dôležitým faktorom je *veľkosť ontológií. Obsah malých je ľahšie pochopiteľný* a naopak. A taktiež je potreba rozlíšiť, či *používateľom ontológií je človek*, alebo *inteligentný agent*.
3. **Personalizácia:** Poskytuje *kritéria pre dekompozíciu* ontológií na menšie moduly.
4. **Znovupoužitie:** predstavuje *základnú motiváciu kompozičného prístupu*. Ale takisto môže byť aplikované do dekompozičného prístupu. Hlavnou úlohou je opätovné použitie vytvorených modulov. Dochádza k maximalizácií možnosti modulov na:
   - pochopenie
   - výber
   - použitie ďalšími službami a aplikáciami [4].

3. Definícia a popis

Je niekoľko definícií (ontologického) modulu.

- Jedna z nich [3] definuje modul ako *opätovne používanú komponentu/prvok väčšej* alebo *komplikovanejšej ontológie*, ktorý je *samostatný/uzavretý*, ale zároveň v sebe zahŕňa vzájomný vzťah vzhľadom k inému modulu. Táto definícia hovorí, že moduly môžu byť znovupoužité buď tak ako už sú, alebo ich rozšírením prostredníctvom nových konceptov a vzájomných vzťahov.
- Ďalšia definícia [5] tvrdí: "aby proces znovupoužitia modulov bolo možné vykonať, je potrebné zabezpečiť, že moduly sú sebeobsažné (bez referencií na ďalšie koncepty). Inými slovami modul je *samostatnou podmnožinou rodičovskej ontológie*".
  Ontologický modul je sebeobsažný/samostatný, ak všetky koncepty modulu sú definované pomocou iných konceptov modulu a nevytvárajú odkazy/referencie na nejaký iný koncept mimo daného modulu [3].
- Podľa [6] je modul definovaný ako objekt predstavujúci minimálnu podmnožinu axióm v ontológii, ktorá dostatočne presne zachycuje význam určitých pojmov.
- Podľa [1] modul M_i(O) ontológie O je množina axióm (podtrieda, rovnocennosť, konkretizácia atď) taká, že platí Sig(M_1(O) $\subseteq$ Sig(O)), kde Sig(O) je signatúra (sada mien vyskytujúcich sa v axiómach ontológie O).
  Podľa tejto definície úlohou dekompozičného prístupu je rozdelenie axióm ontológie na množinu modulov {M_1,...M_k} tak, že každé M_i je modul a zjednotenie všetkých modulov je sémanticky ekvivalentné pôvodnej ontológie O. Pre tento prístup sú vytvorené editory ako napríklad PATO a SWOOP. Okrem **dekompozičného prístupu** [1] uvádza aj **prístup extrakcie modulu**. Jeho úlohou je redukovanie ontológie na modul, ktorý pokrýva konkretný pod-slovník SV (Sub-Vocabulary). Inými slovamie, ak existuje ontológia O a sada SV $\subseteq$ Sig(O) výrazov, mechanizmus extrakcie modulu vráti modul M_{SV}. M_{SV} predstavuje relevantnú časť ontológie, ktorá pokrýva SV, (Sig(M_{SV}) $\supseteq$ SV). Pre tento prístup boli vytvorené editory ako napríklad KMi a Prompt.

Ontológiu je možné chápať ako dvojicu: **O=(C, R)**, kde
O - ontológia
C - množina konceptov: C = {C_1,...C_n}
R - množina rolí: R = {R_1(a, b)...R_n(a, b)}

a modul ontológie ako dvojicu: $\mathbf{O}_M$ **=** ($\mathbf{C}_M$**,** $\mathbf{R}_M$**)**, kde
$C_M \neq \emptyset \wedge C_M \subseteq C$
$R_M \subseteq R$
symbolicky: $O_M \subseteq O$

DRAFT 24. srpna 2009 - 14:39

Modularizácia nutne neznamená, že každý modul ontológie je disjunktný s ostatnými modulmi danej ontológie. Napríklad ak A má potriedy B a C, potom vytvorenie modulu z A by malo obsahovať všetky tri koncepty, ale vytvorenie modulu z triedy B iba jeden a to B. V tomto prípade modul B nie je disjunktný s modulom A [3].

Väčšina ontológií je vyvýjaná za *predpokladu otvoreného sveta* OWA `(Open World Assumption)` [7], to znamená, že sú povolené referencie na koncepty mimo ontologického modulu. Avšak aby bolo možné získať sebeobsažný modul, je potrebný *predpoklad zatvoreného sveta* CWA `(Close World Assumption)` [7], čiže nie sú povolené referencie na koncepty mimo modulu [3].

Podľa [4] je moduly možné deliť na tzv. *zatvorené* a *otvorené*:
Zatvorené: ak modul nie je spojený s ďalším modulom
Otvorené: ak modul je spojený s ďalším modulom [4].

4. Dekompozičný prístup

Cieľom je spracovávať veľké, zložitejšie ontológie, ktoré môžeme nájsť napríklad v medicíne alebo biológií. Ich veľkosť a doména sú ťažšie pochopiteľné a spracovateľné. Z tohto dôvodu sa vyvýjajú metódy, ktoré by boli schopné automaticky rozdeľovať zložité ontológie na menšie moduly [4].

Jednou z týchto metód je metóda rozdeľovania, ktorá využíva techniky zo sieťovej analýzy, je však schopná rozdeliť iba jednoduchšie hierarchické štruktúry. Ale ontológie, ako napríklad v spomínanej medicíne, pozostávajú zo zložitejších hierarchií. Preto využívajú expresívnu silu ontologického jazyka OWL. Cieľom je adaptovať túto metódu rozdeľovania jednoduchej hierarchickej štruktúry do viac expresívnejších ontológií, najmä do ontológií kódovaných v OWL [4].

4.1. Algoritmus dekompozície

Podľa [8] algoritmus dekompozičného prístupu pozostáva z *troch úloh*, ktoré vyplývajú zo závislostí medzi konceptmi. V prvom rade je potrebné vytvoriť *závislostný graf* z definície ontológie. Druhou úlohou je vytvorenie aktuálneho rozdelenia podľa už vytvoreného grafu. A posledným krokom je *optimalizácia rozdelenia* a to na základe zistených izolovaných konceptov, spojenia niektorých modulov a opakovania vybraných axióm.

4.1.1 Vytvorenie závislostného grafu: (Pomocné pojmy potrebné k vytvoreniu grafu:

- *Podtrieda:* v rámci hierarchie tried
- *Vlastnosti:* Keď sa zavádzajú vlastnosti, doména a rozsah každej vlastnosti sú medzi sebou prepojené.
- *Definície:* Relácie definície sú medzi konceptami a výrazmi, ktoré sú obsiahnuté v ich definícii. Využíva sa to pri vytváraní konceptov, ktoré sú závisle na niektorej spoločnej vlastnosti.
- *Podreťazec:* Ďalšia relácia sa týka mien konceptov, ak meno jedného konceptu je obsiahnuté v inom. Relácia reťazca je vhodná v prípade ak výrazy ontológie majú tzv. "kompozičnú štruktúru".)

Ešte pred samotným vytvorením závislostného grafu je potrebná konverzia ontológie v OWL, RDF alebo KIF formáte do váhového grafu - spočíva vo vytvorení grafu a spočítaní váhy.

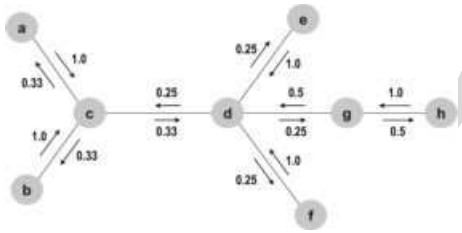
- **Vytvorenie grafu:** Hlavnou myšlienkou je, že elementy (koncepty, relácie, inštancie) sú reprezentované uzlami v grafe. Medzi jednotlivými uzlami sú spojenia v prípade, ak medzi elementami sú určité súvislost [4]. Typy týchto spojení sú práve vyššie uvedené pomocné pojmy.
- **Určenie sily/mohutnosti závislosti:** Určuje sa sila medzi konceptmi. Pomocou algoritmu zo sieťovej analýzy sa vypočíta stupeň príbuznosti medzi konceptmi. Potom sa určujú váhy medzi rôznymi časťami závislosti, napríklad relácie podtriedy majú väčší vplyv ako relácie domény. Na určenie váh závislostí sa použije štruktúra zo závislostného grafu `(DP - Depedency Graf)`. Nakoniec sa vypočíta **proporcionálna sila (PS) w** pre tento graf.
  PS popisuje význam dôležitosti spojenia od jedného uzlu k ďalším na základe počtu spojení, ktorý tento uzol má. Vypočíta sa ako podiel súčtu váh spojení medzi uzlom c_i a uzlom c_j a súčtu váh všetkých spojení c_i k ďalším uzlom.

$$w(c_i, c_j) = \frac{a_{ij}+a_{ji}}{\sum_k a_{ik} + a_{ki}}, \text{ kde}$$

a_{ij} - váha spojenia medzi uzlom c_i a uzlom c_j
$w(c_i, c_j)$ - PS spojenia medzi uzlom c_i a uzlom c_j

DRAFT 14. 5. 2009 - 14:32

V následujúcom obrázku napríklad uzol **d** má 4 spojenia s ďalšími uzlami, čo znamená, že proporcionálna sila k susedným uzlom je **0.25**, teda (1/4).
Iná úroveň závislosti medzi **d** a jeho susedmi vychádza zo vzájomnej závislosti susedov s uzlom **d**, (PS je nesymetrická). Napríklad medzi **e** a **f** ja PS rovná **1**, keďže oba tieto uzli majú len jedno spojenie s uzlom **d**. Sila závislosti medzi **g** a **d** je **0.5**, keďže **g** má dvoch susedov.



Obrázok 1: Príklad grafu s proporcionálnou silou závislosti

4.1.2 Identifikácia modulov - určenie modulu: Pomocou PS sieti sa určia množiny súvisiacich konceptov. Použije sa algoritmus, ktorý vypočíta všetky maximálne "LI" (Line Islands) daného grafu.

Podľa predchádzajúceho obrázku môžeme definovať množinu **{a,b,c,d,e,f}**, ktorá vytvára spojený subgraf. Tzv. "*maximálne rozvetvený strom*" tejto množiny pozostáva z hrán a ich proporcionálnych síl:

- a $\longrightarrow^{(1.0)}$ c
- b $\longrightarrow^{(1.0)}$ c
- c $\longrightarrow^{(0.33)}$ d
- e $\longrightarrow^{(1.0)}$ d
- f $\longrightarrow^{(1.0)}$ d

Avšak táto množina nepredstavuje LI, pretože minimálna váha stromu je **0.33** a to medzi uzlami **c** a **d**, popritom váha spojenia medzi uzlom **g** $\longrightarrow$ **d** je **0.5**, to znamená, že spojenie medzi uzlami **g** a **d** má väčšiu PS.
Zvyšná množina uzlov **{g,h}** spĺňa podmienky LI. Táto množina vytvára spojený subgraf, kde PS **h** $\longrightarrow^{(1.0)}$ **g** a maximálna hodnota vstupných a výstupných spojení je **0.5 (g** $\longrightarrow$ **d)**. No napriek tomu tento rozvetvený strom stále nie je optimálny. Úplne podmienky spĺňa množina **{d,e,f,g,h}**, ktorá predstavuje LI s maximálne rozvetveným stromom:

- e $\longrightarrow^{(1.0)}$ d
- f $\longrightarrow^{(1.0)}$ d
- g $\longrightarrow^{(0.5)}$ d
- h $\longrightarrow^{(1.0)}$ g

4.1.3 Optimalizácia - určenie izolovaných konceptov: V niektorých prípadoch rozdelenia veľkej ontológie na menšie moduly môže nastať, že ostanú samostatné uzly, ktoré nie sú priradené k žiadnej skupine. Preto algoritmus automaticky priradí tieto uzly k modulu a to na základe sily relácie, teda na základe najsilnejšieho spojenia. Ide vlastne o LI susediacich uzlov s najsilejšou reláciou.

4.1.4 Optimalizácia - zlúčenie: Použitím spomínaného algoritmu sa generujú moduly. V niektorých prípadoch podstromy, ktoré sú určené k formovaniu modulu, sú ďalej rozdeľované. A to aj v prípade, ak kompletný podstrom neprevýšil určitú hranicu veľkosti. Môže to byť zapríčinené nesymetrickým vytváraním modulov z ontológie ako podstromov, ktoré majú tendenciu sa ďalej deliť na koncepty.

Počas kontrolovania závislosti v relevantných častiach ontológie sa môžu vyskytovať problematické moduly, ktoré majú silné vnútorné závislosti. Aby sa mohlo predísť tejto situácií, je potrebné merať vnútornú závislosť. Toto meranie je známe ako "`height of island`" a je určené pomocou tzv. "minimálneho položeného stromu" **T** pre identifikáciu modulov.
Celková sila vnútornej závislosti sa rovná sile najslabšieho spojenia v položenom strome **T**.

$$\text{height}(I) = \min w(u,u')$$

5. Hodnotenie výsledku modularizácie

V [9] autor popisuje sadu kritérií, ktoré sú založené na štruktúre ontológií a výsledku modularizácie, a ktoré sú navrhnuté pre údržbu a efektivitu usudzovania, a to prostedníctvom použitia tzv. "distribuovaných modulov".

Medzi tieto kritéria patrí:

1. **Veľkosť:** relatívna veľkosť modulu (počet tried a ich vlastností patrí medzi dôležitejšie ukazovatele efektívnosti modularizačných techník. Veľkosť modulu má vplyv na jeho údržbu a robustnosť aplikácie.

DRAFT 14. júna 2009 - 14:32

2. **Nadmernosť:** v prípade prekrývania modulov pri rozdeľovaní takisto dochádza k zlepšeniu efektívnosti a robustnosti. Na druhej strane s nadbytočnými znalosťami sa zvyšuje aj ich údržba.

3. **Spojitosť:** Vzhľadom k nezávislosti medzi jednotlivými výslednými modulmi sa môže očakávať nespojitosť generovaných modulov. Spojitosť modulov, ktoré sú v grafe reprezentované pomocou uzlov, je hodnotená na základe počtu strán.

4. **Vzdialenosť:** vzdialenosť sa zisťuje pomocou merania, ako sa výrazy popísané v moduli priblížujú ku každému ďalšiemu v porovnaní s pôvodnou ontológiou. Vzdialenosť "intra-modulu" je vyjadrená počtom relácií po najkratšej ceste od jednej entity k druhej. Táto vzdialenosť, čiže spočítanie počtu modulov, ktoré spájajú dva objekty, predstavuje spôsob komunikácie medzi jednotlivými modulmi rozdelenej ontológie.

6. Záver

V príspevku sa zaoberám základnými vlastnosťami modularizácie a dôvodom prečo je potrebné zavádzať modularizáciu. Sú popísané základné ciele a podrobnejšie je rozobratý jeden z prístupov a to dekompozičný prístup, kde dôležitým krokom je algoritmus, ktorý spočíva vo vytvorení závislostného grafu. V grafe sa určuje mohutnosť závislosti a to na základe proporcionálnej sily. Určenie týchto síl je bližšie popísane na príklade závislostného grafu. Pomocou týchto metód sa vyhodnocujú závislosti medzi konceptmi a určuje sa ich dôležitosť a prioritnosť.

Literatura

[1] Mathieu d'Aquin, Anne Schlicht, Heiner Stuckenschmidt, and Marta Sabou, "Ontology Modularization for Knowledge Selection: Experiments and Evaluations"

[2] Camila Bezerra, Fred Freitas, Jérôme Euzenat, Antoine Zimmermann, "ModOnto: A tool for modularizing ontologies"

[3] Paul Doran, "Ontology reuse via ontology modularisation," Department of Computer Science, University of Liverpool, Liverpool, L69 3BF, UK.

[4] Stefano Spaccapietra, "Report on Modularization of Ontologies," Institute of Computer Science, Austria

[5] Heiner Stuckenschmidt and Michel Klein, "Integrity and Change in Modular Ontologies," Vrije Universiteit Amsterdam de Boelelaan 1081a, 1081HV Amsterdam heiner.

[6] Bernardo Cuenca Grau, Bijan Parsia, Evren Sirin and Aditya Kalyanpur, "Modularizing OWL Ontologies," University of Maryland at College Park.

[7] Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens and Chris Wroe, "A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and COODE Tools," Edition 1.0, The University, Of Manchester, Stanford University Manchester, August 2004.

[8] Heiner Stuckenschmidt and Anne Schlicht, "Structure-Based Partitioning of Large Ontologies," Universität Mannheim, Germany.

[9] A. Schlicht and H. Stuckenschmidt, "Towards Structural Criteria for Ontology Modularization," In: Proc. of the ISWC 2006 Workshop on Modular Ontologies (2006).

DRAFT 14. srpna 2009, 14:32

Gradient Learning of Spiking Neural Networks

Post-Graduate Student:
BC. LUKÁŠ HOŠEK
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25
118 00 Prague 1, CZ
lukas.hosek@gmail.com

Supervisor:
ING. RNDR. MARTIN HOLEŇA, CSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
martin@cs.cas.cz

Field of Study:
Theoretical Computer Science

Abstract

This paper discusses two methods of gradient descent learning for Spiking Neural Networks (SNNs). We shortly describe the network architectures and algorithms used in these two particular approaches and discus the properties and limitations of each method. In addition, we describe an approach for coding continuous input variables using a population of receptive fields.

1. Introduction

Neural computational models with sigmoidal transfer function are well established and explored. While inspired by biological neurons, they differ in one significant aspect: in biological neurons, information is not encoded as continuous values, but rather as a series of spikes being propagated across the network. It has been a common belief for many years that the essential information in biological networks is represented as neuron's rate of fire. In that frame of reference, output of sigmoidal neurons can be interpreted as rate of fire. Recent research has, however, shown that phase or precise timing also constitutes a significant portion of information in biological systems. For example, precise timing is crucial for generating a smooth movement in neuroprosthetic systems which aim at producing useful movements of paralyzed limbs [1].

piking neurons have recently emerged as a more biologically plausible alternative to sigmoidal neurons. Since they naturally operate in temporal domain, they are generally recognized to be capable of processing temporally coded information in a much more sophisticated way than typical neural computational models. It also has been proven that networks of spiking neurons can simulate arbitrary feedforward sigmoidal networks [2] and shown theoretically that SNNs are computationally more powerful that networks with sigmoidal activation function [3]. However, finding an efficient mechanism for learning spiking neural networks is still an open problem.

Numerous approaches to supervised learning of SNNs which don't utilize gradient descent have been proposed, such as a strictly mathematical method where authors define algebraic operations on time series and use these in an iterative algorithm for learning spiking patterns [8], an algorithm which utilizes a chaining rule to find links between neurons firing in the desired contiguity [9], a probabilistic algorithm which maximizes the probability of output neurons firing at desired times [10] or an evolutionary algorithm for modifying synaptic weights [11].

In this article, we explore approaches which utilize gradient descent in a fashion similar a classical back-propagation in sigmoidal neural networks: SpikeProp, devised by Bohte et al. in [4] and a method devised by Jiří Šíma in [5].

2. SpikeProp

2.1. Architecture

The network used in SpikeProp can be defined as a set V of spiking neurons connected into an oriented network. Some of these neurons serve as inputs (denoted H) or outputs (denoted J).

Each neuron generates at most one spike during the simulation interval. For input neurons, firing times are given externally. For non-input neurons, they are calculated as follows:

For each neuron $j \in V \setminus H$ we denote as Γ_j the set of its immediate predecessors. For each $i \in \Gamma_j$ the connection between i and j consists of multiple synaptic terminals,

DRAFT 14. April 2009, 14:52

each of them is assigned a *delay* (denoted d_{ij}^k for the k-th terminal) and a *weight* (denoted w_{ij}^k). Firing time t_j of neuron j is calculated from the firing times of its immediate predecessors as the time when its internal state variable reaches threshold ϑ for the first time. The internal state variable x_j is a weighted sum of all pre-synaptic contributions:

$$x_j = \sum_{i \in \Gamma_j} \sum_{k=1}^{m} w_{ij}^k y_{ij}^k(t)$$

The pre-synaptic contribution of a single synaptic terminal is defined as

$$y_i^j(t) = \varepsilon(t - t_i - d^k)$$

where ε is the spike-response function of of the form

$$\varepsilon(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{t}{\tau} e^{1-\frac{t}{\tau}} & \text{if } t \geq 0 \end{cases}$$

modelling a simple α-function for $t > 0$, and τ is a constant determining rise and decay time of the pre-synaptic pulse.

2.2. Learning rule

The learning rule is derived in a fashion similar to classic back-propagation of sigmoidal networks. We supply the algorithm with an input pattern, denoted $P[t_1, \dots t_h]$ and target firing times of output neurons, denoted $\{t_j^d\}$. First, we calculate actual firing times of output neurons for current network settings, denoted $\{t_j^a\}$. Given these, we can define the error function:

$$E = \frac{1}{2} \sum_{j \in J} (t_j^a - t_j^d)^2$$

Each synaptic terminal is treated separately and its weight is modified using gradient descent:

$$\Delta w_{ij}^k = -\eta \frac{\partial E}{\partial w_{ij}^k}$$

η being the learning rate. The derivative can be expanded to:

$$\frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial t_j}(t_j^a) \frac{\partial t_j}{\partial w_{ij}^k}(t_j^a) = \frac{\partial E}{\partial t_j}(t_j^a) \frac{\partial t_j}{\partial x_j(t)} \frac{\partial x_j(t)}{\partial w_{ij}^k}(t_j^a)$$

For a small enough region around $t = t_j^a$, x_j is assumed to be approximable by a linear function of t, hence the local derivative of t_j with respect to $x_j(t)$ is assumed to be constant (which implies that for larger values of η the algorithm will be less effective).

Derived back-propagation equations for a fully connected network are as follows:

$$\frac{\partial E}{\partial w_{ij}^k} = y_{ij}^k(t) \delta_j$$

where for output neurons, δ_j equals

$$\delta_j = \frac{-(t_j^a - t_j^d)}{\sum_{i \in \Gamma_j} \sum_k w_{ij}^k \frac{\partial y_{ij}^k(t)}{\partial t}}$$

and for hidden neurons we have

$$\delta_j = \frac{\sum_{i \in \Gamma_j} \delta_i \sum_k w_{ij}^k \frac{\partial y_{ij}^k(t)}{\partial t}}{\sum_{i \in \Gamma_j} \sum_k w_{ij}^k \frac{\partial y_{ij}^k(t)}{\partial t}}$$

2.3. Encoding of continuous input data

Various approaches have been developed for coding of continuous input variables, the one most at hand being coding one input variable directly into firing time of one input neuron. The effectiveness of this approach, however, decreases with increasing size of the dataset: inputs have to be encoded with increasingly smaller temporal differences and since the network operates in fixed time steps, temporal resolution of the simulation has to be increased to produce sufficiently fine-grained results, which in turn imposes a computational penalty on the whole network.

Another approach devised in [4] works with encoding a single variable n into a population of m neurons. Each of the neurons represents a Gaussian receptive field:

Let the range of n be $[I_{min}^n, I_{max}^n]$. The m neurons, which we will use for encoding n, represent an array of one-dimensional receptive fields. For i-th neuron in the array, the center of its Gaussian receptive field is $E = I_{min}^n + i\frac{I_{max}^n - I_{min}^n}{m-2}$ and width $\sigma = \beta\frac{I_{max}^n - I_{min}^n}{m-2}$. The stimulation of i-th receptive field is then calculated as $G(E, \sigma; n)$. The values of each receptive filed are then converted to firing times, associating the highest response with $t = 0$ and increasingly lower responses with later firing times, up to $t = 10$. Resulting spike times are then rounded to the nearest internal time step. Empiric tests show better results when neurons with very low excitation levels (i.e. $t > 9$)are coded not to fire at all. This approach also has the advantage of producing sparse coding, which allows for optimizations such as event-based network simulation. Accuracy of representation can be controlled by varying the number of neurons and sharpness of receptive fields (experimental results show that optimal values for β lie between 1.0 and 2.0). This encoding has been shown to be statisticaly bias-free [6].

DRAFT Tatranská Lomnica 2009 - 14.-17.9.

2.4. Results

The abilities of SpikeProp have been tested on a set of experiments, including standard and interpolated XOR and various common classification benchmarks (the Iris dataset, the Wisconsin breast cancer dataset and the Statlog Landsat dataset). The results were comparable to that of sigmoidal networks; in addition SpikeProp always converged in experiments on real world datasets, whereas algorithms such as Levenberg-Marquardt ocassionaly failed.

In the original proposition, only positive weights were allowed. Other experiments [7] showed that negative weights could also be allowed and stil lead to succesfull convergence, which is in contradiction to Boohte's original conclusion, according to which allowing mixed weights would cause contributions of single neuron-neuron connections to no longer be a monotonically increasing function.

3. Smoothly spiking neural networks

From the description of SpikeProp, certain shortcomings are immediately apparent. First of all, the architecture allows for only one spike per neuron and subsequently only for time-to-first-spike coding. Secondly, there is no rule for modifying delays. Instead, multiple synaptic terminals for each connection are used, each with a hardcoded delay. The following approach to SNNs proposed by Jiří Šíma in [5] presents a modified version of Spike Response Model SRM_0 with smooth dynamic of spike creation and deletion. This model can naturally cope with multiple spikes per neuron. A non-trivial back-propagation rule is derived for calculating gradients of the error function with respect to both synaptic weights and delays.

3.1. Architecture

The network is defined as a set V of smoothly spiking neurons connected into a directed graph. We denote by $X \subseteq V$ the set of input neurons and by $Y \subseteq V$ the set of output neurons. For each neurons j we define $j_{\leftarrow}$ as the set of all neurons from which a synapse leads to j and $j_{\rightarrow}$ as the set of all neurons to which a synapse leads from j. Each synapse leading from i to j is assigned a weight w_{ij} and delay d_{ij}. We denote as $\mathbf{w}$ and $\mathbf{d}$ the vector of all network weight and delay parameters.

The simulation runs in timeframe $[0, T]$. During the course of the simulation, each neuron j produces a sequence of p_j spikes. The firing times of these spikes are denoted as $0 < t_{j1} < \ldots < t_{jp_j} < T$. Additionally, we formally define $t_{j0} = 0$ and $t_{jp_j+1} = T$. For input neurons, the firing times are given externally. For each noninput neuron $j \in V \setminus X$, firing times are calculated as the time instants when its *excitation*

$$\xi(t) = w_{j0} + \sum_{i \in j_{\leftarrow}} w_{ij} \varepsilon(t - d_{ij} - \tau_i(t - d_{ij})) \quad (1)$$

evolving in time $t \in [0, T]$ crosses 0 from below, i.e.

$$\{t \mid 0 \leq t \leq T \;\&\; \xi_j(t) = 0 \;\&\; \xi_j'(t) > 0\} = \{t_{j1} < t_{j2} < \ldots < t_{jp_j}\}$$

Neuron's excitation is calculated as a weighted sum of delayed responses from its immediate antecedents. ε in (1) is the *response function*, defined as follows:

$$\varepsilon(t) = e^{-(t-1)^2} \cdot \sigma_0(t)$$

σ is an auxiliary function used as a smooth approximation of the stair function:

$$\sigma(\alpha, \beta, \delta; x) = \begin{cases} \alpha & \text{if } x < 0 \\ (\beta - \alpha)((6\frac{x}{\delta} - 15)\frac{x}{\delta} \\ \quad +10) \cdot (\frac{x}{\delta})^3 + \alpha & \text{if } 0 \leq x \leq \delta \\ \beta & \text{if } x > \delta \end{cases}$$

$$\sigma_0(t) = \sigma(0, 1, \delta_0; t)$$

$\tau_j(t)$ is a smooth approximation of the last firing time of neuron j lower than t. First we have to define transformed firing times of neuron j:

$$\widetilde{t_{js}} = \begin{cases} t_{js} & \text{for } j \in X \\ \widetilde{\sigma(t_{j,s-1}, t_{js}, \delta; \xi_j'(t_{js}))} & \text{for } j \in V \setminus X \end{cases}$$

Given these, the function τ itself is then defined as

$$\tau_j(t) = \sum_{s=1}^{p_j+1} (\widetilde{t_{js}} - \widetilde{t_{j,s-1}}) P^C(t - \widetilde{t_{js}}) \quad (2)$$

P is the logstic sigmoid function with a real gain parameter λ:

$$P(\lambda; x) = \frac{1}{1 + e^{-\lambda x}}$$

$C \geq 1$ in (2) is an optional exponent.

3.2. Learning rule

The algorithm is supplied a set of inputs, specifying firing times $0 < t_{i1} < t_{i2} < \ldots < t_{ip_i} < T$ for every input neuron $i \in X$ and desired firing times of output neurons $0 < \rho_{j1} < \rho_{j2} < \ldots < \rho_{jq_j} < T$ for each $j \in Y$. Given these, we can calculate the error function:

$$E(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{j \in Y} \sum_{s=0}^{q_j} (\tau_j(\rho_{j,s+1}) - \rho_{js})^2$$

Each subsequent generation of $\mathbf{w}$ and $\mathbf{d}$ is derived from the previous one using gradient descent method:

$$w_{ij}^{(t)} = w_{ij}^{(t-1)} - \alpha \frac{\partial E}{\partial w_{ij}}(\mathbf{w}^{(t-1)}) \quad \text{for } i \in j_{\leftarrow} \cup \{0\}$$

$$d_{ij}^{(t)} = d_{ij}^{(t-1)} - \alpha \frac{\partial E}{\partial d_{ij}}(\mathbf{d}^{(t-1)}) \quad \text{for } i \in j_{\leftarrow}$$

First, a list P_j of m_j ordered triplets $(\pi_{jc}, \pi'_{jc}, u_{jc}),\ s = 0, \ldots, m_j$ is calculated for each noninput neuron $j \in V \setminus X$. From this list, partial derivatives of E with respect to all weights and delays are calculated:

$$\frac{\partial E}{\partial w_{ij}} = \sum_{c=1}^{m_j} \Big(\pi_{jc} \cdot \frac{\partial}{\partial w_{ji}} \tau_j(u_{jc}) + \pi'_{jc} \cdot \frac{\partial}{\partial w_{ji}} \tau'_j(u_{jc}) \Big) \quad \text{for } i \in j_{\leftarrow} \cup \{0\} \tag{3}$$

$$\frac{\partial E}{\partial d_{ij}} = \sum_{c=1}^{m_j} \Big(\pi_{jc} \cdot \frac{\partial}{\partial d_{ji}} \tau_j(u_{jc}) + \pi'_{jc} \cdot \frac{\partial}{\partial d_{ji}} \tau'_j(u_{jc}) \Big) \quad \text{for } i \in j_{\leftarrow} \tag{4}$$

Here, n_{js} is the smallest index such that $1 \leq n_{js} \leq s-1$ and

$$\frac{\partial t_{j,s-n_{js}}}{\partial t_{j,s-n_{js}-1}} = 0$$

Triples $(\pi_{jc_1}, \pi'_{jc_1}, u_{jc_1})$ and $(\pi_{jc_2}, \pi'_{jc_2}, u_{jc_2})$ corresponding to the same time instant $u_{jc_1} = u_{jc_2}$ can be merged into one $(\pi_{jc_1} + \pi_{jc_2}, \pi'_{jc_1} + \pi'_{jc_2}, u_{jc_1})$. Triples $(\pi_{jc}, \pi'_{jc}, u_{jc})$, where $\pi_{jc} = \pi'_{jc} = 0$ can be omitted.

The algorithm starts with output neurons, ie. $j \in Y$. The list P_j for output neurons is of the form

$$P_j = ((\tau_j(\rho_{j,s+1}) - \rho_{js}, 0, \rho_{j,s+1}); s = 0, \ldots, q_j)$$

and the partial derivatives from equation 3 are:

$$\frac{\partial}{\partial w_{il}} \tau_j(t) = \sum_{s=1}^{p_j} \frac{\partial}{\partial \widetilde{t_{js}}} \tau_j(t) \sum_{r=s-n_{js}}^{s} \Big(\prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \Big) \times \Big(\Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial \tau_i} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial \tau_i} \Big) \frac{\partial \tau_i}{\partial w_{il}} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial \tau'_i} \cdot \frac{\partial \tau'_i}{\partial w_{il}} \Big) \tag{5}$$

$$\frac{\partial}{\partial w_{il}} \tau'_j(t) = \sum_{s=1}^{p_j} \frac{\partial}{\partial \widetilde{t_{js}}} \tau'_j(t) \sum_{r=s-n_{js}}^{s} \Big(\prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \Big) \times \Big(\Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial \tau_i} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial \tau_i} \Big) \frac{\partial \tau_i}{\partial w_{il}} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial \tau'_i} \cdot \frac{\partial \tau'_i}{\partial w_{il}} \Big) \tag{6}$$

For each hidden neuron $i \in V \setminus (X \cup Y)$, P_j can be constructed once the lists P_j for all $j \in i_{\rightarrow}$. have been calculated:

$$P_i = \Big(f_{jcsr} \Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial \tau_i} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial \tau_i} \Big), f_{jcsr} \cdot \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial \tau'_i}, t_{jr} - d_{ij} \Big) \tag{7}$$

where

$$f_{jcsr} = \Big(\pi_{jc} \cdot \frac{\partial}{\partial \widetilde{t_{js}}} \tau_j(u_{jc}) + \pi'_{jc} \cdot \frac{\partial}{\partial \widetilde{t_j s}} \tau'_j(u_{jc}) \Big) \times \prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \tag{8}$$

for all $j \in i_{\rightarrow}$, $c = 1, \cdots, m_j$, $s = 1, \cdots, p_j$ and $r = s - n_{js}, \cdots, s$. The partial derivatives from equation 3 for hidden neurons are:

$$\frac{\partial}{\partial w_{ij}} \tau_j(t) = \sum_{s=1}^{p_j} \frac{\partial}{\partial \widetilde{t_{js}}} \tau_j(t) \sum_{r=s-n_{js}}^{s} \Big(\prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \Big) \times \Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial w_{ij}} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial w_{ij}} \Big) \tag{9}$$

$$\frac{\partial}{\partial w_{ij}} \tau'_j(t) = \sum_{s=1}^{p_j} \frac{\partial}{\partial \widetilde{t_{js}}} \tau'_j(t) \sum_{r=s-n_{js}}^{s} \Big(\prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \Big) \times \Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial w_{ij}} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial w_{ij}} \Big) \tag{10}$$

$$\frac{\partial}{\partial d_{ij}} \tau_j(t) = \sum_{s=1}^{p_j} \frac{\partial}{\partial \widetilde{t_{js}}} \tau_j(t) \sum_{r=s-n_{js}}^{s} \Big(\prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \Big) \times \Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial d_{ij}} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial d_{ij}} \Big) \tag{11}$$

$$\frac{\partial}{\partial d_{ij}} \tau'_j(t) = \sum_{s=1}^{p_j} \frac{\partial}{\partial \widetilde{t_{js}}} \tau'_j(t) \sum_{r=s-n_{js}}^{s} \Big(\prod_{q=r+1}^{s} \frac{\partial \widetilde{t_{jq}}}{\partial \widetilde{t_{j,q-1}}} \Big) \times \Big(\frac{\partial \widetilde{t_{jr}}}{\partial t_{jr}} \cdot \frac{\partial t_{jr}}{\partial d_{ij}} + \frac{\partial \widetilde{t_{jr}}}{\partial \xi'_j} \cdot \frac{\partial \xi'_j}{\partial d_{ij}} \Big) \tag{12}$$

Finally, we enumerate the partial derivatives. For τ we have:

$$\frac{\partial}{\partial \widetilde{t_{sj}}} \tau_j(t) = P^C(t - \widetilde{t_{sj}})(1 - C\lambda(\widetilde{t_{sj}} - \widetilde{t_{j,s-1}}) \times (1 - P(t - \widetilde{t_{sj}}))) - P^C(t - \widetilde{t_{j,s+1}})$$

$$\frac{\partial}{\partial \widetilde{t_{sj}}} \tau'_j(t) = C\lambda(((1 - C\lambda(\widetilde{t_{sj}} - \widetilde{t_{j,s-1}})(1 - P(t - \widetilde{t_{sj}}))) + \lambda(\widetilde{t_{sj}} - \widetilde{t_{j,s-1}})P(t - \widetilde{t_{sj}}))P^C(t - \widetilde{t_{sj}}) \times (1 - P(t - \widetilde{t_{sj}})) - P^C(t - \widetilde{t_{j,s+1}}) \times (1 - P(t - \widetilde{t_{j,s+1}})))$$

DRAFT 14. srpna 2009 - 14:32

for transformed firing times $\widetilde{t}$:

$$\frac{\partial \widetilde{t_{sj}}}{\partial \widetilde{t_{j,s-1}}} = \begin{cases} \frac{\partial}{\partial \alpha}\sigma(\xi'_j(t_{sj})) & \text{for } s > 1 \\ 0 & \text{for } s = 1 \end{cases}$$

$$\frac{\partial \widetilde{t_{sj}}}{\partial t_{sj}} = \Big(\frac{\partial}{\partial \beta} + \xi''_j(t_{sj}) \cdot \frac{\partial}{\partial x}\Big)\sigma(t_{j,s-1}, \widetilde{t_{sj}, \delta}; \xi'_j(t_{sj})$$

$$\frac{\partial \widetilde{t_{js}}}{\partial \xi'_j} = \sigma'(\widetilde{t_{j,s-1}}, t_{sj}, \delta; \xi'_j(t_{sj}))$$

for excitation ξ:

$$\frac{\partial \xi'_j}{\partial \tau_i} = - w_{ij}\varepsilon''(t_{sj} - d_{ij} - \tau_i(t_{sj} - d_{ij})) \times (1 - \tau'_i(t_{sj} - d_{ij}))$$

$$\frac{\partial \xi'_j}{\partial \tau'_i} = - w_{ji}\epsilon'(t_{sj} - d_{ij} - \tau_i(t_{sj} - d_{ij}))$$

for firing times t:

$$\frac{\partial t_{sj}}{\partial \tau_i} = \frac{w_{ij}\varepsilon'(t_{sj} - d_{ij} - \tau_i(t_{sj} - d_{ij}))}{\xi'_j(t_{sj})}$$

$$\frac{\partial t_{jr}}{\partial w_{ij}} = \begin{cases} -\frac{1}{\xi'_j(t_{jr})} & \text{for } i = 0 \\ -\frac{\varepsilon(t_{jr} - d_{ij} - \tau_i(t_{jr} - d_{ij}))}{\xi'_j(t_{jr})} & \text{for } i \in j_{\leftarrow} \end{cases}$$

$$\frac{\partial t_{jr}}{\partial d_{ij}} = w_{ij}\varepsilon(t_{jr} - d_{ij} - \tau_i(t_{jr} - d_{ij})) \times \frac{(1 - \tau'_i(t_{jr} - d_{ij}))}{\xi'_j(t_{jr})}$$

for ξ':

$$\frac{\partial \xi'_j}{\partial w_{ij}} = \begin{cases} 0 & \text{for } i = 0 \\ \varepsilon'(t_{jr} - d_{ij} - \tau_i(t_{jr} - d_{ij})) \times (1 - \tau'_i(t_{jr} - d_{ij})) & \text{for } i \in j_{\leftarrow} \end{cases}$$

$$\frac{\partial \xi'_j}{\partial d_{ij}} = w_{ij}\big(\varepsilon'(t_{jr} - d_{ij} - \tau_i(t_{jr} - d_{ij}))\tau''_i(t_{jr} - d_{ij}) - \varepsilon''(t_{jr} - d_{ij} - \tau_i(t_{jr} - d_{ij})) \times (1 - \tau'_i(t_{jr} - d_{ij}))^2\big)$$

This concludes the gradient calculation for networks of smoothly spiking neurons.

4. Discussion

In this review we presented two approaches to gradient learning of spiking neural networks. In SpikeProp, gradient of the error function with respect to weights is explicitly evaluated, however the derivation of the learning rule makes an assumption about linearity of the threshold function. Instead of adapting weights, this approach makes use of multiple synaptic terminals with hardcoded delays for a single neuron-neuron connection. This essentially makes the network operate on a fixed time step. The architecture allows for only one spike per neuron, fundamentally limiting data encoding options to time-to-first-spike. A population of receptive fields can be used as a biologically plausible way of encoding continuous input variables.

The second approach uses a modified architecture to make network computational dynamic completely smooth. This allows for explicit evaluation of gradient of the error function with respect to both weights and delays and removes the discontinuity of spike creation and deletion. This architecture can naturally process multiple spikes per neuron.

The smooth computational dynamic of the second approach also provides other advantages: in constrast to SpikeProp, a situation where a post-synaptic neuron no longer fires for any input pattern is still recoverable, whereas in SpikeProp such neuron would be degenerated with no way to modify its synaptic weights. In this frame of reference, Smoothly Spiking Networks are also less sensitive to initial parameter initialization.

References

[1] D. Popović and T. Sinkjaer, *Control of Movement for the Physically Disabled.* London, Springer 2000.

[2] W. Maass, “Paradigms for computing with spiking neurons”, *Models of Neural Networks, Vol. 4.* (L. van Hemmen ed.) Berlin, Springer 1999.

[3] W. Maass, “Noisy spiking neurons with temporal coding have more computational power than sigmoidal neurons”, *Advances in Neural Information Processing Systems, Vol. 9* (M. C. Moser, M. I. Jordan, T. Petsche eds.) The MIT Press 1997.

[4] S. Bohte, H. Poultré, and J. Kok, “Error Backpropagation in Temporally Encoded Networks of Spiking Neurons”, *Neurocomputation, Vol. 48*, 2002.

[5] J. Šíma, “Gradient Learning in Networks of Smoothly Spiking Neurons (Revised Version)” *Technical report No. 1045*, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2009.

[6] P. Baldi and W. Heiligenberg, “How sensory maps could enhance resolution through ordered

DRAFT 14.srpna 2009 - 14:52

arrangements of brodaly tuned receivers" *Biological Cybernetics, Vol. 59*, 1988.

[7] S.C. Moore, "Back-Propagation in Spiking Neural Networks" *M.Sc. thesis, University of Bath*, 2002.

[8] A. Carnell and D. Richardson, *Linear algebra for time series of spikes* www.bath.ac.uk/ masdr/inpr.ps , 2004.

[9] J.P. Sougne, "A learning algorithm for synfire chains" *Connectionist Models of Learning, Development and Evolution*, London, Springer 2001.

[10] J.B. Pfister, D. Barber, and W. Gerstner, "Optimal Hebbian Learning: A Probabilistic Point of View" *ICANN/ICONIP 2003, Vol. 2714, Lectire Notes in COmputer Science*, Berlin, Springer 2003.

[11] A. Belatreche, L.P. Maguire, M. McGinnity, and Q.X. Wu, "A Method for Supervised Training of Spiking Neural Networks" *Pros. IEEE Conf. Cybernetics Intelligence - Challenges and Advances*, Reading, UK, 2003.

DRAFT
14. srpna 2009 - 14:32

Syntactic Approach to Fuzzy Modal Logics in MTL

Post-Graduate Student:
MGR. KAREL CHVALOVSKÝ
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ

Department of Logic, Faculty of Arts
Charles University in Prague
Celetná 20

116 42 Prague 1, CZ

chvalovsky@cs.cas.cz

Supervisor:
MGR. MARTA BÍLKOVÁ, PH.D.

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2

182 07 Prague 8, CZ

bilkova@cs.cas.cz

Field of Study:
Logic

This work was supported by GA ČR EUROCORES project ICC/08/E018, GA ČR project 401/09/H007 and GA UK project 73109/2009.

Abstract

We study provability in Hilbert-style calculi obtained by adding standard modal logic axioms to Monoidal T-norm based Logic (MTL) by automated theorem proving methods. The aim of this paper is to present some basic properties of systems K, D, T, S4 and S5 over MTL. These system are defined in the same way as are in classical propositional logic. It is shown that many classically valid formulae become unprovable.

1. Introduction

In logic is quite common to enrich the expressive power of given system by new logical connectives or operators. The most prominent such systems over classical propositional calculus (CPC) are modal logics, which introduce new operators formalising a necessity and a possibility. The practical importance of these logics constantly grows and are studied not only over classical logic but also over non-classical logics. Interesting candidates for such generalisations are mathematical fuzzy logics.

The basic generally studied modal logic is the minimal normal modal logic K. A similar role in mathematical fuzzy logic has, from some point of view, Esteva and Godo's Monoidal T-norm based Logic (MTL) [5], which is the logic of left-continuous t-norms and their residua.

Fuzzy (or more precisely many-valued) modal logics have already been studied in the literature, e.g., [9, 8, 6]. However, in most cases only very strong modal logics like S4 and S5 have been considered. The systematic study of modal logics starting from the minimal normal modal logic K is relatively recent, see, e.g., [4].

In [4] is used a semantic approach to build a minimal normal modal logic over finite residuated lattices. The syntactic problems which this brings are discussed in [2]. Our starting point is completely different, we are interested solely in these syntactic notions. We enrich MTL Hilbert-style calculi by standard modal axioms and by the methods of automated theorem proving we study provability and unprovability in obtained systems.

Similar problems were quite extensively studied in intuitionistic modal logics, for some discussions see, e.g., [11]. In [12] were used automated theorem proving methods, which are very similar to ours, to study dependencies in modal logics over CPC.

We emphasise that in this paper we only touch some basic properties. However, all given material can be proved by automated or semi-automated theorem proving methods. The work on this approach is currently in progress and a much more comprehensive paper is being planned. From these reasons and to make the paper shorter some proofs are omitted.

The paper is organised as follows. In Section 2 we set up terminology and in Section 3 we discuss provability and unprovability of some formulae in K, D, T, S4 and S5 over MTL. The choice of studied systems and formulae is mainly influenced by [10].

DRAFT
Kotrna 2009 - MAS.

2. Preliminaries

2.1. Monoidal T-norm based Logic MTL

We define standard Hilbert style calculus for the Monoidal T-norm based Logic (MTL), which consists of axioms and modus ponens as the only deduction rule. The language of MTL consists of implication ($\rightarrow$), multiplicative ($\&$) and additive ($\wedge$) conjunctions and a constant for falsity ($\overline{0}$).

Definition 2.1 *We define the monoidal t-norm based logic MTL as a Hilbert style calculus with following formulae as axioms*

(A1) $(\varphi \rightarrow \psi) \rightarrow ((\psi \rightarrow \chi) \rightarrow (\varphi \rightarrow \chi))$,

(A2) $(\varphi \,\&\, \psi) \rightarrow \varphi$,

(A3) $(\varphi \,\&\, \psi) \rightarrow (\psi \,\&\, \varphi)$,

(A4a) $(\varphi \,\&\, (\varphi \rightarrow \psi)) \rightarrow (\varphi \wedge \psi)$,

(A4b) $(\varphi \wedge \psi) \rightarrow \varphi$,

(A4c) $(\varphi \wedge \psi) \rightarrow (\psi \wedge \varphi)$.

(A5a) $(\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\varphi \,\&\, \psi) \rightarrow \chi)$,

(A5b) $((\varphi \,\&\, \psi) \rightarrow \chi) \rightarrow (\varphi \rightarrow (\psi \rightarrow \chi))$,

(A6) $((\varphi \rightarrow \psi) \rightarrow \chi) \rightarrow (((\psi \rightarrow \varphi) \rightarrow \chi) \rightarrow \chi)$,

(A7) $\overline{0} \rightarrow \varphi$.

The only deduction rule of MTL is modus ponens

(MP) If φ is derivable and $\varphi \rightarrow \psi$ is derivable then ψ is derivable.

Let us note properties stated by each axiom, following [8, 5]. Axiom (A1) is transitivity of implication. Axiom (A2) states that multiplicative conjunction implies its first member. Axiom (A3) is commutativity of multiplicative conjunction. Axioms (A4c), (A4b) and (A4a) state that additive conjunction is commutative, implies its first member and one implication of the divisibility property. Axioms (A5a) and (A5b) represent residuation. Axiom (A6) is a variant of proof by cases, and states that if both $\varphi \rightarrow \psi$ and $\psi \rightarrow \varphi$ implies χ, then χ. Axiom (A7) states that false implies everything.

Further logical connectives—pseudo-complement negation ($\neg$), disjunction ($\vee$) and equivalence ($\equiv$)—are definable in MTL. Therefore, we read them as following abbreviations

$$\neg\varphi =_{df} \varphi \rightarrow \overline{0},$$
$$\varphi \vee \psi =_{df} ((\varphi \rightarrow \psi) \rightarrow \psi) \wedge ((\psi \rightarrow \varphi) \rightarrow \varphi),$$
$$\varphi \equiv \psi =_{df} (\varphi \rightarrow \psi) \,\&\, (\psi \rightarrow \varphi).$$

For some purposes can be suitable to have an involutive negation which we obtain by adding axiom $\neg\neg\varphi \rightarrow \varphi$ to MTL. The system so obtained is called Involutive Monoidal T-norm based Logic (IMTL). If we add the contraction axiom $\varphi \rightarrow \varphi \,\&\, \varphi$ to MTL we obtain Gödel logic (G). The last two axiomatic extensions of MTL mentioned in the paper are Hájek's Basic Logic (BL) and Łukasiewicz logic (Ł). These logics are obtained by adding the divisibility axiom $\varphi \wedge \psi \rightarrow \varphi \,\&\, (\varphi \rightarrow \psi)$ to MTL and IMTL, respectively.

The following theorems of MTL are very useful for our purposes. An interested reader can find proofs in [8].

Lemma 2.2 *The following formulae are provable in MTL:*

(F1) $(\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow (\psi \rightarrow (\varphi \rightarrow \chi))$,

(F2) $\varphi \rightarrow \varphi$,

(F3) $(\varphi \,\&\, (\varphi \rightarrow \psi)) \rightarrow \psi$,

(F4) $(\varphi \rightarrow (\psi \rightarrow (\varphi \,\&\, \psi))$,

(F5) $(\varphi \wedge \psi) \rightarrow \varphi, (\varphi \wedge \psi) \rightarrow \psi, (\varphi \,\&\, \psi) \rightarrow (\varphi \wedge \psi)$,

(F6) $((\varphi \rightarrow \psi) \wedge (\varphi \rightarrow \chi)) \rightarrow (\varphi \rightarrow (\psi \wedge \psi))$,

(F7) $\varphi \rightarrow (\varphi \vee \psi), \psi \rightarrow (\varphi \vee \psi)$,

(F8) $((\varphi \rightarrow \chi) \wedge (\psi \rightarrow \chi)) \rightarrow ((\varphi \vee \psi) \rightarrow \chi)$,

(F9) $\varphi \rightarrow \neg\neg\varphi$,

(F10) $(\varphi \rightarrow \psi) \rightarrow (\neg\psi \rightarrow \neg\varphi)$,

(F11) $(\varphi \equiv \psi) \rightarrow ((\varphi \rightarrow \chi) \equiv (\psi \rightarrow \chi))$,

(F12) $(\varphi \equiv \psi) \rightarrow ((\chi \rightarrow \varphi) \equiv (\chi \rightarrow \psi))$,

(F13) $(\neg\varphi \vee \neg\psi) \equiv \neg(\varphi \wedge \psi)$.

It is worth pointing out that we will restrict our attention to the syntactic aspects of MTL. To emphasise this approach we completely ignore the semantics of MTL. An interested reader can consult [5].

DRAFT 14. prince 2009 - 14:32

2.2. Modal logics

For our purposes only a very limited introduction to modal logics is needed, for a detail treatment we refer the reader to, e.g., [10, 3, 1]. We obtain modal logics by adding a unary modal necessity operator box ($\Box$) to our language. Another standard modal operator is a possibility operator diamond ($\Diamond$) which is usually defined as an abbreviation for $\neg\Box\neg$. Although in logics with a non-involutive negation ($\neg\neg\varphi\rightarrow\varphi$ is not true) this definition evidently leads to some problems, we use this approach for simplicity and to stress these problems.

$$\Diamond\varphi =_{df} \neg\Box\neg\varphi,$$

The properties of a modal operator box depends on chosen axioms. Some of the most widely studied are these:

(K) $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$,

(4) $\Box\varphi \rightarrow \Box\Box\varphi$,

(T) $\Box\varphi \rightarrow \varphi$,

(D) $\Box\varphi \rightarrow \Diamond\varphi$,

(B) $\varphi \rightarrow \Box\Diamond\varphi$,

(E) $\Diamond\varphi \rightarrow \Box\Diamond\varphi$.

We also need some derivational rules dealing with modalities. The most common is the necessitation rule

(Nec) $\varphi/\Box\varphi$.

In Table 1 are presented some of the most prominent modal logics over CPC. All of them are so called normal modal logics, which means that contain the minimal normal modal logic K.

Logic	Additional axioms and rules
K	(K) and (Nec)
D	(K), (Nec) and (D)
T	(K), (Nec) and (T)
S4	(K), (Nec), (T) and (4)
S5	(K), (Nec), (T) and (E)

Table 1: Modal logics in CPC.

[1] http://www.eprover.org/
[2] http://www.cs.chalmers.se/ koen/folkung/

We construct our systems of modal logics over MTL in the very same way as are in CPC. It, not so surprisingly, turns out that this leeds to some problems.

Let us remark that when proving that some formulae are equivalent in some modal logics over CPC, we can use the interdefinability of logical connectives, which is mostly impossible in MTL.

2.3. Automated theorem proving methods

All given results can be obtained automatically or semiautomatically by automated theorem proving. There is a well known technique for encoding propositional Hilbert-style calculus into classical first-order logic through terms. The key idea is that formula variables are encoded as first-order variables and propositional connectives as first-order function symbols. For details, see, e.g., [13].

We used freely available software—E prover version 1.0-004 Temi[1] and finite-domain model finder Paradox 3.0[2]. No special prover setting is needed for our purposes, but can lead to great speed improvements. However, these aspects are too complex to be discussed here. Moreover, all presented proofs are easy to find for anyone familiar with Hilbert-style calculus for MTL and counterexamples can be find completely automatically even with default setting. More complex problems are not included in this paper.

2.4. Models

The standard way to prove that some formula φ is not provable from the given set of formulae Γ is to present a model M in which all formulae from Γ are true, but formula φ is false. In our case we will present tables with finitely many elements which are labelled by integers starting from 0. We always interpret $\overline{0}$ as 0 in a model and truth in a given model M is the maximal value in this model M, e.g., in a four element model true formulae are these with value 3. A function from atoms or formula variables to elements of model is called a valuation. The definition of a valuation can be easily extended to all formulae in a standard way. To show that Γ is true in M we must show that all formulae from Γ are true in M under all valuations. To show that φ is not true in M it is enough to find a valuation for which φ is not true in M.

We present tables for every connective separately and for better readability even for some defined connectives, but never for negation which corresponds to the first column of implication. Let us note that some formulae

DRAFT 14. srpna 2009 - 14:32

have smaller counterexamples than these presented, but we tried to make the paper more compact.

3. Modal logics in MTL

3.1. K_{MTL}

The basic generally studied modal logic is K. This system is obtained by adding axiom which distribute box over implication and the necessitation rule. In the same way we define K_{MTL} over MTL.

Definition 3.1 *Logic K_{MTL} is obtained by adding axiom* (K) *and the derivational rule* (Nec) *to MTL.*

By an easy application of (Nec) and (K) we immediately obtain that the derivation rule

(DR1) $\varphi \to \psi / \Box\varphi \to \Box\psi$,

is valid in K_{MTL}.

The fundamental property of the classical modal logic K is the distributivity of box over conjunction. However, in K_{MTL} this is not true. Moreover, we cannot interchange box with diamond, because we don't have an involutive negation. From this follows that the same problem is with the distribution of diamond over disjunction, which is a part of popular diamond based definition of K over CPC. Nevertheless, at least some implications can still be proved.

Lemma 3.2 *The following formulae are provable in K_{MTL}:*

(a) $(\Box\varphi \,\&\, \Box\psi) \to \Box(\varphi \,\&\, \psi)$,

(b) $\Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)$,

(c) $(\Diamond\varphi \vee \Diamond\psi) \to \Diamond(\varphi \vee \psi)$,

(d) $\Box\varphi \to \neg\Diamond\neg\varphi$.

Proof:

(a)

1:	$\Box\varphi \to \Box(\psi \to (\varphi \,\&\, \psi))$	(F4), (DR1)
2:	$\Box(\psi \to (\varphi \,\&\, \psi)) \to (\Box\psi \to \Box(\varphi \,\&\, \psi))$	(K)
3:	$(\Box\varphi \,\&\, \Box\psi) \to \Box(\varphi \,\&\, \psi)$	1, 2, (A1), (A5a)

(b)

4:	$\Box(\varphi \wedge \psi) \to \Box\varphi$	(F5), (DR1)
5:	$\Box(\varphi \wedge \psi) \to \Box\psi$	(F5), (DR1)
6:	$\Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)$	(A1), (F4), (F5), (F6)

(c) Let us remark that $\Diamond\varphi$ is an abbreviation for $\neg\Box\neg\varphi$.

7:	$\Box(\neg(\varphi \vee \psi)) \to \Box\neg\varphi$	(F7), (F10), (DR1)
8:	$\Box(\neg(\varphi \vee \psi)) \to \Box\neg\psi$	(F7), (F10), (DR1)
9:	$\Box(\neg(\varphi \vee \psi)) \to (\Box\neg\varphi \wedge \Box\neg\psi)$	(A1), (F4), (F5), (F6)
10:	$(\Diamond\varphi \vee \Diamond\psi) \to \Diamond(\varphi \vee \psi)$	(F10), (F13), (A1)

An alternative slightly shorter proof uses the derivational rule (DR2), which we show later on.

(d)

11:	$\Box\varphi \to \Box\neg\neg\varphi$	(F9), (DR1)
12:	$\Box\neg\neg\varphi \to \neg\neg\Box\neg\neg\varphi$	(F9)
13:	$\Box\varphi \to \neg\Diamond\neg\varphi$	(A1)

■

Lemma 3.3 *The following formulae are not provable in K_{MTL}:*

(a) $\Box(\varphi \,\&\, \psi) \to (\Box\varphi \,\&\, \Box\psi)$,

(b) $(\Box\varphi \wedge \Box\psi) \to \Box(\varphi \wedge \psi)$,

(c) $\Diamond(\varphi \vee \psi) \to (\Diamond\varphi \vee \Diamond\psi)$,

(d) $\neg\Diamond\neg\varphi \to \Box\varphi$.

Proof: For (a) use Table 2 and $\varphi = 0$ and $\psi = 0$.

&	0	1	2
0	0	0	0
1	0	0	1
2	0	1	2

→	0	1	2
0	2	2	2
1	1	2	2
2	0	1	2

□	
0	1
1	1
2	2

Table 2: Truth tables over K_{MTL}.

For (b) and (c) use Table 3 and $\varphi = 1, \psi = 2$ and $\varphi = 2, \psi = 3$, respectively.

DRAFT 14. srpna 2009 - 14:32

$\wedge$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	0	1	1	1
2	0	0	2	0	2	2
3	0	1	0	3	1	3
4	0	1	2	1	4	4
5	0	1	2	3	4	5

&	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	1	0	1
2	0	0	2	0	2	2
3	0	1	0	3	1	3
4	0	0	2	1	2	4
5	0	1	2	3	4	5

$\rightarrow$	0	1	2	3	4	5
0	5	5	5	5	5	5
1	4	5	4	5	5	5
2	3	3	5	3	5	5
3	2	4	2	5	4	5
4	1	3	4	3	5	5
5	0	1	2	3	4	5

$\vee$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	1	4	3	4	5
2	2	4	2	5	4	5
3	3	3	5	3	5	5
4	4	4	4	5	4	5
5	5	5	5	5	5	5

$\Box$	
0	2
1	4
2	4
3	4
4	4
5	5

$\Diamond$	
0	0
1	1
2	1
3	1
4	1
5	3

Table 3: Truth tables over K_{MTL}.

Table 4 and $\varphi = 1$ is a counterexample for (d).

$\rightarrow$	0	1	2
0	2	2	2
1	0	2	2
2	0	1	2

$\Box$	
0	0
1	0
2	2

$\Diamond$	
0	0
1	2
2	2

Table 4: Truth tables over K_{MTL}.

■

It is evident that the models in Table 2 and 3 have an involutive negation and satisfy the divisibility axiom and so are counterexamples to (a), (b) and (c) also in K_{IMTL}, K_{BL} and even $K_{Ł}$. All these systems are obtained in the very same way as K_{MTL} from MTL. The completely different situation is in K_G where $\varphi \,\&\, \psi \equiv \varphi \wedge \psi$ is true and we can prove formulae (a) and (b) similarly to Lemma 3.2. Formula (c) then easily follows from (b).

A different situation is with (d) which is easily provable if we have an involutive negation, but is false in K_G as follows from Table 4.

If we take into account (F10) we can prove similarly to (DR1) that the derivational rule

(DR2) $\varphi \rightarrow \psi / \Diamond\varphi \rightarrow \Diamond\psi$,

is valid in K_{MTL}.

In K we can also prove the partial distribution of box over disjunction and diamond over conjunction which holds even in K_{MTL}.

Lemma 3.4 *The following formulae are provable in K_{MTL}*

(a) $\Diamond(\varphi \wedge \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$,

(b) $(\Box\varphi \vee \Box\psi) \rightarrow \Box(\varphi \vee \psi)$.

Proof: Both proofs are very similar to Lemma 3.2b. In (a) we only use (DR2) instead of (DR1) and for (b) the proof reads as follows:

14:	$\Box\varphi \rightarrow \Box(\varphi \vee \psi)$	(F7), (DR1)
15:	$\Box\psi \rightarrow \Box(\varphi \vee \psi)$	(F7), (DR1)
16:	$(\Box\varphi \vee \Box\psi) \rightarrow \Box(\varphi \vee \psi)$	(F4), (F5), (F8)

■

The following distributivity of diamond over implication remains true in K_{MTL} only partially.

Lemma 3.5 *The following formula is provable in K_{MTL}*

$$\Diamond(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Diamond\psi).$$

Proof:

17:	$\varphi \rightarrow ((\varphi \rightarrow \psi) \rightarrow \psi)$	(F3), (A5b)
18:	$\Box(\varphi \rightarrow (\neg\psi \rightarrow \neg(\varphi \rightarrow \psi)))$	(F10), (Nec)
19:	$\Box\varphi \rightarrow (\Box\neg\psi \rightarrow \Box\neg(\varphi \rightarrow \psi))$	(K), (K)
20:	$\neg\Box\neg(\varphi\rightarrow\psi)\rightarrow(\Box\varphi\rightarrow\neg\Box\neg\psi)$	(F10), (F1)

■

The opposite implication in the previous lemma, which is true in classical logic, is not true in K_{MTL} and has a three element counterexample.

We have shown that some important modal formulae of K are not provable in K_{MTL}. A stronger system can be thus easily obtained by adding these formulae to K_{MTL}. On the other hand, some axiomatics of K are same even over MTL. For example, if we take (DR1) and $\Box(\varphi\rightarrow\varphi)$ instead of the necessitation rule (Nec) we obtain again K_{MTL}. Moreover, we obtain K_{MTL} even if we replace in this system axiom (K) with $(\Box\varphi \,\&\, \Box\psi) \rightarrow \Box(\varphi \,\&\, \psi)$.

DRAFT 14. září 2009 - 1432

3.2. $\mathbf{D_{MTL}}$

Logic D, which has deontic interpretations, is the least standard system we are going to study, but both its standard axiomatics remain equivalent.

Definition 3.6 *Logic D_{MTL} is an axiomatic extension of K_{MTL} by axiom* (D).

Lemma 3.7 *The following formula is provable in D_{MTL}*

$$\Diamond(\varphi \to \varphi).$$

Proof: We obtain $\Diamond(\varphi \to \varphi)$ immediately from (F2) by the necessitation and (D). ■

The previous formula form an alternative axiomatic system of D_{MTL} as we have already noted. If we add $\Diamond(\varphi \to \varphi)$ to K_{MTL} then (D) is provable by Lemma 3.5.

3.3. $\mathbf{T_{MTL}}$

The rest of the paper deals with logics containing axiom (T). This axiom is sometimes called the axiom of necessity.

Definition 3.8 *Logic T_{MTL} is an axiomatic extension of K_{MTL} by axiom* (T).

The following formula well illustrates problems we are facing with our diamond definition over MTL.

Lemma 3.9 *The following formula is not provable in T_{MTL}*

$$\Diamond(\varphi \to \Box\varphi).$$

Proof: Use Table 5 and $\varphi = 1$.

$\&, \wedge$	0	1	2	3
0	0	0	0	0
1	0	1	1	1
2	0	1	2	2
3	0	1	2	3

$\Box$	
0	0
1	0
2	1
3	3

$\to$	0	1	2	3
0	3	3	3	3
1	0	3	3	3
2	0	1	3	3
3	0	1	2	3

$\Diamond$	
0	0
1	3
2	3
3	3

Table 5: Truth tables over T_{MTL}.

■

However, some diamond based formulae are still provable.

Lemma 3.10 *The following formula is provable in T_{MTL}*

$$\varphi \to \Diamond\varphi.$$

Proof: Follows immediately from $\Box\neg\varphi \to \neg\varphi$ by (F10) and (F9). ■

The previous lemma with the transitivity of implication gives that axiom (D) is provable in T_{MTL} and thus T_{MTL} is an extension of D_{MTL}.

In CPC the axiomatic extension of K by the previous formula proves axiom (T), but in K_{MTL} it is not the case. There is a three element counterexample. It is also well known that if we take rule (DR1), axiom (T) and formula $\Box(\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi))$ in CPC we obtain T. It turns out that over MTL we obtain exactly T_{MTL}. On the other hand, if we take another classicaly equivalent axiomatics which has $\varphi \to \Diamond\varphi$ instead of (T) we obtain a weaker system.

Corollary 3.11 *The following formulae are provable in T_{MTL}:*

(a) $\Box\Diamond\varphi \to \Diamond\varphi$,

(b) $\Box\varphi \to \Diamond\Box\varphi$,

(c) $\Diamond\varphi \to \Diamond\Diamond\varphi$,

(d) $\Box\Box\varphi \to \Box\varphi$.

Together with the opposite implications these formulae form so called reduction laws. These opposite implications

(R1) $\Diamond\varphi \to \Box\Diamond\varphi$

(R2) $\Diamond\Box\varphi \to \Box\varphi$

(R3) $\Diamond\Diamond\varphi \to \Diamond\varphi$

(R4) $\Box\varphi \to \Box\Box\varphi$

lead in classical logic to the well known axiomatic extensions of T. If we add (R3) or (R4) to T we obtain S4 and if we add (R1) or (R2) to T we obtain S5 which is a proper extension of S4.

It turns out that over T_{MTL} the situation slightly changes.

DRAFT 14. srpna 2009 - 14:32

Lemma 3.12 *The following provability conditions hold*

(a) $T_{MTL}, R2 \vdash R1$,

(b) $T_{MTL}, R2 \vdash R4$,

(c) $T_{MTL}, R1 \vdash R3$,

(d) $T_{MTL}, R4 \vdash R3$,

(e) $T_{MTL}, R1 \nvdash R2$,

(f) $T_{MTL}, R3 \nvdash R4$,

Proof: For (e) and (f) use Table 5 and $\varphi = 2$. ■

Thus we have two non-equivalent axiomatics of S4 and two non-equivalent axiomatics of S5 over MTL. We will briefly study three of them.

3.4. $S4_{MTL}$

The first system is obtained by adding (R4), called axiom (4), to T_{MTL}. This is the most common definition of axiomatics for S4.

Definition 3.13 *Logic $S4_{MTL}$ is an axiomatic extension of T_{MTL} by axiom* (4).

The following formulae are the direct consequences of (R4) and thus also (R3) over T_{MTL}.

Lemma 3.14 *The following formulae are provable in $S4_{MTL}$:*

(a) $\Diamond\varphi \equiv \Diamond\Diamond\varphi$,

(b) $\Box\varphi \equiv \Box\Box\varphi$,

(c) $\Diamond\Box\Diamond\varphi \to \Diamond\varphi$,

(d) $\Box\Diamond\varphi \equiv \Box\Diamond\Box\Diamond\varphi$,

(e) $\Diamond\Box\varphi \equiv \Diamond\Box\Diamond\Box\varphi$,

Proof: All proofs are the same or very similar as in classical logic. ■

3.5. $S5_{MTL}$

The standard definition of S5 uses (R1), called axiom (E), and we define this system over MTL in the same way.

Definition 3.15 *Logic $S5_{MTL}$ is an axiomatic extension of T_{MTL} by axiom* (E).

However, we already know that this definition leads to unprovability of axiom (4) in such system. Also the following formulae are not provable.

Lemma 3.16 *The following formulae are not provable in $S5_{MTL}$:*

(a) $\Diamond\Box\varphi \to \Box\varphi$,

(b) $\Box\varphi \to \Box\Box\varphi$,

(c) $\Box(\varphi \vee \Box\psi) \to (\Box\varphi \vee \Box\psi)$,

(d) $(\Box\varphi \vee \Box\psi) \to \Box(\varphi \vee \Box\psi)$,

(e) $\Diamond(\varphi \,\&\, \Box\psi) \to (\Diamond\varphi \,\&\, \Box\psi)$,

(f) $(\Diamond\varphi \,\&\, \Box\psi) \to \Diamond(\varphi \,\&\, \Box\psi)$,

(g) $\Diamond(\varphi \wedge \Box\psi) \to (\Diamond\varphi \wedge \Box\psi)$,

(h) $(\Diamond\varphi \wedge \Box\psi) \to \Diamond(\varphi \wedge \Box\psi)$.

Proof: For (a) and (b) use Table 5 and $\varphi = 2$. In all other cases use Table 6. For (c) use $\varphi = 4$ and $\psi = 3$, for (d) use $\varphi = 1$ and $\psi = 3$, for (e) and (g) use $\varphi = 3$ and $\psi = 3$, for (f) and (h) use $\varphi = 4$ and $\psi = 3$.

$\wedge, \&$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	0	1	0	1
2	0	0	2	2	2	2
3	0	1	2	3	2	3
4	0	0	2	2	4	4
5	0	1	2	3	4	5

$\to$	0	1	2	3	4	5
0	5	5	5	5	5	5
1	4	5	4	5	4	5
2	1	1	5	5	5	5
3	0	1	4	5	4	5
4	1	1	3	3	5	5
5	0	1	2	3	4	5

$\vee$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	1	3	3	5	5
2	2	3	2	3	4	5
3	3	3	3	3	5	5
4	4	5	4	5	4	5
5	5	5	5	5	5	5

$\Box$	
0	0
1	0
2	0
3	1
4	0
5	5

$\Diamond$	
0	0
1	5
2	5
3	5
4	5
5	5

Table 6: Truth tables over $S5_{MTL}$.

■

DRAFT 14. srpna 2009 14:32

Another very important equivalence is provable in S5 only partially.

Lemma 3.17 *The following formula is provable in $S5_{MTL}$*

$$\varphi \to \Box\Diamond\varphi.$$

Lemma 3.18 *The following formula is not provable in $S5_{MTL}$*

$$\Diamond\Box\varphi \to \varphi.$$

Proof: Use Table 5 and $\varphi = 2$. ■

We can also present some other alternative axiomatics of $S5_{MTL}$. One standard way is to add axiom (B) (the formula from the previous lemma) to $S4_{MTL}$. An alternative way is to add axiom (R2) to T_{MTL}. We already know that (R4) is provable in T_{MTL} with (R2). It is not difficult to show that both axiomatics lead to the same logic.

Definition 3.19 *Logic $S5{+}_{MTL}$ is an axiomatic extension of T_{MTL} by axiom* (R2).

However, many formulae are not provable even in this stronger system.

Lemma 3.20 *The following formulae are not provable in $S5{+}_{MTL}$:*

(a) $\Box(\varphi \,\&\, \psi) \to (\Box\varphi \,\&\, \Box\psi)$,

(b) $\neg\Diamond\neg\varphi \to \Box\varphi$,

(c) $(\Box\varphi \to \Diamond\psi) \to \Diamond(\varphi \to \psi)$,

(d) $\Diamond(\varphi \to \Box\varphi)$,

(e) $\Box(\varphi \vee \Box\psi) \to (\Box\varphi \vee \Box\psi)$.

Proof: Use Table (7). For (a) use $\varphi = 3$ and $\psi = 2$, for (b) use $\varphi = 3$, for (c) use $\varphi = 3$ and $\psi = 2$, for (d) use $\varphi = 2$, for (e) use $\varphi = 3$ and $\psi = 4$.

$\wedge$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	1	1	1	1
2	0	1	2	2	2	2
3	0	1	2	3	2	3
4	0	1	2	2	4	4
5	0	1	2	3	4	5

&	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	1	0	1
2	0	0	2	2	2	2
3	0	1	2	3	2	3
4	0	0	2	2	4	4
5	0	1	2	3	4	5

$\to$	0	1	2	3	4	5
0	5	5	5	5	5	5
1	4	5	5	5	5	5
2	1	1	5	5	5	5
3	0	1	4	5	4	5
4	1	1	3	3	5	5
5	0	1	2	3	4	5

$\vee$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	1	2	3	4	5
2	2	2	2	3	4	5
3	3	3	3	3	5	5
4	4	4	4	5	4	5
5	5	5	5	5	5	5

$\Box$	
0	0
1	1
2	1
3	1
4	4
5	5

$\Diamond$	
0	0
1	1
2	4
3	5
4	4
5	5

Table 7: Truth tables over $S5{+}_{MTL}$.

■

One more system which is equivalent to $S5{+}_{MTL}$ is K_{MTL} extended by axioms $\Box\Diamond\Box\varphi \to \varphi$ and $\Diamond\Box\varphi \to \Box\Diamond\Box\Box\varphi$. It is worth pointing out that none of these two formulae is provable in $S5_{MTL}$.

4. Summary and future work

Our paper presents a small introduction to the problems of modal Hilbert-style calculi in mathematical fuzzy logics. We only touch some prominent modal systems and their axiomatics.

We also only slightly touch, in case of modal logic K, problems in axiomatic extensions of MTL, where some formulae unprovable in modal logics over MTL become provable. However, it is not difficult to show that all given counterexamples satisfy the divisibility axiom and some of them even contraction. We also do not discuss the difference between additive and multiplicative conjunctions.

We have shown that some important tautologies are not provable in naively constructed modal systems over MTL. On the other hand, the fact that some formulae are not provable in modal logics over MTL can be seen as an advantage and intended property which enable us to have some formulae, which are over CPC equivalent, true and some false if needed.

DRAFT 24. srpna 2009 - 14:32

References

[1] P. Blackburn, M. de Rijke, and Y. Venema *Modal Logic*, Cambridge Tracts in Theoretical Computer Science, 2000.

[2] F. Bou, F. Esteva, and L. Godo, "Exploring a syntactic notion of modal many-valued logics," *Mathware and Soft Computing*, vol. 15, pp. 175–188, 2008.

[3] A. Chagrov and M. Zakharyaschev, *Modal Logic*, Oxford Logic Guides, vol. 35. Oxford: Oxford University Press, 1997.

[4] F. Bou, F. Esteva, L. Godo, and R.O. Rodriguez, "On the Minimum Many-Valued Modal Logic over a Finite Residuated Lattice," *Journal of Logic and Computation*, Accepted.

[5] F. Esteva and L. Godo, "Monoidal t-norm based logic: Towards a logic for left-continuous t-norms," *Fuzzy Sets and Systems*, vol. 124, no. 3, pp. 271–288, 2001.

[6] M. Fitting, "Many-valued modal logics," *Fundamenta Informaticae*, vol. 15, pp. 235–254, 1992.

[7] M. Fitting, "Many-valued modal logics, II.," *Fundamenta Informaticae*, vol. 17, pp. 55–73, 1992.

[8] P. Hájek, *Metamathematics of Fuzzy Logic*, vol. 4 of *Trends in Logic*. Dordercht: Kluwer, 1998.

[9] P. Hájek and D. Harmancová, "A many-valued modal logic," in *Proceedings IPMU'96. Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 1021–1024, Granada, 1996. Universidad de Granada.

[10] G.E. Hughes and M.J. Cresswell, *A New Introduction to Modal Logic*, Routledge, London, 1996.

[11] A.K. Simpson, *The Proof Theory and Semantics of Intuicionistic Modal Logic*, Ph.D.-dissertation, Edinburgh, 1993.

[12] F. Rabe, P. Pudlák, G. Sutcliffe, and W. Shen, "Solving the $100 modal logic challenge," *Journal of Applied Logic*, vol. 7,nno. 1, pp. 113–130, 2009.

[13] L. Wos and G.W. Pieper *The Collected Works of Larry Wos*, In 2 vols. Singapore: World Scientific, 2000.

DRAFT 14. srpna 2009 -- 14:32

Signature Provenance obtained from the Ontology Provenance

Post-Graduate Student:
ING. FRANTIŠEK JAHODA
Department of Mathematics
Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Trojanova 13
120 00 Prague 2, CZ
jahoda@cs.cas.cz

Supervisor:
ING. JÚLIUS ŠTULLER, CSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
stuller@cs.cas.cz

Field of Study:
Ontology Modularisation in the Semantic Web Context

Abstract

The data provenance technology can be modified to describe the provenance of an ontology. The ontology provenance is of the same importance as the provenance of the data described by this ontology. The paper deals with recording the ontology provenance up to the ontology axioms and with deriving the provenance of a signature (set of concepts, relations, and individuals) from the stored ontology provenance. Despite the fact the exact solution is unfortunately undecidable even for simple ontologies, it is possible to give its upper estimate.

1. Introduction

Semantic web applications usually process data from many sources, which can be completely uncontrollable and sometimes even with questionable reliability. Consequently, it is very important to record and process the data provenance. Although the data provenance may be controlled, this does not need to be enough. In the semantic web paradigm [3], data are usually integrated through related ontologies [4] describing them. These ontologies are designed and maintained by specialists, thus the application designer does not have an immediate control of the application ontology, notwithstanding that his application may be influenced by changes in the ontology. Therefore, recording the ontology provenance is equally important as recording the data provenance.

A proper documentation of requirements and design history of a computer program enables to extend the program along the design requirements, to ensure that the design requirements are met and that the future changes will not disrupt implemented properties, and to trace the project progress to fulfill the project plan in time. Similarly, an ontology alone without any relationship to the outside world is not complete. Recording the ontology design process details (ontology annotations) can improve the design process, locating imperfections in the ontology and in the design process, and checking design requirements (e.g. each later change in the ontology should be motivated by a corresponding change in these requirements).

The approach described in [5] proposes to relate the ontology provenance to the ontology axioms. In this approach, user can ask by which events an axiom was influenced. Nevertheless, user can be interested in more complicated questions, e.g. how to transfer the provenance to a new axiom if we replace some axiom(s) in the ontology by (an) axiom(s) derived from this ontology [2]. How to connect the provenance of axioms to an ontology concept or relation meaning is yet another question. The paper presents a partial solution to the latter question.

1.1. Ontology provenance

The word provenance comes from the French *provenir*, which means *to come from*. This word denotes the origin or the source of an object. Sometimes, it has even wider meaning and denotes the whole history of the object and all influences on the object, thus all events related to the object. The provenance has a wide use in the law theory, archives, arts, science, etc. It is also used in computer science, especially in connection with data (details). The data can come from different sources usually with some applied transformations and algorithms. These kind of metadata are called data provenance.

The data provenance is commonly used and it is very important, especially in scenario with many data sources

DRAFT 14. Strnad 2009 - 14:54

(such as semantic web applications). It enables to state the source of information, the applied algorithms, and to transfer the thrust in the data sources and in the algorithms to the input data and the application results.

In the semantic web paradigm, ontologies are used to describe data and the logical relationships between data, and also to derive new properties. An ontology is a shared logical model of some domain, thus it should not be subject to frequent changes. However, the understanding of the domain can change as the world itself is changing. Therefore, it is necessary to project these modifications into the ontology. The ontology modifications may change also the meaning of relationships derived from the data, thus recording the ontology provenance is equally important as recording the data provenance.

To record the ontology provenance, it is important to know the possible changes in an ontology and to which objects of the ontology they are related to. Majority of ontology changes consist in an addition of an axiom to the ontology, in a removal of an ontology axiom, and in a rewrite of an ontology axiom, thus the changes can be bind to ontology axioms.

It also seems useful to describe the ontology provenance by its own stand-alone ontology. Such approach enables to reason about the ontology metadata with the help of an ontology logical model (e.g. it is possible to write out axioms related to some design change). Exact properties of the provenance ontology will be strongly dependent on intended applications, therefore they will not be discussed in this paper. An example of a more elaborate provenance ontology can be found in [6].

1.2. OWL annotations

Any change in an ontology consists of few fundamental types of changes: the addition, the removal, and the rewrite of an axiom. According to the OWL 2.0 draft it will be possible to tag an axiom with its own URI, thus rewriting an axiom without change of its URI will be indistinguishable from the meta-ontology perspective. Therefore, it is appropriate to tag each axiom version with its own URI. This approach leads to the reduction of the fundamental types of changes to an axiom addition and an axiom removal. The axiom rewrite has to be represented as a substitution of the old axiom by the new one with different URI. Of course, it is possible to note an axiom is a logical successor of another one.

The older approach consists in reifying (a transformation of an axiom to a set of new axioms expressing the syntactic structure of the original one) axioms from the original ontology and then referencing the reified axioms. A reified axiom is represented as an individual in a new ontology, consequently the provenance properties can be bind to this individual. A reified ontology is usually few times larger than the original ontology, therefore reasoning on the new ontology is not as effective as in the first approach.

2. Annotations for a Signature

If the ontology provenance is properly recorded, it will be possible to ask for the provenance related to a specific axiom. It will be also interesting to ask for the provenance related to a concept definition. However, a concept need not to be defined by only one axiom, it can be influenced by other axioms such as definition of sub-concepts used in the definition or axioms expressing relationships of the defined concept to other concepts and relations.

To determine provenance atoms related to a concept, an individual, or a relation, it is necessary to connect these symbols with axioms which influence them.

The following definition enables to select such subset of an ontology that defines the same meaning of the symbols from a certain set $\mathbf{S}$ of symbols (named a signature) as the whole ontology. More precisely:

Definition 1 (Model Conservative Extension) *Let O and $O_1 \subseteq O$ be two $\mathcal{L}$-ontologies and $\mathbf{S}$ a signature over $\mathcal{L}$.*
We say that O is a model $\mathbf{S}$-conservative extension of O_1, if for every model $\mathcal{I}$ of O_1, there exists a model $\mathcal{J}$ of O such that $\mathcal{I}|_{\mathbf{S}} = \mathcal{J}|_{\mathbf{S}}$.

Unfortunately, checking this property is highly undecidable (non recursively enumerable) even for $\mathcal{ALC}$. Therefore, we will use the following preposition from [7] and well-known locality property [1] , which has NEXPTIME-complete complexity even for OWL DL.

Preposition 1 *Let O_1, O_2 be two ontologies and $\mathbf{S}$ a signature such that O_2 is local w.r.t. $\mathbf{S} \cup Sig(O_1)$.*
Then $O_1 \cup O_2$ is an $\mathbf{S}$-model conservative extension of O_1.

Thus, it is possible to upper estimate minimal $O_1 \subseteq O$ such that O is a model $\mathbf{S}$-conservative extension of O_1. We compute minimal subset of O filling the locality condition w.r.t $\mathbf{S}$. Let any axiom which is present in such

a subset be called *computed axiom* and the signature provenance for **S** be the union of provenance atoms related to computed axioms.

Such approach has the deficiency that the provenance gained by this approach is based on actual axioms presented in the ontology only.

This obstacle can be overcome by computing the uni on of *essential* axioms through the whole ontology life-cycle. As was noted above, an ontology history can be represented as the addition and the removal of axioms, thus the following approach for obtaining the provenance was drafted.

Firstly, the history for an ontology has to be defined.

Definition 2 (Version History for the Ontology) *Let $O_1, O_2 \ldots O_N$ be a sequence of all states of an ontology O sorted by ascending date of change, with O_1 being the first version of ontology and O_N being the last version (equal to O) and with no intermediate version between O_i and O_{i+1} for $i \in \{1, 2, \ldots, N-1\}$.*

We say that $O_1, O_2 \ldots O_N$ is a version history for ontology O.

It is necessary to consider the whole version history for an ontology, because a removed axiom from the ontology can be a former essential axiom and an added axiom can be an incoming essential axiom. If a removed or an added axiom was ignored, part of the provenance could be lost. If some intermediate version of an ontology was omitted, the resulting provenance would lack provenance atoms related to changes in this version.

Algorithm 1 *Let S be a signature over $\mathcal{L}$, $(O_i)_{i \in \{1 \ldots N\}}$ version history for ontology O over $\mathcal{L}$, and $Prov(axiom)$ a mapping from axioms of O to the set of provenance atoms for such axiom.*

To search provenance atoms for a signature S, compute a computed axioms set E_i for each ontology O_i. Let the

$$E = \bigcup_{i \in \{1, \ldots, N\}} E_i$$

be the union of such sets.

Finally, the provenance atoms for a signature $Prov(S)$ is the union of all provenance atoms for axioms in E.

$$Prov(S) = \bigcup_{\alpha \in E} Prov(\alpha)$$

3. Conclusions

The paper presents aproach, which connects the ontology provenance to symbols used in an ontology. This approach is usefull, when the user wants to know by which changes was a symbol (concept, relation) influenced during the ontology life-cycle. The obtained provenance will not be fully correct, because the model conservative extension is estimated by the locality; however, it can provide an acceptable approximation.

An unanswered question deserving future attention is which optimalizations are possible for computing locality-based modules on two similar version of an ontology.

References

[1] B.C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, "School of Computer Science: Modular Reuse of Ontologies: Theory and Practice", Journal of Artifical Inteligence Research, 2008.

[2] M. Vacura and V. Svatek, In Proc. "Pattern-Based Representation and Propagation of Provenance Metadata in Ontologies", EKAW 2008 Poster and Demo Proceedings, pp.66-68, 2008.

[3] M. Vacura, V. Svatek, G. Antoniou and F. van Hamerlen, "A Semantic Web Primer", The MIT Press, ISBN 0-262-01210-3, 2004.

[4] F. Baader, D. Calvanese, D.L. McGuiness, D. Nardi, and P.F. Patel-Schneider, "The Description Logic Handbook", Cambridge, ISBN 978-0-521-87625-4, 2007.

[5] D. Vrandecic, J. Volker, P. Haase, T. Tran Duc, and P. Cimiano, In Proc. "Metamodel for Annotations of Ontology Elements in OWL DL", Proccedings of the 2nd Workshop on Ontologies and Meta-Modeling. GI Gesellschaft fur Informatik, Karlsruhe, Germany, 2006.

[6] S. Ram, In Proc. "The Active Conceptual Modelling of Learning Workshop", Space and Naval Warfare Systems Center, San Diego, May 9-12, 2006.

[7] B.C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, In Proc. "Just the Right Amount: Extracting Modules from Ontologies", Sixteenth International World Wide Web Conference (WWW2007), 2007.

DRAFT 14. srpna 2009 – 14:32

Cost Functions for Graph Repartitionings Motivated by Factorization

Post-Graduate Student:
MGR. KATEŘINA JURKOVÁ

Faculty of Mechatronics, Informatics and Interdisciplinary Studies
Technical University of Liberec,
Studentská 2

461 17 Liberec 1, CZ

katerina.jurkova@tul.cz

Supervisor:
PROF. ING. MIROSLAV TŮMA, CSC.

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2

182 07 Prague 8, CZ

tuma@cs.cas.cz

Field of Study:
Scientific Computing

This work has been partially supported by the internal grant FM-IG/2009/NTI-02 Faculty of Mechatronics, Informatics and Interdisciplinary Studies, TUL.

Abstract

The paper deals with the parallel computation of matrix factorization using graph partitioning-based domain decomposition. It is well-known that the partitioned graph may have both a small separator and well-balanced domains but sparse matrix decompositions on domains can be completely unbalanced.

In this paper we propose to enhance the iterative strategy for balancing the decompositions from [13] by graph-theoretical tools. We propose the whole framework for the graph repartitioning. In particular, new global and local reordering strategies for domains are discussed more in detail. We present both theoretical results for structured grids and experimental results for unstructured large-scale problems.

1. Introduction

The problem of proper graph partitioning is one of the classical problems of the parallel computing. The actual process of obtaining high-quality partitionings of undirected graphs which arises in many practical situations is reasonably well understood. In addition, the resulting algorithms are sophisticated enough [5], [1]. Such situations are faced, e.g., if standard criteria for partitionings expressed by *balancing sizes of domains* and *minimizing separator sizes* are considered. However, the situation may be different if one needs to balance the time to perform some specific operations. An example can be the time to compute sparse matrix decompositions, their incomplete counterparts, or the time for some auxiliary numerical transformations. It can happen that a partitioning which is well-balanced with respect to the above-mentioned standard criteria may be completely unbalanced with respect to some time-critical operations on the domains. The general framework of multi-constraint graph partitioning may not solve the problem.

The graph partitioning problem is closely coupled with the general problem of *load balancing*. In particular, the partitioning represents a *static* load balancing. In practice, load distribution in a computation may be completely different from the original distribution at the beginning of the computation. Generally, dynamic load balancing strategies can then redistribute the work dynamically. A lot of interest was devoted to analysis of basic possible sources of such problems [4]. Principles of the cure of such problems one can find, e.g., in [7], [14]. In some situations, in order to cover complicated and unpredictably time-consuming operations on the individual domains, one can talk about minimization with respect to *complex objectives* [13], see also [12]. The strategy proposed in [13] consists in improving the partitioning iteratively during the course of the computation.

In some cases it is known much more about such critical operations. This paper aims at exploiting this knowledge. Then the additional information may be included into the graph partitioner, or used to improve the graph partitioning in one simple step providing some guarantees on its quality at the same time. Both these strategies have their own pros and cons. While integration of the additional knowledge into the graph partitioner seems to be the most efficient approach, it may not be very flexible. In addition, the analysis of such approach may not be simple when the typical multilevel character of partitioning algorithms is taken into account. A careful redistribution in one subsequent step which follows the partitioning seems to provide the useful flexibility.

TRADITIONAL ICS PRAGUE 2009

Since the time-critical operation performed on the domains is the sparse matrix factorization, the key to our strategy is to exploit the graph-theoretic tools and indicators for the repartitioning. Let us concentrate on the complete factorization of a symmetric and positive definite (SPD) matrix which is partitioned into two domains. In this case, the underlying graph model of the factorization is the *elimination tree.* Our first goal is to show the whole framework of theoretical and practical tools which may allow post-processing of a given graph partitioning in one simple step. Then the repartitioned graph should be better balanced with respect to the factorization. Further we will discuss one such tool more in detail. Namely, we will mention that we can directly compute number of columns which should be modified in the factorization after changes of the *border* nodes, which are vertices of the separated domains incident to the separator. We confirm both theoretically and experimentally that we can decrease the number of these modifications by carefully chosen reorderings.

Section 2 summarizes some terminology and describes the problem which we would like to solve. Section 3 explains basic ideas of our new framework. Then we discuss the problem of minimizing modifications in factorized matrices on domains both theoretically and experimentally.

2. Basic terminology and our restrictions

Let us first introduce some definitions and concepts related to the complete sparse matrix factorizations and reorderings. For simplicity we assume that the adjacency graph of all considered matrices are connected. Also, we will discuss only the standard graph model. Nevertheless, note that practical strategies for graph repartitioning should be based on blocks or other coarse representations which should be described by factorgraphs or hypergraphs.

The decomposition of an SPD matrix A is controlled by the *elimination tree*. This tree and its subtrees provide most of the structural information relevant to the sparse factorization. Just by traversing the elimination tree, sizes of matrix factors, their sparsity structure, supernodal structure or other useful quantities [3], [10] can be quickly determined. The elimination tree T is the rooted tree with the same vertex set as the adjacency graph G of A and with the vertex n as its root. It may be represented by one vector, typically called PARENT[.], defined as follows:

$$PARENT[j] = \begin{cases} \min\{i > j | \, l_{ij} \neq 0\}, & \text{for } j < n, \\ 0, & \text{for } j = n. \end{cases}$$

where l_{ij} are entries of L. The n-th column is the only column which does not have any offdiagonal entries.

When applying Cholesky factorization to a sparse matrix, it often happens that some matrix entries which were originally zeros become nonzeros. These new nonzero entries are called *fill-in*. High-quality sparse Cholesky factorization strongly minimizes the fill-in. Tools for this minimization are called *fill-in minimizing reorderings*. Basically, there are two categories of these reordering approaches. *Global* reorderings as nested dissection (ND) or generalized nested dissection (GND) consider the graph as one entity and divide it into parts by some predefined, possibly recursive heuristics. *Local* reorderings are based on subsequent minimization of some quantities which represent local estimates of the fill-in. Important representatives of such reorderings are MMD and AMD variations of the basic minimum degree (MD) algorithm.

Many quantities related to the sparse factorization of SPD matrices can be efficiently computed only if the matrix is preordered by an *additional* specific reordering apart from a chosen fill-in minimizing reordering. One such additional reordering useful in practical implementations is the *postordering*. It is induced by a postordering of the elimination tree of the matrix, being a special case of a *topological ordering* of the tree. For a given rooted tree, its topological ordering labels children vertices of any vertex *before* their parent. Note that the root of a tree is always labeled last. Further note that any reordering of a sparse matrix that labels a vertex earlier than its parent vertex in the elimination tree is equivalent to the original ordering in terms of *fill-in* and the operation count. In particular, postorderings are equivalent reorderings in this sense.

3. Framework for the graph-based repartitioning

In this section we will split the problem of repartitioning into several simpler tasks. Based on this splitting we will propose individual steps of our new approach. As described above, the problem arises if we encounter a lack of balance between sizes of the Cholesky factors on the domains. Using the elimination tree mentioned above, we are able to detect this imbalance without doing any actual factorization. This detection is very fast having its time complexity close to linear [10]. Then, the result of the repartitioning step is the new distribution of the graph vertices into domains which also implicitly defines the graph separator.

The repartitioning step can be naturally split into two simpler subproblems. First, one needs to decide

which vertices should be removed from one domain. Second, it should be determined *where these removed vertices should be placed* into the reordering sequence of the other domain. Alternatively, the domains may be reordered and their factorizations recomputed from scratch. In the following two subsections, we will consider these two mentioned subproblems separately. For both of them we present new considerations. The third subsection of this section will present one simpler task more in detail as well as both theoretical and experimental results.

3.1. Removal of vertices

Assume that the matrices on domains were reordered by a fill-in minimizing reordering. Further assume that some vertices should be removed from one domain to decrease the potential fill-in in the factorization. An important task is to determine which vertices should be removed from the domain such that their count would be as small as possible in addition to the further constraints mentioned below. In other words, the removal of chosen vertices should decrease the fill-in as fast as possible.

The following Algorithm 1 offers a tool to solve this problem. It counts the number of row subtrees of the elimination tree in which each vertex is involved. Note that *row subtrees* represent sparsity structures of rows of Cholesky factor and they can be found by traversing the elimination tree. The algorithm is new, but it was obtained by modifying the procedure which determines the leaves of the row subtrees in the elimination tree in [11].

Algorithm 1 *Count number of row subtrees in which the vertices are contained.*

```
for column j=1,n do
    COUNT(j):=n-j+1
    PREV_ROWNZ(j)=0
end for

for column j=1,n do
    for each a_ij ≠ 0, i > j do
        k=PREV_ROWNZ(i)
        if k < j − |T[j]| + 1
            for ξ = c_{t−1}, ..., c_t − 1 do
                COUNT(ξ ) = COUNT(ξ ) -1
            end for
        end if
        PREV_ROWNZ(i)=j
    end for
end for
```

Here T denotes the elimination tree of matrix A, and $T[i]$ denotes the subtree of T rooted in the vertex i. $T[i]$ also represents the vertex subset associated with the subtree, that is the vertex i and all its proper descendants in the elimination tree. $|T[i]|$ denotes the number of vertices in the subtree $T[i]$. Consequently, the number of proper descendants of the vertex i is given by $|T[i]| - 1$. PREV_ROWNZ is an auxiliary vector for tracking nonzeros in previously traversed rows. The computed quantity is denoted by COUNT. A critical assumption here is that the elimination tree is postordered.

Having computed the counts, our heuristic rule for fast decrease of the fill-in is to remove vertices with the largest COUNT. Let us note that the removal of vertices may also change the shape of the elimination tree, and our rule does not take this fact into account. To consider this, recent theory of sparse exact updates which uses multiindices should be taken into account, see the papers by Davis and Hager quoted in [3]. Further note that the removal should also take into account distance of the removed vertices from the border vertices. Therefore, we propose to use the counts from Algorithm 1 as a *secondary cost* for the Fiduccia-Mattheyses improvement of the Kernighan-Lin algorithm [6]. This is an iterative procedure which, in each iteration, looks for a subset of vertices from the two graph domains such that their swapping leads to a partition with smaller size of the edge separator. Our modification of the cost function then seems to enable more efficient repartitionings.

3.2. Insertion of vertices

Having a group of vertices to be inserted into the new domain D_2 we need to determine where these vertices should appear in the new reordering sequence. Here the objective function is to *minimize the effect on the fill-in* in the corresponding Cholesky factor. Note that in the next subsection we will mention another motivation: minimize the number of columns to be recomputed in the Cholesky factor, if it was computed. Shortly, theoretical considerations related to the delayed elimination in [8] motivate the insertion of a vertex to the position of the *parent of the least common ancestor* of its neighbors in the elimination tree T which we have.

Consider a vertex to be inserted, and denote by N the set of its neighbors in D_2. Let α be the least common ancestor of N in T. Denote by $T_r[\alpha]$ the unique subtree of T determined by α and N. Given vertex will connect to all vertices on the path among their neighbors N. Then the *increase of the fill-in* in the new decomposition includes one edge for each vertex of $T_r[\alpha]$ and at most β multiple of the union of adjacency sets of the vertices from $T_r[\alpha]$ where β is the distance from α to the root of

DRAFT 14. srpna 2009 - 14:42

T plus one.

$$|T_r[\alpha]| + (1 + height(\alpha)) \left| \bigcup_{r \in T_r[\alpha]} \{i | \, L_{ir} \neq 0, i > \alpha\} \right|$$

In order to minimize the effect of the insertion on the fill-in, we need to minimize this amount. As in the previous subsection, this criterion may represent an additional cost function for the local heuristic like Kerninghan-Lin and we are about to perform an experimental study of its application.

3.3. Repartitioning for generalized decompositions

Consider the problem of repartitioning when we construct a factorization for which it is difficult to obtain a tight prediction of the fill-in. An example can be the *incomplete* Cholesky decomposition. Similar situation can be faced when solving a nonlinear problem via a sequence of systems of linear equations. Then we face the two following problems: repartitioning as well as the *recomputation of the decomposition*. In this subsection we propose techniques that minimize the effort needed for recomputing the partition by a careful choice of reorderings in advance. The efficiency of the new strategies is measured by the counts of columns or block columns which should be recomputed. For simplicity, here we restrict ourselves to changes in the domain from which the vertices are removed.

The first approach which we propose generalizes the concept of local reorderings with constraints. This concept was introduced in [9] to combine local and global approaches, and recently investigated in [2]. Our procedure exploits the minimum degree reordering and uses the *distance of vertices from the main separator* as the second criterion which breaks the MD ties.

Table 1: *Counts of columns which should be recomputed in Cholesky decomposition if boundary vertices are modified*

matrix	application	dimension	nnz	standard MD	new approach
bmw7st_1	structural mechanics	141,347	3,740,507	7,868	5,039
bodyy6	structural mechanics	19,366	77,057	2,354	476
cfd1	CFD pressure matrix	70,656	949,510	10,924	7,497
cfd2	CFD pressure matrix	123,440	1,605,669	15,021	10,416
hood	car hood	220,542	5,494,489	7,099	2,192
kohn-sham4	quantum chemistry	90,300	270,598	3,564	2,233
m_t1	tubular joint	97,578	4,925,574	9,093	7,095
pwtk	pressurized wind tunnel	217,918	5,926,171	8,218	4,437
x104	beam joint	108,384	5,138,004	4,656	3,842

Table 1 summarizes numerical experiments with the new reordering. All matrices except for the discrete Kohn-Sham equation are from the Tim Davis collection. The counts of factor columns to be recomputed (standard and new strategy, respectively) if a group of border nodes of the size fixed to two hundred is removed are in the last two columns. The counts were computed via the elimination tree.

Here, we will demonstrate the power of our approach on a simple model problem depicted in Figure 1

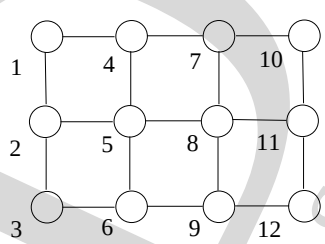


Figure 1: Graph which demonstrates our modified minimum degree reordering. We assume that the main separator contains nodes 10, 11 and 12.

If the border nodes are 10,11 and 12, then our approach provides the reordering sequence: $1, 3, 4, 6, 2, 5, 9, 7, 8, 10, 12, 11$. Note that not only the border nodes are ordered last, but also the nodes which are more distant from the border are ordered sooner. A principal advantage over the concept of constrained minimum degree family of the algorithms with just two sets which are ordered [2] is that here we do not need to know in advance how much the domain will be changed.

Let us present formalized theoretical result for the structured grids. We will show that the choice of the first separator in the case of a $k \times k$ regular grid problem strongly influences the number of columns to be recomputed in case the border is modified by the *removal or insertion*. The situation is depicted in Figure 2 for $k = 7$. The figures represent the separator and the subdomain sets after four steps of ND. The border vertices are on the right and they are filled. The following theorem uses the *separator tree* in which the vertices describe the subdomain sets and separators, and which is a coarsening of the standard elimination tree.

DNTA 14. srpna 2009

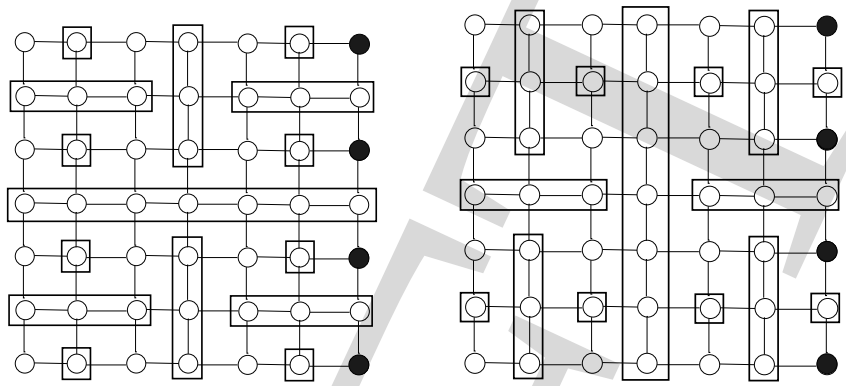


Figure 2: Grids with the ND separator structure related to Theorem 3.1. Type-I-grid on the left and Type-II-grid on the right.

Theorem 3.1 *Consider the matrix A from a $k \times k$ regular grid problem with ND ordering having l levels of separators. Assume that the matrix entries corresponding to the border vertices are modified. Denote by a_l and b_l, respectively, maximum number of matrix block columns which may change in the Cholesky decomposition of A from Type-I-grid or Type-II-grid. Then $lim_{l\to\infty} a_l/b_l = 3/2$ for odd l and $lim_{l\to\infty} a_l/b_l = 4/3$ for even l.*

Proof: Clearly, $a_1 = 3$ since the changes influence both domains. Consequently, all block columns which correspond to the entries of the separator tree have to be recomputed. Similarly we get $b_1 = 2$ since the block factor column which corresponds to the subdomain without the border vertices does not need to be recomputed. Consider the Type-I-grid with $k > 1$ separators. It is clear that the separator structure of this grid we get by doubling Type-II-grid and separating them by a central separator. Consequently, $a_{k+1} = 2 * b_k + 1$, where the additional 1 corresponds to the central separator. Similarly we get the relation $b_{k+1} = a_k + 1$, since its separator structure is the same as if we would add to the considered Type-II-grid with $k > 1$ separators of another Type-II-grid and separate them by the central separator. The block columns of the new Type-II-grid do not need to be recomputed. Putting the derived formulas together we get $a_{k+2} = 2 * a_k + 3$ and $b_{k+2} = 2 * b_k + 2$. This gives $a_l = 3(2^{l+1} - 1)$ and $b_l = 2(2^{l+1} - 1)$ for $k = 2l + 1$, and $a_l = 4(2^l) - 3$ and $b_l = 3.(2^l) - 2$ for $k = 2l$, and we are done. ■

Similar result we can get for GND reordering

Theorem 3.2 *Consider the matrix A from a $k \times k$ regular grid problem with generalized nested dissection (GND) ordering having l levels of separators. Assume that the matrix entries corresponding to the border vertices are modified. Denote by a_l and b_l, respectively, maximum number of matrix entries which may change in the Cholesky decomposition of A from Type-I-grid or Type-II-grid. Then $lim_{l\to\infty} a_l/b_l = 4/3$.*

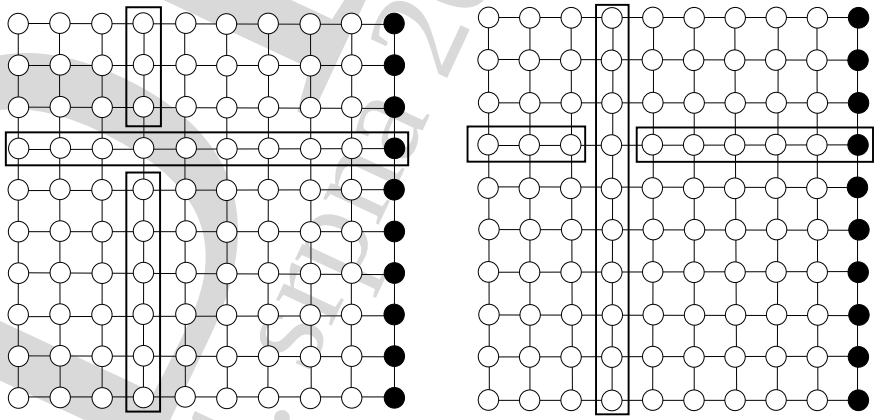


Figure 3: Grids with the GND separator structure related to Theorem 3.2. Type-I-grid on the left and Type-II-grid on the right.

DATA 14. ročník 2009 14:52

Proof: Clearly, $a_1 \leq k^2 + \beta k$ since the changes influence both domains and separator. Consequently, all matrix entries have to be recomputed. Similarly we get $b_1 \leq \alpha k^2 + \beta k$. Consider the Type-I-grid with $l > 1$ separators. Consequently,

$$a_l \leq \alpha^{\lfloor \frac{l}{2} \rfloor} k^2 + \beta k l.$$

Similarly we get the relation

$$b_l \leq \alpha^{\lceil \frac{l}{2} \rceil} k^2 + \beta k + \frac{3}{2}\beta k \lfloor \frac{l}{2} \rfloor, \qquad b_l \leq \alpha^{(\frac{l}{2})} k^2 + \frac{3}{2}\beta k \frac{l}{2}.$$

If we consider odd $l = 2i + 1$, we get

$$\lim_{l\to\infty} \frac{a_l}{b_l} = \lim_{i\to\infty} \frac{a_i}{b_i} = \lim_{i\to\infty} \frac{\alpha^i k^2 + (2i+1)\beta k}{\alpha^{i+1} k^2 + \beta k + \frac{3}{2}\beta k i} = \lim_{i\to\infty} \frac{\frac{\alpha^i k^2}{i} + \frac{\beta k}{i} + 2\beta k}{\frac{\alpha^{i+1} k^2}{i} + \frac{\beta k}{i} + \frac{3}{2}\beta k} = \frac{4}{3}$$

For even $l = 2i$ we get same result.

$$\lim_{l\to\infty} \frac{a_l}{b_l} = \lim_{i\to\infty} \frac{a_i}{b_i} = \lim_{i\to\infty} \frac{\alpha^i k^2 + 2i\beta k}{\alpha^i k^2 + \frac{3}{2}\beta k i} = \frac{4}{3}$$

■

Clearly, the choice of the first separator of ND and GND plays a decisive role. Further, there exist accompanying results for the generalized ND and one-way dissection. The counts of modified vertices were obtained from the separator tree [10].

4. Conclusion

We considered new ways to find proper and fast graph repartitioning if our task is to decompose matrices on the domains. In this case it is possible to propose efficient and theoretically sound new ways refining the general-purpose concept of complex objectives. The approach goes beyond a straightforward use of symbolic factorization. After describing a comprehensive framework of the whole approach we presented theoretical and experimental results for one particular problem. The explained techniques can be generalized to more domains and for general LU decomposition.

References

[1] U.V. Catalyürek and C. Aykanat, "Hypergraph-partitioning based decomposition for parallel sparse-matrix vector multiplication". IEEE Transactions on Parallel and Distributed Systems **20** (1999) 673–693.

[2] Y. Chen, T.A. Davis, W.W. Hager, and S. Rajamanickam, "Algorithm 887: CHOLMOD", Supernodal sparse Cholesky factorization and update/downdate. ACM Trans. Math. Softw., 35 (2008), 22:1–22:14.

[3] T.A. Davis, "Direct Methods for Sparse Linear Systems". SIAM, Philadelphia (2006).

[4] B. Hendrickson, "Graph partitioning and parallel solvers: Has the emperor no clothes?" LNCS **1457**, Springer (1998) 218–225.

[5] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs". SIAM J. Sci. Comput. **20** (1999) 359–392.

[6] B.W. Kernighan and S.Lin, "An efficient heuristic procedure for partitioning graphs". The Bell System Technical Journal **49** (1970) 291–307.

[7] V. Kumar, A. Grama, A. Gupta, and G. Karypis, "Introduction to Parallel Computing". Benjamin-Cummings (1994).

[8] J.W.H. Liu, "A tree model for sparse symmetric indefinite matrix factorization". SIAM J. Matrix Anal. Appl. **9** (1988) 26–39.

[9] J.W.H. Liu, "The minimum degree ordering with constraints". SIAM J. Sci. Comput. **10** 1989 1136–1145.

[10] J.W.H. Liu, "The role of elimination trees in sparse factorization". SIAM J. Matrix Anal. Appl. **11** (1990) 134–172.

[11] J.W.H. Liu, E.G. Ng, and B.W. Peyton, "On finding supernodes for sparse matrix computations". SIAM J. Matrix Anal. Appl. **14** (1993) 242–252.

[12] A. Pinar and B. Hendrickson, "Combinatorial Parallel and Scientific Computing". in: Parallel Processing for Scientific Computing, M. Heroux, P. Raghavan, and H. Simon, eds., SIAM (2006) 127–141.

[13] A. Pinar and B. Hendrickson, "Partitioning for complex objectives". Parallel and Distributed Processing Symposium **3** (2001) 1232–1237.

[14] K. Schloegel, G. Karypis, and V. Kumar, "A unified algorithm for load-balancing adaptive scientific simulations". No. 59 in: Proceedings of the ACM/IEEE Symposium on Supercomputing, ACM (2000).

DRAFT 14. srpna 2009 - 14:32

Parallel Mining of Frequent Itemsets

Post-Graduate Student:
ING. ROBERT KESSL
CTU FEE, Department of Computers
Karlovo náměstí 13
121 35 Praha 2
kesslr@cs.cas.cz

Supervisor:
PROF. PAVEL TVRDÍK, DRSC.
CTU FEE, Department of Computers
Karlovo náměstí 13
121 35 Praha 2
tvrdik@fel.cvut.cz

Field of Study:
Computer Science

Abstract

This paper presents the Parallel-FIMI-Seq and Parallel-FIMI-Par methods for static load balancing of mining of frequent itemsets on a distributed-memory parallel computer. The method partitions all frequent itemsets into partitions of approximatelly the same size. We experimentaly show the speedup of the method for up to 10 processors. The method achieves the speedup $\approx$ 6 on 10 processors and the speedup is linear in the number of processors for a reasonably structure database.

1. Introduction

Due to the growth of the computational power and the cheap storage media the companies store huge amount of data. The companies would like to analyze the data and use them to grow revenue. The process of analysis of the data uses the so called *data mining*.

One of the important data mining task is the so called *association rule mining* or *market basket analysis*. The customers are visiting a supermarket and the owner of the supermarket is storing the basket of each customer as a *transaction* in a database. We are searching for rules like {bread, butter} $\Rightarrow$ {milk}, i.e. if a customer buys bread and butter he will likely buy milk. The association rules are generated from the so called *frequent itemsets* (FIs in short). A frequent itemset can be for example {bread, butter, milk}.

In this paper we will discuss the parallel algorithms for mining of frequent itemsets.

The paper is organized as follows: in Section 2 we give a brief theoretical overview of the mathematics used in the algorithms; in Section 3 we show how to use the theory for mining of FIs; In Section 5 and in Section 4 we described the parallel algorithm for mining of FIs; and in Section 6 we experimentally evaluate the parallel algorithm.

2. Mathematical foundation

Let B be a *base set* of items (items can be numbers, symbols, strings, goods, etc.). A *transaction* $t = (id, U)$, where U is any subset $t \subseteq B$ and id is a unique transaction identification. If $W \subset U$ for a transaction $t = (id, U)$ we simply write $W \subset t$. A superset of a transaction will be denoted similarily, i.e. $t \subset V$. Further, we need to view the baseset B as an ordered set. The items are therefore ordered using an order $<$: $b_1 < b_2 < \ldots < b_n, n = |B|$. The order can be changed dynamicaly during the execution of a depth-first search algorith for mining of FIs.

A transaction is a set of items. However, in most algorithms, we need to view it as an ordered set. So, $t[i]$ denotes the ith item in transaction t ordered using the same relation $<$ we use for ordering the baseset B (does not matter how the order is chosen). A *database* $\mathcal{D}$ on B (or database $\mathcal{D}$ if B is clear from context) is a sequence of transactions $t \subseteq B$. Other subsets, not necessarily in the database, of the baseset B will be further called *itemsets*.

Definition 2.1 (Itemset cover and support) *[1] Let $U \subseteq B$ be an itemset. Then the* cover *of U is the subset of transactions from database $\mathcal{D}$ that contain U as a subset. This subset is denoted by $\mathcal{T}(U, \mathcal{D})$. The number of transactions in $\mathcal{T}(U, \mathcal{D})$ is called the support of U in $\mathcal{D}$, $Supp(U, \mathcal{D}) = |\mathcal{T}(U, \mathcal{D})|$.*

We will also use the word *tidlist* (an abbreviation of the transaction ID list) for the list of transaction IDs $\mathcal{T}(U, \mathcal{D})$. We define the support as the number of transactions containing U, but in some literature, the relative support is defined by $Supp^*(U) = Supp(U)/|\mathcal{D}|$.

DRAFT
14. SIPR 2009 - 14.4.2009

Definition 2.2 (Frequent itemset) *Let $\mathcal{D}$ be a database on B, $U \in B$ an itemset, and $min_support \in Z$ a natural number. We call U* frequent *in the database $\mathcal{D}$ if $Supp(U, \mathcal{D}) \geq min_support$.*

We will denote the set of all frequent itemsets as $\mathcal{F}$. In the text, we use $\mathcal{D}$ and $min_support$ generally. If neccessary, the database $\mathcal{D}$ and the value of minimal support $min_support$ will be clear from the context.

In our algorithms we need to sample the set $\mathcal{F}$. A sample of frequent itemsets is denoted by $\mathcal{F}_{smpl}$.

Definition 2.3 (Maximal Frequent Itemset) *Let $\mathcal{D}$ be a database on B, $U \subset B$ an itemset, and $min_support \in Z$ a natural number. We call U* maximal frequent itemset *if $Supp(U, \mathcal{D}) \geq min_support$, and for all $V, U \subsetneq V$, $Supp(V, \mathcal{D}) < min_support$.*

We denote the set of all maximal frequent itemsets (MFIs in short) as $\mathcal{M} = \{m_i\}, 1 \leq i \leq n$. The MFIs delimit the set $\mathcal{F}$ from above in the sense of the set inclusion.

Let $X, Y \subseteq B$ be frequent itemsets such that $Y \cap X = \emptyset$. The ordered pair (X, Y), written $X \Rightarrow Y$, is called an *association rule*. The itemset X is called an *antecedent* and the itemset Y is called a *consequent*. The strength of the association rule is measured by the support $Supp(X \cup Y)$ and by the confidence $Conf(X, Y) = \frac{Supp(X \cup Y)}{Supp(X)}$

The association rules are mined in a two step process: 1) mine all FIs $X = U \cup W, W \cap U = \emptyset$; 2) create association rules $U \Rightarrow W$ from the FIs mined in the first step. The values of $min_support$ and $min_confidence$ and a database $\mathcal{D}$ are inputs for algorithms for mining of association rules. These algorithms first find all frequent itemsets, using the $min_support$, and then generate association rules, using $min_confidence$. In this paper, we concentrate on the first step.

For the purpose of the description of our parallel algorithm, we denote the number of processors by P, the ith processor by p_i. At the start of the parallel algorithm, each processor has a database partition. A database partition is denoted by D_i. Our parallel algorithms partitions the database at the beggining into disjoint database partitions D_i, D_j such that $\cup_i D_i = \mathcal{D}$, $D_i \cap D_j = \emptyset$, and $|D_i| \approx |\mathcal{D}|/P$.

2.1. The monotonicity of support

The basic property of frequent itemsets is the so called *monotonicity of support*. The monotonicity of support is important for all algorithms for mining of FIs:

Lemma 2.4 (Monotonicity of support) *Let $U \subseteq B$ be an itemset with support $Supp(U)$ in a database $\mathcal{D}$. For every superset V of U holds: $Supp(U) \geq Supp(V)$.*

Proof: It is clear that if a set U is contained in transactions $\mathcal{T}(U)$ then a superset $V \supset U$ is contained in transactions $\mathcal{T}(V) \subseteq \mathcal{T}(U)$. ■

Corollary 2.5 *Let V be a frequent itemset, then all subsets $U \subseteq V$ are also frequent.*

2.2. The lattice of all itemsets

Zaki[2] use the set of all items, $\mathcal{P}(B)$, and the underlying lattice for description of depth-first search (DFS in short) algorithms.

Definition 2.6 *Let P be an ordered set, and let $S \subseteq P$. An element $X \in P$ is an* upper bound *(*lower bound*) of S if $s \leq X$ ($s \geq X$) for all $s \in S$. The least upper bound is called* join *and is denoted by $\bigvee S$, and the greatest lower bound, also called* meet*, of S is denoted $\bigwedge S$. The greatest element of P, denoted $\top$, is called the* top element*, and the least element of P, denoted $\bot$, is called* bottom element

Definition 2.7 *Let $\mathcal{L}$ be an ordered set, $\mathcal{L}$ is called a* join *(*meet*)* semilattice *if $X \vee Y$ ($X \wedge Y$) exists for all $X, Y \in \mathcal{L}$. $\mathcal{L}$ is called a* lattice *if it is both a join and meet semilattice. $\mathcal{L}$ is* complete lattice *if $\bigvee S$ and $\bigwedge S$ exist for all subsets $S \subseteq \mathcal{L}$. An ordered set $M \subset \mathcal{L}$ is a* sublattice *of $\mathcal{L}$ if $X, Y \in M$ implies $X \vee Y \in M$ and $X \wedge Y \in M$.*

It is well known that for a set S the powerset $\mathcal{P}(S)$ is a complete lattice. The *join* operation is the *set union operation* and *meet* the *set intersection operation*.

For any $\mathcal{S} \subseteq \mathcal{P}(B)$, $\mathcal{S}$ forms a lattice $(\mathcal{S}; \subseteq)$ of sets if it is closed under finite number of unions and intersections.

Lemma 2.8 *The set of all* frequent itemsets *forms a* meet semilattice.

DRAFT 12.2.2009 14:32

Proof: The result follows from the corollary 2.5 and the fact that $V \wedge W = V \cap W$. ■

Proposition 2.9 *The set of maximal frequent itemsets bounds the set of all frequent itemsets from above in the lattice.*

3. Using the lattice of frequent itemsets in algorithms

For parallelization of the FIM algorithms we need to partition the lattice of all itemsets into disjoint partitions. An equivalence relation partitions the set $\mathcal{P}(B)$ into disjoint subsets called *prefix-based equivalence classes*:

Definition 3.1 (prefix-based equivalence class) *Let $U \subset B, |U| = n$ be an itemset. We use the order of the set B and hence view $U = (u_1, u_2, \ldots, u_n), u_i \in B$ as an ordered set. A* prefix-based equivalence class *of U, denoted by $[U]_l$, is a set of all itemsets that have the same prefix of length l, i.e. $[U]_l = \{W = (w_1, w_2, \ldots, w_m) | u_i = w_i, i \leq l, m \geq l, W \subseteq B\}$*

To simplify the notation, we use $[U]$ for the prefix-based equivalence class $[U]_l, l = |U|$.

Proposition 3.2 *Let $U \subseteq B$ be an itemset and $l \leq |U|$ a natural number.* The prefix-based relation $[U]_l$ *is an equivalence for fixed l.*

Lemma 3.3 *Let $W \subseteq B$ be an itemset. Each equivalence class $[W]$ is a sublattice of the lattice $(\mathcal{P}(B), \subseteq)$.*

Proof: Let U, V be itemsets in class $[W]_l$, i.e., U, V share common prefix W. $W \subseteq U \cup V$ implies that $U \wedge V \in [W]_l$, and $W \subseteq U \cap V$ implies that $U \vee V \in [W]_l$. Therefore, $[W]_l$ is a sublattice of $(\mathcal{P}(B), \subseteq)$. ■

Definition 3.4 *Let $U, W \subseteq B$ and $[U], [W]$ be prefix-based equivalence classes. We call $[W]$ a* prefix-based equivalence subclass *of $[U]$ if and only if $[W] \subsetneq [U]$.*

Proposition 3.5 *Let $W, U \subseteq B$. If $[W]$ is a prefix-based equivalence subclass of $[U]$ then $U \subsetneq W$.*

The lattice $(\mathcal{P}(B), \subseteq)$ can be partitioned into disjoint prefix-based equivalence classes. The partitioning depends on the order of B, as described in Section 2. Because the prefix-based equivalence classes form a hierarchy, we partition the lattice into sublattices recursively. The recursive partitioning forms a tree, where each node corresponds to one itemset. Let U_i, W and $1 \leq i \leq n$ be itemsets, where the nodes labeled by U_i correspond to the successors of the node labeled by W. That is: $W \subset U_i$, $l = |U_i| = |W| + 1$, and $[U_i]$ is a prefix-based equivalence subclass of $[W]$. The set of items $\bigcup_i U_i \setminus W$ is called the *extensions of* W. The partitioning of $(\mathcal{P}(B), \subseteq)$ into the prefix-based equivalence classes $[U_i]$ implies partitioning of $\mathcal{F}$ into classes of the form $\mathcal{F} \cap [U_i]$. For the purpose of our parallel algorithm, we need to partition $\mathcal{F}$ into k disjoint sets, denoted $(F_1, \ldots, F_k)$, satisfying $F_i \cap F_j = \emptyset, i \neq j$, and $\bigcup_i F_i = \mathcal{F}$. Each partition F_i (union of prefix-based equivalence classes) is a meet sublattice of $(\mathcal{P}(B), \subseteq)$.

The prefix-based equivalence classes decompose the lattice into smaller parts for which computing supports can be done independently in main memory. That is, for the computation of supports of itemsets in one prefix-based class, we start with the tidlists of the atoms and recursively construct the tidlists of itemsets belonging to that class by intersecting them. Due to this, the computation of support in different prefix-based classes is done independently. This is important, because this independence makes parallelization easier. Moreover, we can recursively decompose each equivalence class into smaller prefix-based equivalence subclasses.

For computation of supports of an itemset $U \subseteq B$ we use the tidlists $\mathcal{T}(\{b_i\}), b_i \in B$. The support of U, $|\mathcal{T}(U)|$, can be computed using the tidlists $\mathcal{T}(\{b_i\})$:

Lemma 3.6 *Let B be a baseset and $U \subseteq B, U = \bigcup_{b_i \in U} \{b_i\}$. Then the support of U can be computed by $Supp(U) = |\bigcap_{b_i \in U} \mathcal{T}(b_i)|$.*

Proof: The support of $U = \{u_i | 1 \leq i \leq n, u_i \in B\}$ is defined by $Supp(U) = |\mathcal{T}(U)|$, i.e. the number of transaction containing all the items u_i. Hence, the set of all transactions containing U is $\mathcal{T}(U) = \cap_i \mathcal{T}(u_i)$. ■

Corollary 3.7 *Let B be a baseset and $U, W_i \subset B, 1 \leq i \leq n$ and $U = \bigcup_i W_i$ then $Supp(U) = |\bigcap_i \mathcal{T}(W_i)|$.*

It follows that for a prefix Π and the extensions Σ, we can compute support of $\Pi \cup U, U \subset \Sigma$ using the tid lists of items in Σ and the tidlist $\mathcal{T}(\Pi)$.

DRAFT 14. srpna 2009 - 14:32

4. Proposal of a new DM parallel method

We have created a method for *Parallel Frequent Itemset MIning* (Parallel-FIMI in short). The basic idea is to partition all FIs into disjoint sets using the prefix-based equivalence classes of relative sizes $\approx \frac{1}{P}$. The prefix-based classes are then assigned to processors and each processor computes the FIs from the assigned prefix-based classes. This procedure statically balance the computational load. The size of a prefix-based equivalence classes is estimated using a sample of FIs, denoted by $\mathcal{F}_{smpl}$, computed from a sample of the database, denoted by $\mathcal{D}_{smpl}$. The prefix-based equivalence classes are then assigned to processors, so each processor computes approximately the same number of FIs. The method consists of four phases:

Phase 1 (sampling of FIs): generally the purpose of the first phase is to compute a sample $\mathcal{F}_{smpl}$ of all frequent itemsets $\mathcal{F}$. We sample the database $\mathcal{D}$ making the database sample $\mathcal{D}_{smpl}$. The algorithm then computes the sample of FIs $\mathcal{F}_{smpl}$ using the database sample $\mathcal{D}_{smpl}$. To make the whole process more efficient, we can create $\mathcal{F}_{smpl}$ in parallel. The parallel computation of $\mathcal{F}_{smpl}$ is balanced dynamically.

Phase 2 (lattice partitioning): we use the $\mathcal{F}_{smpl}$ for constructing the prefix-based equivalence classes. The classes are collated and assigned to processors.

Phase 3 (data distribution): is only a communication phase. It serves only for exchanging the input database among the processors.

Phase 4 (computation of FIs): in this phase, we run an arbitrary sequential algorithm that computes frequent itemsets in the assigned prefix-based equivalence classes.

4.1. Detailed description of Phase 1

In this phase, we need to compute a sample $\mathcal{F}_{smpl}$ of all frequent itemsets. Because the whole database can be quite large, we compute $\mathcal{F}_{smpl}$ using the database sample $\mathcal{D}_{smpl}$ using the MFIs. The details of this process are described below.

Toivonen[3] presented an analysis of the sampling of the database used for mining of FIs. Using the database sample $\mathcal{D}_{smpl}$, we can efficiently estimate support of a particular itemset U. The error of the estimate of $Supp(U, \mathcal{D})$ from a database sample $\mathcal{D}_{smpl}$ is defined by:

$$E(U, |\mathcal{D}_{smpl}|) = |Supp^*(U, \mathcal{D}) - Supp^*(U, \mathcal{D}_{smpl})|$$

The error can be analyzed using the sampling with replacement with no other constraint on the database. The error analysis then holds for a database of arbitrary size and properties. From the following theorem, we can estimate $Supp(U, \mathcal{D})$ with error ϵ that occurs with probability δ:

Theorem 4.1 *[3] Given an itemset $U \subseteq B$ and a random sample $\mathcal{D}_{smpl}$ drawn from database $\mathcal{D}$ of size:*

$$|\mathcal{D}_{smpl}| \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta},$$

then the probability that $E(U, |\mathcal{D}_{smpl}|) > \epsilon$ is at most δ.

MFI based sampling: Let $\mathcal{M} = \{m_i\}$ be the set of all MFIs. The set of all FIs is given by $\mathcal{F} = \cup\mathcal{P}(m_i)$. The approximation of MFIs $\mathcal{M}_{apprx} = \{m'_i\}$ is the set of all MFIs computed from $\mathcal{D}_{smpl}$. To create the sample $\mathcal{F}_{smpl}$, we first create $\mathcal{M}_{apprx}$. The set of all FIs represented by $\mathcal{M}_{apprx}$ is denoted by $\mathcal{F}_{apprx} = \cup\mathcal{P}(m'_i)$. Because $\mathcal{M}_{apprx}$ represents $\mathcal{F}_{apprx}$, $\mathcal{F}_{smpl}$ is created using $\mathcal{M}_{apprx}$.

The uniform sampling of $\mathcal{F}_{apprx}$ could be performed by Monte Carlo method: the *coverage algorithm* [4]. The coverage algorithm can be quite slow, because it makes $O(|\mathcal{M}_{apprx}|)$ checks for each sample and the size of $\mathcal{M}_{apprx}$ can be quite large. Therefore, we give up uniform sampling of $\mathcal{F}_{apprx}$ and use the following procedure: pick i with probability $Pr[i] = \frac{|\mathcal{P}(m'_i)|}{\sum|\mathcal{P}(m'_i)|}$ and then select $v \in \mathcal{P}(m'_i)$ uniformly at random. This makes the sampling non-uniform because it prefers itemsets contained as a subset of many MFIs. Therefore, the estimate of the prefix-based subspace is just a heuristic.

To mine the MFIs, we have used the *fpmax** [5] algorithm. To make the mining of MFIs faster, we can execute the *fpmax** in parallel. Therefore, we have two versions of the first phase: a) MFIs computed sequentialy on single processor; b) MFIs computed in parallel on multiple processors using dynamic load-balancing.

The dynamic load-balancing of mining of MFIs works this way: because $\mathcal{D}_{smpl}$ is much smaller than the whole database $\mathcal{D}$, the processors create $\mathcal{D}_{smpl}$ from $\mathcal{D}$, i.e. every processor has its copy of the database sample $\mathcal{D}_{smpl}$ and knows the items that are frequent in the database $\mathcal{D}$ (note that $\mathcal{D}$ is distributed among the processors). We assume that all items $b_i \in B$ are frequent. All processors partition the base set

DRAFT 14. April 2009 - 11:52

$B, |B| = N$ on P parts of size N/P. Processor p_i runs a sequential MFI algorithm in the i-th part of B, where the items are interpreted as 1-prefixes. The 1-prefixes $\{b_i\}$ are prefix-based equivalence classes $[\{b_i\}]$.When a processor finishes its assigned items, it asks other processors for work. The computation is terminated using the Dijkstra's token termination detection algorithm.

Let $\mathcal{M}^i_{apprx} = \{m'_i\}$ be the set of all MFIs, computed by the processor p_i from $\mathcal{D}_{smpl}$. The sampling of $\mathcal{F}_{apprx}$ is then performed in the following way: every processor p_i broadcasts the sum $s_i = \sum_{m \in \mathcal{M}^i_{apprx}} |\mathcal{P}(m)|$ of sizes of powersets of its local MFIs (hence, an all-to-all broadcast takes place) and then it gets $\frac{s_i}{\sum s_i}$ fraction of the samples $\mathcal{F}_{smpl}$.

Because the computation of the approximationn to the MFIs is done in parallel using the dynamic load-balancing, the output of the algorithm is a superset of all MFIs, as shown below. For computation of MFIs, we use the DFS *fpmax** algorithm. *fpmax** uses optimalizations that at each step checks every currently processed candidate MFI against the already computed MFIs. If the candidate MFI is found, the algorithm removes the current itemset from processing. Because the computation is distributed, the algorithm is unable to check the candidate against all MFIs resulting in a superset of all MFIs.

For example: let $B = \{b_1, b_2, b_3, b_4, b_5, b_6\}$ and $P = 2$ and assume that processor p_1 is scheduled to processes prefix-based equivalence classes $[b_1], [b_2], [b_3]$ and p_2 is scheduled to process prefix-based equivalence classes $[b_4], [b_5], [b_6]$. Processor p_1 processes only prefixes $\{b_1\}, \{b_2\}, \{b_3\}$, but use all items B as extensions, e.g. processor p_i uses for prefix $\{b_1\}$ extensions b_2, b_3, b_4, b_5, b_6, for prefix $\{b_2\}$ extensions b_3, b_4, b_5, b_6, etc. Let the itemset $U = \{b_2, b_3, b_5, b_6\}$ be an MFI. The processor p_1 computes U correctly, but processor p_2 computes also the itemset $\{b_5, b_6\}$ as an MFI. The reason is that p_2 does not know that the MFI $\{b_2, b_3, b_5, b_6\}$ was already computed by processor p_1.

Despite the problem, the computed itemsets still delimit $\mathcal{F}_{apprx}$. The *fpmax** algorithm runs in parallel and aside computing all the MFIs computes some additional non-MFI itemsets. However, the additional itemsets are always subsets of some MFIs. The reason is that every processor has the same database sample $\mathcal{D}_{smpl}$ and the *fpmax** always correctly computes the support of an itemset that is an candidate on MFI.

4.2. Detailed description of the phase 2

The phase 2 is responsible for partitioning of $\mathcal{F}$. As an input of the partitioning we use the samples $\mathcal{F}_{smpl}$ from the phase 1.

We partition $\mathcal{F}$ into $F_i, 1 \leq i \leq P$ such that $\mathcal{F} = \bigcup_i F_i$ and $|F_i| \approx |\mathcal{F}|/P$. Each F_i is a union of some prefix-based classes intersected with $\mathcal{F}$. Hence, each F_i can be solved independently on processor p_i. First, we create a list of prefix-based equivalence classes $[U_k]$ small enough, so that we can create set of indexes L_i such that $|F_i|/|\mathcal{F}| \approx 1/P$, where $F_i = \bigcup_{k \in L_i} [U_k] \cap \mathcal{F}$.

Prefix-based classes $[U_k]$ are created so that the relative size satisfies $\frac{|[U_k] \cap \mathcal{F}|}{|\mathcal{F}|} \leq \alpha \cdot \frac{1}{P}$, where $0 < \alpha < 1$ is a parameter of the computation. We initially set $U_k = \{b_i\}$ and estimate the size of $[U_k]$ using $\mathcal{F}_{smpl}$. If some of the prefix-based class $[U_k]$ is too big, i.e. $\frac{|[U_k] \cap \mathcal{F}|}{|\mathcal{F}|} > \alpha \cdot \frac{1}{P}$, we recursively break $[U_k]$ into smaller prefix-based subclasses.

The problem of creating L_i such that $F_i = \bigcup_{k \in L_i} ([U_k] \cap \mathcal{F})$ and $\max_i |F_i|/|\mathcal{F}|$ minimized is known to be NP-complete problem with known approximation algorithms. We will use the LPT-SCHEDULE algorithm (see [6] for the proofs). The LPT-SCHEDULE algorithm is a best-first algorithm, see Algorithm 4.2.

Lemma 4.2 *[6]*LPT-SCHEDULE *is* $4/3$*-approximation algorithm.*

The schedule is then broadcasted to the processors.

Algorithm 1 LPT-SCHEDULE

1: Sort all prefixes U_k in decreasing order given by the relative size $|[U_k]|/|\mathcal{F}|$.
2: Assign $[U_k]$ in greedy manner to processor p_i, creating the index sets L_i.

A DFS algorithm (like Eclat and FPGrowth) expands every prefix Π using the extensions Σ sorted by the support in ascending order. Using a different order can significantly reduce the speedup of the algorithm. This allows for efficient computation of intermediate steps. This optimalization is used by other DFS algorithms, e.g. the FPGrowth algorithm. The order of the extensions is estimated from the database sample $\mathcal{D}_{smpl}$.

While experimenting with the Eclat algorithm, we have observed that the run of the sequential Eclat algorithm in

DATA 14. ročník 2009 - 1.59

the phase 4 can be very slow. Each processor is assigned a set of prefixes together with the set of extensions for every prefix. The reason of the slow run of the Eclat algorithm in phase 4 is the different ordering of extensions used for creating prefix-based classes in the sequential and the parallel version of the algorithm.

4.3. Detailed description of the phase 3

Every processor has to send its database partition D_i to all other processors. The broadcast is done in $\lfloor \frac{P}{2} \rfloor$ steps. We can consider the broadcast as a tournament of P players. If P is odd, a dummy processor can be added, whose scheduled opponent waits for the next round.

4.4. Detailed description of the phase 4

Every processor has been assigned some prefix-based equivalence classes. The sequential algorithm is run for every prefix-based equivalence class assigned to the processor, i.e. the processors must prepare the sequential algorithm for each processed prefix. For example, if we want to run the Eclat algorithm, we have to prepare the tidlists for every assigned prefix and the prefix extensions. In the rest of the section we describe how to run the Eclat algorithm on the assigned prefix-based equivalence classes.

At the start of this phase, processor p_i creates tidlists $\mathcal{T}(b_i), b_i \in B$. We denote the extensions of the prefix U_i by $\Sigma_i = \{(b_k, \mathcal{T}(b_k))\}$, where $b_k \in B$ is the extension and $\mathcal{T}(b_k)$ is its tidlist. The sequential algorithms reuses the datastructures used for the computation of the supports during the recursive depth-first search of the lattice. To make the parallel execution of a DFS algorithm fast, we need to cache the datastructures in the same way as done by a DFS algorithm, i.e. we simulates the execution of a DFS algorithm.

Let U_k be the prefixes assigned to processor p_i. The Eclat algorithm uses the tidlists for computation of supports. The cache of the tidlists is an array $\pi_{tidlists}$ of pairs $(item, \Sigma)$ at position i. The items $\pi_{tidlists}[j].item, j < i$ correspond to the prefix of the prefix-based equivalence class $[U_k]_i$. $\pi_{tidlists}[j].\Sigma$ at position j corresponds to the possible branches of a DFS algorithm for the prefix-based class $[U_k]_i$.

To prepare the tidlists efficiently for each assigned prefix $\Pi_j = (a_1^j, \ldots, a_{n_j}^j)$, we reuse the tidlists. That is, we construct the tidlists for each prefix of the first assigned prefix Π_1, i.e. we construct the tidlists for each itemset in $\{(a_1^1, \ldots, a_k^1) | k < n_1\}$ and store the tidlists of the itemset $(a_1^1, \ldots, a_k^1)$ in an array $\pi_{tidlists}$ at position k. After processing the prefix Π_j, we need to modify the array for the next processed prefix Π_{j+1}. For the prefix $\Pi_{j+1} = (a_1^{j+1}, \ldots, a_{n_{j+1}}^{j+1})$ we reuse m array elements such that $a_i^j = a_i^{j+1}, i \leq m$ and $a_{m+1}^j \neq b_{m+1}^{j+1}$.

To make the proccess more efficient, we sort the prefixes using the lexicographical order.

Algorithm 2 Prepare-Tidlists

PREPARE-TIDLISTS(**In/Out:** Tidlists $\pi_{tidlists}$, **In:** Prefix π)

```
n ← −1
for i = 0, ..., |π| − 1 do
    if π_tidlists[i].item ≠ π[i] then
        n ← i
        break
    end if
end for
for i = n, ..., |π| − 1 do
    π_tidlists[i] ← create new array element from π_tidlists[i − 1]
end for
for i = |π|, ..., |B| − 1 do
    π_tidlists[i] ← ∅
end for
```

Algorithm 3 Execution of the Eclat algorithm in the scheduled prefix based equivalence classes.

EXEC-ECLAT(**In:** Set of prefixes π, **In:** Database D)

```
sort π lexicographically by the prefix
π_prev ← 0
π_tidlists ← array of size |B| with π_tidlists[i] ← ∅
π_tidlists[0] ← (∅, {(b_i, T(b_i)) | b_i ∈ B})
for all p ∈ π do
    PREPARE-TIDLISTS(π_tidlists, p)
    run the Eclat algorithm with prepared tidlists π_tidlists[|p.Π|]
end for
```

The PREPARE-TIDLISTS algorithm summarizes the preparation of the tidlists for the sequential run of the Eclat algorithm, see Algorithm 2. The execution of the Eclat algorithm is summarized in Algorithms 3.

5. The parallel FIMI algorithms

We have described a method for mining of FIs that can be parametrized using some algorithms. As the algorithm for mining of MFIs we use the *fpmax** algorithm and as the algorithm for mining of FIs, we use the Eclat algorithm. Because we can execute the *fpmax** algorithm in parallel or sequentially, we have the two

PhD Conference 2009 - 14:32

following algorithms:

1. The PARALLEL-FIMI-SEQ algorithm computes the MFIs sequentially, for details see Algorithm 4.
2. The PARALLEL-FIMI-PAR algorithm computes the MFIs in parallel, see Algorithm 5.

Algorithm 4 PARALLEL-FIMI-SEQ

// *Phase 1*
create a random sample of its database part D_i.
broadcast its database sample $\mathcal{D}_{smpl}$. (all-to-all broadcast)
Compute approximation of MFIs.
sample $\mathcal{F}_{apprx}$
Phase 2: Processor p_1 samples $\mathcal{F}$ and divides $\mathcal{F}$ to prefix-based classes and uses $\mathcal{D}_{smpl}$ for estimation of the dynamic items ordering. Then using the LPT-MAKESPAN algorithm, p_1 joins the prefix-based classes into partitions $F_i, i = 1, \ldots, P$ such that $|F_i|/|\mathcal{F}| \approx 1/P \pm \epsilon$ (computed on p_1).
// *Phase 3*
Exchange database partitions and work assignment among all processors (a one-to-all broadcast followed by an all-to-all scatter)
// *Phase 4*
Compute support for every itemset in $\mathcal{F}$ (all processors in parallel)

Algorithm 5 PARALLEL-FIMI-PAR

Phase 1: Perform local sampling in database parts, collect database samples from all processors (all-to-all broadcast), partition the items in the base set as 1-prefixes among processors and compute in parallel approximations of MFIs with dynamic load-balancing (all processors in parallel).
Phase 2: Every processor computes its portion of samples of $\mathcal{F}$ and sends them to processor p_1. It divides F to prefix-based classes and uses D_S for estimation of the dynamic items ordering. Then using the LPT-MAKESPAN algorithm, p_1 joins the prefix-based classes into partitions $F_i, i = 1, \ldots, P$ such that $|F_i|/|\mathcal{F}| \approx 1/P \pm \epsilon$ (all-to-one gather followed by a sequential computation).
Phase 3: see the Parallel-FIMI-Seq algorithm.
Phase 4: see the Parallel-FIMI-Seq algorithm.

6. Evaluation of the speedup

We have evaluated the PARALLEL-FIMI-SEQ and PARALLEL-FIMI-PAR algorithms on a cluster of workstations interconnected with the Infiniband network. Every node in the cluster has two dual-core AMD Opteron processors at 2.6GHz with 8GB of main memory.

In Phases 1 of the PARALLEL-FIMI-PAR, we use the parallelization of the *fpmax** algorithm with dynamic load-balancing of computation of MFIs and in the Phases 1 of the PARALLEL-FIMI-SEQ, we use the sequential *fpmax** algorithm (run on processor p_1). The Phase 4 of the two methods is parametrized with the ECLAT algorithm.

We have used datasets generated by the IBM database generator [7] with 500k transactions and set the supports for each experiment such that the sequential run of the Eclat algorithm is between 700 and 12000 seconds. The IBM generator is parametrized by the average transaction length `TL` (in thousands), the number of items `I` (in thousands), by the number of patterns `P` used for creation of the parameters, and by the average length of the patterns `PL`. To clearly differentiate the parameters of a database we are using the string `T[number in thousands]I[items count in 1000]P[number]PL[number]TL[number]`, e.g. the string `T500I0.4P150PL40TL80` labels a database with 500000 transactions 400 items, 150 patterns of average length 40 and with average transaction length 80. All experiments were performed with various values of the support parameter on 2, 4, 6, and 10 processors. The databases used for evaluation of our algorithm is summarized in Table 6.

Dataset	Supports
T500I0.1P100PL20TL50	0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18
T500I0.4P250PL10TL120	0.2, 0.25, 0.26, 0.27, 0.3
T500I1P100PL20TL50	0.09, 0.07, 0.05

Table 1: Databases used for measuring of the support and the supports used for measuring.

The PARALLEL-FIMI-SEQ and the PARALLEL-FIMI-PAR method achieved speedup up to $\approx$ 6 with 10 processors.

Figures 1, 2 demonstrate that for reasonably large and reasonably structured datasets, the speedup is linear with speedup $\approx$ 6 on 10 processors. It follows from the graphs that the PARALLEL-FIMI-PAR is sometimes faster then the PARALLEL-FIMI-SEQ.

DRAFT 16. ICS PhD Conference 2009

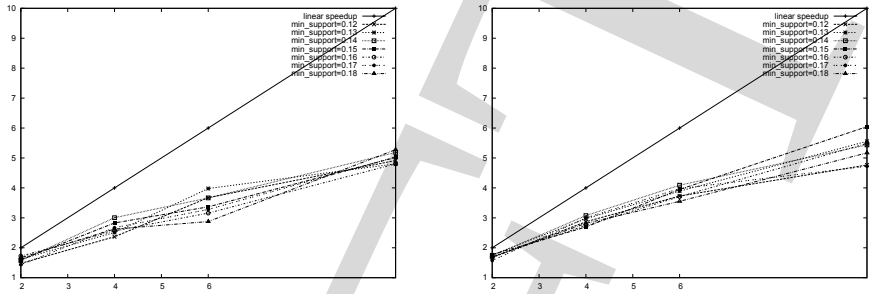


Figure 1: Speedups of the PARALLEL-FIMI-SEQ and PARALLEL-FIMI-PAR algorithms (from left to right) on the `T500I0.1P100PL20TL50` dataset.

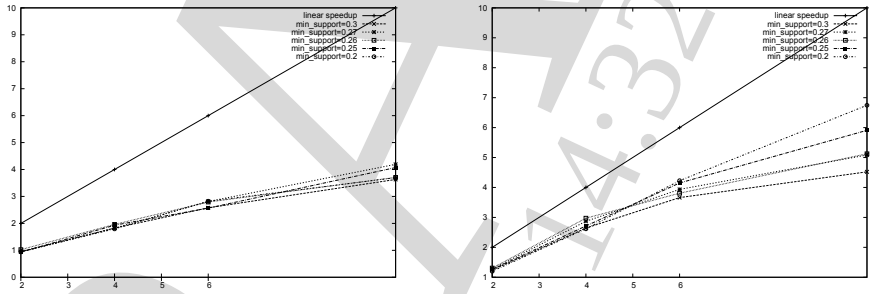


Figure 2: Speedups of the PARALLEL-FIMI-SEQ and PARALLEL-FIMI-PAR algorithms (from left to right)on the `T500I0.4P250PL10TL120` dataset.

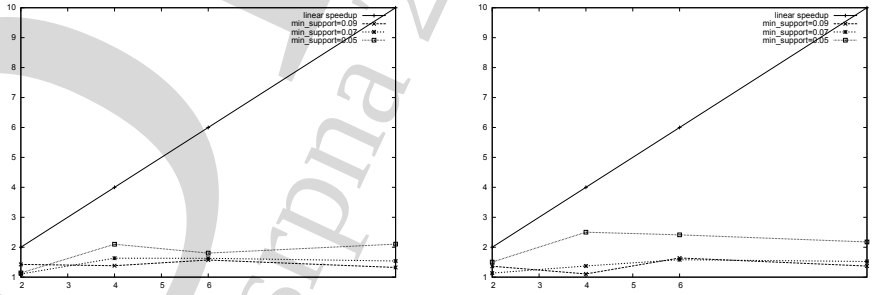


Figure 3: Speedups of the PARALLEL-FIMI-SEQ and PARALLEL-FIMI-PAR algorithms (from left to right) on the `T500I1P100PL20TL50` dataset.

DRAFT 14 srpna 2009 - 14:32

The complicated cases are the datasets with 1000 items in the Figure 3. Our hypothesis explaining the bad speedup is that the database contains MFIs m_i, m_j such that $|m_i \cap m_j|$ (i.e. $|\mathcal{P}(m_i) \cap \mathcal{P}(m_j)|$) is small and the number of MFIs is large. Therefore, the number of MFIs is very large and the set of frequent itemsets given by a particular MFI has very few common frequent itemsets with set of itemsets given by other MFIs. Since there is large number of sets with very small intersections and the number of sets is much larger than the number of samples, the sizes of the prefix-based classes cannot be well approximated. The nepresnost causes the zhorseni of the parallel speedup.

7. Acknowledgment

I would like to thank to Petr Savický for reading draft of this paper and for his help with formulations in it.

References

[1] T. Calders and B. Goethals, "Mining all non-derivable frequent itemsets," in *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Computer Science, pp. 74–85, Springer-Verlag, 2002.

[2] M.J. Zaki, "Scalable algorithms for association mining," *Knowledge and Data Engineering*, pp. 372–390, 2000.

[3] H. Toivonen, "Sampling large databases for association rules," in *In Proc. 1996 Int. Conf. Very Large Data Bases* (T. M. Vijayaraman, A. P. Buchmann, C. Mohan, and N. L. Sarda, eds.), pp. 134–145, Morgan Kaufman, 09 1996.

[4] R. Motwani and P. Raghavan, *Randomized algorithms*. Cambridge university press, 1995.

[5] G. Grahne and J. Zhu, "Efficiently using prefix-trees in mining frequent itemsets," in *FIMI*, 2003.

[6] R.L. Graham, "Bounds on multiprocessing timing anomalies.," *SIAM Journal of Applied Mathematics*, vol. 17, no. 2, pp. 416–429, 1969.

[7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499, Morgan Kaufmann, 1994.

DRAFT 14. srpna 2009 -- 14:32

Virtual Distributed Environment for Exchange of Medical Images

Post-Graduate Student:
MGR. TOMÁŠ KULHÁNEK
CESNET z.s.p.o.,
Zikova 4,
160 00 Praha 6
tomaton@centrum.cz

Supervisor:
ING. MILAN ŠÁREK, CSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
sarek@euromise.cz

Field of Study:
Biomedical Informatics

This work was supported by CESNET z.s.p.o. and grants MSM6383917201 and IS 1ET200300413.

Abstract

The exchange of medical images within a PACS system depends on high capacity of communication channels and high performance of computational resources. We introduce pilot project utilizing grid technology to distribute functionality of PACS system to several machines located in distant places which allows economizing utilization of network channels. We also discuss benefits and disadvantages of virtualization techniques allowing to separate physical machine capabilities from the operating system. We compare this pilot project utilizing high speed CESNET 2 network with similar mature projects based mainly on P2P secure connection, centralized system and proprietary protocols.

1. Introduction

The Digital Imaging and Communications in Medicine (DICOM) standard is widely used in medical devices and applications. PACS (Picture archiving and Communication Systems) systems to archive DICOM is currently used in information systems within hospitals and todays effort is focused on connecting the systems among hospitals. The additional security and authorization mechanism must be kept with respect of data privacy and safety as DICOM itself doesn't provide such features [1]. DICOM series represents also usually large amount of data, which has specific requirements of capacity of communication channels.

Dostal et al. [2] introduced the client-server message brokering system with a centrally located server cluster and client application on user computer, the MeDiMed project. It was primary used on national education network CESNET2; however other clients may connect via public Internet channels too. Client application may retrieve DICOM series from the client's local or institutional PACS and send it via proprietary protocol using SSL encryption to server. Client application may identify the receiver and may set some other metadata regarding the message. The receiver must have the same client application and get the DICOM series from the server later. This solution based on the central point of the system architecture may become a bottleneck or single point of failure. There are other commercial solution using SSL encryption and authentication which are based on establishing VPN connection between peer endpoints.

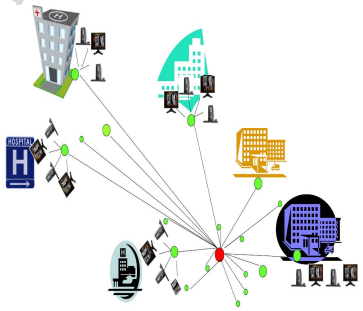


Figure 1: Centralized access to medical image exchange

Erberich et al. [3] utilize grid technology and open standards and protocols to process DICOM images securely in distributed environment to prevent some issues coming from VPN and proprietary protocols. They introduced project named Globus MEDICUS which integrates DICOM interface as a service of a grid infrastructure. Montagnat et al. [4] use similar approach in their Medical Data Manager which integrates grid middleware gLite with a DICOM interface providing strong security and encryption mechanism to preserve patient's privacy.

DRAFT 14. STRIKA 2009 - 15:55

Different systems and technologies have different requirements on hardware and software environment. Virtualization techniques allow providing separation between software and underlying hardware. However virtualization introduces some overhead when translating isolated application instruction to lower level of a system. Current virtualization techniques allow full operating system isolation. Youseff et al. [5] shows that XEN paravirtualization doesn't impose an onerous performance penalty comparing to non-virtualized OS configuration.

The grid middleware can be deployed to the virtual environment of paravirtualized system machines which are within the physical servers geographically spread throughout various institutions. We exchanged the DICOM series between the DICOM grid interface and the existing participant from the MeDiMed project infrastructure.

2. Methods

The Globus Medicus project [3] provides a DICOM grid interface service (DGIS), metacatalog service and storage service provider. Each of the service may be deployed independently on the grid middleware Globus Toolkit and form a grid based storage of PACS system. DGIS is able to communicate in DICOM standard and is a bridge to grid infrastructure which hides the fact that the data are processed throughout a grid from the client side to metacatalog service or storage service. Each service may run on independent host.

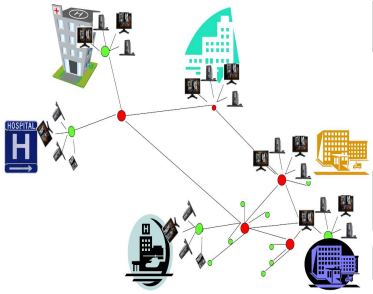


Figure 2: Grid based access to medical image exchange

The opensource XEN paravirtualization implementation adds a modification to the kernel of a guest system to be able to be executed and monitored by the host machine. Modification of the host system is, however, not required on hardware with virtualization support. We installed the services of Globus Medicus within the virtual grid nodes on the paravirtualized guest systems Centos 5.2 Linux, kernel version 2.6 which are hosted on 64-bit Intel XEON running XEN 3.0.3.

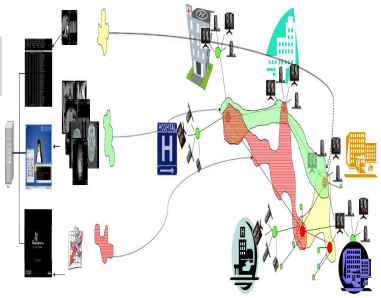


Figure 3: Virtual Grids for medical image exchange

DICOM standard uses two independent direct connection to the user's location to send the results of the user's request via IP protocol. The consequence is that the DGIS service must have access to the user's application or DICOM device via backward IP connection established independently. This communication channel is not secured by itself and the security task is up to higher level of network protocols. Thus the DGIS might be usually installed behind the institutional firewall. The DGIS connects to the other local or remote services of the grid infrastructure via HTTP and gridFTP protocol. The communication between nodes and services is by default secured by asymmetric encryption and x.509 certificates.

3. Results and Discussion

We deployed nodes of the pilot grid infrastructure into three pilot location: CESNET association, First Faculty of Medicine of Charles University and Central Military Hospital in Prague. All three locations are in Prague and are connected via high speed national educational and research network CESNET2 operated by CESNET association. We plan to use the pilot grid infrastructure also for another purpose and the XEN paravirtualization allows us to deploy and test another isolated projects next to this one. We configured the guest virtual machines to share the same IP connection with the host system. We configured the transport to virtual machine via network address translation (NAT) and we use Linux ipfiltering 'iptables' ruleset to forward incoming connection to the grid services.

Some institutions follow strict security policy, so they require the installation and execution of the grid services in demilitarized zone next to the institutional firewall with restricted access to local resources. With administrators of the institutional firewall we explicitly agreed and configured the firewall exception for the gridFTP protocol as the transport of such protocol uses TCP port number usually restricted by default.

We uploaded initial DICOM studies with about 1300 DICOM images for demonstration purposes. The DGIS must be configured to be able to communicate with the client application of MeDiMed project and accept the DICOM images exchanged in this project. We successfully exchanged and processed the DICOM studies and demonstrated that connection and DICOM studies exchange is possible between the MeDiMed project and the Globus Medicus. We used the client application of the MeDiMed project to retrieve and send selected DICOM series from the grid Globus MEDICUS to the participant connected in the MeDiMed project successfully and vice versa.

We had to solve problems that comes from usage of virtualization and relates with sharing one IP connection between multiple virtual machines on the same host. Explicit setting of the NAT and IP filtering rules must be done on each physical machine. On top of that the access to the functionality of the DGIS was restricted by some institutional policy and explicit exception had to be implemented on the institutional firewall. In contrast the client application of the MeDiMed project doesn't need such network configuration. The solutions based on VPN need similar effort on network configuration.

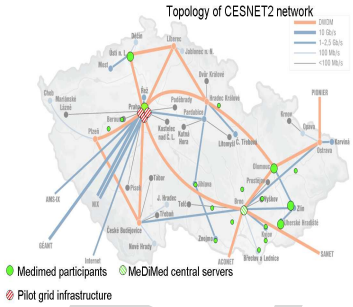


Figure 4: Grid nodes and MeDiMed participants in CESNET 2 network

The grid technology is able to serve medical image processing in secure and reliable way as well as current systems. The only unsecured communication is between DGIS and DICOM compliant client, which is same for other types of solution (MeDiMed or VPN based) and is not usually recognized as security issue if unsecured connections are within trusted local network.

The DICOM grid service interface behaves as another DICOM compliant device and the whole system with the utilizing grid services may be considered as another PACS system e.g. as a remote backup or an external PACS for exchanging e.g. educational DICOM studies. In contrast the MeDiMed client's application doesn't allow to be controlled via DICOM protocol thus cannot be accessed by institutional application and the proprietary MeDiMed client application must be used to process DICOM studies from MeDiMed project.

4. Conclusion

The MeDiMed project will have become to face with problems of scalability and single point of failure. The grid technology and virtualization might be an answer to such problem for future enhancement and development as it can benefit from live network topology and doesn't need to maintain virtual topology established by VPN based solution. The MeDiMed client uses the proprietary protocol to communicate with server in contrast to the pilot grid infrastructure which is based on open standards.

Virtualization techniques allow dynamic allocation and management of physical resources. On top of the pilot infrastructure the physical servers might be utilized to deploy another application or services. This benefit is currently considered by the other participated institutions.

References

[1] M. Sarek, "New Aspects of Pacs in Dwdm Network.", *World Congress on Medical Physics and Biomedical Engineering*, 417-420, 2006 (2007).

[2] O. Dostal, M. Javornik, and P. Ventruba, "Collaborative Environment Supporting Research and Education in The Area of Medical Image Information", *INTERNATIONALJOURNAL OF COMPUTER ASSISTED RADIOLOGY AND SURGERY 1*, 98, 2006.

[3] S. Erberich, J. Silverstein, A. Chervenak, R. Schuler, M. Nelson, and C. Kesselman, "Globus MEDICUS-Federation of DICOM Medical Imaging Devices into Healthcare Grids", *STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS 126*, 269, 2007.

[4] J. Montagnat, A. Frohner, D. Jouvenot, C. Pera, P. Kunszt, B. Koblitz, N. Santos, C. Loomis, R. Texier, D. Lingrand, P. Guio, R.B.D. Rocha, A.S. de Almeida, and Z. Farkas, "A Secure Grid Medical Data Manager Interfaced to The Glite Middleware", *J. Grid Comput. 6* , 45-59., 1 (2008).

[5] L. Youseff, R. Wolski, B. Gorda, and C. Krintz, "Paravirtualization for Hpc Systems. ", *Frontiers of High Performance Computing and Networking a ISPA 2006 Workshops* , 474-486, 2006.

14. 9. 2009 14:32

Clinical Contents Harmonization of EHRs and its Relation to Semantic Interoperability

Post-Graduate Student:
MGR. MIROSLAV NAGY
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
nagy@euromise.cz

Supervisor:
RNDR. ANTONÍN ŘÍHA, CSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
riha@euromise.cz

Field of Study:
Biomedical Informatics

This work was partially supported by the project of the Institute of Computer Science of Academy of Sciences of the Czech Republic AV0Z10300504 and by the project of Ministry of Education of Czech Republic No. 1M06014.

Abstract

This paper describes solutions proposals in the field of clinical content harmonization of electronic health records and semantic interoperability establishment. First the Czech national project "Information Technologies for the Development of Continuous Shared Healthcare" will be mentioned and its approach to creation of semantic interoperability platform. Afterwards an approach using openEHR architecture will be described. Finally a technique of creation of electronic health records with harmonized clinical content will be stated.

1. Introduction

Sharing and reusing the data among different institutions in the Czech healthcare environment is at relatively low level. The majority of healthcare information systems in the Czech Republic communicate with each other using a national communication standard called DASTA [1], which is based on the national nomenclature called National code-list of laboratory items (NCLP) [1]. These standards are developed and administered by the developers of healthcare information systems that are either specialized companies, university IT centers or research institutions in the Czech Republic. The development of the standard is supported by the Czech Ministry of Health. DASTA is specialized mainly in transfer of requests and results of laboratory analyses. The current version of DASTA is XML based and provides also the functionality for sending statistical reports to the Institute of Health Information and Statistics of the Czech Republic [2] and limited functionality of free text clinical information exchange. Unfortunately, the DASTA has almost no relation to international communication standards such as HL7 [3] or European standards like EN13606 [4].

The use of international standards such as HL7 v2, HL7 v3, EN13606 or DICOM [5] is induced mainly by the local requirements within healthcare institutions to communicate with modern instruments and modalities. Here, the major role is played by the HL7 version 2 and DICOM standards; however, it represents only a minor part of the overall communication within the Czech healthcare system.

Many papers, especially in last few years, deal with the problem how to establish semantic interoperability among various EHR systems [6], [7], [8]. As stated in [9] the semantic interoperability has 4 prerequisites. They are a standardized EHR reference model, standardized service interface models, a standardized set of domain-specific concept models and standardized terminologies. The problem of clinical content harmonization has similar objectives – unambiguous semantics of common information model connected to international nomenclatures and ontologies. We can say that having EHRs with harmonized clinical content (HCC) and message development process we could achieve semantic interoperability.

Achievement of semantic interoperability and especially an EHR with HCC in our work will be based on results of projects "Information Technologies for the Development of Continuous Shared Healthcare" (ITDCSH) [10], ARTEMIS [11], openEHR foundation [12].

DRAFT
Prague 2009

2. Materials

A national research project of the Academy of Sciences of the Czech Republic, ITDCSH had in its main goals the creation of an interoperability platform for structured healthcare data exchange, serving as a basis for lifelong healthcare support, which would be based on international communication standards. For this purpose the HL7 standard v3 was chosen from the set of HL7, DICOM [5], openEHR [12] and ENV 13606 [13].

This unique project (in the context of Czech healthcare environment) served primarily as a demonstration of possibilities, tasks and issues. It was not possible to cover the whole area of medicine as an interoperability domain. Our department has a long history of interdisciplinary research oriented on the field of cardiology. Therefore, the cardiology was chosen as the medical domain for the pilot realization of the semantic interoperability platform. A set of important medical concepts in the field of cardiology named Minimal Data Model of Cardiology (MDMC) [14] was prepared by the representatives of Czech Society of Cardiology and statisticians specialized in medical data processing. This set of concepts served as a basis for information models of two EHR systems used in our solution.

2.1. Terms and definitions

At this point it is advisable to summarize some terms of key importance for the rest of the text:

electronic health record (EHR) is defined as "a repository of information regarding the health status of a subject of care, in computer processable form" [9].

EHR system is defined as "a system for recording, retrieving, and manipulating information in electronic health records" [4].

clinical content is a part of set of concepts that underline the EHR and which refers to medical domain concepts such as physical examination, laboratory, medication, rather than demographic information, billing or bed management.

archetype according to [9] (from the technical point of view) is "a computable expression of a domain-level concept in the form of structured constraint statements, based on some reference information model".

semantic interoperability according to [15] is "the ability of information systems to exchange information on the basis of shared, pre-established and negotiated meanings of terms and expressions".

openEHR template is a directly, locally usable definition which composes archetypes into a larger structure logically corresponding to a screen form. Templates may add further local constraints on the archetypes it mentions, including removing or mandating optional sections, and may define default values.

HL7 template are used to apply constraints on R-MIMs (refined message information models) generated from generic reference information model.

HL7 v3 message is an instance of classes of R-MIM which are composed in a hierarchy defined by hierarchy definition model (HMD).

LIM template is a pattern defining tree-like structure of instances of LIM (Local Information Model) [16] classes. Each LIM template represents one integrated part of the EHR system the LIM describes, e.g. physical examination, medication and ECG data.

2.2. Possible content stored in an EHR

In the following text we put an example of concept groups that may appear in an EHR as described in [17].

1. A collection of concepts that together form fixed attributes of a higher level concept that is not recorded as its component parts alone - e.g.:
   - a blood pressure measurement with its two pressure measurements, patient position, cuff size etc.
   - a body weight with details about the baby's state of undress and the device used for measurement
2. A generic concept (with other fixed attributes) that is a value or a collection of values which form a subset of a larger (or very large) known set - e.g.:
   - a diagnosis - the value - with fixed attributes such as the date of onset, the stage of the disease etc
   - a laboratory battery result which includes an arbitrary set of values - the collection - with fixed attributes such as the time of sampling, or a challenge applied to the patient at the time the sample was taken (e.g. fasting).

DRAFT 14. 5. 2009 - 14:35

3. A collection of these higher level concepts that are usually measured together and might be considered themselves concepts - e.g.:

- Vital signs - with temperature, blood pressure, pulse and respiratory rate
- Physical examination - with for example observation, palpation and auscultation (and other findings)

4. A collection of these aggregations which might form a record composition or a document - e.g.:

- A clinic progress note containing symptoms, physical examination, an assessment and a plan
- A laboratory report that contains the results as well as interpretation and details about any notifications and referrals that have been made
- An operation report detailing the participants and their roles, a description of the operation, any complications and followup monitoring and care required

2.3. Archetypes

Archetypes play a key role in development of future proof EHR systems [18]. As defined in the section 2.1 archetypes are structured constraint statement based on some reference model (RM). In the paper [19] we can find example of archetype that represents "weight at birth" based on HL7 RIM as well as on openEHR reference model. The archetype binding to a RM is realized in archetype definition that is formalized in Archetype Definition Language (ADL) [20] particularly in the part called *definition*. Language that is used for this binding is called contraint ADL – cADL.

The openEHR foundation presents an application called Clinical Knowledge Manager accessible from their web site [21]. Its purpose is to concentrate archetypes in one repository in order to be reviewed by the community and create a repository of archetypes that could serve as a basis for development of EHRs with HCC.

In the Figure 1, the structure of archetype representing blood pressure concept is depicted. The part *data* contains values of the actual pressure, i.e. systolic, diastolic, mean arterial pressure, pulse pressure and textual comment on blood pressure reading. *State* is a list of information describing conditions of the measurement, e.g. the position of the patient at the time of measuring blood pressure. *History* covers separate measurements and adds temporal data in the implicit form, i.e. base measurement in the history, another reading after 5 minutes of rest, 10 minutes etc. Finally, the *protocol* holds technical data such as size of a sphygmomanometer's cuff if it is used or a specification of an instrument used to measure the blood pressure. For the sake of further computerized processing, archetypes are defined in ADL.

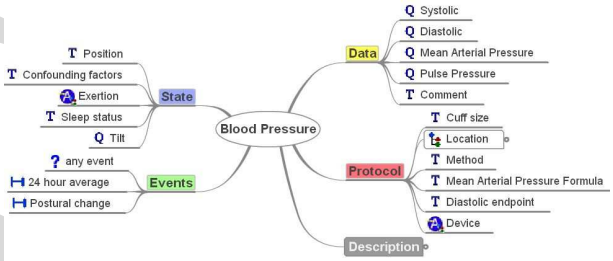


Figure 1: Blood pressure archetype example.

2.4. Communication standards

Communication among EHRs can be understood as data exchange in form of messages which have well defined syntax that is supported by all participants. This ensures so called *syntactical interoperability* where the structure and provenance of information or knowledge is understood by a clinical system – data are in machine readable form.

In Czech healthcare environment two kinds of communication could be recognised – passive and active. *Passive communication* is realized between healthcare institution and registries gathering data of patient with particular diagnosis (e.g. joint replacement, organ transplantation and oncology). *Active communication* is actively initiated by a request or query. Messages in active communication process have typically form of application forms, various documents (e.g. medical treatment summary), structured forms (e.g. laboratory results) etc.

Despite the long term effort in the field of communication standardization there still does not exist one universally accepted communication standard. There are two commonly used standards: HL7 v3 (international) and EN 13606 (European). The HL7 standard served as a basis for the solution described in section 4.1 and EN 13606 is indirectly connected with the proposal in the section 4.2 since it defines archetypes and templates for messaging as well as the reference model originating from openEHR.

2.5. HL7 v3 RIM

The Reference Information Model (RIM) [22] is used to express the information content for the collective work of the HL7 Working Group. It is the information

model that encompasses the HL7 domain of interest as a whole. The RIM is a coherent, shared information model that is the source for the data content of all HL7 messages. As such, it provides consistent data and concept reuse across multiple information structures, including messages.

3. Methods

3.1. Semantic interoperability and Semantically enriched Web Services

In order to achieve semantic interoperability of EHR information, there are four prerequisites, with the first two of these also being required for functional interoperability [9]:

- a standardized EHR reference model, i.e. the EHR information architecture, between the sender (or sharer) and receiver of the information,
- standardized service interface models to provide interoperability between the EHR service and other services such as demographics, terminology, access control and security services in a comprehensive clinical information system,
- a standardized set of domain-specific concept models, i.e. archetypes and templates for clinical, demographic, and other domain-specific concepts, and
- standardized terminologies which underpin the archetypes. Note that this does not mean that there needs to be a single standardized terminology for each health domain but rather, terminologies used should be associated with controlled vocabularies.

An elaborate work regarding semantic interoperability can be found in [6], where the development framework (not the implementation itself) for semantically interoperable health information systems is described. However, this paper will orient on the realization and validation of the interoperability platform.

Procedure of semantic interoperability achievement among EHR systems storing clinical information in various proprietary formats was studied in project ARTEMIS. The resulting solution contained an idea of wrapping and exposing the existing healthcare applications as Web Services [19]. The semantic interoperability was achieved by using OWL [23] (Web Ontology Language) mappings of archetypes based on reference models of, possibly, different standards (openEHR, HL7 RIM). These archetypes semantically enrich the Web Services messages. The interoperability was realized through a mediator that transformed the source message using mapping definitions into appropriate form to be accepted by the destination system and its Web Service.

3.2. Clinical content harmonization

EHR systems with harmonized clinical content are the most appropriate ones to achieve semantic interoperability. Their clinical content that refers the same domain is ready for exchange with minimum transformations and mappings during its delivery. However, the actual implementation of communication is out of scope of this area.

4. Results

This section presents results of various approaches to semantic interoperability approaches; one based on HL7 v3 messaging and the other one on openEHR architecture especially the construct called templates.

Some results in process of clinical content harmonization of EHR are shown. Particularly the clinical concepts mapping to coding systems (table 1), modeling these concepts using standardized reference model (here HL7 RIM; see figure 6) and finally finding archetypes that cover modeled concepts.

4.1. Semantic interoperability platform based on HL7 v3 messages

Primary result of the project ITDCSH is a proposal of semantic interoperability platform based on international communication standard, which is shown in Figure 2.

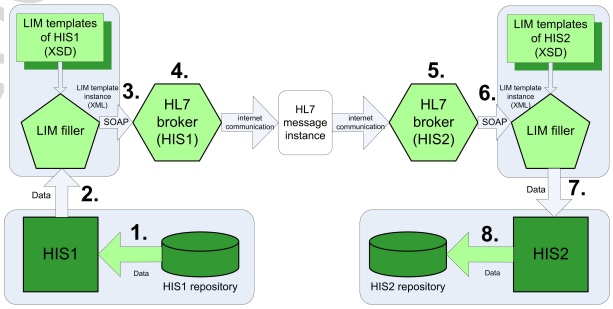


Figure 2: Proposal of semantic interoperability platform based on international communication standard.

The proposal consists of LIM filler module, HL7 broker and original HISes. Numbers in the Figure 2 represent the data flow in a situation when HIS1 sends data to HIS2. First of all, the requested data are gathered from HIS1. This is done by the LIM filler that has a

DRAFT Accepted 2009-11-32

connection to the HIS repository. Next, the LIM filler takes the suitable LIM template which contains the correct concepts to represent the communicated data. LIM filler adds data values to empty classes in the LIM template, thus creating a LIM message. HL7 broker receives the LIM message via the SOAP protocol used by the LIM filler module. Again, another transformation is performed; in this case the HL7 broker produces appropriate HL7 message instances, which are sent in a secure way to the receiving HL7 broker. Now the process of data transformation runs backwards. The HL7 broker attached to HIS2 creates LIM messages recognized by the LIM filler of HIS2 and sends them via SOAP. The receiving LIM filler recognizes the incoming LIM message and extracts the data into form suitable for storage in HIS2 repository. Finally, the data are stored in the HIS2. In this example we left out the requesting and confirmation mechanisms for simplicity reasons.

LIM fillers and HL7 brokers components were developed to support data transformations of a given HIS. Both components will be described in more detail in the following text.

4.1.1 LIM filler module: For reasons mentioned earlier, it is necessary to convert data from local EHR into LIM message, which is an instance of LIM template, on each side of communication. This task is performed by the module called LIM filler. LIM filler is adjusted for each local EHR to produce LIM messages according to local EHR structure.

LIM filler module can be EHR plug-in or standalone application, which takes the data from local EHR and fills them into LIM template. It works in two modes. In the first one, it creates the LIM messages on user's demand and sends them to the HL7 broker. In the second mode, it polls the HL7 broker for new messages. In case of new message it downloads it, extracts the data from the message and acts according to particular storyboard or just stores the data of the patient in the local EHR.

LIM filler must respect security aspects of the communication protocol. It communicates with the HL7 broker through the secured HTTP channel using SOAP protocol. The LIM messages must be digitally signed by both parties involved and the signatures must be checked before extracting the data from the LIM message.

4.1.2 HL7 broker: Fundamental part of the solution is a component called HL7 broker. The HL7 broker serves as a configurable communication interface to the "world of HL7" for each of EHR systems. The configuration is performed via an XML file containing the LIM model of a particular EHR and mapping of classes from this model to the actual HL7 messages. The configuration says how to convert data represented in the form of LIM message into the form of HL7 v3 messages. Only the prepared artifacts from current HL7 v3 ballot were used, no new HL7 messages were created.

Communication between EHR system and HL7 broker is implemented using Web Services (33) based on SOAP (34) over HTTPS protocol. The HL7 broker provides several methods (`sendLimMsg()`, `ackLimMsg()`, `getLimMsg()`) for transfer of the data between EHR system and HL7 broker. The data are transported in the form of a LIM message described by the LIM template. Several LIM templates are defined, e.g. administrative data, ECG or laboratory results. There are two communication modes - querying and passive one.

In the query mode the EHR system receives the LIM message (the query) from the HL7 broker. The query LIM message contains only several values assigned to concepts, which serve as parameters of the query. After the information is retrieved from the local database of EHR, it is sent back to the HL7 broker in the form of LIM message.

The passive mode is used to import the content of the LIM message (with all the required data) into the target EHR. Such data should be flagged as external - received using HL7 standard.

The result of a query in the EHR, initiated by the received query LIM message, could consist of several LIM messages according to the query specification. In this case the individual messages will be sent to the HL7 broker in sequence with the last message marked as the final one.

4.1.3 Implemented storyboard: The HL7 storyboards were used to implement querying mentioned above. For example we can mention a storyboard from the "Patient Administration" domain called "Patient Registry Find Candidates Query" (artifact code PRPA_ST201305) that was implemented in order to search for patient administrative data. UML sequence diagram representing activities performed according to the "patient query" storyboard with added HIS1 query and HIS2 responses is shown in Figure 3.

Queries that are produced by incorporated HISes and passed to HL7 broker are composed of empty LIM templates with only some attributes containing values, which are recognized as parameters of a query. Using wildcards like the * symbol is allowed in parameter values to denote arbitrariness. For example, when you search for patient with name "Wil*" you get all patients whose name starts with "Wil".

DATA Agenta 2009-11-32

Our solution enables composing queries to all domains covered by the LIM templates. As mentioned above, the query is done by using specific LIM template which is partially filled in. That means one

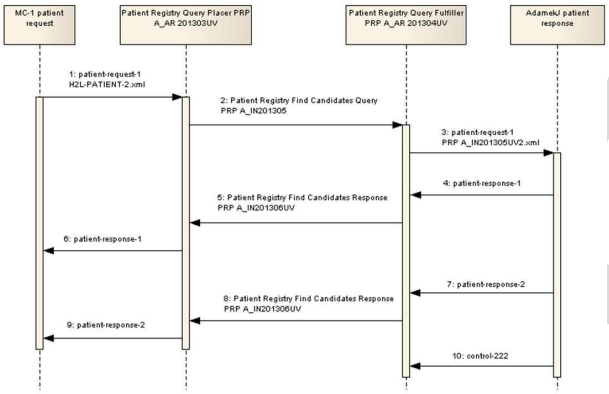


Figure 3: Sequence diagram of the Patient Registry Find Candidates Query.

can use the "Administrative information LIM template" to query administrative data of a patient or using the "Physical examination LIM template" to get data referring the physical examination of a patient. Finally, the HL7 broker executes the appropriate storyboard that leads to acquisition of data queried by the LIM filler module.

4.1.4 HL7 message instance: In order to consolidate reader's apprehension of HL7 messaging and queries described in previous text we put an example of "Patient Registry Find Candidates Response" (artifact code PRPA_IN201306) in the XML form. It holds data acquired after search for Mr. John Smith and is depicted in Figure 4.

```
<?xml version="1.0"?>
<hl7:PRPA_IN101306UV02 xmlns:hl7="urn:hl7-org:v3">
  <hl7:id root="48f5eb3d5d7399.80032341" extension=
    "48f5eb3d5d7462.58285204"/>
  <hl7:creationTime>151008150608</hl7:creationTime>
  <hl7:versionCode code="V3PR1"/>
  <hl7:interactionId root="1234.1234.1234" extension=
    "PRPA_IN101306UV02"/>
  <hl7:processingCode code="D"/>
  <hl7:processingModeCode code="T"/>
  <hl7:acceptAckCode code="AL"/>

  <hl7:Acknowledgement>
    <hl7:typeCode code="AA"/>
    <hl7:TargetMessage>
      <hl7:id root="48f5eb0f8d9773.81740782" extension=
        "48f5eb0f8d9887.41094918"/>
    </hl7:TargetMessage>
  </hl7:Acknowledgement>

  <hl7:ControlActProcess>
    <hl7:classCode code="CACT"/>
    <hl7:moodCode code="EVN"/>
    <hl7:priorityCode code="R" codeSystem=
        "2.16.840.1.113883.5.7" codeSystemName=
        "ActPriority"/>
    <hl7:QueryAck>
      <hl7:queryResponseCode code="OK" codesystem=
        "2.16.840.1.113883.5.1067" codeSystemName=
        "QueryResponse"/>
    </hl7:QueryAck>
    <hl7:Subject1>
      <hl7:typeCode code="SUBJ"/>
      <hl7:contextConductionInd value="true"/>
      <hl7:RegistrationEvent>
        <hl7:classCode code="REG"/>
        <hl7:moodCode code="EVN"/>
        <hl7:id root="2.16.840.1.113883.19.420.2"
            extension="cust1"/>
        <hl7:statusCode code="active"/>
        <hl7:Custodian>
          <hl7:typeCode code="CST"/>
          <hl7:contextControlCode code="AP"/>
        </hl7:Custodian>
        <hl7:Subject2>
          <hl7:typeCode code="SBJ"/>
          <hl7:IdentifiedPerson>
            <hl7:classCode code="IDENT"/>
            <hl7:id root="2.16.840.1.113883.19.420.1"
                extension="6501010001"/>
            <hl7:statusCode code="active"/>
            <hl7:Person>
              <hl7:classCode code="PSN"/>
              <hl7:determinerCode code="INSTANCE"/>
              <hl7:name>
                <hl7:prefix>Ing. CSc.</hl7:prefix>
                <hl7:prefix>Doc.</hl7:prefix>
                <hl7:given>John</hl7:given>
                <hl7:family>Smith</hl7:family>
              </hl7:name>
              <hl7:telecom use="MO" value="Tel:
                (+420) 377259020"/>
              <hl7:telecom use="H" value="Tel:
                (+420) 737151760"/>
            </hl7:Person>
            <hl7:Subject>
              <hl7:typeCode code="SBJ"/>
              <hl7:QueryMatchObservation>
                <hl7:classCode code="OBS"/>
                <hl7:moodCode code="EVN"/>
                <hl7:code code="V16847" codeSystem=
                    "2.16.840.1.113883.11.19723"/>
                <hl7:value>100</hl7:value>
              </hl7:QueryMatchObservation>
            </hl7:Subject>
          </hl7:IdentifiedPerson>
        </hl7:Subject2>
      </hl7:RegistrationEvent>
    </hl7:Subject1>
  </hl7:ControlActProcess>

  <hl7:Receiver>
    <hl7:typeCode code="RCV"/>
    <hl7:Device>
      <hl7:classCode code="DEV"/>
      <hl7:determinerCode code="INSTANCE"/>
      <hl7:id root="1.2.203.25666011.99.1.2"
        extension="UI-1"/>
    </hl7:Device>
  </hl7:Receiver>

  <hl7:Sender>
    <hl7:typeCode code="SND"/>
    <hl7:Device>
      <hl7:classCode code="DEV"/>
      <hl7:determinerCode code="INSTANCE"/>
      <hl7:id root="1.2.203.25666011.99.1.2"
        extension="MC-1"/>
    </hl7:Device>
  </hl7:Sender>
</hl7:PRPA_IN101306UV02>
```

Figure 4: Dump of communication according to storyboard PRPA_ST201305 - XML representation of Patient Registry Find Candidates Response.

DRAFT 14. srpna 2009 - 14:25

4.2. Semantic interoperability platform based on openEHR archetypes and templates

Archetypes are distinct, structured models of domain concepts, such as "blood pressure". They sit between lower layers of knowledge resources in a computing environment, such as clinical terminologies and ontologies, and actual data in production systems. Their primary purpose is to provide a reusable, interoperable way of managing generic data so that it conforms to particular structures and semantic constraints. Consequently, they bind terminology and ontology concepts to information model semantics, in order to make statements about what valid data structures look like. ADL provides a solid formalism for expressing, building and using these entities computationally. Every ADL archetype is written with respect to a particular information model, often known as a "reference model", if it is a shared, public specification.

Archetypes are applied to data via the use of *templates*, which are defined at a local level. Templates generally correspond closely to screen forms, and may be re-usable at a local or regional level. Templates do not introduce any new semantics to archetypes, they simply specify the use of particular archetypes, and default data values.

A third artifact which governs the functioning of archetypes and templates at runtime is a local palette, which specifies which natural language(s) and terminologies are in use in the locale. The use of a palette removes irrelevant languages and terminology bindings from archetypes, retaining only those relevant to actual use. Figure 5 illustrates the overall environment in which archetypes, templates, and a locale palette exist.

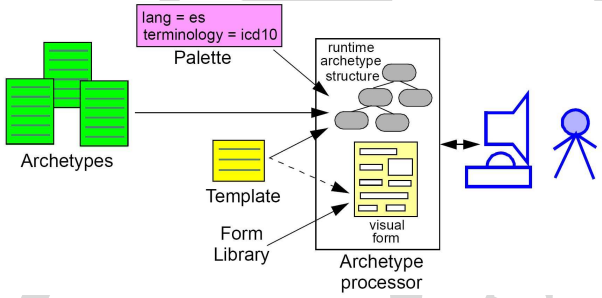


Figure 5: Archetypes, templates and palletes.

According to [24] templates include the following semantics:

- archetype 'chaining': choice of archetypes to make up a larger structure, specified via indicating identifiers of archetypes to fill slots in higher-level archetypes;
- local optionality: narrowing of some or all 0..1 constraints to either 1..1 (mandatory) or 0..0 (removal) according to local needs;
- tightened constraints: tightening of other constraints, including cardinality, value ranges, terminology value sets, and so on;
- default values: choice of default values for use in templated structure at runtime.

At runtime, templates are used with archetypes to create data and to control its modification.

The main advantages [25] of the openEHR approach are the functional and semantical interoperability. The functional interoperability represents the correct communication between two or more systems. This is also covered by other approaches like the HL7 v2.x. The openEHR approach also offers the semantic interoperability. It is the ability of two or more computer systems to exchange information which can be comprehended unambiguously by both, humans and computers.

4.3. Harmonizing clinical content of EHRs using international nomenclatures, openEHR architecture and HL7 v3 RIM

Concepts of clinical content of an EHR are usually "hiden" in object model, database schema or in meta-models developed during the information system creation. The process of enabling the creation of EHRs witch HCC has following steps:

1. *map clinical concepts* to an international coding system or ontology (SNOMED CT, LOINC, etc.)
2. *find archetypes* in a repository or knowledge base, that sufficiently cover encoded concepts
3. *underlying reference model* may be openEHR RM or HL7 v3 RIM thanks to OWLmt Mapping Engine [26] that is capable of transforming one to the other by using pre-defined mappings

In [17] the controlled and uncontrolled archetype development is described as well as techniques for ensuring maximal reusability of created constructs, curtailing their complexity and minimizing their number are proposed. Such practice would perfectly support the second step mentioned in previous enumeration.

The example of partial implementation of the step 1 is shown in the Table 1.

Description of encoded concept	Code	Coding system
Measurement of the breath frequency in one minute	9279-1	LOINC
Measurement of the heart beats in one minute	8893-0	LOINC
Measurement of blood temperature	8328-7	LOINC
Measurement of intravascular diastolic pressure	8462-4	LOINC
Amount of proteins in blood sample	2885-2	LOINC
Subjective complaints of the patient are described	10154-3	LOINC
Treatment of Ischemic Heart Disease	C0585894	UMLS CUI
Detection of Left ventricular hypertrophy	C0149721	UMLS CUI
Coughing after administration of ACE inhibitors	C0740723	UMLS CUI
Sequelae of cerebrovascular disease	I61	ICD10
Angina Pectoris	I20	ICD10
Hyperplasia of prostate	N40	ICD10

Table 1: Mapping concepts of MDMC to LOINC, UMLS and ICD-10 coding systems.

The step 2 was accomplished using the archetype repository [21] and some found archetypes are put down in Table 2 together with matching classes from the model partially depicted on Figure 6.

LIM class	Archetype ID
Subjective Complaints Description (Observation-cls)	openEHR-EHR-CLUSTER.issue.v1
Patient Height Measurement (Observation-cls)	openEHR-EHR-OBSERVATION.height.v1
Body temperature measurement (Observation-cls)	openEHR-EHR-OBSERVATION. body_temperature.v1
Heart rate measurement (Observation-cls)	openEHR-EHR-OBSERVATION. heart_rate-pulse.v1
Breath frequency measurement (Observation-cls)	openEHR-EHR-OBSERVATION. respiration.v1
Waist circumference measurement (Observation-cls)	openEHR-EHR-OBSERVATION.waist_hip.v1
Laboratory examination (Act-cls)	openEHR-EHR-OBSERVATION.lab_test.v1
Smoking state determination (Observation-cls)	openEHR-EHR-OBSERVATION. substance_use-tobacco.v1

Table 2: Some archetypes matching the concepts modeled in a reference model.

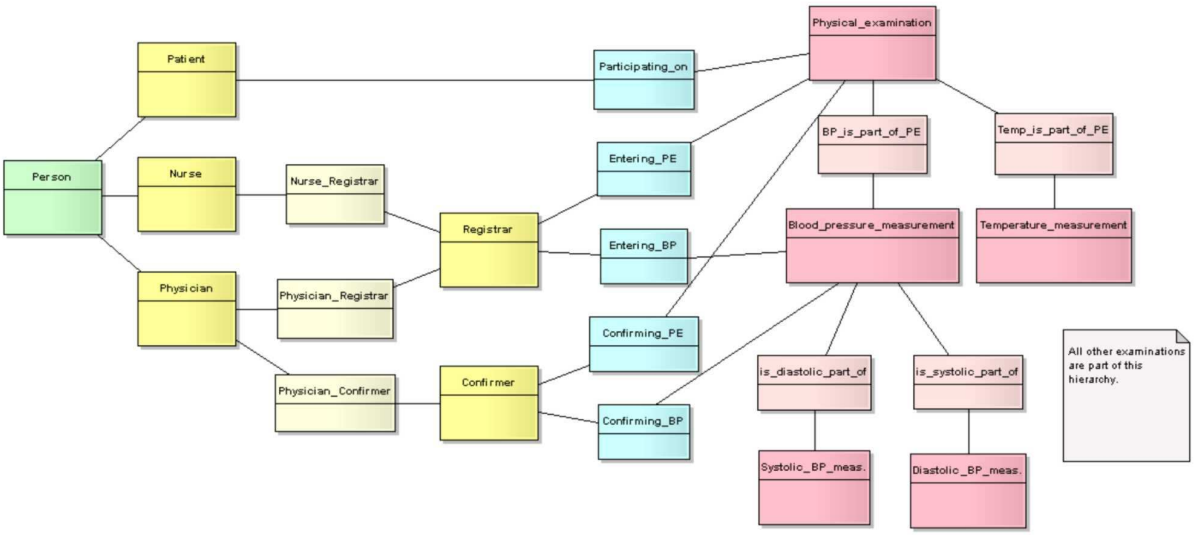


Figure 6: Selected part of HIS1 LIM.

5. Discussion

The development process of message interchange, recommended by HL7 v3 (see Figure 7), was altered by splitting the implementation effort between HIS developers and HL7 standard implementers. This new approach might help developers to overcome the initial frustration which is caused by the overwhelming size of the HL7 standard (RIM, amount of artifacts etc.). The development of individual LIM, closely related to internal information structure of the particular HIS, with the simple communication interface between HIS and HL7 broker based on LIM messages, SOAP and Web Services, seems to be more manageable for smaller developer teams than a strict adherence to HL7 v3 methodology.

Message interchange based on openEHR templates is a very interesting and relatively unexplored field, as the openEHR approach is primarily oriented to describe the development of future-proof EHR systems. This kind of messaging is based on a simple idea – instead of rendering the definitions contained in templates as screen forms, it is used to carry structured data

in a form of a message. Such concept is close to document interchange via HL7 CDA, but with major one difference – a higher degree of data structuring.

The communication via HL7 v3 messaging standard or openEHR templates is in real life result of huge effort of many people – domain experts, developers, medical stuff etc. Therefore, having EHRs with HCC available would reduce the complexity of communication frameworks and various translator and mapping modules. The data interchange would be much straightforward and that is worth studying rigorously.

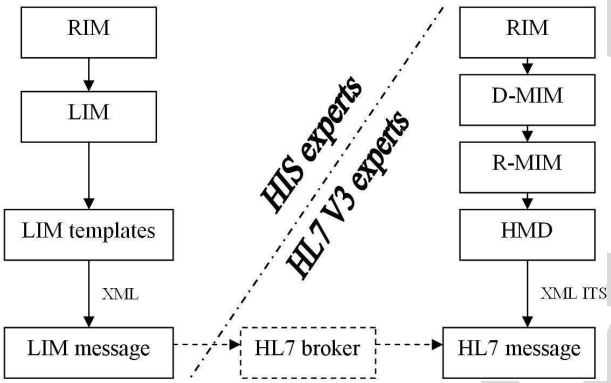


Figure 7: The messaging development process, recommended by HL7 v3 on the right and our solution on the left side.

6. Conclusion

During the development and implementation of the platform for semantic interoperability it was necessary to use the simulated patient data as the use of real patient data is not allowed for such a purpose due to legislative reasons. Results of performed tests were not affected by the fact that the data were simulated and are valid for real patient data as well.

Using LIM models and LIM fillers resulted in considerable universality of the solution, which does not depend on communication standards being used, although the LIM is based on HL7 v3 RIM. This independency is supported by the fact that contemporary modern communication standards have some important characteristics in common: basic reference model, user defined models derived from that reference model using strict methodology, and finally, some kind of templates helping in creating a new message or document. Comparison of contemporary communication standards can be found in [27].

The HL7 v3 implementation process was divided between HIS developers and HL7 implementers by utilization of LIM models. This approach resulted in better distribution of the experts' and developers' tasks.

The UMLS Knowledge Source Server was used to find the appropriate mappings of MDMC concepts to international nomenclatures and evaluate the applicability of international nomenclatures in the Czech medical terminology. During the analysis, we found that approximately 85% of MDMC concepts are included in at least one classification system. We managed to map most of MDMC concepts to LOINC and more than 50% are included in SNOMED Clinical Terms [28]. During the mapping we had to cope with some problematic concepts with too small or too big granularity, concepts with different synonyms differing slightly in their meaning or concepts which cannot be found in any available classification system [29].

After evaluation of the outcomes of the project ITDCSH we can say that the HL7 v3 is usable in a restricted form in the Czech healthcare environment. It has no support by the governmental institutions and only a limited support by the software vendors. The main step for wider use of HL7 v3 in the Czech Republic should be the implementation of functionality, which is currently provided by the DASTA national standard, the inclusion of NCLP on the list of HL7-supported code systems, or better the mapping of the NCLP to an established international nomenclature like SNOMED CT. The next fundamental step would be obtaining the translation of the international nomenclature in the Czech language.

References

[1] Ministry of Health of the Czech Republic (homepage on the internet), Data Standard of MH CR - DASTA and NCLP. http://ciselniky.dasta.mzcr.cz.

[2] Institute of Health Information and Statistics of the Czech Republic (homepage on the internet). http://www.uzis.cz.

[3] Health Level Seven, Inc. (homepage on the internet) Health Level 7. http://www.hl7.org.

[4] European Committee for Standardization (CEN), Technical Committee CEN/TC 251: European Standard EN 13606, "Health informatics - Electronic health record communication"

[5] NEMA - Medical Imaging & Technology Alliance (homepage on the internet), DICOM. http://dicom.nema.org.

[6] D.M. Lopez and G.M.E. Blobel, "A development framework form semantic interoperable health

14. ročník 2009 - 1432

information systems". Int. J. of Medical Informatics 2009; 78:83-103.

[7] B.G. Blobel, K. Engel, and P. Pharow, "Semantic interoperability - HL7 Version 3 compared to advanced architecture standards". Methods Inf Med. 2006; 45(4):343-53.

[8] D. Kalra and B.G. Blobel, "Semantic interoperability of EHR systems". Stud Health Technol Inform. 2007;127:231-45.

[9] Technical report. ISO/TR 20514 – Health informatics – Electronic health record – Definition, scope, and context. ISO. 2005.

[10] EurMISE.org. The project of the "Information Society" programme. http://www.euromise.org/research/news.html.

[11] Software R&D Center, Middle East Technical University. Artemis Project Homepage. http://www.srdc.metu.edu.tr/webpage/projects/artemis/home.html.

[12] openEHR (homepage on the internet), openEHR – future-proof and flexible EHR specifications. http://www.openehr.org.

[13] European Committee for Standardization (CEN), Technical Committee CEN/TC 251: European Standard ENV 13606-1, "Health informatics - Electronic healthcare record communication"

[14] M. Tomeckova et al., "Minimal data model of cardiological patient". (in Czech). Cor et Vasa 2002; 4: 123.

[15] K.H. Veltman, "Syntactic and Semantic Interoperability: New Approaches to Knowledge and the Semantic Web". The New Review of Information Networking 2001; 7: 159-84.

[16] M. Nagy et al., "Applied Information Technologies for Development of Continuous Shared Health Care". CESNET08 Conference: security, middleware, virtualization; CESNET,z.s.p.o; 2008. p. 131-38.

[17] S. Heard et al., "Templates and Archetypes: how do we know what we are talking about?" http://www.openehr.org/publications/archetypes/templates_and_archetypes_heard_et_al.pdf

[18] T. Beale, "Archetypes: Constraint-based domain models for future-proof information systems". In: Baclawski K, Kilov H, editors. Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer. Northeastern University, Boston, 2002, pp. 16-32.

[19] V. Bicer et al., "Archetype-Based Semantic Interoperability of Web Service Messages in the Health Care Domain". Int'l Journal on Semantic Web & Information Systems. 2005; 1(4): 1-22.

[20] T. Beale and S. Heard, "Archetype Definition Language (ADL)". The openEHR foundation. Rev. 1.3.1, 2004.

[21] openEHR foundation. Clinical Knowledge Manager. http://www.openehr.org/knowledge/

[22] HL7 Inc. (homepage on the internet), HL7 Version 3 - January 2009. http://www.hl7.org/v3ballot/html/welcome/environment.

[23] Ed.: D.L. McGuinness and F. van Harmelen, "OWL Web Ontology Language". W3C Recommendation. 2004. http://www.w3.org/TR/owl-features.

[24] T. Beale and S. Heard, "The Template Object Model (TOM)". http://www.openehr.org/releases/1.0.1/architecture/am/tom.pdf, 2007.

[25] M. Goek, "Introducing an openEHR-Based Electronic Health Record System in a Hospital". Masters Thesis. Department of Medical Informatics, University of Goettingen, Germany. 2008.

[26] Artemis projectt. OWLmt – OWL mapping tool. 2005. http://www.srdc.metu.edu.tr/artemis/owlmt.

[27] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G.B. Laleci, "A Survey and Analysis of Electronic Healthcare Record Standards". ACM Comp Surv, 2005 Dec; 37(4): 277 - 315.

[28] College of American Pathologists (homepage on the internet), SNOMED Terminology Solutions. http://www.cap.org/apps/cap.portal?_nfpb=true&_pageLabel=snomed_page.

[29] P. Hanzlicek, P. Preckova, and J. Zvarova, "Semantic Interoperability in the Structured Electronic Health Record". Ercim News 2007, 69:52-3.

DRAFT 14. 5. 2009 - 14:32

Preference Handling in Relational Query Languages

Post-Graduate Student:
RADIM NEDBAL
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ ,

Department of Mathematics
Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Trojanova 13

120 00 Prague 2, CZ

radned@seznam.cz

Supervisor:
ING. JÚLIUS ŠTULLER, CSC.

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2

182 07 Prague 8, CZ

stuller@cs.cas.cz

Field of Study:
Mathematical Engineering

This work was supported by the project 1ET100300419 of the Program Information Society (of the Thematic Program II of the National Research Program of the Czech Republic) “Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realization", and by the Institutional Research Plan AV0Z10300504 “Computer Science for the Information Society: Models, Algorithms, Applications".

Abstract

The paper outlines an approach to preference handling in relational query languages. The approach is based on the assumption that the information on possible outcomes is represented in the relational data model.

1. Introduction

Being one of the basic paradigms of human decision making, preferences are inherently a multi-disciplinary topic, of interest to philosophers, psychologists, political scientists, economists, mathematicians and other people coming from different human-centered disciplines, but facing similar questions. Recently, preferences have been studied in operations research, game theory, and several other areas related to computer science.

The main added value computer science has brought into the research on user preferences is an attempt to automate the whole process of preference handling. The goal of such automation is to make logical and mathematical foundations usable in systems that act on behalf of users or simply support their decisions. These could be (a) decision-support systems dealing with the situation where both the number of choice alternatives is huge, and no professional analyst is available to help a user, e.g., information search and retrieval engines that attempt to provide users with the most preferred pieces of information or web-based recommender systems such as shopping sites that attempt to help users identify the most preferred items, (b) automated problem solvers such as configurators, (c) sophisticated autonomous systems such as personal assistants, robots (e.g., Mars rovers), etc. Consequently, the preference handling has become a flourishing topic in many fields related to computer science (see Fig. 1 on the following page) such as database systems, electronic commerce, human-computer interaction, and numerous areas of artificial inteligence dealing with “choice situations”, e.g., knowledge representation, planning and scheduling, configuration and design, multiagent systems, algorithmic decision theory, computational social choice, and other tasks concerning intelligent decision support or autonomous decision making. In brief, preference-based systems allow finer-grained control over decision making automation and new ways of interactivity, and therefore provide more satisfactory results. In particular, explicit preference modeling provides a declarative way to choose among alternatives, whether these are answers to database queries, solutions of problems to solve, decisions of an autonomous agent, plans of a robot, and so on. Moreover, preference models may provide a clean understanding, analysis, and validation of heuristic knowledge used in existing systems such as heuristic orderings, dominance rules, heuristic rules, etc.

2. Preference Handling Meta-Model

The meta-model of preference handling provides a conceptualization consisting of six basic concepts capturing the most important aspects of preference handling:

DRAFT
Prague 2009-11-12

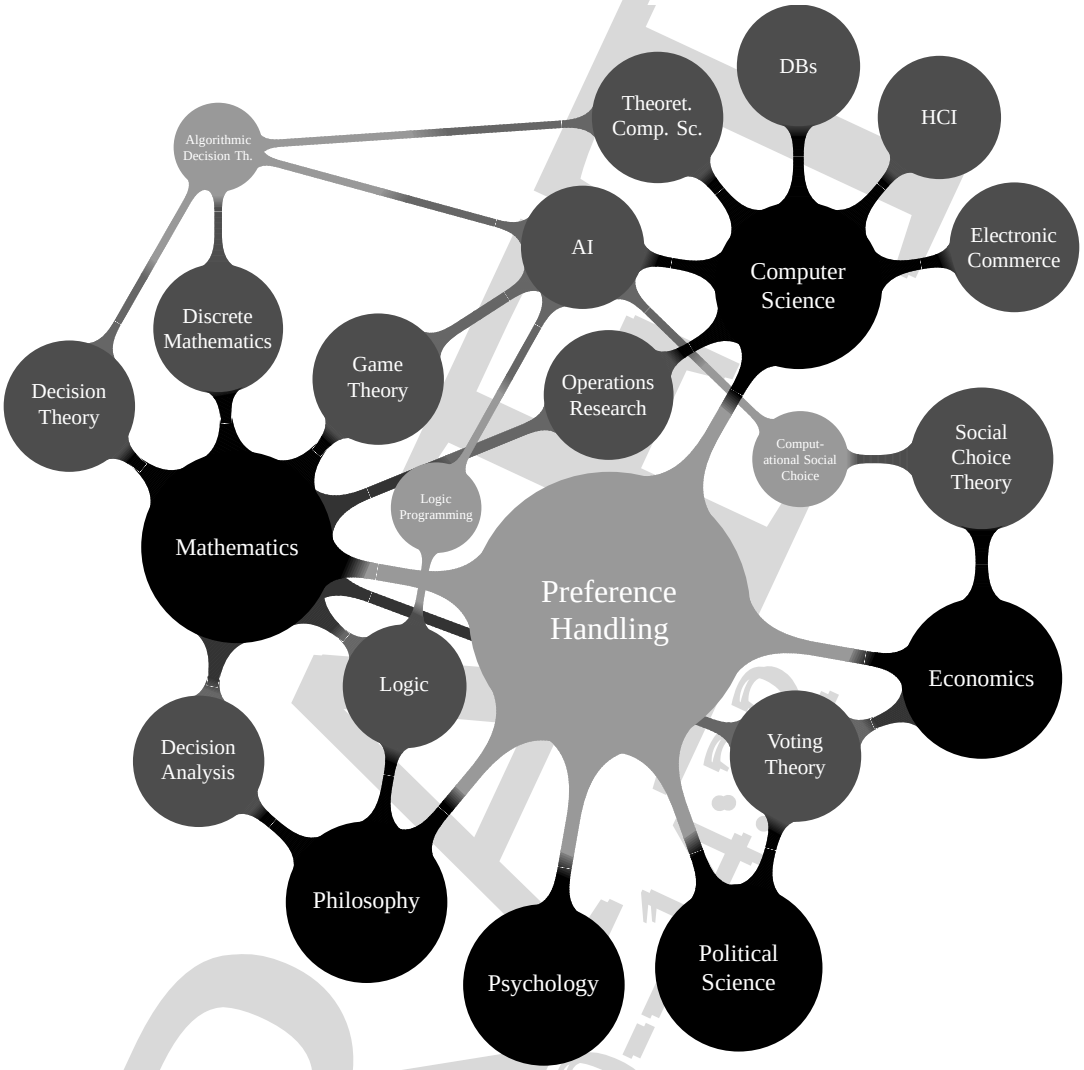


Figure 1: Preference handling mindmap

1. *Preference model* – a suitable mathematical (algebraical) structure that captures properties of specified preferences. (It is the structure we really care about.)
2. *Language* to specify models (ideally in an intuitive, concise manner).
3. *Interpretation* to give the exact meaning to language expressions. (It provides the mapping of the language expressions into a preference model.)
4. *Representation* to capture language expressions in a framework suitable for efficient query-answering algorithms.
5. *Queries* – questions about the models (the questions of interest).
6. *Algorithms* to evaluate answers to queries.

These concepts are depicted and interconnected graphically in Fig. 2 on the next page (adapted from [1]), in which the semantics of directed edges is "choice dependence", and the dashed directed edges picture the interpretation mapping language expressions to preference models and to instances of representation structure.

To explain the "choice dependance," note the two key questions that arise when modeling preference handling: What is the model? What queries do we want to ask about this model? Once we have a model and queries, we need algorithms to evaluate these queries about the model. However, algorithms for handling queries about preferences are typically tailored down to the specifics of the representation structure, which captures the language expressions specifying the model. The choice

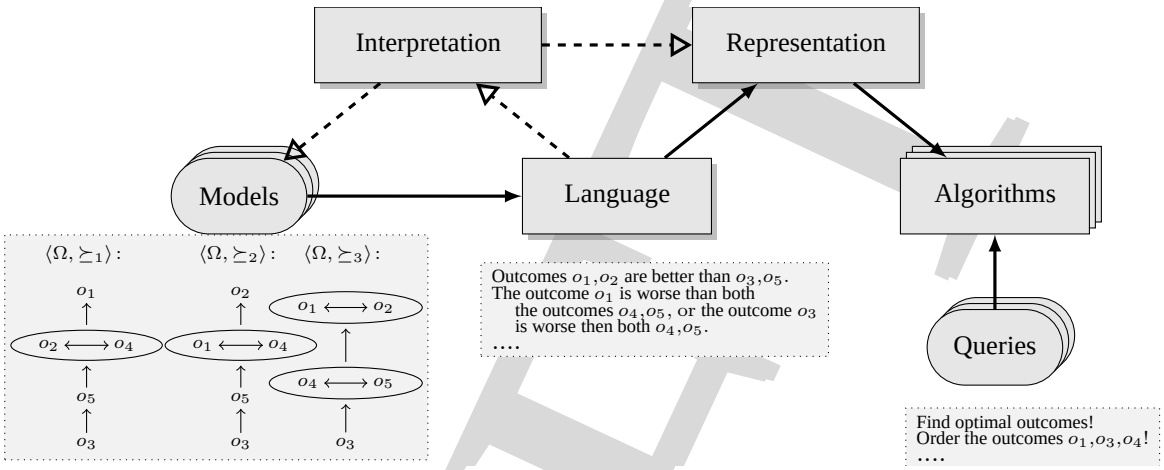


Figure 2: The meta-model of preference handling

of a language, in turn, depends on the assumptions about the preference models.

Observe that the language, its interpretation, and representation are closely related because an interpretation gives a meaning to expressions in a given language, which can be possibly compactly represented. However, a compact representation is possible only when our preferences can be communicated to the system at hand in terms of concise expressions of the language.

3. The Goal, the Objective, Addressed Questions, and Targeted Activities

Our **goal** is to embed the concept of preference into relational query languages (RQLs).

Accordingly, the **objective** is to provide database users with a *language* that:

1. can express *heterogenous preferences* in an easy *declarative* manner,
2. *compactly* specifies the *preference model*,
3. is based on information that is
   (a) *cognitively easy to express and reflect upon* and
   (b) reasonably *easy to interpret*,
4. has *intuitive*, well defined *semantics* allowing for *conflicting preferences*,
5. allows *representation* that supports *efficient* query-answering *algorithms* for finding optimal matches with respect to *preference models*.

Primarily, the following **questions** have to be **addressed**:

I. Haw can all the capabilities of such a language be embedded into RQLs?
   A) What are the suitable *algebraic operators*[1]?
   B) What are the *algebraic properties* of such operators to lay foundation for algebraic optimization of database queries?
II. What *kinds of preferences* can be expressed by such a language?
III. How can *semantics*
   A) of possibly conflicting preferences be *defined*?
   B) be *computed* effectively?

Consequently, the following **activities** have also to be **targeted** to bring the results into a practice:

⋆ Development of efficient *algorithms for evaluating new algebraic operators*.
⋆ Proposal and analysis of novel *optimization strategies* and their integration with the existing ones.

All these steps are necessary to make the notion of preference a practical concept in RQLs.

4. The Proposed Preference Handling Meta-Model and its Key Concepts

4.1. Models

In general, preferences are expressed over a particular set W of possible worlds. In the relational data model (RDM) context, a possible world can be viewed as a tuple over a finite set $\boldsymbol{A}$ of attributes. Consequently, W can be abstracted to the Cartesian product of the domains of attributes from $\boldsymbol{A}$.

[1] We base ourselves on the algebraic paradigm

DRAFT 14. srpna 2009 - 14:35

We propose to define the *preference model* as a single *preference relation* $\langle W, \succeq \rangle$ – a *partial pre-order* $\succeq$ over the set W of possible worlds (outcomes). In fact, the partial pre-order is introduced in order to capture possible conflicts in preferences in terms of incomparability among worlds.

4.2. Language

As the quantitative type of information is usually cognitively difficult to express and reflect upon, we propose to introduce a declarative language that is based on the qualitative type of information. That is to say, we suggest applying the *qualitative approach* to preference handling. Moreover, the language should enable an easy way to express *various kinds of preferences*.

To lift the propositional approach developed by [2] to the first-order case required by the RDM context, we propose to substitute propositional formulae in the language by *first order queries*. Accordingly, a user preference will be expressed by an appropriate *preference formula* of the form $\varphi \rhd \psi$, where φ, ψ are first order queries and $\rhd$ denotes a distinct kind of preference. These preference formulae constitute a simple declarative language that allows to capture complex, heterogenous preferences.

4.3. Interpretation

Interpretation of preferences (soft requirements) over a set W of possible worlds depends both on the information and mandatory requirements we have on W. This dependence is captured in terms of the so-called *forcing relation*, which represents relationships between individual possible worlds and preference formulae. Thus forcing relation is a parameter of interpretation, which ultimately is formalized by means of the *interpretation function* $\mathscr{I}(x, y)$ of two variables: x for forcing relation and y for a set of preference formulae.

We propose interpretation under ceteris paribus semantics in the sense of "all other things being *similar*", as formalized by [2] in terms of contextual equivalence relation. Moreover, we base ourselves on [3]'s proposal of a minimal logic of preference, in which *any* set of preferences is interpreted in a consistent way. We extend their approach so that *any* set of (possibly heterogenous) preferences, i.e., any set of preference formulae of our proposed language, can be represented by a first-order theory that is satisfiable.

In general, a set of preference formulae has no unique preference model under the proposed interpretation. Therefore, it is necessary to apply non-monotonic reasoning (NMR) mechanisms to identify the *distinguished models* with desired properties. Specifically, we suggest that the distinguished models are those that are maximal with respect to the set inclusion of the preference relation.

4.4. Representation

We want to prove that each set of preference formulae is logically equivalent to a set of disjunctive logic programs (DLPs) that are isomorphic: these DLPs are identical up to a renaming of constants from their Herbrand universes. Most importantly, it can be shown that the cardinality of these Herbrand universes is bounded by a function exponential in the cardinality of the set of preference formulae.

As isomorphic first order formulae have isomorphic models [4], it can be proved that a set of preference formulae is logically equivalent to a set of preference models, each of which is isomorphic to a particular model of a single DLP. Finally, these models are to be used to determine the most preferred possible worlds.

4.5. Queries and Algorithms

The most fundamental type of queries over preference models with the view of embedding the notion of preference in the RQLs is to find the most preferred matches with respect to user preferences.

It can be shown that the proposed distinguished model semantics (refer to Subsect. 4.3) and minimal model semantics of DLP agree. Consequently, the machinery of logic programming can be employed to compute the suggested declarative semantics of a set of preference formulae.

The overall concretization of the meta-model of the proposed approach to preference handling in the database context is depicted in Fig. 3 on the next page.

5. Embedding Preference into Relational Query Languages

5.1. Preference Operator

To filter out bad tuples, database users express a *selection condition*, which is embedded by a *selection* operator of the relational algebra (RA). This selection operator is parameterized by a logical condition that serves as a *hard constraint*. The user gets a perfect match if it is fulfilled. However, not every wish can become true.

DRAFT 14. mar 2009 - 14:32

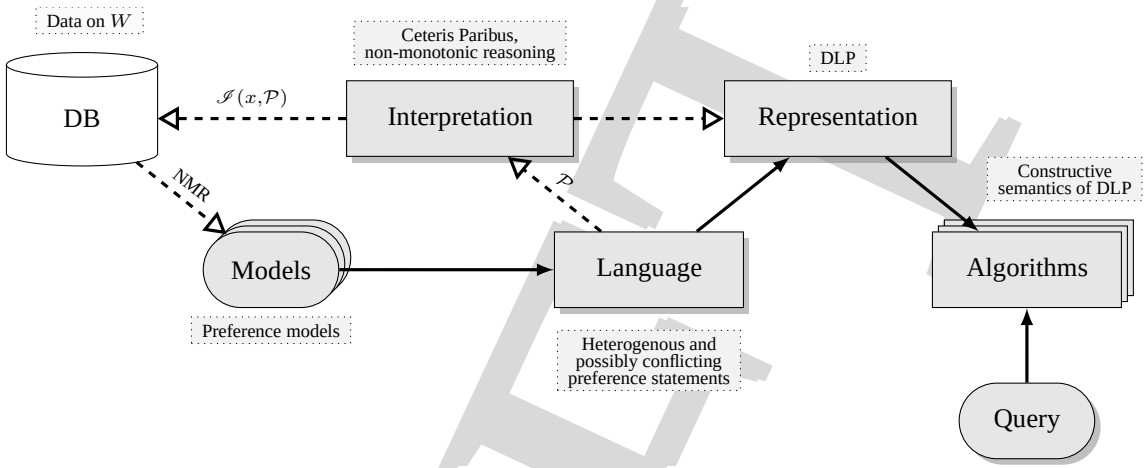


Figure 3: The meta-model of the proposed approach

To filter out not all the bad tuples, but only worse tuples than the best matching alternatives, we will introduce a new, *preference operator*, parameterized by user preferences. It selects from its argument relation the most preferred tuples according to its parameter – a set of preference formulae.

5.2. Algebraic Optimization

In general, the algebraic optimization aims at minimizing the data flow during the query execution. Basically, it utilizes various optimization strategies such as pushing *selection* and *projection* operators down the *query execution tree*. These strategies, in turn, are based on the assumption that early application of the selection or projection operator reduces intermediate results. As input relations are usually too big to fit into *main memory*, using the number of the *secondary storage I/O's* as our measure of cost for an operation, it is easily seen that this reduction of intermediate results has a remarkable positive impact on the performance of query processing.

To provide a formal foundation for algebraic optimization, the focus should be on abstract properties of the preference operator. These abstract properties include algebraic rules that describe the interaction of the preference operator with other RA operators. By considering the preference operator on its own, we should be able, on one hand, to focus on the abstract properties of user preferences and, on the other hand, to study special evaluation and optimization techniques for the preference operator itself.

We propose a new, analogical optimization strategy of *pushing the preference operator* down the query execution tree. Most importantly, sufficient conditions under which the preference operator commutes with selection or projection, or can be distributed over *cartesian product* or *union* are identified.

6. Related Work – Preference in Database Systems

The study of preference in the context of database queries has been originated by [5]. They, however, don't deal with algebraic optimization. Following their work, *preference datalog* was introduced in [6], where it was shown that the concept of preference provides a modular and declarative means for formulating optimization and relaxation queries in deductive databases.

Nevertheless, only at the turn of the millennium this area has attracted broader interest again. [7, 8, 9, 10] and [11, 12, 13, 14] pursued independently a similar (*qualitative*) approach within which preferences between tuples are specified directly, using binary *preference relations*. The embedding into RQL they have used is similar to ours: they have defined an operator returning only the best preference matches. In particular, they provided rewriting rules for the operator to lay foundation for algebraic optimization of database queries with preferences. Their optimization framework extends established query optimization techniques: preference queries can be evaluated by extended – preference RA. While some transformation laws for queries with preferences have been presented in [15, 16], the results presented in [11] are mostly more general.

A special case of the same embedding represents *skyline operator* introduced by [17]. Some examples of possible rewritings for skyline queries were given, but no general rewriting rules were formulated.

DRAFT 14. April 2009, 14:32

Building on the recent advances in logic of preference, [18] suggested a framework within which preferences between tuples are specified indirectly, using a declarative language based on the qualitative type of information. His language captures various kinds of preferences and allows for comfortable specification of preferences. The embedding of the concept of preference into RQLs is similar to that of [7] and [11]: it is realized by means of the *preference operator* returning only the best preference matches. By contrast, the best preference matches, in general, are sets of tuples. Basing himself on this framework, [19] aims at algebraic optimization of RQLs with preferences. In particular, he identifies the algebraic properties governing the interaction of the preference operator with the other operators RA. However, the semantics of the preference operator is unnatural in the sense that it is not based on the *closed world assumption* (CWA) – an implicit hypothesis standardly used in the realm of database systems.[2]

[20] addressed the issue of extending the RDM to incorporate partial orderings into data domains. Partially ordered data domains, in turn, are the leitmotiv of the approach to preference queries over web repositories [21]. Also in [22], actual values of an arbitrary attribute domain are allowed to be partially ordered according to user preferences. Accordingly, RA operations, aggregation functions and arithmetic are redefined. However, some of their properties are lost, and the query optimization issues are not discussed. Finally, [23] proposed a data structure for an effective representation of information representable by a partial order.

A comprehensive work on partial order in databases, presenting the partially ordered sets as the basic construct for modeling data, is [24]. Other contributions aim at exploiting linear order inherent in many kinds of data, e.g., time series: in the context of statistical applications systems SEQUIN [25], SRQL [26], Aquery [27, 28]. Various kinds of ordering on power-domains have also been considered in the context of modeling incomplete information: an extensive and general study is provided in [29].

By contrast, preference is specified indirectly using *scoring functions* within the *quantitative* approach [30, 31, 32, 33, 34, 35, 36]. A scoring function associates a numeric score with every tuple.

[2] CWA basically states that all the facts not in the database are false.

7. Conclusions

We propose a framework for embedding preferences into RQLs. The framework relaxes assumptions that are inherent in traditional approaches to preference handling in the database systems. Specifically, various kinds of preferences are taken into account. Most importantly, the proposed approach ensures that any set of user preferences (preference specification) specified in our language can be interpreted in a consistent way. Another distinctive feature of the framework is the utilization of logic programming machinery to efficiently compute preference models. Building on recent leading ideas that have contributed to remarkable advances in the field, the framework also deals with the optimization of relational queries:

- Preferences are embedded into relational query languages by means of a single preference operator returning only the best tuples in the sense of user preferences.
- An optimization strategy is based on the assumption that early application of a selective operator reduces intermediate results and thus reduces data flow during the query execution.

Consequently, we propose "pushing the preference operator strategy", which is based on its algebraic properties.

References

[1] R. I. Brafman and C. Domshlak, "Preference handling – an introductory tutorial," Tech. Rep. 08-04, Computer Science Department, Ben-Gurion University, Negev Beer-Sheva, Israel 84105, December 2007.

[2] J. Doyle and M. P. Wellman, "Representing preferences as ceteris paribus comparatives," in *Decision-Theoretic Planning: Papers from the 1994 Spring AAAI Symposium*, pp. 69–75, AAAI Press, Menlo Park, California, 1994.

[3] G. Boella and L. W. N. van der Torre, "A non-monotonic logic for specifying and querying preferences.," in *IJCAI* (L. P. Kaelbling and A. Saffiotti, eds.), pp. 1549–1550, Professional Book Center, 2005.

[4] V. Švejdar, *Logika: neúplnost, složitost a nutnost (Logic: Incompleteness, Complexity, and Necessity)*. Praha: Academia, 2002. In Czech.

DRAFT 14. SIGMA 2009 – 1432

464 pages. With a section on Gödel-Dummett logic written by Petr Hájek.

[5] M. Lacroix and P. Lavency, "Preferences; Putting More Knowledge into Queries.," in *VLDB* (P. M. Stocker, W. Kent, and P. Hammersley, eds.), pp. 217–225, Morgan Kaufmann, 1987.

[6] K. Govindarajan, B. Jayaraman, and S. Mantha, "Preference datalog," Tech. Rep. 95-50, 1, 1995.

[7] W. Kießling, "Foundations of Preferences in Database Systems," in *Proceedings of the 28th VLDB Conference*, (Hong Kong, China), pp. 311–322, 2002.

[8] W. Kießling, "Preference constructors for deeply personalized database queries," Tech. Rep. 2004-07, Institute of Computer Science, University of Augsburg, March 2004.

[9] W. Kießling, "Optimization of Relational Preference Queries," in *Conferences in Research and Practice in Information Technology* (H. Williams and G. Dobbie, eds.), vol. 39, (University of Newcastle, Newcastle, Australia), Australian Computer Society, 2005.

[10] W. Kießling, "Preference Queries with SV-Semantics.," in *COMAD* (J. Haritsa and T. Vijayaraman, eds.), pp. 15–26, Computer Society of India, 2005.

[11] J. Chomicki, "Preference Formulas in Relational Queries," *ACM Trans. Database Syst.*, vol. 28, no. 4, pp. 427–466, 2003.

[12] J. Chomicki, "Semantic optimization of preference queries.," in *CDB* (B. Kuijpers and P. Z. Revesz, eds.), vol. 3074 of *Lecture Notes in Computer Science*, pp. 133–148, Springer, 2004.

[13] J. Chomicki and J. Song, "Monotonic and nonmonotonic preference revision," 2005.

[14] J. Chomicki, S. Staworko, and J. Marcinkowsk, "Preference-driven querying of inconsistent relational databases," in *Proc. International Workshop on Inconsistency and Incompleteness in Databases*, (Munich, Germany), March 2006.

[15] W. Kießling and B. Hafenrichter, "Algebraic optimization of relational preference queries," Tech. Rep. 2003-01, Institute of Computer Science, University of Augsburg, February 2003.

[16] B. Hafenrichter and W. Kießling, "Optimization of relational preference queries," in *CRPIT '39: Proceedings of the sixteenth Australasian conference on Database technologies*, (Darlinghurst, Australia), pp. 175–184, Australian Computer Society, Inc., 2005.

[17] S. Börzsönyi, D. Kossmann, and K. Stocker, "The skyline operator," in *Proceedings of the 17th International Conference on Data Engineering*, (Washington, DC, USA), pp. 421–430, IEEE Computer Society, 2001.

[18] R. Nedbal, "Non-monotonic reasoning with various kinds of preferences in the relational data model framework," in *ITAT 2007, Information Technologies – Applications and Theory* (P. Vojtáš, ed.), pp. 15–21, PONT, September 2007.

[19] R. Nedbal, "Algebraic optimization of relational queries with various kinds of preferences," in *SOFSEM* (V. Geffert, J. Karhumäki, A. Bertoni, B. Preneel, P. Návrat, and M. Bieliková, eds.), vol. 4910 of *Lecture Notes in Computer Science*, pp. 388–399, Springer, 2008.

[20] W. Ng, "An Extension of the Relational Data Model to Incorporate Ordered Domains," *ACM Transactions on Database Systems*, vol. 26, pp. 344–383, September 2001.

[21] S. Raghavan and H. Garcia-Molina, "Complex queries over web repositories," tech. rep., Stanford University, February 2003.

[22] R. Nedbal, "Relational Databases with Ordered Relations," *Logic Journal of the IGPL*, vol. 13, no. 5, pp. 587–597, 2005.

[23] R. Nedbal, "Model of preferences for the relational data model," in *Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realisation* (J. Štuller and Z. Linková, eds.), (Prague), pp. 70–77, Institute of Computer Science Academy of Sciences of the Czech Republic, October 2006.

[24] D. R. Raymond, *Partial-order databases*. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1996. Adviser-W. M. Tompa.

[25] P. Seshadri, M. Livny, and R. Ramakrishnan, "The design and implementation of a sequence database system," in *VLDB '96: Proceedings of the 22th International Conference on Very Large Data Bases*, (San Francisco, CA, USA), pp. 99–110, Morgan Kaufmann Publishers Inc., 1996.

[26] R. Ramakrishnan, D. Donjerkovic, A. Ranganathan, K. S. Beyer, and M. Krishnaprasad, "Srql: Sorted relational query language," in *SSDBM '98: Proceedings of the 10th International Conference on Scientific and Statistical Database Management*, (Washington, DC, USA), pp. 84–95, IEEE Computer Society, 1998.

[27] A. Lerner, *Querying Ordered Databases with AQuery*. PhD thesis, ENST-Paris, France, 2003.

DRAFT 14 April 2009 - 14:52

[28] A. Lerner and D. Shasha, "Aquery: Query language for ordered data, optimization techniques, and experiments," in *29th International Conference on Very Large Data Bases (VLDB'03)*, (Berlin, Germany), pp. 345–356, Morgan Kaufmann Publishers, September 2003.

[29] L. Libkin, *Aspects of partial information in databases*. PhD thesis, University of Pensylvania, Philadelphia, PA, USA, 1995.

[30] R. Agrawal and E. Wimmers, "A Framework for Expressing and Combining Preferences.," in *SIGMOD Conference* (W. Chen, J. F. Naughton, and P. A. Bernstein, eds.), pp. 297–306, ACM, 2000.

[31] A. Eckhardt, "Methods for finding best answer with different user preferences," Master's thesis, 2006. In Czech.

[32] A. Eckhardt and P. Vojtáš, "User preferences and searching in web resoursec," in *Znalosti 2007, Proceedings of the 6th annual conference*, pp. 179–190, Faculty of Electrical Engineering and Computer Science, VŠB-TU Ostrava, 2007. In Czech.

[33] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," in *Symposium on Principles of Database Systems*, 2001.

[34] R. Fagin and E. L. Wimmers, "A formula for incorporating weights into scoring rules," *Theor. Comput. Sci.*, vol. 239, no. 2, pp. 309–338, 2000.

[35] P. Gurský, R. Lencses, and P. Vojtáš, "Algorithms for user dependent integration of ranked distributed information," in *Proceedings of TED Conference on e-Government (TCGOV 2005)* (M. Böhlen, J. Gamper, W. Polasek, and M. Wimmer, eds.), pp. 123–130, March 2005.

[36] S. Y. Jung, J.-H. Hong, and T.-S. Kim, "A statistical model for user preference," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, pp. 834–843, June 2005.

DRAFT
14. srpna 2009 -- 14:32

Databáze biomedicínských informačních zdrojů

doktorand:
MUDR. VENDULA PAPÍKOVÁ
Oddělení medicínské informatiky
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
papikova@euromise.cz

školitel:
DOC. PHDR. RUDOLF VLASÁK
Ústav informačních studií a knihovnictví
Filozofická fakulta Univerzity Karlovy
U Kříže 8
158 00 Praha 5
rudolf.vlasak@ff.cuni.cz

obor studia:
Informační věda

Práce byla částečně podpořena výzkumným záměrem AV0Z10300504 a projektem 1M06014 MŠMT ČR.

Abstrakt

Práce popisuje spektrum biomedicínských informačních zdrojů a předkládá návrh jejich klasifikace, která bere v úvahu fakt, že množství a rozmanitost zdrojů informací se v biomedicínských oborech velmi rozšířily. Podle tradiční typologie databází používané v rámci informační a knihovní vědy lze informační zdroje třídit do čtyř základních skupin: zdroje bibliografické; plnotextové; faktografické a zdroje typu registu, katalogu, adresáře. Mnohdy se také setkáváme se zdroji, které jsou kombinací výše uvedených typů (zdroje hybridní). S přibývajícím množstvím informací v medicíně však dochází nejen k rozšiřování nabídky databází v rámci výše uvedených kategorií, ale také k vytváření zcela nových informačních zdrojů, které nelze zařadit do žádné z výše uvedených skupin. S ohledem na jejich zaměření a způsob vzniku je lze nazvat jako „prospektivně-exploratorní“ (generativní) a postpublikačně evaluované informační zdroje. Kromě zdrojů textových informací navíc nabývají na významu sbírky obrazových záznamů, zvukových nahrávek a videozáznamů (zdroje multimediální), stejně jako kolekce několika databází (zdroje agregované). Vedle těchto kategorií zdrojů, které jsou založeny na informacích pocházejících z výzkumu a z vědecké literatury („*evidence-based*“), lze v rámci medicíny stanovit ještě další skupinu zdrojů založených na informacích získaných z praxe monitorováním událostí („*event-based*“), nezbytných například pro oblast farmakovigilance či epidemiologického dohledu. Klasifikace informačních zdrojů popsaná v rámci této práce byla použita pro sestavení databáze biomedicínských informačních zdrojů.

1. Úvod

Rozhodneme-li se studovat problematiku vědeckých informací, prakticky jistě se hned na počátku setkáme s pojmem „informační exploze“ nebo s jiným obdobným vyjádřením označujícím fakt, že množství publikovaných informací přesáhlo lidskou kapacitu pojmout je a zpracovat přirozeným způsobem. Medicína pochopitelně není výjimkou. Naopak, v medicíně je fenomén informační exploze ve srovnání s jinými obory umocněn dynamickým rozvojem jednak v oblasti genomického výzkumu, jednak na poli klinického výzkumu. Technologie genových čipů umožňují provádět rozsáhlé studie, jejichž výsledkem jsou ohromná množství dat. Nutnost pečlivě prověřit bezpečnost a účinnost veškerých léčebných postupů před jejich zavedením do rutinní medicínské praxe je zase důvodem pro jejich pečlivé klinické testování. Nárůst objemu nových informací je pak zřetelný i během krátkého období. Například pouze za rok 2004 stoupl počet databází v oblasti molekulární biologie a genomiky ze 171 na více než 700[1] [2]. Počet publikací klinických studií uložených jen v databázi MEDLINE/PubMed ve stejném roce převyšoval 30 000, přičemž toto číslo se pro každý další rok zvyšuje[2].

Z hlediska zpracování informací a jejich přeměny v nové poznání tak vzniká mnoho výzev. Je nutné budovat nová, specializovaná úložiště pro data, informace i poznatky. Je nutné stále znovu a lépe řešit otázku, jak informace efektivně vyhledávat. Relativně novým a pro medicínu naléhavým úkolem je otázka, jak velké množství publikovaných článků účinně zpracovat tak, aby byly využitelné jednak pro další bádání, jednak pro aplikaci v klinické praxi. Rovněž průběžné sledování nových vědecko-výzkumných poznatků je stále obtížnějším úkolem, a to i přes vysokou míru specializace, někdy až dokonce tzv. atomizace medicíny.

[1] Uvedený údaj navíc zahrnuje pouze volně dostupné on-line zdroje.
[2] 2004: 31 806 záznamů; 2005: 35 511 záznamů; 2006: 36 101 záznamů; 2007: 38 105 záznamů. (zdroj dat: www.pubmed.gov)

DRAFT 4. srpna 2009, 23:32

2. Cíl práce

Cílem této práce bylo vytvořit ucelenou klasifikaci informačních zdrojů pro biomedicínské obory, která by reflektovala výše uvedený vývoj, a na základě této klasifikace sestavit databázi biomedicínských informačních zdrojů.

3. Metodika

Jak bylo zmíněno v úvodu, množství a spektrum zdrojů vědeckých informací se v biomedicínských oborech velmi rozšířilo. Jako základ pro jejich klasifikaci byla využita **typologie databází** používaná v rámci informační a knihovní vědy, která rozlišuje čtyři základní kategorie: *zdroje bibliografické; zdroje plnotextové; zdroje faktografické a zdroje typu registru, katalogu, adresáře* [9]. Mnohdy se také setkáváme se zdroji, které jsou kombinací výše uvedených typů (*zdroje hybridní*). Vznikají rovněž zdroje zcela nové, pro medicínu mnohdy specifické, které nelze bez výhrad zařadit do žádné z uvedených kategorií. Pro tyto zdroje byly vytvořeny nové kategorie (*postpublikačně evaluované informační zdroje a „prospektivně-exploratorní“ databáze*). Kromě textových informací nabývají na významu také databáze obrazových a zvukových záznamů (*zdroje multimediální*), stejně jako kolekce několika databází (*zdroje agregované*). Zvláštní skupinu tvoří informační zdroje pro sledování událostí významných z hlediska epidemiologického dohledu a farmakovigilance (*zdroje monitorovací*). Charakteristiky jednotlivých typů informačních zdrojů jsou popsány v následujícím textu.

Pro **správu obsahu** databáze biomedicínských informačních zdrojů byl zvolen redakční a publikační systém Blogger[3]. Záznamy jsou vkládány do databáze se stručným popisem, relevantními webovými odkazy a jsou označené štítkem podle typu informačního zdroje. Jednotlivé zdroje jsou vyhledávány jak ve vědecké literatuře, tak na volném internetu. Postupně jsou doplňovány jednak stávající zdroje, jednak jsou průběžně přidávány nově vznikající zdroje.

3.1. *Zdroje bibliografické*

Do této skupiny informačních zdrojů patří databáze, jejichž datovou základnu tvoří bibliografické informace, vymezené obsahově, typem popisovaných zdrojů nebo jejich lokací. Slouží především k vyhledávání bibliografických informací; mohou být propojeny i se systémem dodávání původních dokumentů (DDS) [10] nebo mohou být vybaveny webovými odkazy na plné texty dokumentů volně dostupných na internetu. V některých případech obsahují současně také souhrny nebo abstrakty článků (tzv. referátové databáze). Obsah bibliografických databází je uložen v jednotně strukturovaných bibliografických záznamech, umožňujících vyhledávání podle hodnoty obsažených položek. Pravidla popisu i jeho podrobnost se mohou v různých databázích lišit. Základní typy bibliografických databází v současné době představují elektronické katalogy knihoven a archivů, oborové databáze zpřístupňované databázovými centry a seznamy zdrojů internetu [10].

3.2. *Zdroje plnotextové*

Plnotextové (fulltextové) informační zdroje jsou textové databáze, jejichž datovou základnu tvoří plné texty dokumentů [10]. V případě plnotextových bází dat je tedy k dispozici kompletní text primárního dokumentu již v přímé dialogové komunikaci [9]. Obvykle se jako plnotextová označuje databáze umožňující plnotextové vyhledávání podle textových řetězců za pomoci invertovaného souboru [10]. Při vyhledávání v plnotextových bázích dat je vhodné použití speciálních vyhledávacích prostředků a nástrojů (proximitní operátory), neboť jinak mohou výsledky vyhledávání obsahovat velké procento nerelevantních výsledků [9]. Plnotextové zdroje zahrnují časopisy, sborníky z konferencí, knihy, webové stránky, dizertační práce, výzkumné zprávy, patenty, návody a učební texty.

Z hlediska plnotextových informačních zdrojů zaměřených na medicínskou praxi jsou významné nové typy dokumentů, které lze vyhledávat ve specializovaných **databázích pro podporu klinického rozhodování**, jako jsou *systematické přehledy*[4], *kriticky posouzená témata* (Critically Appraised

[3] www.blogger.com

[4] **Systematické přehledy** jsou strukturované literární přehledy, které řeší otázky pomocí analýzy důkazů. Vyžadují objektivní způsoby vyhledávání informací, kritické posouzení relevantní literatury, aplikaci předem stanovených kritérií pro začlenění jednotlivých studií do systematického přehledu, extrakci dat z vybraných dokumentů a jejich sloučení do finálního dokumentu [3]. Systematické přehledy často bývají doplněny statistickým zpracováním dat (tzv. metaanalýzou).

[5] **Kriticky posouzená témata (CAT)** jsou dokumenty vznikající primárně pro účely studijní, obvykle na základě reálné klinické situace lékaře hledajícího odpověď na otázku, která se týká onemocnění konkrétního pacienta. CAT jsou strukturované, v původním pojetí jednostránkové souhrny výsledků vyhledávání a kritického posouzení důkazů k dané problematice [12].

[6] **Klinická doporučení** jsou systematicky vyvíjená oficiální vyjádření pro jednu či více klinických situací, jejichž úkolem je pomáhat lékařům a pacientům v rozhodování o patřičné zdravotní péči. Tyto dokumenty mohou být nahlíženy jako určitý typ HTA (viz níže) nebo naopak mohou z HTA vycházet [3].

Topics, CAT)[5], *klinická doporučení* (Clinical Practice Guidelines, CPG)[6], *hodnocení zdravotnických technik* (Health Technology Assessments, HTA)[7] a *ekonomické analýzy*[8].

Další významnou skupinu informačních zdrojů, které mají často plnotextový charakter, tvoří **zdroje medicínských informací pro pacienty**. Vytváření těchto zdrojů napomáhá v realizaci konceptu sdíleného klinického rozhodování („shared clinical decision making“)[9] a potažmo (ve spojení s medicínou založenou na důkazech, „evidence-based medicine“)[10] k poskytování zdravotnické péče, která je v souladu s principy medicíny založené na hodnotách („value(s)-based medicine“)[11].

3.3. *Zdroje faktografické*

Údajovou základnu faktografických databází tvoří faktografické informace [10]. Faktografické databáze uvádějí konkrétní údaje a mohou mít numerický, textový nebo kombinovaný charakter. Není nutné dodávat primární pramen, neboť jde v podstatě již o primární informaci. Některé faktografické systémy však mohou odkazovat na další literaturu a mít bibliografickou součást. Za jistých okolností je možné zahrnout do této kategorie i většinu statistických informací [9].

V případě numerických databází převažují v datové základně číselná vyjádření parametrů různých předmětů a jevů (např. ceníky, kurzovní lístky, kalendária, jízdní a letové řády, matematické, fyzikální, chemické aj. tabulky, výsledky laboratorních a vědeckých měření) nebo ukazatele různých vývojových procesů (např. statistiky, časové řady) [10].

Faktografické databáze jsou tradičně velmi rozšířené v chemii. Nicméně jejich význam narůstá [9] a jsou stále častější také v jiných oborech. V medicíně patří k typickým databázím faktografického charakteru nejen databáze **chemické a toxikologické**, ale především databáze **lékové a epidemiologické**.

Za faktografické zdroje informací lze považovat také zcela specifickou skupinu databází vznikajících jako výsledek výzkumu na poli genomiky, proteomiky a bioinformatiky, které bývají označovány jako **databáze molekulárně-biologické**. Tyto databáze tvoří skupinu informačních zdrojů, které jsou základem pro výzkum v oblasti genomické medicíny. Mezníkem v rozvoji genomiky, proteomiky a potažmo databází shromažďujících poznatky z těchto vědních disciplín byl rok 2001, kdy byla publikována pracovní verze kompletní sekvence lidského genomu [7], [13]. Bylo tak zveřejněno ohromné množství dat a informací, které se navíc každým rokem exponenciálně navyšuje. Díky technologii tzv. DNA čipů je dnes možné provádět velmi rozsáhlé studie genových expresí a funkční aktivity DNA, které jsou východiskem pro výzkum genetických základů nemocí, stejně jako geneticky podmíněných reakcí na léky (farmakogenomika), individuálních nutričních a metabolických charakteristik (nutrigenomika) a dalších vlastností každého jedince.

Genetické a proteomické databáze dnes tvoří velmi širokou a neustále se zvětšující skupinu informačních zdrojů. Aktuální přehled těchto databází je publikován každý rok v časopise Nucleic Acids Research[12]. S ohledem na ukládaný obsah lze genomické a proteomické databáze rozdělit do následujících skupin:

- **databáze sekvencí nukleových kyselin**: zahrnují sekvence párů bází deoxyribonukleových a ribonukleových kyselin;
- **databáze proteinových sekvencí**: zahrnují sekvence aminokyselin v jednotlivých bílkovinách;
- **databáze proteinových struktur**: obsahují trojrozměrné modely bílkovinných struktur;
- **databáze expresních profilů**: obsahují informace o stupni exprese jednotlivých genů;

[7] **Hodnocení zdravotnických technik (HTA)** zahrnují systematické zhodnocení vlastností, účinků a/nebo dopadů zdravotnických postupů. Mohou se zabývat jak přímými, zamýšlenými výsledky hodnocených technik, tak jejich nepřímými, neplánovanými důsledky. HTA jsou vytvářeny interdisciplinárními týmy a při jejich tvorbě jsou používány explicitní analytické nástroje vycházející z různých metod [3].

[8] **Ekonomické analýzy** srovnávají pomocí formálních kvantitativních metod alternativní postupy z hlediska nákladů a výsledků („cost-benefit analyse“, „cost-effectiveness analyse“) [8].

[9] **Sdílené klinické** (nebo **medicínské**) **rozhodování** je model, který klade zvýšený důraz na pacientovu účast v medicínském rozhodování a je alternativou k tradičnímu paternalistickému modelu, v němž veškerá léčebná rozhodnutí dělá sám lékař. Během tohoto procesu dvojice lékař-pacient bere v úvahu všechny relevantní léčebné možnosti a s nimi související důsledky a přezkoumává, do jaké míry se předpokládané výhody a důsledky léčby slučují s pacientovými preferencemi [4].

[10] **Medicína založená na důkazech (EBM)** je svědomité, jednoznačné a kritické uplatňování nejnovějších a nejlepších důkazů při rozhodování o péči o jednotlivé pacienty. Vykonáváním EBM v praxi je myšlena integrace individuální klinické odbornosti lékařů s nejkvalitnějšími objektivními důkazy pocházejícími ze systematicky prováděného výzkumu [11].

[11] **Medicína založená na hodnotách** [1] posuzuje zdravotnické intervence nejen podle jejich vlivu na objektivní parametry, jako je délka života, ale také podle jejich dopadu na kvalitu života pacienta, která úzce souvisí s jeho individuálním vnímáním významu a důsledků těchto intervencí.

[12] http://nar.oxfordjournals.org

MATFYZPRESS 14. srpna 2009 – 14:52

- **databáze genomů a genových map**: zahrnují informace o lokalizaci genů na chromozomech, které poskytují ve formě popisných přehledů nebo prostřednictvím speciálních prohlížečů.

3.4. *Zdroje typu registru, katalogu, adresáře*

Registrem se v kontextu informačních zdrojů rozumí jakýkoli seznam, soupis, evidence, katalog či přehled, tj. množina záznamů s jednotnou strukturou uspořádaná podle nějakého kritéria. Označení registr se obvykle používá pro seznamy obsahující záznamy pořízené jako úřední dokumenty, většinou v nějaké míře upravené právními předpisy. Jako registr může být označován také typ informačního systému, jehož účelem je zaznamenávat, uchovávat a zpřístupňovat informace o určitých objektech nebo jevech [10].

Katalog je sekundární informační zdroj obsahující soubor katalogizačních záznamů o dokumentech, které daná instituce uchovává ve svých fondech nebo které trvale nebo dočasně zpřístupňuje, vytvářený podle předem stanovaných zásad a umožňující zpětné vyhledávání dokumentů. K základním funkcím katalogu patří lokační funkce (katalogizační záznam informuje o umístění dokumentu a o organizaci fondu), bibliografická funkce (katalogizační záznam informuje o existenci dokumentu), rešeršní funkce (katalogizační záznam umožňuje efektivní vyhledání dokumentu), propagační funkce (katalogizační záznam informuje o nově vydaných dokumentech) [10].

Jako **adresář** je označována příruční publikace s výčtem osob, organizací, produktů i jiných položek sestaveným dle stanoveného hlediska (tematického, chronologického, teritoriálního aj.), uspořádaným abecedně nebo systematicky, uvádějící kontextově významné informace (identifikační, kontaktní, o struktuře, aktivitách apod.). Často je vydáván jako seriál [10].

V případě medicínských informačních zdrojů mají velký význam **registry klinických studií**. Klinické studie jsou podstatou klinického výzkumu. Jejich cílem je experimentálně ověřit, zda daná léčba je bezpečná a účinná. Publikace výsledků klinických studií jsou zase podstatou tzv. medicíny založené na důkazech (EBM). Primární vědecká literatura v této oblasti je však zvlášť citlivá na fenomén zvaný „publikační bias“. Ta vzniká na základě tendence autorů studií i vědeckých časopisů publikovat spíše kladné výsledky klinického výzkumu, což následně může vést k mylným interpretacím výsledků týkajících se účinnosti léčebných postupů. Ve snaze předejít tomuto zkreslení je vyvíjen tlak na pořadatele klinických studií, aby své výzkumné záměry zveřejnili ještě před dokončením samotných klinických zkoušek. Vznikají tak klinické registry, které z hlediska zaměření mohou být jednak všeoborové, jednak oborově specifické. Vedle nezávislých registrů udržovaných například vládními organizacemi existují také korporátní registry poskytované některými farmaceutickými společnostmi jako projev snahy umožnit transparentnost klinických dat a výsledků z nich odvozených.

Klinické registry jsou zdrojem cenných informací, díky kterým lze (kromě snížení rizika publikační bias) navíc předejít duplikování výzkumu a zrychlit přenos nejnovějších výsledků z klinického výzkumu do medicínské praxe. Jsou nezbytným doplňkem při vyhledávání existujících informací pro tvorbu tzv. systematických přehledů a metaanalýz klinických studií, jakožto klíčových zdrojů informací pro podporu klinického rozhodování. Některé registry poskytují kromě protokolů probíhajících studií také výsledky již ukončených klinických zkoušek a/nebo odkazy na související publikace o daném léku či daném léčebném postupu.

Relativně méně známou skupinou databází z kategorie medicínských katalogů jsou **katalogy genů a geneticky podmíněných nemocí**. První databáze tohoto druhu vznikla v šedesátých letech dvacátého století pod názvem *Mendelian Inheritance in Man (MIM)* [5]. Toto kompendium veškerých známých lidských genů a s nimi souvisejících fenotypů je inspirací i výchozím bodem pro budování řady specializovaných genových katalogů, zaměřených například na onkologická, kardiovaskulární a další onemocnění.

Další skupinu databází spadající do popisované kategorie tvoří **webové katalogy**. Obecně lze říci, že webové katalogy jsou databáze seznamů ručně vybraných internetových stránek. Kvalitní katalogy nabízejí kromě podrobného členění také hodnocení užitečnosti či kvality zdrojů, byť obvykle pouze semikvantitativní, znázorňované pomocí tří až pětistupňové škály. Nevýhodou webových katalogů je zanikání a přemisťování webových stránek. Katalogizované odkazy tak mnohdy nevedou uživatele k obsahu, který hledá, ale na zcela nerelevantní nebo zaniklé stránky. Aktualizace všech odkazů u rozsáhlých katalogů je náročná a vezmeme-li navíc v úvahu rychlost, se kterou přibývají na internetu nové informace, je logické, že řada katalogů v průběhu let přestala plnit svoji funkci a postupně zanikla. V případě medicínských webových katalogů to například znamená, že z katalogů uvedených v publikaci Internet a medicína [6] jich po osmi letech zůstalo jen 54 procent. Jsou-li však webové katalogy pravidelně aktualizovány, jsou koncentrovaným a velmi cenným zdrojem informací.

3.5. *Zdroje hybridní*

Řada informačních zdrojů sdružuje více typů dokumentů. Patří mezi ně jednak rozsáhlé polytematické a víceoborové databáze, jednak čistě medicínské databáze, které **zahrnují informace jak bibliografické, tak určité procento plných textů**. Mezi hybridní databáze patří také řada informačních zdrojů zaměřených na podporu klinického rozhodování, a to především těch, které jsou určeny pro použití bezprostředně během léčebně-preventivní péče („point-of-care“). Tyto zdroje jsou nejčastěji **kombinací plnotextových a faktografických databází** (zahrnují například plné texty lékařských doporučení, informace o lécích, epidemiologická data ap.).

3.6. *Postpublikačně evaluované zdroje*

Novou, pro medicínu specifickou skupinu tvoří postpublikačně evaluované informační zdroje (neboli tzv. informační zdroje „druhého řádu“). Tyto zdroje vznikly především pro potřeby klinické praxe poté, co se ukázalo, že prohledávání tradičních bibliografických či plnotextových biomedicínských databází přináší příliš mnoho klinicky nerelevantních a metodologicky nespolehlivých výsledků.

Pro postpublikačně evaluované informační zdroje je typické, že na jejich vzniku spolupracuje tým specialistů, kteří nejdříve vyhledávají články v tradičních biomedicínských informačních zdrojích a poté je hodnotí z hlediska kvality, klinické relevance a potenciálního dopadu v klinické praxi. Jsou tedy úzce spjaty s tzv. *informačními službami s přidanou hodnotou*, která je v tomto případě daná výběrem článků z primární biomedicínské literatury a jejich hodnocením specialisty z klinických oborů. Záznamy zahrnují kromě bibliografických údajů a event. abstraktu navíc hodnocení a v řadě případů také komentáře expertů v daném oboru.

3.7. *Zdroje „prospektivně-exploratorní“ (generativní)*

Databáze, které lze pracovně nazvat jako „prospektivně-exploratorní“ nebo také generativní, jsou novou skupinou informačních zdrojů, která vznikla z potřeby zkoumat souvislosti mezi různými druhy dat a informací především faktografického charakteru, jež jsou rozeseta v nepřeberném a stále rychleji rostoucím množství článků publikovaných v oblasti biomedicíny. Tyto informace jsou průběžně vybírány specialisty (kurátory databáze) z tradičních biomedicínských databází. V jistém ohledu tedy lze na tyto databáze nahlížet také jako na faktografické informační zdroje s postpublikačně evaluační složkou. Od nich se však podstatně liší tím, že data a fakta nejen shromažďují a popisují, ale současně umožňují jejich zkoumání. Umožňují tedy informace jednak vyhledávat, jednak vzájemně porovnávat z hlediska možných souvislostí. Výsledky vyhledávání je obvykle možné zobrazit nejen jako lineární seznam, ale také ve formě přehledných tabulek a názorných grafů zobrazujících vztahy mezi hledanými prvky. Na rozdíl od výše uvedených genomických a proteomických databází, které informace o genech a proteinech primárně shromažďují a popisují, tyto databáze umožňují hledání nových vztahů mezi geny, příznaky nemocí, chemikáliemi, léky ap. Jejich cílem je hledání příčin a následků, identifikace souvislostí a objevování hypotéz v množství dat publikovaných v rámci nestrukturovaného textu. S ohledem na tento záměr bývají orientovány na úzkou vědní disciplínu (např. výzkum v oblasti farmakogenomiky, toxikogenomiky nebo „jen“ konkrétního mikroorganizmu či některých z jeho produktů).

3.8. *Multimediální zdroje*

Multimediální informační zdroje zahrnují databáze **obrazových informací, zvukových nahrávek a videonahrávek**. Informační zdroje založené na zvukových nahrávkách a videozáznamech se velmi rozšířily především díky aplikacím a službám Webu 2.0. Řada recenzovaných biomedicínských časopisů poskytuje pravidelně nebo příležitostně zvukové nahrávky ve formátu mp3 (tzv. „podcasts“). Služby pro sdílení videí jsou zase základem pro vznik databází videozáznamů s nejrůznějším medicínským obsahem od laboratorních technik přes obrazová vyšetření až po operační postupy.

3.9. *Agregované zdroje*

Agregované informační zdroje se vyznačují seskupením dvou a více databází do jednoho celku. Někteří databázoví vendoři nabízejí tematicky nebo účelově související informační zdroje ve formě **databázových kolekcí**. V rámci internetu patří k agregovaným informačním zdrojům **webové portály**, které poskytují informace z odlišných webových stránek a sídel prostřednictvím jednoho rozhraní.

3.10. *Monitorovací zdroje*

Kromě výše uvedených kategorií zdrojů, které jsou založeny na informacích pocházejících z výzkumu a z vědecké literatury („*evidence-based*“) [13], lze v rámci medicíny stanovit ještě jinou a také velmi podstatnou

[13] Pojem „evidence“ je zde chápán ve smyslu „svědectví“, „známka“, „empirický důkaz“.

skupinu zdrojů založených na informacích získaných z praxe monitorováním událostí („*event-based*“) [14]. Monitorovány přitom mohou být nežádoucí účinky léků (**farmakovigilance**), onemocnění vzniklá po požití závadných potravin (**hygiena výživy**) nebo ohniska a šíření infekčních nemocí (**epidemiologický dohled**). Tyto zdroje nabývají na důležitosti zvláště dnes, v době globalizace a častého cestovaní.

4. Závěr

Uvedená práce předkládá klasifikaci informačních zdrojů pro medicínu a související obory, která zohledňuje exponenciální nárůst v množství i rozmanitosti biomedicínských informací. Tato klasifikace je základem pro nově vytvořenou databázi medicínských informačních zdrojů. Databáze obsahuje v době sepisování této práce více než 70 záznamů, přičemž průběžně jsou doplňovány jednak záznamy o nově vznikajících informačních zdrojích, jednak jsou postupně doplňovány již existující zdroje. Jednotlivé záznamy zahrnují stručný popis zdroje, literární citace, v případě, že existují, a relevantní webové odkazy. Databázi je možné prohledávat plnotextově nebo prohlížet v rámci kategorií. Databáze je dostupná prostřednictvím webového rozhraní na adrese http://medizdroje.blogspot.com .

Literatura

[1] M.M. Brown, G.C. Brown, and S. Sharma, “Evidence-based to Value-based Medicine”, AMA Bookstore, 2005.

[2] K. Davies, “The 2005 Database Explosion”, *Bio-IT World [online]*, Feb. 2005. Dostupné na www: http://www.bio-itworld.com/archive/021105/itin_explosion.html [cit. 2009-04-30].

[3] C.S. Goodman, “HTA 101: Introduction to Health Technology Assessment”, Aug. 2004. Dostupné na www: http://www.nlm.nih.gov/nichsr/hta101/ta10103.html [cit. 09-06-30].

[4] [D.L. Frosch and R.M. Kaplan, “Shared decision making in clinical medicine: past research and future directions”, *American Journal of Preventive Medicine*, vol. 17, pp. 285–294, Nov. 1999.

[5] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, and V.A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders”, *Nucleic Acids Research*, vol. 33, pp. D514–D517, Jan. 2005.

[6] P. Kasal and Š. Svačina, “Internet a medicína”, Praha: Grada Publishing, 2001.

[7] E.S. Lander, L.M. Linton, B. Birren, et al., “Initial sequencing and analysis of the human genome”, *Nature*, vol. 409, pp. 860–921, Feb. 2001.

[8] K.A. McKibbon, A. Eady, and S. Marks, “PDQ: evidence-based principles and practice” B.C. Decker, Inc., 1999.

[9] R. Papík, “Vyhledávání informací III. Dialogové služby světových databázových center”, *Národní knihovna - knihovnická revue*, č. 1, pp. 20–30, 2002.

[10] M. Ressler, ed., “Informační věda a knihovnictví: Výkladový slovník české terminologie z oblasti informační vědy a knihovnictví. Výběr z hesel v databázi TDKIV (Elektronická verze 1.0)”, Praha: VŠCHT / NK, 2006.

[11] D. Sackett, W.M.C. Rosenberg, M.J.A. Gray, B.R. Haynes, and W.S. Richardson, “Evidence based medicine: what it is and what it isn’t”, *BMJ*, vol. 312, pp. 71–72, Jan. 1996.

[12] J. Sauve, H. Lee, M. Farkouh, and D.L. Sackett, “The critically appraised topic: a practical approach to learning critical appraisal”, *Ann R Coll Physicians Surg Can*, vol. 28, pp. 396-8,1995.

[13] J.C. Venter, M.D. Adams, E.W. Myers, et al., “The sequence of the human genome”, *Science*, vol. 291, pp. 1304–51, Feb. 2001.

[14] Pojem „event“ je zde chápán jako „případ“ či „událost“.

DRAFT 14. srpna 2009, 14:52

Properties of Fuzzy Logical Operations

Post-Graduate Student:
ING. MILAN PETRÍK
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
petrik@cs.cas.cz

Supervisor:
PROF. ING. MIRKO NAVARA, DRSC.
Center for Machine Perception, Department of Cybernetics,
Faculty of Electrical Engineering, Czech Technical University,
Technická 2, 166 27 Prague 6, CZ
navara@cmp.felk.cvut.cz

Field of Study:
Artificial Intelligence and Biocybernetics

The author was supported by the Grant Agency of the Czech Republic under Project 401/09/H007.

Abstract

We deal with geometrical and differential properties of triangular norms (t-norms for short), i.e. binary operations which implement logical conjunctions in fuzzy logic. The first part discusses the problem of a visual characterization of the associativity of t-norms. The results given by web geometry are adopted, mainly the concept of the Reidemeister closure condition, in order to characterize the shape of level sets of t-norms. This way, a visual characterization of the associativity is provided for general, continuous, and continuous Archimedean t-norms. The second part deals with differential properties of continuous Archimedean t-norms. It is shown that partial derivatives of such a t-norm on a particular subset of its domain correspond directly to the generator (or to the derivative of the generator) of the t-norm. As the result, several methods which reconstruct multiplicative and additive generators of continuous Archimedean t-norms are introduced. The presented results contribute to a partial solution of an open problem whether a non-trivial convex combination of two t-norms can be a triangular norm again.

1. Introduction

The fuzzy logic has been proposed as an alternative to the classical Boolean logic. The notion "fuzzy" was firstly introduced in 1965 by Zadeh in his paper [40] where he defined fuzzy logic and fuzzy sets.

The main idea of the fuzzy logic is to enlarge the set of truth values, i.e. 0 and 1 (false and true), to the real unit interval $[0, 1]$. In comparison to the classical logic where a statement can be either true or false, the generalization to the fuzzy logic allows to express also a partial truth of a statement as it admits degrees of truth.

Generalization of the set of truth values hangs together with a generalization of the logical operations. The logical conjunction is usually implemented by a *triangular norm* (shortly, a *t-norm*). Although the notion of a t-norm was originally introduced within the framework of probabilistic metric spaces [37], it has found a successful application in fuzzy logic. The currently studied fuzzy logics, as will be described in the sequel, are primarily based on t-norms. Another important logical connective, the implication, is usually implemented by a *residuum* (also *residuated implication*) which is derived from a t-norm in order to form an adjoint pair and work correctly in the generalized Modus Ponens rule.

The logical calculus which is able to cope with partially true statements is called a fuzzy or many-valued logic. The beginning of many-valued reasoning dates back to 1920 when Łukasiewicz proposed his three-valued logic [23] and to the work of Post [36] in 1921. Now, one of the most successful fuzzy logics is the *Basic Fuzzy Logic* (BL for short) which has been introduced by Hájek [15] and fully described in his monograph [16]. We remark that BL includes the fuzzy logics, known so far at the time of its introduction, as its special cases. The semantical counterpart of BL is represented by BL-algebras which play an analogous role as Boolean algebras for the classical Boolean logic. An example of a BL-algebra is the real unit interval $[0, 1]$ endowed with a continuous t-norm which represents a conjunction and the corresponding residuum which represents an implication. Such a BL-algebra is called a standard BL-algebra. Hájek proved that BL is sound and complete with respect to the class of BL-algebras. This means that a formula is provable in BL if and only if it is a tautology in all BL-algebras. BL is complete even with respect to standard BL-algebras. This fact is known as the *Standard Completeness Theorem* of BL [11].

DRAFT
Prague 2009-11-25

2. Preliminaries

We present here some basic facts about triangular norms. The proofs and more details can be found e.g. in the monographs on triangular norms [7, 20]. Another good introduction to triangular norms can also be given by monographs on fuzzy sets and fuzzy logic [22, 30].

Definition 2.1 *A* triangular norm *(a* t-norm *for short) is a binary operation $T\colon [0,1]\times[0,1]\to[0,1]$ such that for all $x,y,z\in[0,1]$ the following axioms are satisfied:*

(T1) $T(x,y)=T(y,x)\,,$ *(commutativity)*

(T2) $T(x,T(y,z))=T(T(x,y),z)\,,$ *(associativity)*

(T3) $x\le y\Rightarrow T(x,z)\le T(y,z)\,,$ *(monotonicity)*

(T4) $T(x,1)=x\,.$ *(neutral element)*

The three most common t-norms are the *minimum t-norm*, $T_{\mathbf{M}}(x,y)=\min\{x,y\}$, the *Łukasiewicz t-norm*, $T_{\mathbf{L}}(x,y)=\max\{x+y-1,0\}$, and the *product t-norm*, $T_{\mathbf{P}}(x,y)=x\cdot y$.

A continuous t-norm T is called *Archimedean* if $T(x,x)<x$ for all $x\in\,]0,1[$. A t-norm which is continuous and strictly increasing on the half-open square $]0,1]^2$ is said to be *strict*; such a t-norm is always Archimedean. A continuous Archimedean t-norm is called *nilpotent* if it is not strict. Thus every continuous Archimedean t-norm is either strict or nilpotent. For example, the product t-norm is strict, the Łukasiewicz t-norm is nilpotent, and the minimum t-norm is an example of a continuous t-norm which is not Archimedean.

Every continuous Archimedean t-norm can be represented by a one-dimensional real function called *generator*. This result is formalized by the *Representation Theorem* [1, 14, 21, 27]:

Theorem 2.2 (Representation Theorem) *For a function $T\colon [0,1]^2\to[0,1]$ the following statements are equivalent:*

1. *T is a continuous Archimedean t-norm.*
2. *T has a continuous* additive generator*, i.e., there exists a continuous strictly decreasing function $t\colon[0,1]\to[0,\infty]$ with $t(1)=0$ such that $T(x,y)=t^{(-1)}\left(t(x)+t(y)\right)$ holds for all $(x,y)\in[0,1]^2$. Here, $t^{(-1)}$ denotes the pseudo-inverse of t which is (in this case) defined as:*
$$t^{(-1)}(y)=\begin{cases}0 & \text{if}\quad y>t(0)\,,\\ t^{-1}(y) & \text{if}\quad y\le t(0)\,.\end{cases}$$
3. *T has a continuous* multiplicative generator*, i.e., there exists a continuous strictly increasing function $\theta\colon[0,1]\to[0,1]$ with $\theta(1)=1$ such that $T(x,y)=\theta^{(-1)}\left(\theta(x)\cdot\theta(y)\right)$ holds for all $(x,y)\in[0,1]^2$. Here, $\theta^{(-1)}$ denotes the pseudo-inverse of θ which is (in this case) defined as:*
$$\theta^{(-1)}(y)=\begin{cases}0 & \text{if}\quad y<\theta(0)\,,\\ \theta^{-1}(y) & \text{if}\quad y\ge\theta(0)\,.\end{cases}$$

The *support* of a binary operation $T\colon [0,1]^2\to[0,1]$, denoted by $\operatorname{supp}T$, is the closure of the set
$$\left\{(x,y)\in[0,1]^2\mid T(x,y)>0\right\}.$$

3. Current situation of the studied problem

3.1. Convex combinations of t-norms

This work has been primarily inspired by the long standing open problem of convex combinations of triangular norms and summarizes the results which have been achieved while solving this problem. This problem has been formulated, for example, in the list of open problems by Alsina, Frank, and Schweizer [6]:

Problem 3.1 *Is the arithmetic mean, or for that matter any convex combination, of two distinct t-norms ever a t-norm?*

We recall that a convex combination of two t-norms T_1, T_2 is a function $F=\alpha T_1+(1-\alpha)T_2$ where $\alpha\in[0,1]$. It is immediate that for trivial convex combinations, i.e. for $\alpha\in\{0,1\}$ or for $T_1=T_2$, the answer is positive. A positive example can be given even for non-trivial convex combinations of non-continuous t-norms [17, 34, 39]. For example, let T_1 be an ordinal sum of the product t-norm $T_{\mathbf{P}}$ on the carrier $[0,\frac{1}{2}]$. Let T_2 be a binary operation on $[0,1]$ such that $T_2(x,y)=0$ for $x,y\in[0,\frac{1}{2}]$ and $T_2(x,y)=\min\{x,y\}$ otherwise. It is easy to check that T_2 is a left-continuous t-norm. Observe now that any convex combination of T_1 and T_2 is a left-continuous t-norm. However, for continuous t-norms the problem still has not been answered completely although it is conjectured that for the continuous t-norms the answer to the question posed in Problem 3.1 is “never” [6].

DRAFT 14. srpna 2009 - 14:32

Thus, in order to exclude the trivial cases mentioned above, whenever we write "convex combination" we mean a function $\alpha T_1 + (1-\alpha) T_2$ where $\alpha \in]0,1[$, $T_1 \neq T_2$, and both t-norms are continuous.

In the rest of this section we briefly outline the results related to the convex combinations of t-norms which have been done so far. In the historically first paper dealing with this problem, Tomás [38] has given a result on strict t-norms under additional (and rather restrictive) constraints. In the papers by Ouyang, Fang and Li [31, 32], the whole class of continuous t-norms is treated under no additional assumptions. For example, they prove [31] that a convex combination of a continuous Archimedean t-norm and a continuous non-Archimedean t-norm is never a t-norm. In other words, if a convex combination of two continuous t-norms is a t-norm again, then both combined t-norms are ordinal sums with the same structure of summand carriers. By this result, in order to clarify the convex structure of the class of continuous t-norms it is sufficient to clarify the convex structure of the class of continuous Archimedean t-norms. By another result of theirs [31], a convex combination of a strict and a nilpotent t-norm is never a t-norm. Thus even the latter task can be subdivided into solving the convex structure of the nilpotent class and of the strict class separately. Another result is due to Jenei [17] and applies to all pairs of left-continuous t-norms with an additional property that both t-norms share an involutive level set. An immediate consequence of this result is that a convex combination of two nilpotent t-norms, T_1 and T_2, such that $\operatorname{supp} T_1 = \operatorname{supp} T_2$, is never a t-norm. Let us mention also the recent result by Mesiar and Mesiarová-Zemánková [26] where it is stated that a convex combination of two continuous t-norms with the same diagonal is never a t-norm. (We recall that a *diagonal* of a t-norm T is the function $x \mapsto T(x,x)$.)

Two new, recently published [33, 34], results on this topic are presented here. Using a web-geometrical approach to describe associativity of t-norms, it is proven that any convex combination of two nilpotent t-norms is never a t-norm. Furthermore, using an idea of reconstruction of generators according to partial derivatives of t-norms, several new results on the problem of convex combinations of strict t-norms are presented.

3.2. Associativity of t-norms

The commutativity, the non-decreasingness and the existence of a neutral element have an easy graphical interpretation. However, the question how to visually interpret the associativity is a long-standing open problem within the community of people dealing with t-norms. Some results have been done, mainly thanks to the effort of Jenei [18], and Maes and De Baets [24, 25], yet a satisfactory answer to the question still has not been given.

The theory of *web geometry* [9, 2, 3, 4] has come with results which answer such, and similar, kinds of questions in a rather intuitive way. In particular, associative loops are characterized by the *Reidemeister closure condition*. These results were, however, done to characterize algebraic properties of loops. Although t-norms do not form loops, there are, fortunately, some similarities between t-norms and loops (monotonicity, neutral element, ...). We will show that some modifications of the Reidemeister closure condition can still be applied to t-norms in order to characterize their associativity.

Motivation 3.2 *Consider the Łukasiewicz t-norm, $T_{\mathbf{L}}(x,y) = \max\{x+y-1, 0\}$. The structure of its level sets is extremely simple as they are formed by parallel lines.*

Notice the following easy property of these sets: draw a rectangle (by vertical and horizontal lines) anywhere in the support of the operation and denote the level sets passing through the vertices of the rectangle. Now draw another rectangle such that three of its vertices match the three distinct denoted level sets. The fourth vertex of the rectangle shall, naturally, match the fourth denoted level set.

The property described in Motivation 3.2 characterizes associativity and corresponds to the Reidemeister closure condition introduced by web geometry [9, 2, 3, 4].

3.3. Reconstruction of generators

When a continuous (multiplicative or additive) generator is defined, it is easy to construct the corresponding (continuous Archimedean) t-norm. The reverse task, however, is not so trivial. One way how to obtain a generator of a continuous Archimedean t-norm is to use the proof of the Representation Theorem. This proof is constructive, however, it does not need to result in an explicit formula of the generator. This significantly reduces the usability of this method. Another possibility is to use the results given by Pi-Calleja [5, 35] and by Craigen and Páles [12]. Both these results give explicit formulas for additive generators of strict t-norms. However, the computations of formulas are rather non-intuitive and non-straightforward which disallows an

DRAFT 24. srpna 2009 - 14:32

easy usage. The formulas also show no direct relation between t-norms and their generators.

In this work, an alternative [28, 29] is presented. It is shown that partial derivatives of t-norms admit to obtain formulas for generators in a closed form. As the partial derivatives need not exist, this approach cannot be applied to all continuous Archimedean t-norms, but it seems general enough for all practical applications. It is even shown that every continuous t-norm can be approximated (with an arbitrary precision) by a t-norm from the class of strict t-norms on which one of the introduced methods is applicable. An advantage of this approach is that it relates (the shape of) the generator directly to (the shape of) the t-norm and that it is based on the basic differential calculus which makes the computational procedure straightforward. Benefiting from the fact that computation with the first derivatives is well described and can be well algorithmized, these methods can be easily applicable both by a manual computation and by computational systems. Furthermore, a simplified proof of the Representation Theorem for a subclass of strict t-norms is given as one of the results based on this approach.

4. Results

4.1. Associativity of t-norms

Let $F\colon [0,1]^2 \to [0,1]$ be a commutative and non-decreasing binary operation satisfying $F(x,1) = x$ for all $x \in [0,1]$.

By a *rectangle* we mean a set of four points $\mathbf{R} = \{x_1^{\mathbf{R}}, x_2^{\mathbf{R}}\} \times \{y_1^{\mathbf{R}}, y_2^{\mathbf{R}}\} \subset [0,1]^2$ where $x_1^{\mathbf{R}} \le x_2^{\mathbf{R}}$ and $y_1^{\mathbf{R}} \le y_2^{\mathbf{R}}$. Let $\mathbf{P}, \mathbf{R} \subset [0,1]^2$ be two rectangles. We say that $\mathbf{P} \approx_F \mathbf{R}$ if and only if $F\left(x_i^{\mathbf{P}}, y_j^{\mathbf{P}}\right) = F\left(x_i^{\mathbf{R}}, y_j^{\mathbf{R}}\right)$ for all $i, j \in \{1,2\}$; $\mathbf{P} \sim_F^{k,l} \mathbf{R}$ if and only if the equality $F\left(x_i^{\mathbf{P}}, y_j^{\mathbf{P}}\right) = F\left(x_i^{\mathbf{R}}, y_j^{\mathbf{R}}\right)$ is violated for at most $i = k$ and $j = l$; $\mathbf{P} \sim_F \mathbf{R}$ if and only if the equality $F\left(x_i^{\mathbf{P}}, y_j^{\mathbf{P}}\right) = F\left(x_i^{\mathbf{R}}, y_j^{\mathbf{R}}\right)$ is violated for at most one combination of i and j. Clearly, $\approx_F$, $\sim_F^{k,l}$, and $\sim_F$ are equivalences, $\approx_F$ is a subrelation of $\sim_F^{k,l}$, and $\sim_F^{k,l}$ is a subrelation of $\sim_F$ for any $k, l \in \{1,2\}$.

Theorem 4.1 *Let $T\colon [0,1]^2 \to [0,1]$ be a non-decreasing, commutative binary operation which satisfies $T(x,1) = x$ for every $x \in [0,1]$.*

- *T is associative if and only if $\mathbf{P} \sim_T^{1,1} \mathbf{R}$ implies $\mathbf{P} \approx_T \mathbf{R}$ for every pair of rectangles, $\mathbf{P}$ and $\mathbf{R}$, such that $\mathbf{P} = \{x_1^{\mathbf{P}}, x_2^{\mathbf{P}}\} \times \{y_1^{\mathbf{P}}, 1\} \subset [0,1]^2$ and $\mathbf{R} = \{x_1^{\mathbf{R}}, 1\} \times \{y_1^{\mathbf{R}}, y_2^{\mathbf{R}}\} \subset [0,1]^2$.*
- *If T is continuous then it is associative if and only if $\mathbf{P} \sim_T^{1,1} \mathbf{R}$ implies $\mathbf{P} \approx_T \mathbf{R}$ for every pair of rectangles, $\mathbf{P}, \mathbf{R} \subset [0,1]^2$.*
- *If T is continuous and Archimedean then it is associative if and only if $\mathbf{P} \sim_T \mathbf{R}$ implies $\mathbf{P} \approx_T \mathbf{R}$ for every pair of rectangles, $\mathbf{P}, \mathbf{R} \subset \operatorname{supp} T \cap]0,1]^2$.*

4.2. Reconstruction of generators

We denote by t', θ' the derivatives of generators t, θ, respectively. We denote by

$$\begin{aligned} \mathrm{D}T(x,y) &= \lim_{h \to 0} \frac{T(x+h,y) - T(x,y)}{h} \\ &= \lim_{z \to x} \frac{T(z,y) - T(x,y)}{z - x}\,. \end{aligned}$$

the partial derivative of a t-norm T with respect to the first variable.

Assumption 4.2 *The partial derivative $\mathrm{D}T$ will be considered only in the support $\operatorname{supp} T$. In particular,*

$$\mathrm{D}T(1,y) = \lim_{x \to 1_-} \frac{y - T(x,y)}{1 - x}$$

is the left *partial derivative with respect to the first variable. If T is strict, then*

$$\mathrm{D}T(0,y) = \lim_{x \to 0_+} \frac{T(x,y)}{x}$$

is the right *partial derivative. For T nilpotent, we require the second argument $y > 0$; then $\mathrm{D}T(x,y)$ is defined for all $x \in [N_T(y), 1]$, in particular,*

$$\mathrm{D}T(N_T(y), y) = \lim_{z \to N_T(y)_+} \frac{T(z,y)}{z - N_T(y)} \qquad (1)$$

is the right *partial derivative. Since T is nilpotent, the negation N_T is involutive. Therefore, substituting $x = N_T(y)$, we can write (1) as*

$$\mathrm{D}T(x, N_T(x)) = \lim_{z \to x_+} \frac{T(z, N_T(x))}{z - x}\,.$$

For T nilpotent and $y = 0$, the line $\{(x,0) \mid x \in \mathbb{R}\}$ intersects $\operatorname{supp} T$ only at a single point $(1,0)$ and $\mathrm{D}T(x,0)$ is undefined for any $x \in [0,1]$.

We say that a strict t-norm T is *annihilator-differentiable* if the function $\mathrm{D}T(0,y)$ is defined for all $y \in [0,1]$.

DRAFT Spring 2009 - 14:32

Theorem 4.3 (Reconstruction along annihilator) *Let T be a strict annihilator-differentiable t-norm and let $\xi\colon [0,1] \to [0,1]\colon y \mapsto \mathrm{D}T(0,y)$. Then $\xi(0) = 0$, $\xi(1) = 1$, and the restriction of ξ to $]0,1[$ is either (1) the constant 0, (2) the constant 1, or (3) a bijection on $]0,1[$. Moreover, in case (3) the function ξ is a multiplicative generator of T.*

Theorem 4.4 (Reconstruction along level set) *Let T be a continuous Archimedean t-norm. Suppose that T has an absolutely continuous additive generator with a non-zero finite derivative at some point $a \in]0,1]$. (We take the left derivative at 1.) Let $\mathrm{D}T$ be the partial derivative of T with respect to the first variable in the support $\operatorname{supp} T$. Suppose that $\mathrm{D}T(z, I_T(z,a))$ exists for almost all $z \in [a,1]$. Suppose further that $\mathrm{D}T(a, I_T(a,z))$ exists and is in $]0,\infty[$ for almost all $z \in [0,a[$. Then T has an additive generator*

$$t^*(x) = \int_x^1 v(z)\,\mathrm{d}z\,,$$

where

$$v(z) = \begin{cases} \mathrm{D}T(z, I_T(z,a)) & \text{if} \quad z \geq a\,, \\ \frac{1}{\mathrm{D}T(a, I_T(a,z))} & \text{if} \quad z < a \end{cases}$$

for almost all $z \in [0,1]$. Explicitly, if $x \geq a$ then

$$t^*(x) = \int_x^1 \mathrm{D}T(z, I_T(z,a))\,\mathrm{d}z$$

and if $x < a$ then

$$\begin{aligned} t^*(x) \quad &= \int_x^a \frac{1}{\mathrm{D}T(a, I_T(a,z))}\,\mathrm{d}z \\ &+ \int_a^1 \mathrm{D}T(z, I_T(z,a))\,\mathrm{d}z\,. \end{aligned}$$

Remark 4.5 *We admit that the function v may attain zero or infinite value at some points. Then we obtain an infinite value of t'. However, this may happen only in countably many points and this does not influence the integral defining t. The assumption of absolute continuity includes also the convergence of the integral.*

As a special case of Theorem 4.4, we obtain:

Theorem 4.6 (Reconstruction along unit) *Let T be a continuous Archimedean t-norm and let t be an additive generator of T such that t is absolutely continuous at $]0,1]$ and $t'(1) = b_{t,1} \in]-\infty,0[$. Suppose that $\mathrm{D}T(1,y) \in]0,\infty[$ for almost all $y \in]0,1]$. Then*

$$t'(y) = \frac{b_{t,1}}{\mathrm{D}T(1,y)} \qquad \textit{(almost everywhere in }]0,1]\textit{)}$$

and

$$t(y) = \int_y^1 \frac{-b_{t,1}}{\mathrm{D}T(1,u)}\,\mathrm{d}u$$

for all $y \in]0,1]$.

Theorem 4.4 allows us to reconstruct an additive generator when a non-negative constant $a \in]0,1]$ is given. The following theorem shows that even $a = 0$ can be used. However, this works for nilpotent t-norms only.

Theorem 4.7 *Let T be a nilpotent t-norm. Suppose that T has an absolutely continuous additive generator with a non-zero finite (right) derivative at the point 0. Let $\mathrm{D}T$ be the right partial derivative of T with respect to the first variable in the support $\operatorname{supp} T$. Suppose that $\mathrm{D}T\,(z, N_T(z))$ exists for almost all $z \in [0,1]$. Then T has an additive generator*

$$t^*(x) \quad = \quad \int_x^1 \mathrm{D}T\,(z, N_T(z))\,\mathrm{d}z\,.$$

4.3. Convex combinations of t-norms

With the help of web geometry, the following result can be achieved:

Theorem 4.8 *Let T_1 and T_2 be two continuous Archimedean t-norms such that $\operatorname{supp} T_1 \neq \operatorname{supp} T_2$. Then no non-trivial convex combination of T_1 and T_2 is a t-norm.*

According to the result by Jenei [17], a convex combination of two nilpotent t-norms with the same support is never a t-norm. Therefore Theorem 4.8 brings the following result:

Corollary 4.9 *A non-trivial convex combination of two distinct nilpotent t-norms is never a t-norm.*

Theorem 4.8 also gives an alternative proof of the result by Ouyang and Fang [31]:

Corollary 4.10 *A non-trivial convex combination of a strict and a nilpotent t-norm is never a t-norm.*

DRAFT 14 srpna 2009 – 14:32

Now, we present some results on convex combinations of strict t-norms based on the reconstruction methods. Let T be a strict annihilator-differentiable t-norm and let $\xi\colon [0,1] \to [0,1]\colon y \mapsto \mathrm{D}T(0,y)$. Then T is said to be

- *annihilator-weak* (and we write $T \in \mathcal{T}_{\mathrm{AW}}$) if $\xi(x) = 0$ for all $x \in]0,1[$,
- *annihilator-strong* (and we write $T \in \mathcal{T}_{\mathrm{AS}}$) if $\xi(x) = 1$ for all $x \in]0,1[$,
- *annihilator-reconstructible* (and we write $T \in \mathcal{T}_{\mathrm{AR}}$) if ξ is a bijection.

The set of all strict t-norms which are not annihilator-differentiable will be denoted by $\mathcal{T}_{\mathrm{N}}$.

Let T be a continuous Archimedean t-norm with a multiplicative generator θ such that θ' is continuous at 1 and $\theta'(1) \in]0,\infty[$. Then we say that T belongs to the class $\mathcal{T}_{\mathrm{UR}}$.

Proposition 4.11 *Let T_1 and T_2 belong to two distinct classes from $\mathcal{T}_{\mathrm{AR}}, \mathcal{T}_{\mathrm{AW}}, \mathcal{T}_{\mathrm{AS}}$. Then no non-trivial convex combination of T_1 and T_2 is a t-norm.*

Proposition 4.12 *Let $T_1, T_2 \in \mathcal{T}_{\mathrm{AR}} \cap \mathcal{T}_{\mathrm{UR}}$ be strict t-norms. Let $\theta_1\colon y \mapsto \mathrm{D}T\,(0,y)$ and $\theta_2\colon y \mapsto \mathrm{D}T\,(0,y)$ be multiplicative generators of T_1 and T_2, respectively.*

If a non-trivial convex combination of T_1 and T_2 is a t-norm then for each $y \in [0,1]$ at least one of the following conditions is satisfied:

$$\theta_2'(y) = \frac{\theta_2'(1)}{\theta_1'(1)}\,\theta_1'(y)\,,$$

$$\frac{\theta_1'(y)}{\theta_1(y)} = \frac{\theta_2'(y)}{\theta_2(y)}\,.$$

Corollary 4.13 *Let $T_1, T_2 \in \mathcal{T}_{\mathrm{AR}} \cap \mathcal{T}_{\mathrm{UR}}$ be two distinct strict t-norms such that their multiplicative generators, $\theta_1\colon y \mapsto \mathrm{D}T\,(0,y)$ and $\theta_2\colon y \mapsto \mathrm{D}T\,(0,y)$, are absolutely continuous. If there exists $a \in]0,1[$ such that $\theta_1(a) = \theta_2(a)$ then no non-trivial convex combination of T_1 and T_2 is a t-norm.*

5. Summary

We summarize here briefly the contributions of the thesis:

- Some results of web geometry, namely the Reidemeister closure condition, have been generalized also for algebras which do not form loops. (T-norms can be considered as commutative integral monoids on $[0,1]$.)
- A tool which visually characterizes associativity of general t-norms has been given.
- It has been shown that the generators or their derivatives correspond in many cases directly to the partial derivatives of continuous Archimedean t-norms. These results contribute to both practical applications (they allow a straightforward computation) and theoretical research (they give a new insight into the subject). A theoretical contribution has been, furhermore, illustrated by the results on convex combinations of strict t-norms and by the alternative proof of the Representation Theorem.
- The question of convex combinations of t-norms has been answered negatively for all nilpotent t-norms. In the case of strict t-norms, the problem has been divided into several subclasses and a possible further research has been outlined.
- We remark that the thesis also contributes to the question: “Which subsets of its domain uniquely determine an Archimedean t-norm?” Several results [8, 10, 13, 19] (and a summarization [20]) have been published giving concrete types of subsets of the unit square. Knowing functional values on the points of such a subset, an Archimedean t-norm is determined uniquely. Here a similar result is given yet the first partial derivatives are considered instead of functional values.

References

[1] J. Aczél, "Sur les opérations definies pour des nombres réels". *Bulletin de la Société Mathématique de France*, 76:59–64, 1949.

[2] J. Aczél, "Quasigroups, nets and nomograms". *Advances in Mathematics*, 1:383–450, 1965.

[3] M.A. Akivis and V.V. Goldberg, "Algebraic aspects of web geometry". *Commentationes Mathematicae Universitatis Carolinae*, 41(2):205–236, 2000.

[4] M.A. Akivis and V.V. Goldberg, "Local algebras of a differential quasigroup". *Bulletin of the American Mathematical Society*, 43(2):207–226, 2006.

DRAFT 14. srpna 2009 14:54

[5] C. Alsina, "On a method of Pi-Calleja for describing additive generators of associative functions". *Aequationes Mathematicae*, 43:14–20, 1992.

[6] C. Alsina, M.J. Frank, and B. Schweizer, "Problems on associative functions". *Aequationes Mathematicae*, 66(1–2):128–140, 2003.

[7] C. Alsina, M.J. Frank, and B. Schweizer, *"Associative Functions: Triangular Norms and Copulas"*. World Scientific, Singapore, 2006.

[8] J.P. Bézivin and M.S. Tomás, "On the determination of strict t-norms on some diagonal segments". *Aequationes Mathematicae*, 45:239–245, 1993.

[9] W. Blaschke and G. Bol, *"Geometrie der Gewebe, topologische Fragen der Differentialgeometrie"*. Springer, Berlin, Germany, 1939.

[10] C. Burgués, "Sobre la sección diagonal y la región cero de una t-norma". *Stochastica*, 5:79–87, 1981.

[11] R. Cignoli, F. Esteva, L. Godo, and A. Torrens, "Basic fuzzy logic is the logic of continuous t-norms and their residua". *Soft Computing*, 4:106–112, 2000.

[12] R. Craigen and Z. Páles, "The associativity equation revisited". *Aequationes Mathematicae*, 37:306–312, 1989.

[13] W. Darsow and M. Frank, "Associative functions and Abel-Schroeder systems". *Publicationes Mathematicae Debrecen*, 31:253–272, 1984.

[14] W.M. Faucett, "Compact semigroups irreducibly connected between two idempotents". *Proceedings of the American Mathematical Society*, 6:741–747, 1955.

[15] P. Hájek, "Basic fuzzy logic and BL-algebras". *Soft Computing*, 2:124–128, 1998.

[16] P. Hájek, *"Metamathematics of Fuzzy Logic"*. Kluwer, Dordrecht, 1998.

[17] S. Jenei, "On the convex combination of left-continuous t-norms". *Aequationes Mathematicae*, 72(1–2):47–59, 2006.

[18] S. Jenei, "On the Geometry of Associativity". *Semigroup Forum*, 74(3):439–466, 2007.

[19] C. Kimberling, "On a class of associative functions". *Publicationes Mathematicae Debrecen*, 20:21–39, 1973.

[20] E.P. Klement, R. Mesiar, and E. Pap, *"Triangular Norms"*, vol. 8 of *Trends in Logic*. Kluwer Academic Publishers, Dordrecht, Netherlands, 2000.

[21] C.M. Ling, "Representation of associative functions". *Publicationes Mathematicae Debrecen*, 12:189–212, 1965.

[22] R. Lowen, *"Fuzzy Set Theory". Basic Concepts, Techniques, and Bibliography.* Kluwer Academic Publishers, Dordrecht, Netherlands, 1996.

[23] J. Łukasiewicz, "O logice tró jwartościowej (On Three-valued Logic)". *Ruch Filozoficzny*, 5:170-171, 1920, (in Polish).

[24] K.C. Maes and B. De Baets, "On the structure of left-continuous t-norms that have a continuous contour line". *Fuzzy Sets and Systems*, 158(8):843–860, 2007.

[25] K.C. Maes and B. De Baets, "The triple rotation method for constructing t-norms". *Fuzzy Sets and Systems*, 158:(15)1652–1674, 2007.

[26] R. Mesiar and A. Mesiarová-Zemánkov á, "Convex combinations of continuous t-norms with the same diagonal function". *Nonlinear Analysis: Theory, Methods & Applications*, 69(9):2851–2856, 2008.

[27] P.S. Mostert and A.L. Shields, "On the structure of semigroups on a compact manifold with boundary". *Annals of Mathematics*, 65:117–143, 1957.

[28] M. Navara and M. Petrík, "Two methods of reconstruction of generators of continuous t-norms". *12th International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Málaga, Spain, 2008.

[29] M. Navara, M. Petrík, and P. Sarkoci, "Explicit formulas for generators of triangular norms". 2009. Submitted.

[30] V. Novák, I. Perfilieva, and J. Močkoř, *"Mathematical Principles of Fuzzy Logic"*. Kluwer Academic Publishers, Dordrecht, Netherlands, 1999.

[31] Y. Ouyang and J. Fang, "Some observations about the convex combinations of continuous triangular norms". *Nonlinear Analysis*, 2007.

[32] Y. Ouyang, J. Fang, and G. Li, "On the convex combination of T_D and continuous triangular norms". *Information Sciences*, 177(14):2945–2953, 2007.

[33] M. Petrík, "Convex combinations of strict t-norms". *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 2009. Accepted.

DRAFT 14. srpna 2009, 14:32

[34] M. Petrík and P. Sarkoci, "Convex combinations of nilpotent triangular norms". *Journal of Mathematical Analysis and Applications*, 350:271–275, 2009. DOI: 10.1016/j.jmaa.2008.09.060

[35] P. Pi-Calleja, "Las ecuacionas funcionales de la teoría de magnitudes". *Segundo Symposium de Matemática, Villavicencio, Mendoza*, Coni, Buenos Aires, 199–280, 1954.

[36] E. Post, "Introduction to a general theory of elementary propositions". *American Journal of Mathematics*, 43:163–185, 1921.

[37] B. Schweizer and A. Sklar, *"Probabilistic Metric Spaces"*. North-Holland, Amsterdam 1983; 2nd edition: Dover Publications, Mineola, NY, 2006.

[38] M.S. Tomás, "Sobre algunas medias de funciones asociativas". *Stochastica*, XI(1):25–34, 1987.

[39] T. Vetterlein, "Regular left-continuous t-norms". *Semigroup Forum*, 77(3):339–379, 2008.

[40] L.A. Zadeh, "Fuzzy sets". *Information and Control*, 8:338–353, 1965.

DRAFT
14. srpna 2009 - 14:32

Mezinárodní klasifikace nemocí a její využití v Minimálním datovém modelu pro kardiologii

doktorand:
MGR. PETRA PŘEČKOVÁ
Oddělení medicínské informatiky
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
preckova@euromise.cz

školitel:
PROF. RNDR. JANA ZVÁROVÁ, DRSC.
Oddělení medicínské informatiky
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
zvarova@euromise.cz

obor studia:
Biomedicínská informatika

Článek vzniknul s podporou projektu 1M06014 MŠMT ČR.

Abstrakt

Práce popisuje Mezinárodní klasifikaci nemocí, její historii, obsah a uspořádání. Dále se tento příspěvek věnuje Minimálnímu datovému modelu pro kardiologii (MDMK) a využití Mezinárodní klasifikace nemocí (MKN) v tomto modelu. Závěrem se zaměřuje na možnosti klasifikačního systému SNOMED CT a MKN verze 10 pro sémantickou interoperabilitu v českém jazykovém prostředí.

Klíčová slova: Mezinárodní klasifikace nemocí, Minimální datový model pro kardiologii, sémantická interoperabilita

1. Úvod

Jak již bylo zmíněno v [1], [2], [3], vymezení, pojmenování a třídění lékařských pojmů není optimální. Pro jeden termín existuje často mnoho synonym. Tato synonymie v odborné terminologii vede k nepřesnostem a nedorozumění. Z tohoto důvodu začaly vznikat klasifikační a kódovací systémy, které této variabilitě vyjadřování zamezují tak, že každý termín má svůj pevně stanovený formální kód.

Podmínkou spolehlivosti informací je co nejdokonalejší klasifikace jevů. Složitost uspořádání klasifikace, zvláště pak mezinárodní, spočívá v tom, že jiné požadavky mají lékaři nebo odborní lékaři (specialisté) působící v ambulantní péči, jiné lékaři v nemocnicích, zcela jiné pak pracovníci vysoce specializovaných pracovišť a výzkumných ústavů. Některé požadavky mohou vycházet i od nezdravotnických organizací a institucí. A tak se stalo, že v současné době existuje více než 100 různých klasifikačních medicínských systémů a mezi jeden z nejstarších patří i dále popisovaná Mezinárodní klasifikace nemocí.

2. Mezinárodní klasifikace nemocí

Mezinárodní klasifikace nemocí a přidružených zdravotních problémů (MKN) [4], [5], [6] je českým překladem International Classification of Diseases and Related Health Problems (ICD). Jedná se o klasifikaci kódující lidská onemocnění, příčiny smrti, zdravotní problémy a další příznaky. MKN se používá k převodu diagnóz nemocí a jiných zdravotních problémů ze slovní podoby do alfanumerického kódu. Její základ byl položen již v roce 1893 při klasifikaci příčin úmrtí s cílem umožnit mezinárodní porovnání. V roce 1948 převzala tuto klasifikaci Světová zdravotnická organizace WHO (World Health Organisation) a rozšířila ji o další diagnózy. Postupně tak začala vznikat všestranná pomůcka pro řízení zdravotnické politiky a pro výkaznictví ve vztahu ke zdravotnickým pojišťovnám a obdobným platebním systémům. Obsah MKN umožňuje systematické zaznamenávání, analýzu, výklad a porovnávání dat o úmrtnosti a nemocnosti, která jsou shromážděna v různých zemích nebo oblastech a v rozdílném čase.

MKN se musí přizpůsobovat vývoji požadavků současné lékařské vědy, aby mohla poskytovat odpovídající informace. Na druhé straně se však od ní požaduje, aby byla stabilizovaná v dostatečně dlouhém časovém období, aby byla jednotná pro celý svět, protože jen tak může sloužit jako základ pro srovnávání nemocnosti populačních skupin i geneticky odlišných, žijících v různých podmínkách a poskytovat i informace o dlouhodobých trendech vývoje. Kompromisem mezi těmito protichůdnými požadavky bylo přijetí zásady revizí. V současné době se využívá již Desáté revize.

2.1. Historie

Jak již bylo zmíněno výše, předchůdcem MKN byl Mezinárodní seznam příčin úmrtí (International List

DRAFT

of Causes of Death), který v roce 1893 prosadil francouzský lékař Jacques Bertillon na konferenci Mezinárodního statistického institutu (International Statistical Institute) v Chicagu v USA. Tento statistický systém začalo využívat mnoho států a v roce 1898 ho Americká asociace veřejného zdraví (American Public Health Association) (APHA) doporučila k oficiálnímu používání matrikářům v Kanadě, Mexiku a Spojených státech amerických. Zároveň tato asociace doporučila, aby docházelo k pravidelným revizím vždy po deseti letech.

V roce 1900 svolala francouzská vláda první mezinárodní konferenci, jejímž cílem byla revize Klasifikace příčin úmrtí. V této době se jednalo o jednu, ne příliš objemnou knihu, která byla doplněna abecedním rejstříkem. Další konference byly svolány v roce 1910, 1920, 1929 a 1938. Až do páté revize byly prováděny pouze dílčí změny v obsahu, bez zásadního zásahu do struktury. Po smrti Bertillona v roce 1922 byla ustanovena "Smíšená komise", která byla složena ze zástupců Mezinárodního statistického institutu a Zdravotní organizace Společnosti národů (Health Organization of the League of Nations), která připravovala podklady a návrhy k jednání konferencí.

V průběhu let vzniklo v jednotlivých zemích mnoho doplňků a rozšíření, z nichž některé rozšiřovaly klasifikaci příčin úmrtí i o klasifikaci nefatálních nemocí, ale do mezinárodní verze nebyly dlouho přijaty. V roce 1938 ale mezinárodní konference přijala rezoluci, která obsahovala doporučení, aby byly různé národní seznamy v maximální možné míře zapracovány do Mezinárodní klasifikace příčin úmrtí.

V roce 1948 převzala za klasifikaci zodpovědnost Světová zdravotnická organizace a šestou revizí, o níž jednala mezinárodní konference v Paříži, započala přeměna systému v univerzální seznam diagnóz. Název byl změněn na "Manual of International Statistical Classification of Diseases, Injuries and Causes of Death" v českém překladu "Mezinárodní statistická klasifikace nemocí, úrazů a příčin úmrtí" (MKN). Klasifikace byla vydána ve dvou dílech a obsahovala už i klasifikaci duševních poruch. Další konference se konaly v letech 1955, 1965 a 1975. Od sedmé revize zaujaly nefatální nemoci v tomto seznamu rovnocenné místo a MKN zahrnula i kódy dalších okolností, které ovlivňují kontakt se zdravotnickými službami.

V současné době se využívá desátá revize MKN (MKN-10). V České republice je tato klasifikace v platnosti od roku 1994. Ukázalo se ale, že stanovený desetiletý interval mezi revizemi byl příliš krátký. Práce na revizním procesu musely být zahájeny dříve, než byla platná verze MKN používána tak dlouho, aby mohla být důkladně zhodnocena. Potřeba konzultovat s mnoha zeměmi a organizacemi činí tento proces velmi zdlouhavým. První koncept jedenácté revize MKN (MKN-11) je tedy očekáván až kolem roku 2010 a vydání MKN-11 asi o 5 let později.

2.2. Obsah a uspořádání MKN-10

MKN má podobu číselníku. Ve verzi MKN-9 byly kódy diagnóz trojciferná čísla. Jednotlivé výseky číselné řady odpovídaly skupinám nemocí a stavů. Rozšířená verze, ICD-9-CM obsahovala navíc E-kódy vyjadřující vnější příčiny úrazů a jejich čísla byla ze stejné části číselné řady jako kódy pro úrazy a V-kódy označovaly další faktory ovlivňující zdravotní stav nebo kontakt se zdravotnickými službami. Tyto kódy odpovídají Z-kódům v MKN-10.

Jádrem klasifikace MKN-10 je třímístný kód, který je povinnou úrovní kódování pro mezinárodní hlášení o úmrtnosti pro databázi Světové zdravotnické organizace a pro všeobecné mezinárodní srovnávání. V MKN-10 je prvním znakem zleva vždy velké písmeno latinské abecedy, které udává hlavní kategorii. Znaky na druhém a třetím místě určují hlavní skupinu diagnóz. Za tečkou na čtvrtém, případně i dalším místě, následuje podrobnější členění. Výsledkem je více než zdvojnásobení kódovacích možností ve srovnání s devátou revizí. Z 26 možných písmen bylo použito 25. Písmeno U bylo ponecháno volné pro doplňky a změny a pro možné prozatímní klasifikace k vyřešení potíží, které mohou vzniknout mezi revizemi. Kódy U00-U49 se mohou používat pro prozatímní přidělení novým nemocem nejisté etiologie. Kódy U50-U99 mohou být použity ve výzkumech, například zkouší-li se možnosti alternativního podtřídění pro zvláštní projekt.

2.3. Kategorie MKN-10

Mezinárodní klasifikace se člení do těchto kategorií:

- Infekční a parazitální nemoci (A, B),
  - např. A84.1 – československá encefalitida přenášená klíšťaty,
  - B17.1– hepatitida typu C,
- novotvary (C),
  - C15.5 – zhoubný novotvar dolní třetiny jícnu,
- novotvary, nemoci krve a imunity (D),
  - D52.1 – anémie z nedostatku kyseliny listové, vyvolaná léky,

- nemoci endokrinní a metabolické (E),
  - E66.1 – obezita způsobená léky,
- nemoci duševní a poruchy chování (F),
  - F20.0 – paranoidní schizofrenie,
- nemoci nervové soustavy (G),
  - G47.1 – poruchy nadměrné spavosti,
- nemoci oka a očních adnex, nemoci ucha (H),
  - H11.2 – jizvy spojivky,
- nemoci oběhové soustavy (I),
  - I13.0 – hypertenzní nemoc srdce a ledvin s (městnavým) selháním srdce,
- nemoci dýchací soustavy (J),
  - J37.0 – chronická laryngitida,
- nemoci trávící soustavy (K),
  - K70.4 – alkoholická cirhóza jater,
- nemoci kůže a podkožního vaziva (L),
  - L70.0 – acne vulgaris,
- nemoci svalové a kosterní soustavy (M),
  - M24.2 – poruchy vazů,
- nemoci močové a pohlavní soustavy (N),
  - N21.1 – kámen v močové trubici,
- těhotenství, porod, šestinedělí, perinatální stavy, vrozené vady, deformace (O, P, Q),
  - O30.2 – těhotenství čtyřčetné,
  - P05.1 – malý plod vzhledem k délce těhotenství,
  - Q12.0 – vrozený základ,
- příznaky, znaky a nálezy nezařazené jinde (R),
  - R78.2 – nález kokainu v krvi
- poranění, otravy, následky působení vnějších příčin (S, T),
  - S42.0 – zlomenina klíční kosti,
  - T18.2 – cizí těleso v žaludku,
- zevní příčiny nemocí a úmrtí (V, W, X, Y),
  - V86.0 – řidič zcela terénního nebo jiného mimosilničního motorového vozidla, zraněný při provozní (silniční) nehodě,
  - X34 – pád ze skály - W15; oběť zemětřesení,
  - Y06.1 – zanedbání a opuštění rodičem,
- faktory ovlivňující zdravotní stav (Z),
  - Z54.2 – rekonvalescence po chemoterapii.

3. Minimální datový model pro kardiologii zakódovaný v MKN-10

V rámci Centra biomedicínské informatiky navazujeme na výzkum z našich předchozích projektů. V letech 2000-2004 bylo jedním z cílů výzkumného centra EuroMISE – Kardio sestavení Minimálního datového modelu pro kardiologii (MDMK) [7], [8], [9].

Jelikož je kardiologie velice rozsáhlý obor, byl MDMK zaměřen pouze na aterosklerotická kardiovaskulární onemocnění. Cílem tohoto datového modelu bylo vytvoření minimálního souboru znaků, které je potřeba sledovat u pacientů z hlediska aterosklerotického kardiovaskulárního onemocnění, aby mohl být pacient následně zařazen mezi osoby nemocné či rizikové. MDMK se skládá z několika skupin znaků. První část tvoří administrativní údaje, které jsou potřebné pro identifikaci pacienta. Další částí je rodinná anamnéza, zahrnující informace o matce, otci a libovolném počtu sourozenců. Dále následuje sociální anamnéza a toxikománie, která se zaměřuje na rodinný stav, fyzickou zátěž, psychickou zátěž, fyzické aktivity, míru kouření a míru požívání alkoholu. Část MDMK je věnována alergiím pacienta, zejména alergiím na léky. V části osobní anamnézy je zjišťována přítomnost diabetu mellitu, hypertenze, hyperlipoproteinémie, ischemické choroby srdeční a její konkrétní formy, je zjišťováno, zda pacient prodělal cévní mozkovou příhodu, zda se léčí s ischemickou chorobou periferních tepen, jsou zde atributy týkající se aneurysma aorty, ostatních relevantních chorob a u žen menopauzy. V části MDMK nazvané Současné obtíže možného kardiálního původu se lékaři zaměřují na dušnost, bolest na hrudi, palpitace, otoky, synkopu, kašel, hemoptýzu a klaudikaci. Další část MDMK zjišťuje, jakou léčbu pacient podstupuje, jaký má předepsaný druh diety a jaké užívá léky. V části fyzikálních vyšetření se zjišťuje pacientova hmotnost, výška, tělesná teplota, obvod boků, BMI, WHR, krevní tlak, tepová a dechová frekvence a patologické nálezy. Laboratorní vyšetření se zaměřují na glykémii, kyselinu močovou, celkový cholesterol, HDL-cholesterol, LDL-cholesterol a triaglyceroly. Poslední část MDMK tvoří atributy vztahující se k EKG, kde se zjišťuje rytmus, frekvence, průměrné intervaly PQ a QRS a je zde prostor pro celkový popis EKG.

DRAFT 14. srpna 2009, 14:32

1. část – Alergie				
Atributy z MDMKP	**Termín v MKN 10**	**Kód MKN 10**	**English equivalent**	**SNOMED CT** (Concept ID)
alergie přítomna	alergie	T78.4	allergy manifested	nenalezeno
alergie na léky	alergie na lék	T88.7	drug allergy (disorder)	416098002
			allergic reaction to drug (disorder)	416093006
2. část – Osobní anamnéza				
diabetes mellitus	diabetes typu I	E10.-	diabetes mellitus type 1 (disorder)	46635009
	inzulin dependentní	E10.-	insulin-treated non-insulin-dependent diabetes mellitus (disorder)	237599002
	těhotenský	O24.4	pregnancy and insulin-dependent diabetes mellitus (disorder)	237626009
hypertenze	Esenciální (primární) hypertenze	I10	essential hypertension (disorder)	59621000
hyperlipoproteinémie	hyperliprototeinemie	E78.5	hyperlipoproteinemia (disorder)	3744001
	Fredericksonova typu IV	E78.1	Fredrickson type IV hyperlipoproteinemia (disorder)	238085009
	Fredericksonova typu I	E78.3	Fredrickson type I hyperlipoproteinemia (disorder)	238086005
	Fredericksonova typu IIa	E78.0	Fredrickson type IIa hyperlipoproteinemia (disorder)	397915002
ischemická choroba srdeční - ICHS	ischemie koronární	I25.9	ischemic heart disease (disorder)	414545008
ICHS - němá ischemie	ischemie němá (asymptonická)	I25.6	silent myocardial ischemia (disorder)	233823002
ICHS - infarkt myokardu	infarkt myokardu, myokardiální (akutní nebo s dobou trvání 4 týdny nebo méně)	I21.9	myocardial infarction (disorder)	22298006
ICHS - srdeční selhání	selhání srdce akutní (náhlé)	I50.9	acute heart failure (disorder)	56675007
	selhání srdce městnavé	I50.0	acute congestive heart failure (disorder)	10633002
ICHS - arytmie	arytmie (srdeční)	I49.9	abnormal pulse rate (finding)	111972009
aneurysmata aorty	aneuryzma aorty	I71.9	aneurysm of aorta	nenalezeno
	aneuryzma aorty syfilitické	A52.0 + I79.0*	syphilitic aneurysm of aorta (disorder)	12232008
	aneuryzma aorty kongenitální	Q25.4	congenital aneurysm of aorta (disorder)	16972009
	aneuryzma aorty hrudní (oblouku)	I71.2	chronic dissecting aneurysm of thoracic aorta (disorder)	428326005
	aneuryzma aorty břišní	I71.4	repair of aneurysm of abdominal aorta (procedure)	405525004
	aneuryzma sestupné aorty	I71.9	aneurysm of descending aorta (disorder)	426948001
menopauza od	menopauza	N95.1	menopause present (finding)	289903006
	menopauza umělá	N95.3	artificial menopause (qualifier value)	67886002
	menopauza předčasná	E28.3	premature menopause NOS (qualifier value)	237789005
	menopauza chirurgická	N95.3	postsurgical menopause (disorder)	371036001
3. část – Současné potíže možného kardiovaskulárního původu				
dušnost	dušnost	R06.8	asthma (disorder)	187687003
bolest na hrudi	bolest hrudníku	R07.4	dull chest pain (finding)	3368006
palpitace	palpitace (srdce)	R00.2	(palpitations) or (awareness of heartbeat) or (fluttering of heart)	161965005
synkopa	synkopa srdeční	R55	syncope (disorder)	271594007
kašel	kašel	R05	cough	158383001
hemoptýza	hemoptýza	R04.2	haemoptysis	158384007

Tabulka 1: Vybrané atributy MDMK zakódované pomocí MKN-10 a SNOMED CT

DRAFT 14. SEPTEMBER 2009 – 14:32

Na základě MDMK byla vytvořena softwarová aplikace ADAMEK (Aplikace Datového Modelu EuroMISE centra – Kardio). Po jejím dokončení byl od března 2002 zahájen sběr dat v ambulanci preventivní kardiologie EuroMISE centra, která je spravována Městskou nemocnicí Čáslav. V současné době jsou v databázi ADAMEK zaznamenána data o 1289 pacientech.

Jelikož je Mezinárodní klasifikace nemocí jednou z mála mezinárodních medicínských klasifikací, které jsou přeložené do českého jazyka, pokusila jsem se zakódovat termíny Minimálního datového modelu právě pomocí této klasifikace, které uvádí tabulka 1. Pro srovnání jsou uvedeny rovněž kódy atributů MDMK v systému SNOMED CT.

Jak ze samotného názvu Mezinárodní klasifikace nemocí vyplývá, je možné tuto klasifikaci použít zejména pro zakódování nemocí, syndromů, patologických stavů, poranění, obtíží a jiných důvodů pro styk se zdravotnickými službami, tj. toho typu informací, které bývají registrovány lékařem. Bohužel, pomocí této klasifikace tedy nemůžeme zakódovat řadu atributů Minimálního datového modelu pro kardiologii, jako např. rodinný stav, vzdělání, psychickou zátěž, fyzickou zátěž, tělesnou aktivitu, kouření, pití alkoholu, fyzikální vyšetření (hmotnost, výška, tělesná teplota, obvod pasů, obvod boků, BMI, WHR, atd.), laboratorní vyšetření (celkový cholesterol, HDL-cholesterol) a ani popis EKG. MNK se hodí pouze pro části Minimálního datového modelu pro kardiologii týkající se osobní anamnézy a pro současné potíže možného kardiovaskulárního původu (viz tabulka 1).

4. Závěr

Základem sémantické interoperability heterogenních zdravotnických informačních systémů je mapování atributů těchto systémů na mezinárodně používané klasifikační systémy. Nespornou výhodou Mezinárodní klasifikace nemocí je její oficiální překlad do českého jazyka. Velikou nevýhodou je ale její omezení pouze na diagnózy a příznaky nemocí a tudíž nemožnost zakódovat všechny problémy nebo příčiny styku se zdravotnickými službami. Proto se jako výhodnější pro naše účely jeví mezinárodní klasifikační systém SNOMED CT, což je komplexní klinická terminologie, detailně popsána v [3], jejíž největší nevýhodou ale je její neexistence v českém jazyce a proto ji nelze použít ve zdravotnické praxi.

Literatura

[1] P. Přečková, Mezinárodní nomenklatury a metatezaury ve zdravotnictví. Doktorandský den 2005. MATFYZPRESS 2005, ISBN 80-86732-56-8. s. 109-116

[2] P. Přečková, Jazyk lékařských zpráv. Doktorandský den 2007. MATFYZPRESS 2007, ISBN 978-80-7378-019-7, s. 75-79

[3] P. Přečková, SNOMED CT a jeho využití v Minimálním datovém modelu pro kardiologii. Doktorandský den 2008. MATFYZPRESS 2008. ISBN 978-80-7378-054-8. s. 99-105

[4] Mezinárodní statistická klasifikace nemocí a přidružených zdravotních problémů. Desátá revize. Instruktážní příručka. ÚZIS ČR. 1996.

[5] http://www.uzis.cz/cz/mkn/

[6] http://www.who.int/classifications/icd/en/

[7] J. Adášková, Z. Anger, M. Aschermann, V. Bencko, P. Berka, J. Filipovský, L. Goláň, T. Grus, H. Grünfeldová, T. Haas, P. Hanuš, P. Hanzlíček, I. Holcátová, K. Hrach, R. Jiroušek, E. Kejřová, D. Kocmanová, J. Kolář, P. Kotásek, E. Králíková, M. Krupařová, M. Kyloušková, M. Malý, R. Mareš, M. Matoulek, I. Mazura, V. Mrázek, L. Novotný, Z. Novotný, L. Pecen, J. Peleška, M. Prázný, P. Pudil, J. Rameš, J. Rauch, J. Reissigová, H. Rosolová, B. Rousková, A. Říha, P. Sedlak, A. Slámová, P. Somol, Svačina, V. Svátek, D. Šabík, S. Šimek, J. Škvor, J. Špidlen, J. Štochl, M. Tomečková, V. Umnerová, K. Zvára, J. Zvárová: Návrh minimálního datového modelu pro kardiologii a softwarová aplikace ADAMEK. Interní výzkumná zpráva EuroMISE Centra – Kardio. Praha, říjen 2002.

[8] M. Tomečková: Minimální datový model kardiologického pacienta – výběr dat. Cor et Vasa, 2002, Vol. 44, No. 4 Suppl., s. 123

[9] R. Mareš, M. Tomečková, J. Peleška, P. Hanzlíček, J. Zvárová: Uživatelská rozhraní pacientských databázových systémů – ukázka aplikace určené pro sběr dat v rámci Minimálního datového modelu kardiologického pacienta. Cor et Vasa, 2002, Vol. 44, No. 4 Suppl., s. 76

DRAFT 14. Září 2009, 14:32

Experimenty s RDF úložištěm dat a reputacemi zdrojů

doktorand:
ING. MARTIN ŘIMNÁČ
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
rimnacm@cs.cas.cz

školitel:
ING. JÚLIUS ŠTULLER, CSC.
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
stuller@cs.cas.cz

obor studia:
Databázové systémy

Práce byla částečně podpořena projektem 1M0554 Ministerstva školství, mládeže a tělovýchovy ČR "Pokročilé sanační technologie a procesy" a záměrem AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

Vize sémantického webu se v dnešní době ponejvíce uplatňuje při popisu struktury a sémantiky dat ve formě ontologií. Prezentovaná práce se snaží ukázat, že je vhodné ji použít nejen pro popis dat, ale i pro jejich samotné sdílení. Za tímto účelem navrhuje samoseorganizující úložiště RDF dat - trojic (předmět, vlastnost, hodnota). Data poskytovaná různými zdroji mohou být vzájemně nekonzistentní (tak, jak je běžné u klasických webových zdrojů). Tato nekonzistence může být způsobena různorodými okolnostmi, počínaje chybným či nepřesným přiřazením, přes aktualizaci dat až po záměrné vkládání lživých informací.

RDF úložiště může být součástí většího propojeného celku - prostředí. Toto prostředí umožňuje vzájemně sdílet jak popis dat, tak identifikátory (URL) samotných objektů popisovaných těmito daty. V takovém prostředí je možné jednoduše vytvářet inverzní indexy (které úložiště prezentuje informace o daném objektu) a na jejich základě při dotazování poskytnout co možná nejúplnější možnou odpověď. Ta přirozeně může obsahovat i nekonzistetní části.

Neboť není žádoucí automaticky řešit nekonzistenci dat, práce navrhuje ohodnotit jednotlivé nekonzistetní části pomocí nepřímé míry. Tato míra je založena na reputacích dílčích úložišť prezentující hodnocenou část informace. Reputace úložiště se odvijí od mnohých faktorů, například:

- průniku prezentovaných dat,
- potvrzování dat různými zdroji (aktualizace dat),
- nekonzistence dat,
- aktualizace dat monotonně se vyvíjejích procesů.

Zatímco úložiště poskytující identická data, respektive úložiště, jejihž časté a včasné aktualizace dat jsou později potvrzovány ostatními (pomalejšími) úložištěmi, budou mít vysoké reputace, úložiště běžně nabízející údaje zcela odlišné od ostatních budou mít reputaci minimální. Speciálním příspěvkem k reputaci může být i aktuálnost prezentování dynamicky se vyvíjejících dat. Ta však může být definována jen pro procesy, které se vyvíjejí monotonně (existuje uspořádání stavů takové, že proces může přejít pouze do stavů v uspořádání následujících za aktuálním stavem). Tento parametr může razatně ovlivnit celkovou reputaci daného úložiště, neboť jakákoliv nově aktualizovaná informace je vždy nekonzistetní proti zastaralému stavu.

Cílem příspěvku je podrobné představení experimentální části z kapitoly [1]. Kapitola se obecně zaměřuje na inkrementální odhad struktury dat a věnuje se problematice sdílení RDF trojic v prostředí úložišť včetně diskuze použití reputačních systémů pro hodnocení kvality úložišť. Další experimenty včetně sledování monotonních procesů jsou součástí nové připravované kapitoly [2] rozšiřující poznatky z předchozí kapitoly do aplikace v prostředí sémantického webu.

Literatura

[1] M. Řimnáč a R. Špánek, Automated Incremental Building of Weighted Semantic Web Repository In *Studies in Computational Intelligence*, vol. 6, pp. 265-296, ISBN 978-3-642-01090-3, Springer Berlin, 2009.

[2] M. Řimnáč a R. Špánek, Experimental Framework for Self-Organized Incrementally Built RDF Repository In *Object-Oriented Data Modelling and Conceptual Design: Instance-Level Approaches*, IGI Global, (submitted).

DRAFT 14. srpna 2009 - 14:35

Pose Estimation Algorithms Based on Particle Filters

Post-Graduate Student:
MGR. STANISLAV SLUŠNÝ
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
slusny@cs.cas.cz

Supervisor:
MGR. ROMAN NERUDA, CSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
roman@cs.cas.cz

Field of Study:
Software Systems

Abstract

The robot localization problem is a fundamental and well studied problem in robotics research. Algorithms used to estimate pose on the map are usually based on Kalman or particle filters. These algorithms are able to cope with errors, that arise due to inaccuracy of robot sensors and effectors. The performance of the localization algorithm depends heavily on their quality.

This work shows performance of localization algorithm based on particle filter with small miniature low-cost E-puck robot. Information from VGA camera and eight infrared sensors are used to correct estimation of the robot's pose.

1. Introduction

The robot localization problem is a fundamental and well studied problem in robotics research. Several algorithms are used to estimate pose on the known map and cope with with errors, that arise due to inaccuracy of robot sensors and effectors. Their performance depends heavily on quality of robot's equipment: the more precise (and usually more expensive) sensors, the better results of localization procedure.

This work deals with localization algorithm based on particle filter with small miniature low-cost E-puck robot. Information from cheap VGA camera and eight infrared sensors are used to correct estimation of the robot's pose. To achieve better results, several landmarks are put into the environment. We assume, that robot knows the map of the environment in advance (distribution of obstacles and walls in the environment and position of the landmarks). We do not consider the more difficult simultaneous localization and mapping (SLAM) problem in this work (the case, when robot does not know it's own position in advance and does not have the map of the environment available).

E-puck is a widely used robot for scientific and educational purposes - it is open-source and low-cost. Despite its cheapness and limited sensor system, localization can be successfully implemented, as will be shown in this article.

2. Introducing E-puck robot

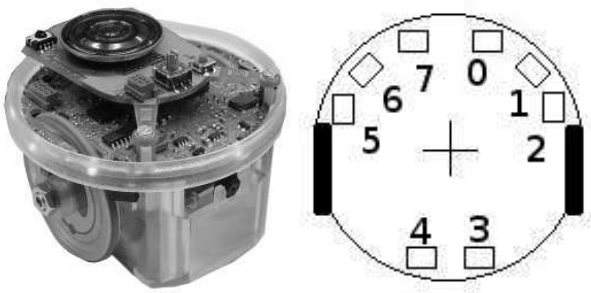


Figure 1: Miniature e-puck robot has eight infrared sensors and two motors.

E-puck (Figure 1) is a mobile robot with a diameter of 70 mm and a weight of 50 g. The robot is supported by two lateral wheels that can rotate in both directions and two rigid pivots in the front and in the back. The sensory system employs eight "active infrared light" sensors distributed around the body, six on one side and two on other side. In "passive mode", they measure the amount of infrared light in the environment, which is roughly proportional to the amount of visible light. In "active mode" these sensors emit a ray of infrared light and measure the amount of reflected light. The closer they are to a surface (the e-puck sensors can detect a white paper at a maximum distance of approximately 8 cm), the higher is the amount of infrared light measured. Unfortunately, because of their imprecision and characteristics (see Figure 2), they can be used as bumpers only. As can be seen, they provide

DRAFT
Version 2009-12-32

high resolution only within few milimeters. They are very sensitive to the obstacle surface, as well. Besides infrared sensors, robot is equipped with low-cost VGA camera. The camera and image processing will be described in the following section.

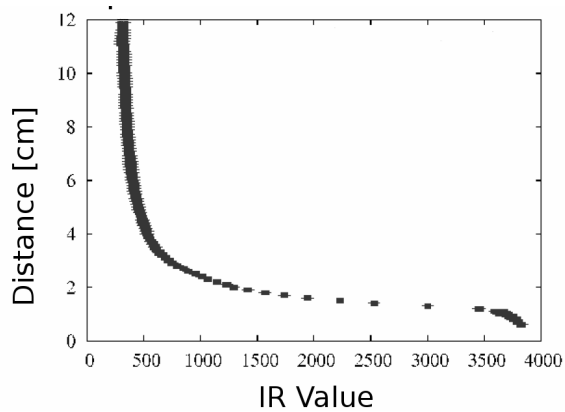


Figure 2: Multiple measurements of front sensor. E-puck was placed in front of the wall at a given distance and average IR sensor value from 10 measurements was drown into the graph.

Two stepper motors support the movement of the robot. A stepper motor is an electromechanical device which converts electrical pulses into discrete mechanical movements. It can divide a full rotation into a 1000 steps, the maximum speed corresponds to about a rotation every second.

3. Related work

Algorithms for robot localization are described in [1]. Most popular approaches are based on on Kalman filter (or some of its variants) or particle filter (PF). Both approaches have their pros and cons. Particle filters are very easy to implement, but approximate posterior probability by random sample of states. Algorithms based on Kalman filter rely on a fixed functional form of the posterior, but tend to work only if the position uncertainty is small.

4. Dead reckoning

Dead reckoning (derived originally from deduced reckoning) is the process of estimating robot's current position based upon a previously determined position. For shorter trajectories, position can be estimated using shaft encoders and precise stepper motors.

E-puck is equipped with a differential drive (Figure 3) - a simplest method to control robot. For a differential drive robot the position

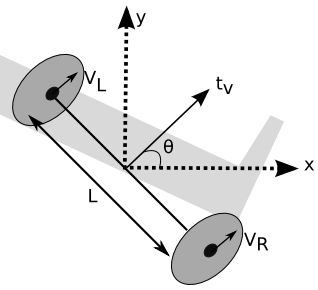


Figure 3: Differential drive robot schema.

of the robot can be estimated by looking at the difference in the encoder values Δs_R and Δs_L. By estimating the position of the robot, we mean the computation of tuple x, y, Θ as a function of previous position $(x_{OLD}, y_{OLD}, \Theta_{OLD})$ and encoder values (Δs_R and Δs_L).

$$\begin{pmatrix} x \\ y \\ \theta \end{pmatrix} = \begin{pmatrix} x_{OLD} \\ y_{OLD} \\ \theta_{OLD} \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta\theta \end{pmatrix} \tag{1}$$

$$\Delta\theta = \frac{\Delta s_R - \Delta s_L}{L} \tag{2}$$

$$\Delta s = \frac{\Delta s_R + \Delta s_L}{2} \tag{3}$$

$$\Delta x = \Delta s.cos(\theta + \frac{\Delta\theta}{2}) \tag{4}$$

$$\Delta y = \Delta s.sin(\theta + \frac{\Delta\theta}{2}) \tag{5}$$

The major drawback of this procedure is error accumulation. At each step (each time you take an encoder measurement), the position update will involve some error. This error accumulates over time and therefore renders accurate tracking over large distances impossible (see Figure 4). Tiny differences in wheel diameter will result in important errors after a few meters, if they are not properly taken into account.

Parameters	Value
Maximum translational velocity	12.8 cm / sec
Maximum rotational velocity	4.86 rad / sec
Stepper motor maximum speed	+- 1000 steps / sec
Distance between tires	5.3 cm

Table 1: Velocity parameters of E-puck mobile robot.

DRAFT 14. srpna 2009 - 14:32

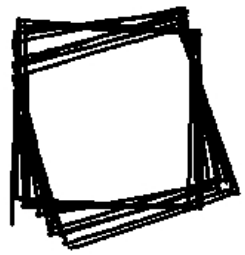


Figure 4: Illustration of error accumulation. Robot was ordered to make 10 squares of size 30 cm. Odometry errors are caused mostly by rotation movement.

5. Image Processing

The robot has a low-cost VGA camera with resolution of 480x640 pixels. Unfortunately, the Bluetooth connection supports only a transmission of 2028 colored pixel. For this reason a resolution of 52x39 pixels maximizes the Bluetooth connection and keeps a 4:3 ratio. This is the resolution we have used in our experiments (see Figure 5). Another drawback of the camera is that it is very sensitive to the light conditions.

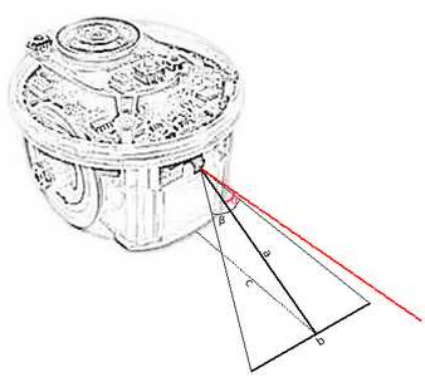


Figure 5: The physical parameters of the real camera (picture taken from [2]). Camera settings used in experiments corresponds to parameters $a = 6$ cm, $b = 4.5$ cm, $c = 5.5$ cm, $\alpha = 0.47$ rad, $\beta = 0.7$ rad.

Despite these limitations, camera can be used to detect objects or landmarks. However, the information about distance to the landmark extracted from the camera is not reliable (due to the noise), and we do not use it in following section.

Landmarks are objects of rectangular shape of size 5x5 cm and three different colors - red, green and blue. We implemented image processing subsystem, that detects relative position of the landmark from the robot. Following steps are included:

- Gaussian filter is used to reduce camera noise.
- Color segmentation into the red, blue and green color.
- Blob detection is used to detect position and size of the objects on the image.
- Object detection is used to remove objects from image, that have non-rectangular shape.

Output from the image processing is the relative position and color of the detected landmarks (for example - I see red landmark by angle 15 degrees).

6. Monte-Carlo Localization

As was shown, pose estimation based on dead reckoning is possible for short distances only. For longer trajectories, more clever methods are needed.

The PF possesses three basic steps - state prediction, observation integration and resampling. It works with quantity $p(x_t)$ - the probability, that robots is located at the position x_t in time t. In the case of PF, the probability distribution is represented by the set of particles. Such a representation is approximate, but can represent much broader space of distributions that, for example, Gaussians, as it is nonparametric.

Each particle $x_t^{[m]}$ is a hypothesis, where the robot can be at time t. We have used M particles in our experiment. The input of the algorithm is the set of particles X_t, most recent control command u_t and the most recent sensor measurements z_t.

1. State prediction based on odometry.

   The first step is the computation of temporary particle set X from X_t. It is created by applying odometry model $p(x_t|u_t, x_{t-1})$ to each particle $x_t^{[m]}$ from X_t.

DRAFT 14. srpna 2009, 14:32

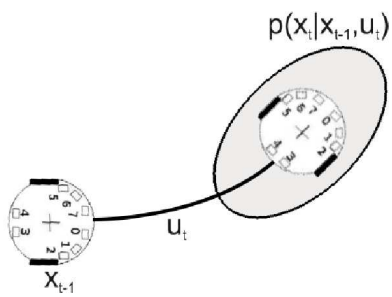


Figure 6: First step in PF algorithm - to each position hypothesis x_{t-1} is applied odometry model based on movement u_{t-1} and new hypothesis x_t is sampled from distribution $p(x_t|x_{t-1}, u_{t-1})$.

2. Correction step - Observation integration

   The next step is the computation of *importance factor* $w_t^{[m]}$. It is the probability of the measurement z_t under particle $x_t^{[m]}$, given by $w_t^{[m]} = p(z_t|x_t^{[m]})$.

   Two types of measurements were considered:

   - Measurement coming from distance sensors

     Distance sensor (one averaged value for front, left, right and back direction) were used as bumpers only. In case of any contradiction between real state and hypothesis, importance factor was decreased correspondingly.

   - Measurement obtained from image processing

     Output from image processing was compared with expected position of the landmarks. In case of any contradiction (colors and relative angle of landmarks were checked), importance factor was decreased. The bigger mismatch, the smaller importance factor was assigned to the hypothesis.

3. Re-sampling

   The last step incorporates so-called *importance sampling*. The algorithm draws with replacement M particles from temporary set X and creates new particle set X_{t+1}. The probability of drawing each particles is given by its importance weight. This principle is sometimes called *survival of the fittest*.

7. Experiments

Experiments were carried out in an arena of size 1x0.75 meters. Three landmarks (red, blue and green, one of each color) were placed into the arena, as shown on Figure 7 Robot was controlled by commands sent from

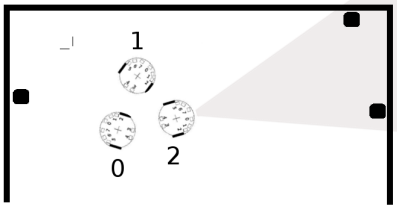


Figure 7: Second step in PF algorithm - each particle is assigned a importance factor, corresponding to the probability of observation z_t. If image processing detects two landmarks on the actual camera image, particles 0 and 1 will be assigned small weight.

computer, values from sensors were sent back to computer by using Bluetooth. Execution of each command took 64 miliseconds. The experiment started by putting robot into the arena and randomly distributing 2000 particles. After several steps, the PF algorithm relocated the particles into real location of the robot. The robot was able to localize itself. The convergence of the algorithm depends on the fact, if robot is moving near the wall or in the middle of the arena. The impact of infrared sensors was obvious. Algorithms were verified in the simulator ([11]) and in reality, as well. The video demonstration can be found at ([12]). The localization algorithm was able to cope with even bigger areas, up to the size of three meters. However, we had to add more landmarks to simplify the localization process. Localization algorithm showed satisfiable performance, relocating hypothesis near real robot pose.

8. Conclusions

Localization and pose estimation is an opening gate towards more sophisticated robotics experiments.As we have shown, the localization process can be carried out even with low-cost robot. Experiments were executed both in simulation and real environment. A lot of work remains to be done. The experiments in this work considered static environment only. Addition of another robot will make the problem much more difficult. As we have mentioned already, there are certain areas in

DRAFT 14 srpna 2009 - 14:32

the environment, where convergence of the localization algorithm is very fast - in corners or near walls. Sensor fusion is the process of combining sensory data from disparate sources such that the resulting information is in some sense better than would be possible when these sources were used individually.We are dealing with sensor fusion of infrared sensors and input from camera. As a future work, we would like to implement path planning, that takes into account performance of the localization algorithm. Suggested path (generated by path planning algorithm) should be safe (the chance to get lost should be small) and short. Multi-criterial path planning will be based on dynamic programming([13]). The idea is to learn areas with high loss probability from experience.

References

[1] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, MA: MIT Press, 2005.

[2] http://en.wikibooks.org/wiki/Cyberbotics_Robot_Curriculum/.

[3] E-puck, online documentation. http://www.e-puck.org.

[4] R.C. Arking, *Behavior-Based Robotics*. The MIT Press, 1998.

[5] L.G. Shapiro and G.C. Stockman, "Computer Vision", page 137, 150. Prentence Hall, 2001.

[6] J. Bruce, T. Balch, and M. Veloso, *Fast and Inexpensive Color Image Segmentation for Interactive Robots*. In Proceedings of IROS-2000, 2000, 2061–2066.

[7] http://www.v3ga.net/processing/BlobDetection/index-page-home.html.

[8] R.E. Kalman, *A new approach to linear filtering and prediction problems*. Trans. ASME, Journal of Basic Engineering 82:35-45.

[9] R.Y. Rubinstein, *Simulation and the Monte Carlo Method.*, John Wiley and Sons, Inc.

[10] K. Kanazawa, D. Koller, and S.J. Russel, *Stochastic simulation algorithms for dynamic probabilistic networks*. In Proceedings of the 11th Annual Conference on Uncertainty in AI, Montreal, Canada.

[11] Webots simulator. http://www.cyberbotics.com.

[12] Video demonstration. http://www.cs.cas.cz/slusny.

[13] R.S. Sutto and A. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, 1998.

DRAFT 14. srpna 2009 - - 14:52

Metody modularizace rozsáhlých ontologií

doktorand:
ING. PETRA ŠEFLOVÁ
Fakulta mechatroniky, informatiky a mezioborových studií,
Technická univerzita Liberec,
Hálkova 6,
461 17 Liberec 1
seflova@cs.cas.cz

školitel:
ING. JÚLIUS ŠTULLER, CSC.
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
stuller@cs.cas.cz

obor studia:
Technická kybernetika

Tento projekt je realizován za finanční podpory prostředků státního rozpočtu ČR prostřednictvím projektu *Pokročilé sanační technologie a procesy* č.1M0554 programu Výzkumná centra PP2-DP01 MŠMT.

Abstrakt

Ontologie jsou srdcem sémantického webu. S rostoucím využíváním ontologií v nejrůznějších oblastech vědy a průmyslu vznikají rozsáhlé ontologie (např. GALEN, DICE,..). S tím jak roste velikost ontologií je stále obtížnější tyto ontologie spravovat. V důsledku toho vznikl požadavek na možnost rozdělení rozsáhlých monolitických ontologií na menší ucelené části. Schopnost extrahovat smysluplnou část z rozsáhlé ontologie je základem pro znovupoužití ontologií při návrhu nových ontologií.

V tomto článku jsou prezentovány základy problematiky modularizace ontologií – co je chápáno pod pojmem modularizace, jaké jsou její cíle a jaké jsou v současné době používané metody.

1. Úvod

Modularita je klíčovým požadavkem pro mnoho úloh týkajících se *návrhu, údržby a integrace* ontologií, zejména *rozsáhlých ontologií*, při kterém obyčejně *spolupracuje mnoho návrhářů*, nebo při slučování nezávisle vy-vinutých ontologií do jedné. Bohužel, ve srovnání s jinými disciplínami, jako je např. softwarové inženýrství, kde jsou pojem a techniky modularizace již zavedeny a hojně využívány – v ontologiích je modularizace relativně nová.

Moderní ontologické jazyky, jako OWL, jsou založeny na logice (speciálně na deskripční logice), následkem čehož se jeví jako výhodné vzít v úvahu u pojmu modularizace zejména *sémantiku* ontologie danou příslušnou deskripční logikou a její důsledky.

Na modularizaci může být nahlíženo třemi rozdílnými způsoby [1].

- modularizace je proces, který *rozdělí rozsáhlou ontologii na menší části* (*moduly*). Tomuto procesu se říká *dekompozice* - výchozím bodem je celá ontologie; cílem jsou moduly.
- modularizace může být chápána jako *proces skládání menších ontologií (modulů) do jedné větší ontologie*. Výchozím bodem tohoto procesu je sada modulů; cílem je nová ontologie. Tento typ modularizace vyžaduje specifikaci mechanismů pro stavbu nové ontologie z jednotlivých modulů, jako jsou např. mapovací pravidla.
- alternativní vnímání modularizace je na úrovni návrhu. Základní myšlenkou je, že při návrhu nové ontologie již známe cílové moduly a proto současně s definicí jednotlivých prvků ontologie již definujeme do kterého modulu daný prvek patří. Tento způsob modularizace je vykonáván tzv. *za letu* jako vedlejší produkt návrhu ontologie.

V tomto článku se budeme hlavně zabývat procesem rozdělení rozsáhlé ontologie do sady menších ontologií – modulů, tedy dekompozicí.

Představme si, že chceme navrhnout ontologii O1, která bude popisovat výzkumné projekty. V této ontologii použijeme termíny jako je např. *Cystická Fibrosa* a *Vývojové poruchy* pro popis určitého lékařského projektu. Abychom zvýšili přesnost naší ontologie, chceme přidat další detaily o významech těchto pojmů, o kterých předpokládáme, že jsou již definované v jiné ontologii O2. Nechť je tato ontologie příliš velká, abychom ji mohli importovat jako celek. Proto v praxi potřebujeme z této rozsáhlé ontologie extrahovat pouze tu část ontologie (modul) M, která zahrnuje související pojmy. Ideálně, modul má být co nejmenší, ale tak, aby ještě zachytil význam použitého pojmu.

DRAFT 14. prosince 2009, 14:32

V úvodu článku je seznámení se základy problematiky modularizace ontologií, v další části jsou vymezeny cíle modularizace (část 2) a definice modulu (část 3). Část 4 seznámí s problematikou kritérií modularity, část 5 dává základní přehled používaných metod.

2. Cíle modularizace

Porozumění tomu, co modularizace přesně znamená a jaké jsou její výhody a nevýhody, které můžeme od modularizace ontologií očekávat, závisí na cílech [1] modularizace.

2.1. Rozšiřitelnost

Rozšiřitelnost je všeobecným cílem, který vidí modularizaci jako způsob jak udržet *'výkonnost'* návrhářů na rozumné úrovni. Základní myšlenkou je, že návrháři jsou *'dobře'* výkonní při návrhu menších ontologií, přičemž *s rostoucí velikostí* ontologií jejich *výkonnost klesá* a *zvyšuje se chybovost*. Tento cíl modularizace je většinou svázán s *dekompozičním přístupem*.

Z pohledu dekompozice můžeme rozšiřitelnost rozdělit do dvou podtémat :

- **Rozšiřitelnost pro získávání znalostí:**
  Tento cíl modularizace má lokalizovat vyhledávací prostor pro získávání znalostí uvnitř ohraničeného modulu.

- **Rozšiřitelnost pro vývoj a údržbu:**
  Tento cíl modularizace se soustředí na *dopad aktualizace* uvnitř ohraničeného modulu.

2.2. Opětovné použití

Opětovné použití je dobře známý cíl v softwarovém inženýrství. Opětovné použití je viděno jako základní motivace pro přístup skládání ontologií z menších modulů. Nicméně, lze ho aplikovat i na dekompoziční přístup, kde by měl vést k dekompozičním kritériím založených na předpokládané opětovné použitelnosti modulů.

2.3. Srozumitelnost

Značným problémem při zkoumání ontologie je schopnost porozumět jejímu obsahu. Je to jednodušší pokud je ontologie malá. Velikost, nicméně, ale není jediným kritériem, které má vliv na pochopení obsahu ontologie.

2.4. Personalizace

Vlastník informací je znám jako důležitý faktor, který musí být brán v úvahu při tvorbě kooperujících systémů. Toto můžeme aplikovat i na ontologii a to i přesto, že mnoho ontologií je viděno jako veřejně dostupné zdroje. Vlastník v těchto případech poskytuje kritéria pro dekompozici ontologie na menší části (moduly).

3. Definice modulu

Ačkoli nástroje pro návrh a správu ontologií mohou pracovat s ontologií skládající se z jednotlivých axiomů, z hlediska užitečnosti modul nemůže být libovolnou podmnožinou ontologie.

Modul je definován jako část ontologie, která *'dává smysl'* [1].

Může se jednat o smysl z *aplikačního hlediska*, t.j. modul je schopný poskytnout rozumnou odpověď alespoň na jeden dotaz, pro který je navržen. Nebo se může jednat o smysl z hlediska systému, t.j. modulární organizace je schopna *zlepšit výkon* alespoň jedné služby, které systém poskytuje. Neurčitost této definice se odráží na *subjektivní* povaze rozhodování o tom, co je a co není považováno za modul.

Definování modulu jako podontologie (*sub-ontology*) vysvětluje skutečnost, že ontologie je rozdělena do jednotlivých modulů. Obráceně modul může být považován za samostatnou ontologii pro účely, kdy není vyžadován přístup k jiným modulům v dané sadě.

Modulem mohou být *nezávisle vyvinuté ontologie*, které spolu tvoří novou ontologii. To je kompoziční přístup k modularizaci ontologií. Nebo na druhou stranu modul může být *vytvořen rozdělením existujíci ontologie*. Toto rozdělení může být „ruční“ nebo automatické pomocí některého z nástrojů pro správu ontologií.

4. Kritéria modularity

Najít vhodná kritéria pro dekompozici je značná výzva. Spoléhat se na člověka je nejjednodušší řešení, ale obecně není vždy uspokojující a navíc velmi závisí na zkušenostech daného člověka.

V dekompozici založené pouze na člověku je vhodné, aby se již při návrhu ontologie identifikovaly skupiny komponent, které mají být drženy pohromadě (v jednom modulu), než se poté dotazovat na umístění jednotlivých komponent ontologie (např. relací, axiomů,...).

Implementace automatické či poloautomatické dekompoziční strategie v aplikaci vyžaduje znalosti

DRAFT 14. srpna 2009 - 14:42

o požadavcích dané aplikace. Tato znalost může být získána, např. analýzou dotazů, které jsou adresovány ontologii a ukládání cest uvnitř ontologie, které jsou použity při odpovědi na dotaz. Četnost cest a jejich překrývání mohou vést k určení pravidel pro optimální dekompozici.

Přístup založený na výkonu může být viděn jako strategie, která pouze uvažuje systémové aspekty a ignoruje požadavky aplikace. Příklady dekompozice založené na výkonu mohou být algoritmy pro dekompozici grafu.

5. Metody modularizace

Základní rozdělení metod podle [3] :

1. metody založené na výběru
2. metody využívající síťové algoritmy
3. extrakce modulu založená na četnosti průchodů
4. segmentace ontologií

5.1. Metody založené na výběru

Mnoho prací je inspirováno oblastí databází k definování ontologických dotazů v syntaxi podobné SQL.

Metody založené na výběru poskytují dotazy jejichž vzhled je podobný dotazům SQL. Toto dělá tyto metody intuitivně blízké pro lidi pracující v oblasti databází.

Nedostatky těchto přístupů jsou, že poskytují pouze nízkoúrovňový přístup k sémantice ontologie v dotazování a neřeší otázky aktualizace originální ontologie v okamžiku, kdy je změněna část ontologie vyjmutá jako modul.

Tento přístup je vhodný pro jednorázové získání velmi malých částí ontologie, které jsou zaměřené na několik konceptů.

5.1.1. SparQL

Dotazovací jazyk SparQL [2] definuje jednoduchý dotazovací mechanismus pro RDF. SparQL může být dobrý nízkoúrovňový nástroj pro implementaci rozdělení ontologií, ale není řešením sám o sobě.

5.1.2. KAON pohledy

Volz a kolegové definovali mechanismus založený na RQL dotazovacím jazyku [19]. Zdůrazňují RQL pouze jako RDF dotazovací jazyk, který bere v úvahu sémantiku RDF schématu. Jejich systém pohledu má schopnost umístit každý koncept na odpovídající místo v kompletní RDF hierarchii.

5.1.3 RVL

Magkanaraki a kolegové presentují podobný přístup jako Volz, jejich systém dovoluje dotazy přebudovat v RDFS hierarchii, když vytvářejí daný pohled [11]. To dovoluje přizpůsobit pohledy za chodu aplikace podle specifických požadavků aplikace. Pohledy jsou kolekcí ukazatelů na aktuální koncept a přestávají existovat po té, co splní svůj učel.

5.2. Metody využívající síťové algoritmy

V oblasti sítí je využíván algoritmus pro uspořádání uzlů sítě do souvisejících oblastí [4]. Někteří výzkumníci v oblasti ontologií analyticky navrhli použití podobné metodiky pro rozdělení ontologií.

Ontologie může být z tohoto hlediska viděna jako síť vzájemně spojených uzlů. Třída hierarchie může být interpretována jako orientovaný acyklický graf (directed acyclic graph DAG) a každý vztah mezi třídami může být interpretován jako spojení mezi uzly.

5.2.1. Strukturové rozdělení

Stuckemschmidt a Klein presentují v [18] metodu rozdělení hierarchie tříd do modulu. Využívají hierarchickou strukturu tříd a omezení vlastností domény k rozložení ontologie do modulů dané velikosti. Tato metoda nebere v úvahu OWL omezení, která mohou být činná jako přidaná spojení mezi koncepty. Místo toho se spoléhá na globální tvrzení.

Tato metoda primárně slouží k rozdělení ontologie do balíčků nebo modulů, aby se ontologie mohla snáze udržovat a zveřejnit. Nicméně tento proces zruší původní ontologii, východiskem je rozložení do modulů vhodným algoritmem. Navíc ontologie modelované v OWL mají sklon být sémanticky bohatší než bude zachyceno jednoduchou síťovou abstrakcí.

5.2.2. Automatické rozdělení pomocí ϵ-spojení

Grau a jeho kolegové [7] představili metodu modularizace OWL ontologií podobnou přístupu Stuckemschidta a Kleina. Princip jejich přístupu spočívá v rozložení originální ontologie použitím ϵ-spojení [9] a po té drží jednotlivé moduly vzájemně propojené. Moduly získané pomocí tohoto algoritmu jsou formálně způsobilé k získání minimální sady atomických axiomů nezbytných pro udržení logických vazeb.

DRAFT 14. srpna 2009

Tato metodologie se nejeví jako vhodná pro modularizaci ontologií využívající vyšší ontologii (*upper-ontology*) [8]. Hodně rozsáhlých ontologií se spoléhá na *vyšší ontologii* k udržování vysoké úrovně organizování struktury. Z toho vyplývá, že přístup Graua a jeho kolegů má pouze omezene využití v reálném světě.

5.3. Extrakce založená na četnosti průchodů

Rozdělení ontologie pomocí četnosti průchodů, podobně jako síťové rozdělení, vidí ontologii jako síť/graf. Liší se ale v tom, že místo rozdělení celého grafu do modulu, tato metodologie začíná v jednotlivých uzlech (konceptech) a sleduje jejich spojení. Tím buduje seznam uzlů (konceptů) pro extrakci. Klíčový rozdíl je v tom, že zanechává strukturu originální ontologie nedotčenou.

Tento způsob využívají dvě metody.

5.3.1. PROMT

Noy - Musenova extrakční metoda je zaměřena na četnosti průchodů [14], která definuje, jak má být ontologie procházena. Soubor příkazů kompletně a jednoznačně definuje pohled na ontologii a může být sám uložen jako ontologie.

Noy a kolegové navrhli mechanismus extrakce ontologie, ale již neuvedli jak jejich systém může být použit k tvorbě vhodných segmentů.

Tento přístup implementovali jako plug-in modul do systému Protégé.

5.3.2. MOVE

Bhatt, Wouters a kolegové prezentují systém zhmotněného pohledu ontologické extrakce (Materialized Ontology View Extraktor - MOVE) pro distribuovanou extrakci podontologie (subontology) [20]. Jedná se o všeobecně použitelný systém, který může pracovat s libovolným formátem ontologie. Systém extrakce náhradní ontologie je založen na uživatelském označení, které pojmy z ontologie zahrne a které vyloučí. Také má schopnost optimalizovat extrakci založenou na sadě uživatelsky volitelných optimalizačních schématech. Tato schémata mohou být získána buď jako nejmenší možný extrakt nebo mohou zahrnovat tolik detailů, jak jen to je možné. Extrakce může být také omezena přidáním sady doplňujících omezení.

5.4. Segmentace ontologie

Základní segmentační algoritmus [3] začíná na jedné nebo více třídách vybraných uživatelem a vytváří extrakt založený na těchto třídách a souvisejících konceptech. Tyto související třídy jsou identifikovány pomocí struktury spojení ontologie.

Tato metoda je založena na 4 krocích:

1. Průchod směrem nahoru

Tato strategie se jeví jako vhodná, když konstruujeme pohled ontologie, ale není vhodná pro extrakci, která má být použita v aplikaci, protože každá nadtřída může obsahovat kritické informace.

2. Průchod směrem dolů

Tento algoritmus prochází ontologii směrem dolů od zvolené třídy, zahrnuje všechny její podtřídy.

3. Sourozenecké třídy v hierarchii

Sourozenecké třídy nejsou zahrnuty v extraktu. Je rozumné předpokládat, že nejsou dostatečně relevantní, aby byly zahrnuty standardně. Uživatel ale může vždy explicitně určit výběr tříd pro zahrnutí do extraktu.

4. Četnost průchodů vzestupně a sestupně podle spojení

V této chvíli již máme vybrané třídy podle cílové třídy, jejich omezení, průnik, spojení a ekvivalentní třídy. Nyní je potřeba zvážit, zda průnik a sjednocení tříd mohou být rozděleny do jednotlivých typů tříd a zda mohou být zpracovány odděleně.

6. Závěr

S rostoucí velikostí ontologií roste potřeba používat principy modularity k reprezentaci ontologických znalostí, aby se usnadnilo vytváření, údržba a využití těchto znalostí.

Tento článek se věnoval základnímu popisu toho co se skrývá pod pojmem modularizace ontologií, definicí jednotlivých pojmů a metod.

Plán na nejbližší období počítá s bližším seznámením s jednotlivými metodami modularizace, provedením jejich srovnání a případně s navržením vylepšení.

Literatura

[1] S. Spaccapietra and A. Tamilin, D2.1.3.1 Report on Modularization of Ontologies *KWEB/2004/D2.1.3.1/v1.1 July 30, 2005.*

DRAFT 4. srpna 2009 - 14:32

[2] A. Seaborne and E. Prud´hommeaux, "SparQL Query Languange for RDF", *Website reference: http://www.w3.org/TR/rdf-sparsql-query/, February 2005.*

[3] J. Seidenberg and A.Rector, "Web Ontology Segmentation: Analysis, Classification and Use", *WWW2006, Edinburgh,Scotland.*

[4] V. Bagatejl, "Analysis of large network islands", *Dagstuhl Semina 03361, University of Ljubljana, Slovenia, August 2003. Algorithmic Aspects of Large and Complex Network.*

[5] T. Bray, "What is RDF?", *Website reference: http://www.xml.com/pub/a/2001/01/24/rdf.html, January 2001.*

[6] S. Brin a L.Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Szstems, 30 (1-7):1007-117,1998.*

[7] B.C. Grau, B. Parsia, E. Sirin, and A. Kalyanpur, "Automatic Partitioning of OWL Ontologies Using E-Connections", *In International Worshop on Description Logics, 2005.*

[8] B.C. Grau, B. Parsia, E. Sirin, and A. Kalyanpur, "Modularizing OWL Ontologies", *In K-CAP 2005 Worshop on Ontology Management, October 2005.*

[9] I. Horrocks, P.F. Patel-Schneider, and F. Van Harmelen, "From SHIQ and RDF to OWL: The making of a web ontology language", *In Journal of Web Semantics, volume 1, pages 7-26, 2003.*

[10] O. Kutz, C. Lutz, F. Wolter, and M. Zakharyaschev, "E-connections of abstract description systems", *In Artifical Intelligence, volume 156, strana 1-73, 2004.*

[11] A. Magkanaraki, V. Tannen, V. Christophides, and D. Plexousakis, "Viewing the Semantic Web through RVL Lenses", *Journal of Web Semantics, 1(4):29, October 2004.*

[12] D.L. McGuinness and F. Van Harmelen, "OWL Web Ontology Language Overview", *February 2004, W3C Recommendation.*

[13] N. Noy and M.A. Musen, "The PROMPT Suite: Interactive Tools for Ontology Merging And Mapping", *International Journal of Human-Computer Studies, 59(6):983-1024, 2003.*

[14] N. Noy and M.A. Musen, "Specifying ontology views by traversal", *In S. A. McIlraith, D. Plexousakis, and F. Van Harmelen, editors, International Semantic Web Conference, volume 3298 of Lecture Notes in Computer Science, pages 713-725. Springer, November 2004.*

[15] A. Pease, I. Niles, and J. Li, "The suggested upper merged ontology: A large ontology for the semantic web and its applications", *In Working Notes of the AAAI-2002 worshop on Ontologies and the Semantic Web, July 28 – August 1, 2002.*

[16] A.L. Rector, "Normalisation of ontology implementations: Towards modularity, re-use, and maintainability", *In EKAW Worshop on Ontologies for Multiagent Systems, 2002.*

[17] H.A. Simon, "The Sciences of the Artifical", *chapter 7, pages 209-217. MIT Press, 1969.*

[18] H. Stuckemschmidt and M. Klein, "Structure-based partitioning of Large Class Hierarchies", *In Proceedings of the 3rd International Semantic Web Conference, 2004.*

[19] R. Volz, D. Oberle, and R. Studer, "Views for light-weight web ontologies", *In Proceedings of the ACM Symposium on Applied Computing (SAC), 2003.*

[20] M. Bhatt, C. Wouters, A. Flahive, W. Rahayu, and D. Taniar, "Semantic completeness in sub-ontology extraction using distributed methods", *In A. Lagana, M.L. Gavrilova, and V. Kumar, editors, Computational Science and Its Applications (ICCSA), volume 3045, pages 508 - 517. Springer-Verlag GmbH, May 2004.*

DRAFT 14. srpna 2009, 14:53

Assessing Classification Confidence Measures in Dynamic Classifier Systems

Post-Graduate Student:
ING. DAVID ŠTEFKA
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
stefka@cs.cas.cz

Supervisor:
ING. RNDR. MARTIN HOLEŇA, CSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
martin@cs.cas.cz

Field of Study:
Mathematical Engineering

The research reported in this paper was partially supported by the Program "Information Society" under project 1ET100300517 and by the grant ME949 of the Ministry of Education, Youth and Sports of the Czech Republic.

Abstract

Classifier combining is a popular technique for improving the classification quality. Common methods for classifier combining can be further improved by using dynamic classification confidence measures. In this paper, we provide a general framework of dynamic classifier systems, which use dynamic confidence measures to adapt the aggregation to a particular pattern. We also introduce methods for assessing classification confidence measures, and we experimentally show that there is a correlation between the feasibility of a confidence measure for a given dataset and a given classifier type, and the improvement of classification quality in dynamic classifier systems.

1. Introduction

Classification is a process of dividing objects (called *patterns*) into disjoint sets called *classes* [1]. A comonly used technique for improving classification quality is *classifier combining* [2] – instead of using just one classifier, a team of classifiers is created and trained; each classifier in the team predicts independently, and the classifier outputs are aggregated into a final prediction. It can be shown that such a team of classifiers can perform better than any of the individual classifiers.

A common drawback of classifier aggregation methods is that they are static, i.e., they are not adapted to the particular pattern to classify. However, if we use the concept of dynamic classification confidence (i.e., the extent to which we can "trust" the output of a particular classifier for the currently classified pattern), the aggregation algorithms can take into account the fact that "this classifier is/is not good *for this particular pattern*".

There has already been some research done in the field of dynamic classifier aggregation. Classifier selection methods [3, 4, 5] try to find out which classifier in the team is locally better than the other classifiers, and this classifier only is used for the prediction. The weakness of these methods is that much of the information is discarded, which can lead to unstability. In classifier aggregation [6, 7], where all the classifiers are used for the prediction, most of the commonly used methods are static. However, for example Robnik-Šikonja [8] and Tsymbal et al. [9] study aggregation of Random Forests with classification confidences, and Avnimelech and Intrator use dynamic aggregation of neural networks [10].

In the wider fields of classification, pattern recognition, and case-based reasoning, the classification confidence has also been studied, e.g. in [11, 12, 13]. The goal of such approaches is usually to refuse to classify a given "hard" pattern and to leave the decision to a human expert. However, in classifier combining, where we have a battery of different classifiers if one classifier refuses to classify a pattern, the classification confidence can be used more exhaustively.

It is although common that the concept of dynamic classification confidence is tightly bound with the aggregation method, or with the particular classifier type used. In this case, it is not clear whether the reported improvements are obtained due to a particular aggregation scheme, or because a dynamic classification confidence was involved in the aggregation process. Moreover, the way a classifier classifies a pattern, the way we measure confidence of a classifier, and the way we aggregate a team of classifiers, are independent on each other, so they should be studied separately.

DRAFT
14. 9. 2009 - 18:52

In this paper, we provide a general framework of dynamic classifier systems, based on three independent aspects – the classifiers in the team, the confidence measures of the individual classifiers, and the aggregation strategy. This allows us to study possible benefits of using classification confidence in classifier combining, regardless of a particular classifier type, or a particular confidence measure. The confidence measures and the aggregation strategy give us three important classes of classifier systems – confidence-free (i.e., systems that do not utilize classification confidence at all), static (i.e., systems that use only "global" confidence of a classifier), and dynamic (i.e., systems that adapt to the particular pattern submitted for classification).

Apart from that, we introduce methods for assessing confidence measures, which can be used for predicting whether a dynamic classifier system will perform better than a confidence-free or static classifier system. We define two heuristics for assessing confidence measures, and we experimentally show that there is a correlation between the feasibility of a confidence measure and the improvement in the classification quality when used in a dynamic classifier system.

The paper is structured as follows. In Section 2, we present the formalism of classification itself and classification confidence, and we introduce the framework of dynamic classifier systems. In Section 3, we deal with methods how the feasibility of classification confidence measures can be measured, and we introduce two heuristics how the assessment can be done. Section 4 experimentally studies the correlation between the feasibility of a confidence measure, and the improvement in classification when used in a dynamic classifier system. Section 5 summarizes the paper and uncovers our plans for the future research.

2. Formalism of Dynamic Classifier Systems

Throughout the rest of the paper, we use the following notation. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a n-dimensional *feature space*, let $C_1, \ldots, C_N \subseteq \mathcal{X}$, $N \geq 2$ be sets called *classes*. A *pattern* is a tuple $(\mathbf{x}, c_\mathbf{x})$, where $\mathbf{x} \in \mathcal{X}$ are *features* of the pattern, and $c_\mathbf{x} \in \{1, ..., N\}$ is the index of the class the pattern belongs to. The goal of classification is to determine to which class a given pattern belongs, i.e., to predict $c_\mathbf{x}$ for unclassified patterns. We assume that for every $\mathbf{x} \in \mathcal{X}$, there is a unique classification $c_\mathbf{x}$ (e.g., provided by some expert), but when we are classifying a pattern, we do not know it – due to this fact, we will sometimes refer to a pattern only as $\mathbf{x} \in \mathcal{X}$.

Definition 1 *Let $[0, 1]$ denote the unit interval. We call a* classifier *every mapping $\phi : \mathcal{X} \to [0, 1]^N$, where for $\mathbf{x} \in \mathcal{X}$, $\phi(\mathbf{x}) = (\mu_1(\mathbf{x}), \ldots, \mu_N(\mathbf{x}))$ are* degrees of classification (d.o.c.) *to each class.*

The d.o.c. to class C_j expresses the extent to which the pattern belongs to class C_j (if $\mu_i(\mathbf{x}) > \mu_j(\mathbf{x})$, it means that the pattern $\mathbf{x}$ belongs to class C_i rather than to C_j). Depending on the classifier type, it can be modelled by probability, fuzzy membership, etc.

Remark 1 *This definition is of course not the only way how a classifier can be defined, but in the theory of classifier combining, this one is used most often [2].*

The prediction of $c_\mathbf{x}$ for an unknown pattern $\mathbf{x}$ is done by converting the continuous d.o.c. of the classifier into a crisp output.

Definition 2 *Let ϕ be a classifier, $\mathbf{x} \in \mathcal{X}$, $\phi(\mathbf{x}) = (\mu_1(\mathbf{x}), \ldots, \mu_N(\mathbf{x}))$.* Crisp output *of ϕ on $\mathbf{x}$ is defined as $\phi^{(cr)}(\mathbf{x}) = \arg\max_{i=1,\ldots,N} \mu_i(\mathbf{x})$ if there are no ties (i.e., $|\arg\max_{i=1,\ldots,N} \mu_i(\mathbf{x})| = 1$), defined arbitrarily as $\phi^{(cr)}(\mathbf{x}) \in \arg\max_{i=1,\ldots,N} \mu_i(\mathbf{x})$ in the case of ties.*

2.1. Classification Confidence

In addition to the classifier output (the d.o.c.s), which predicts to which class a pattern belongs to, we will work with *confidence* of the prediction, i.e., the extent to which we can "trust" the output of the classifier.

Definition 3 *Let ϕ be a classifier. We call a* confidence measure *of classifier ϕ every mapping $\kappa_\phi : \mathcal{X} \to [0, 1]$. Let $\mathbf{x} \in \mathcal{X}$. $\kappa_\phi(\mathbf{x})$ is called* classification confidence *of ϕ on $\mathbf{x}$.*

Classification confidence expresses the degree of trust we can give to a classifier ϕ when classifying a pattern $\mathbf{x}$. $\kappa_\phi(\mathbf{x}) = 0$ means that the classification does not need to be correct, while $\kappa_\phi(\mathbf{x}) = 1$ means the classification is probably correct.

A confidence measure can be either *static*, i.e., it is a constant of the classifier, or *dynamic*, i.e., it adjusts itself to the currently classified pattern.

Definition 4 *Let ϕ be a classifier and κ_ϕ its confidence measure. We call κ_ϕ* static, *iff it is constant in $\mathbf{x}$, we call κ_ϕ* dynamic *otherwise.*

DRAFT Ústavná 2009 - 14.4.2009

Remark 2 *Since static confidence measures are constant, independent on the currently classified pattern, we will omit the pattern* $\mathbf{x}$ *in the notation, i.e., we will denote their values just as* κ_ϕ.

In the rest of the paper, we will use the indicator operator I, defined as $I(\text{true}) = 1$, $I(\text{false}) = 0$.

2.1.1 Static Confidence Measures: After the classifier has been trained, we can use a validation set (i.e., a set of patterns the classifier has not been trained on; we could also use training patterns, but in that case, the results would be biased) to assess its predictive power as a whole (from a global view). These methods include accuracy, precision, sensitivity, resemblance, etc. [1, 14], and we can use these measures as static confidence measures. In this paper, we will use the Global Accuracy measure.

Global Accuracy (GA) of a classifier ϕ is defined as the proportion of correctly classified patterns from the validation set:

$$\kappa_\phi^{(GA)} = \frac{\sum\limits_{(\mathbf{y},c_\mathbf{y})\in\mathcal{M}} I(\phi^{(cr)}(\mathbf{y}) \stackrel{?}{=} c_\mathbf{y})}{|\mathcal{M}|}, \quad (1)$$

where $\mathcal{M} \subseteq \mathcal{X} \times \{1, \ldots, N\}$ is the validation set and $\phi^{(cr)}(\mathbf{y})$ is the crisp output of ϕ on $\mathbf{y}$.

2.1.2 Dynamic Confidence Measures: An easy way how a dynamic confidence measure can be defined is to compute some property on patterns neighboring $\mathbf{x}$. Let $N(\mathbf{x})$ denote a set of neighboring patterns from the validation set. In this paper, we define $N(\mathbf{x})$ as the set of k patterns nearest to $\mathbf{x}$ under Euclidean metric. Now we will define two dynamic confidence measures which use $N(\mathbf{x})$:

Euclidean Local Accuracy (ELA), used in [5], measures the local accuracy of ϕ in $N(\mathbf{x})$:

$$\kappa_\phi^{(ELA)}(\mathbf{x}) = \frac{\sum\limits_{(\mathbf{y},c_\mathbf{y})\in N(\mathbf{x})} I(\phi^{(cr)}(\mathbf{y}) \stackrel{?}{=} c_\mathbf{y})}{|N(\mathbf{x})|}, \quad (2)$$

where $\phi^{(cr)}(\mathbf{y})$ is the crisp output of ϕ on $\mathbf{y}$.

Euclidean Local Match (ELM), based on the ideas in [12], measures the proportion of patterns in $N(\mathbf{x})$ from the same class as ϕ is predicting for $\mathbf{x}$:

$$\kappa_\phi^{(ELM)}(\mathbf{x}) = \frac{\sum\limits_{(\mathbf{y},c_\mathbf{y})\in N(\mathbf{x})} I(\phi^{(cr)}(\mathbf{x}) \stackrel{?}{=} c_\mathbf{y})}{|N(\mathbf{x})|}, \quad (3)$$

where $\phi^{(cr)}(\mathbf{x})$ is the crisp output of ϕ on $\mathbf{x}$. The difference between (2) and (3) is that in the latter case, there is $\phi^{(cr)}(\mathbf{x})$ instead of $\phi^{(cr)}(\mathbf{y})$ in the indicator.

The dynamic confidence measures defined in this section have one drawback – they need to compute neighboring patterns of $\mathbf{x}$, which can be time-consuming, and sensitive to the similarity measure used. There are also dynamic confidence measures, which compute the classification confidence directly from the degrees of classification [10, 11], e.g., the ratio of the highest degree of classification to the sum of all degrees of classification. However, our preliminary experiments with such measures with quadratic discriminant classifiers and random forests show that such confidence measures give very poor results [15].

2.1.3 The Oracle Confidence Measure: For reference purposes, we also define a so-called *Oracle confidence measure*, which represents the "best-we-can-do" approach.

Oracle (OR) confidence is equal to 1 iff the pattern is classified correctly, 0 otherwise:

$$\kappa_\phi^{(OR)}(\mathbf{x}) = I(\phi^{(cr)}(\mathbf{x}) \stackrel{?}{=} c_\mathbf{x}) \quad (4)$$

Of course, in practical applications, we cannot use the Oracle confidence measure, because we do not know the actual class the pattern belong to ($c_\mathbf{x}$). However, the Oracle confidence measure can give us upper bound for performance of a classifier system using classification confidence, and it can also be used to assess the feasibility of a given confidence measure.

2.2. Classifier Teams

In classifier combining, instead of using just one classifier, a team of classifiers is created, and the team is then aggregated into one final classifier. If we want to utilize classification confidence in the aggregation process, each classifier must have its own confidence measure defined.

Definition 5 *Let* $r \in \mathbf{N}$, $r \geq 2$. Classifier team *is a tuple* $(\mathcal{T}, \mathcal{K})$, *where* $\mathcal{T} = (\phi_1, \ldots, \phi_r)$ *is a set of classifiers, and* $\mathcal{K} = (\kappa_{\phi_1}, \ldots, \kappa_{\phi_r})$ *is a set of corresponding confidence measures.*

DRAFT 14 srpna 2009 - 14:36

If a classifier team consists only of classifiers of the same type, which differ only in their parameters, dimensionality, or training sets, the team is usually called an *ensemble of classifiers*. The restriction to classifiers of the same type is not essential, but it ensures that the outputs of the classifiers are consistent. Well-known methods for ensemble creation are *bagging* [16], *boosting* [17], *random forests* [18], *error correction codes* [2], or *multiple feature subset* methods [19].

Remark 3 *The goal of these methods is to create an ensemble of classifiers which are both* accurate *and* diverse *[20]. Here we cite only some of the basic papers about ensemble methods – in the literature, modified and improved versions of the methods can be found. In our framework, any method for creating a team (or ensemble) can be used – i.e., ensemble methods are not competitive to our approach, but they are more or less supplementary. After the classifier team has been created, the aggregation rule is totally independent of the method by which the team has been created.*

If a pattern $\mathbf{x}$ is submitted for classification, the team of classifiers gives us information of two kinds – outputs of the individual classifiers (a *decision profile*), and classification confidences of the classifiers on $\mathbf{x}$ (a *confidence vector*).

Definition 6 *Let $(\mathcal{T}, \mathcal{K})$, where $\mathcal{T} = (\phi_1, \ldots, \phi_r), \mathcal{K} = (\kappa_{\phi_1}, \ldots, \kappa_{\phi_r})$, be a classifier team, and let $\mathbf{x} \in \mathcal{X}$. Then we define* decision profile $\mathcal{T}(\mathbf{x}) \in [0,1]^{r \times N}$

$$\mathcal{T}(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_r(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mu_{1,1}(\mathbf{x}) & \mu_{1,2}(\mathbf{x}) & \ldots & \mu_{1,N}(\mathbf{x}) \\ \mu_{2,1}(\mathbf{x}) & \mu_{2,2}(\mathbf{x}) & \ldots & \mu_{2,N}(\mathbf{x}) \\ & & \ddots & \\ \mu_{r,1}(\mathbf{x}) & \mu_{r,2}(\mathbf{x}) & \ldots & \mu_{r,N}(\mathbf{x}) \end{pmatrix}, \tag{5}$$

and confidence vector $\mathcal{K}(\mathbf{x}) \in [0,1]^r$

$$\mathcal{K}(\mathbf{x}) = \begin{pmatrix} \kappa_{\phi_1}(\mathbf{x}) \\ \kappa_{\phi_2}(\mathbf{x}) \\ \vdots \\ \kappa_{\phi_r}(\mathbf{x}) \end{pmatrix} \tag{6}$$

Remark 4 *Here we use the notation $\mathcal{T}$ for both the set of classifiers, and for the decision profile, and similarly for $\mathcal{K}$. To avoid any confusion, the decision profile and confidence vector will always be followed by $(\mathbf{x})$.*

2.3. Classifier Systems

After the pattern $\mathbf{x}$ has been classified by all the classifiers in the team, and the confidences have been computed, these outputs have to be aggregated using a *team aggregator*, which takes the decision profile as its first argument, the confidence vector as its second argument, and returns the aggregated degrees of classification to all the classes.

Definition 7 *Let $r, N \in \mathbf{N}$, $r, N \geq 2$. A* team aggregator *of dimension (r, N) is any mapping $\mathcal{A} : [0,1]^{r \times N} \times [0,1]^r \to [0,1]^N$.*

A classifier team with an aggregator will be called a *classifier system*. Such system can be also viewed as a single classifier.

Definition 8 *Let $(\mathcal{T}, \mathcal{K})$ be a classifier team, and let $\mathcal{A}$ be a team aggregator of dimension (r, N), where r is the number of classifiers in the team, and N is the number of classes. The triple $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A})$ is called a* classifier system. *We define an* induced classifier *of $\mathcal{S}$ as a classifier Φ, defined as*

$$\Phi(\mathbf{x}) = \mathcal{A}(\mathcal{T}(\mathbf{x}), \mathcal{K}(\mathbf{x})).$$

Depending on the way how a classifier system utilizes the classification confidence, we can distinguish several types of classifier systems.

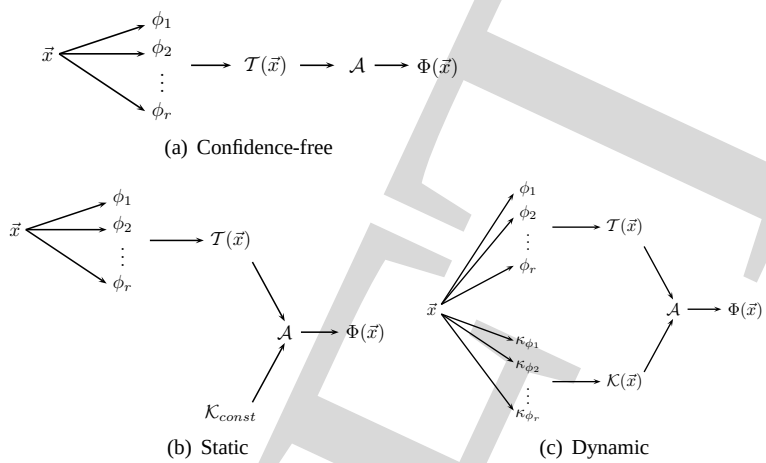
Definition 9 *Let $(\mathcal{T}, \mathcal{K})$ be a classifier team. $(\mathcal{T}, \mathcal{K})$ is called* static, *iff $\forall \kappa \in \mathcal{K} : \kappa$ is a static confidence measure. $(\mathcal{T}, \mathcal{K})$ is called* dynamic, *iff $\exists \kappa \in \mathcal{K} : \kappa$ is a dynamic confidence measure.*

Definition 10 *Let $\mathcal{A}$ be a team aggregator of dimension (r, N). We call $\mathcal{A}$* confidence-free, *iff it is constant in the second argument.*

Definition 11 *Let $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A})$ be a classifier system. We call $\mathcal{S}$* confidence-free, *iff $\mathcal{A}$ is confidence-free. We call $\mathcal{S}$* static, *iff $(\mathcal{T}, \mathcal{K})$ is static, and $\mathcal{A}$ is not confidence-free. We call $\mathcal{S}$* dynamic, *iff $(\mathcal{T}, \mathcal{K})$ is dynamic, and $\mathcal{A}$ is not confidence-free.*

Confidence-free classifier systems do not utilize the classification confidence at all. Static classifier systems utilize classification confidence, but only as a global property (constant for all patterns). Dynamic classifier systems utilize classification confidence in a dynamic way, i.e. the aggregation is adapted to the particular pattern submitted for classification. The different approaches are shown in Fig. 1.

DRAFT 4. června 2009, 14:32



(a) Confidence-free

(b) Static

(c) Dynamic

Figure 1: Schematic comparison of confidence-free, static, and dynamic classifier systems.

2.3.1 Classifier Selection: Classifier selection methods [3, 4, 5] use some criterion to determine which classifier is most suitable for the current pattern $\mathbf{x}$, and the output of this classifier is taken as the final result – outputs of the other classifiers are entirely discarded.

These methods are a special case of dynamic classifier systems – the selection criterion can be viewed as a dynamic confidence measure evaluated on all the classifiers in the team, and the team aggregator $\mathcal{A}$ corresponding to the classifier selection method is defined as $\mathcal{A}(\mathcal{T}(\mathbf{x}), \mathcal{K}(\mathbf{x})) = \Phi(\mathbf{x}) = \phi_i(\mathbf{x})$, where $i \in \arg\max_{i=1,\ldots,r} \kappa_{\phi_i}(\mathbf{x})$.

The weakness of classifier selection methods is that they discard much potentially useful information, which can lead to unstable results in the induced classifier's predictions [21]. In the rest of the paper, we do not deal with classifier selection.

2.3.2 Classifier Aggregation: Many methods for aggregating a team of classifiers into one final classifier have been proposed in the literature [2, 6, 7]. The simplest methods use only some simple arithmetic operation to aggregate the team's output (e.g., voting, sum, maximum, minimum, mean, weighted mean, weighted voting, product, etc.). More advanced methods use for example probability theory (e.g., behavior knowledge space [22], product rule [6], Dempster-Shafer fusion [6]), fuzzy logic (e.g., fuzzy integral [23, 24], decision templates [6, 23]), or second-level classifiers [6].

To emphasize the difference between confidence-free, static, and dynamic classifier systems, we will not consider complex aggregation algorithms, and we will define three simple aggregation algorithms, based on mean value, each representing confidence-free, static, or dynamic classifier system. This will allow us to compare the different classifier systems without bias.

We will use the notation from Def. 6 and Def. 8. Let $\Phi(\mathbf{x}) = \mathcal{A}(\mathcal{T}(\mathbf{x}), \mathcal{K}(\mathbf{x})) = (\mu_1(\mathbf{x}), \ldots, \mu_N(\mathbf{x}))$, and let $j = 1, \ldots, N$.

Mean value aggregation (MV) is the most common (confidence-free) aggregation technique. Its aggregator is defined as

$$\mu_j(\mathbf{x}) = \frac{\sum\limits_{i=1,\ldots,r} \mu_{i,j}(\mathbf{x})}{r}. \tag{7}$$

Static weighted mean aggregation (SWM) computes aggregated d.o.c. as weighted mean of d.o.c. given by the individual classifiers, where the weights are static classification confidences:

$$\mu_j(\mathbf{x}) = \frac{\sum\limits_{i=1,\ldots,r} \kappa_{\phi_i} \mu_{i,j}(\mathbf{x})}{\sum\limits_{i=1,\ldots,r} \kappa_{\phi_i}}. \tag{8}$$

Dynamic weighted mean aggregation (DWM) has the same aggregator as SWM, with the difference that the weights are dynamic classification confidences:

$$\mu_j(\mathbf{x}) = \frac{\sum\limits_{i=1,\ldots,r} \kappa_{\phi_i}(\mathbf{x}) \mu_{i,j}(\mathbf{x})}{\sum\limits_{i=1,\ldots,r} \kappa_{\phi_i}(\mathbf{x})}. \tag{9}$$

Remark 5 *If we aggregate a team of classifiers with the Oracle confidence measure using the DWM aggregator, we obtain an* Oracle classifier *– a common reference*

DRAFT 14. June 2009 - 14:32

classifier system, which gives us correct prediction if and only if any of its classifiers gives correct prediction. The Oracle classifier serves as the "best how classifier combining can be done" approach.

3. Assessing Confidence Measures

In [15, 25], we have experimentally shown that dynamic classifier systems of Random Forests [18] and Quadratic Discriminant Classifiers [1] using the ELA and ELM confidence measures can significantly improve the quality of classification, compared to confidence-free, or static classifier systems.

However, in these experiments, the performance of the dynamic classifier systems varied from dataset to dataset. For some datasets, the ELM confidence measure obtained better results, for others the ELA was more successful, and for some datasets, neither of them improved the classification. In other words, the performance of a dynamic classifier system is heavily influenced by the particular confidence measure used.

Given a particular dataset to classify, and given a set of classifiers which form a classifier team, there are several questions which come into one's mind:

- Will a dynamic classifier system yield improvement in the classification quality compared to confidence-free or static classifier system?
- Which confidence measure will perform the best for the given classifiers and the given dataset?
- Are the benefits of a dynamic classifier system worth the higher computational complexity?

To answer these questions, we could of course build the classifier systems and compare their performance using crossvalidation or other standard machine learning technique. However, it would be more convenient if we had some criterion of feasibility of a given confidence measure, which could answer these questions *prior* to building and crossvalidating the models. In this paper, we introduce two such criteria. Before that, we summarize the properties which should hold for a "good" confidence measure. Intuitively, if $\kappa_\phi(\mathbf{x})$ estimates the degree of trust we can give to the classifier ϕ when classifying a pattern $\mathbf{x}$, the following should be satisfied:

- If the classification confidence $\kappa_\phi(\mathbf{x})$ is high (close to 1), the classifier's prediction $\phi^{(cr)}(\mathbf{x})$ should be correct.
- If the classifier's prediction $\phi^{(cr)}(\mathbf{x})$ is not correct, the classification confidence $\kappa_\phi(\mathbf{x})$ should be low (close to 0).

For example, if $\kappa_\phi(\mathbf{x})$ is an estimate of the probability of correct classification of $\mathbf{x}$ by ϕ (for example the ELA confidence measure), both these implications are satisfied, if the estimate is good enough. According to these two properties, the ideal confidence measure is the Oracle confidence measure.

In this paper, we propose an approach in which the feasibility of a confidence measure is measured empirically, on a set of validation patterns. Let ϕ be a classifier, κ_ϕ a confidence measure, and $\mathcal{M} \subseteq \mathcal{X} \times \{1, \ldots, N\}$ the validation set. The feasibility of κ_ϕ for classifier ϕ, measured empirically on data $(\mathbf{x}, c_\mathbf{x}) \in \mathcal{M}$ will be denoted to as $\mathcal{F}(\phi, \kappa_\phi, \mathcal{M}) \in [0, 1]$. The particular methods how $\mathcal{F}(\phi, \kappa_\phi, \mathcal{M})$ can be defined will be shown in Sec. 3.2 and 3.3.

However, in classifier combining, we do not have a single classifier and its corresponding confidence measure – we have a set of classifiers $\mathcal{T}$, and a set of corresponding confidence measures $\mathcal{K}$. Therefore, we define $\mathcal{F}(\mathcal{T}, \mathcal{K}, \mathcal{M}) \in [0, 1]$ as the average feasibility of $\kappa_\phi \in \mathcal{K}$ for the corresponding classifier $\phi \in \mathcal{T}$, measured on $\mathcal{M}$:

$$\mathcal{F}(\mathcal{T}, \mathcal{K}, \mathcal{M}) = \frac{\sum\limits_{\phi \in \mathcal{T}} \mathcal{F}(\phi, \kappa_\phi, \mathcal{M})}{|\mathcal{T}|}. \tag{10}$$

3.1. Restricting the Validation Set

There is one more important aspect in which assessing the feasibility of a confidence measure differs in the context of classifier systems. If we measure $\mathcal{F}(\phi, \kappa_\phi, \mathcal{M})$ on the whole validation set $\mathcal{M}$, we have an estimate how κ_ϕ predicts the classification confidence *for a single classifier*. However, if we want to assess a confidence measure's performance in the context of dynamic classifier systems, we need to know something different: can this particular confidence measure improve the prediction of the classifier system?

What is the difference between these two information? A typical situation in classifier aggregation is as follows: for most patterns, the crisp outputs of the individual classifiers in a classifier system show consensus on a certain class (i.e., a vast majority of the classifiers predicts one particular class), and the team aggregator is not able to break this consensus, even when incorporating the classification confidences. Therefore, the behavior of the confidence measures on such patterns

DRAFT 14. prosince 2009 -- 12:32

is totally irrelevant. On the other hand, for patterns where there is no such consensus, the behavior of the confidence measure is *much* more important. Therefore, we need to identify such patterns, and restrict $\mathcal{M}$ to a such subset.

Let $0 \leq s \leq r$, where $r = |\mathcal{T}|$. Let $U(s) \subseteq \mathcal{M}$ be the set of patterns $(\mathbf{x}, c_{\mathbf{x}})$, for which for all classes C_j, $j = 1, \ldots, N$, we have

$$|\{i; i = 1, \ldots, r,\ \phi_i^{(cr)}(\mathbf{x}) = j\}| \leq s. \qquad (11)$$

$U(s)$ denotes set of patterns, for which at most s classifiers vote for any particular class. For lower s, this means that there is no consensus on a particular class, and so the team aggregator can easily use the classification confidence to improve the prediction – this suggests that restricted validation set for lower s are more important for the analysis. However, the smaller s, the smaller $|U(s)|$, which leads us to the fact that we need s big enough so the feasibility is measured on enough data. To solve the dilemma, we use the following heuristic: choose smallest s, for which $U(s)$ covers a given portion (5-10%) of the validation data, i.e., $|U(s)| \geq \alpha|\mathcal{M}|$, where $\alpha \in (0, 1)$.

3.2. Similarity to OR

The first approach how $\mathcal{F}(\phi, \kappa_\phi, \mathcal{M})$ can be measured is to compute the similarity of values $\kappa_\phi(\mathbf{x})$ to the values of the Oracle confidence $\kappa_\phi^{(OR)}(\mathbf{x})$ for patterns $(\mathbf{x}, c_{\mathbf{x}}) \in \mathcal{M}$, where $\mathcal{M}$ is the (restricted) validation set. This can be done by taking the average absolute value of the differences of the confidences:

$$\mathcal{F}^{(SOR)}(\phi, \kappa_\phi, \mathcal{M}) = 1 - \frac{\sum\limits_{(\mathbf{x}, c_{\mathbf{x}}) \in \mathcal{M}} |\kappa_\phi(\mathbf{x}) - \kappa_\phi^{(OR)}(\mathbf{x})|}{|\mathcal{M}|}. \qquad (12)$$

3.3. AUC for OK/NOK Histogram

The second approach how $\mathcal{F}(\phi, \kappa_\phi, \mathcal{M})$ can be measured is to analyze histograms of $\kappa_\phi(\mathbf{x})$ for patterns classified correctly by ϕ (*OK patterns*) and for patterns classified incorrectly by ϕ (*NOK patterns*). Values of $\kappa_\phi(\mathbf{x})$ for the OK patterns should be concentrated near 0, while for the NOK patterns, $\kappa_\phi(\mathbf{x})$ should concentrate near 1. Moreover, these two distributions should not overlap.

Let $\mathcal{M}$ be the (restricted) validation set, and let $\mathcal{M}_i \subseteq \mathcal{M}$ for $i = 1, \ldots, N$ denote the sets of validation patterns from class C_i. For two arbitrary classes C_k, C_j, we define the multiset

$$H_{kj} = \{\kappa_\phi(\mathbf{x}) | (\mathbf{x}, c_{\mathbf{x}}) \in \mathcal{M}_k,\ \phi^{(cr)}(\mathbf{x}) = j\}, \qquad (13)$$

as a multiset of classification confidence values for all validation patterns from class C_k, which have been classified to class C_j by ϕ. Using this notation, we can define the *OK histogram* as the histogram computed from $\bigcup_k H_{kk},\ k = 1, \ldots, N$ and the *NOK histogram* as the histogram computed from $\bigcup_{k \neq j} H_{kj},\ k, j = 1, \ldots, N$.

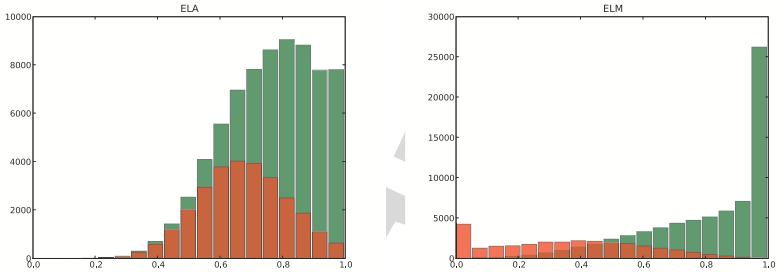
The OK and NOK histograms of the ELA and ELM confidence measures for a Random Forest ensemble for the Waveform dataset (non-restricted) are shown in Fig. 2. Fig. 3 shows the evolution of the histograms for the restricted validation set. Observe that for lower s, the histograms are very different from the histograms for higher values of s.

Altough the OK/NOK (restricted) histograms give us visual information, we need to evaluate the degree of overlapping using a single number. This is possible, if we represent the OK/NOK confidence values by a ROC curve, and then we compute the area under the ROC curve.

Remark 6 *Receiver operating characteristic (ROC) curves [26] are a standard tool in data mining and machine learning. ROC is basically a plot of the fraction of true positives vs. the fraction of false positives of a binary classifier, as some parameter is being varied (e.g., the discrimination threshold of the classifier). If a classifier assigns patterns to classes entirely at random, its ROC curve is the diagonal. On the other hand, for an ideal classifier, the ROC curve consist only of one point* $(0, 1)$. *The closer we are to the ROC of the ideal classifier (i.e., the farther the ROC curve is from the diagonal (above the diagonal)), the better discrimination of the classifier. The strong point of the ROC curve aprroach is that we can summarize the ROC curve into a single number – area under ROC curve (AUC) – which can be used as a criterion of quality of a binary classifier. For a random classifier, AUC=0.5, for an ideal classifier, AUC=1. The higher the AUC, the better discrimination of the classifier. Classifiers with AUC below 0.5 are actually* worse *than a random classifier.*

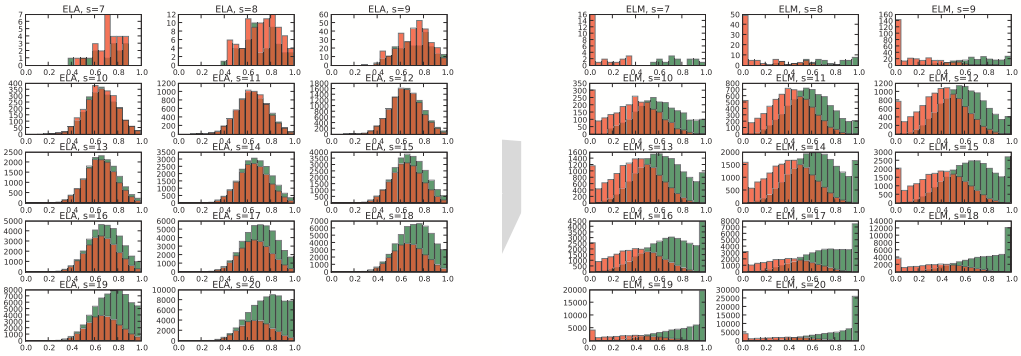
In the context of classification confidence, we will study the AUC of a so-called *OK/NOK classifier*, which assigns a pattern to the class "correctly classified" if the classification confidence is higher than some threshold T, and to the class "incorrectly classified" instead. By varying T between 0 and 1, we obtain the ROC curve. The AUC of the OK/NOK classifier measured on a validation set $\mathcal{M}$ (or, on a restricted set $U(s)$) can be used as an empirical property expressing the degree of

DRAFT 14. SINT 2009 14:32



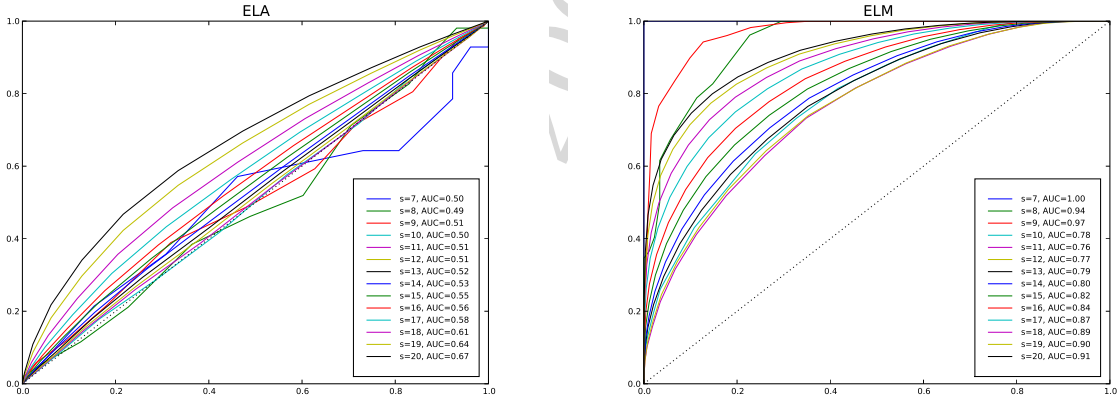
(a) ELA – bad separation (b) ELM – relatively good separation

Figure 2: The OK (green) and NOK (red) histograms of κ_ϕ of a Random Forest ensemble for the Waveform dataset.



(a) ELA (b) ELM

Figure 3: The restricted OK (green) and NOK (red) histograms of κ_ϕ of a Random Forest ensemble for the Waveform dataset for $s = 7, \ldots, 20$.



(a) ELA (b) ELM

Figure 4: The ROC curves and the AUCs of the OK/NOK classifiers for the Waveform dataset, measured on $U(s)$, $s = 7, \ldots, 20$, for a Random Forest ensemble.

overlapping of the OK and NOK distributions. Now we can define $\mathcal{F}^{(AUC)}(\phi, \kappa_\phi, \mathcal{M})$ as the AUC of the OK/NOK classifier for the confidence κ_ϕ, measured on $\mathcal{M}$. Fig. 4 shows an example of the ROCs for the ELA and ELM confidence measures for a Random Forest ensemble for the Waveform dataset.

Remark 7 *Receiver operating characteristic (ROC) curves [26] are a standard tool in data mining and machine learning. ROC is basically a plot of the fraction of true positives vs. the fraction of false positives of a binary classifier, as some parameter is being varied (e.g., the discrimination threshold of the classifier). If a classifier assigns patterns to classes entirely at random, its ROC curve is the diagonal. On the other hand, for an ideal classifier, the ROC curve consist only of one point* $(0, 1)$. *The closer we are to the ROC of the ideal classifier (i.e., the farther the ROC curve is from the diagonal (above the diagonal)), the better discrimination of the classifier. The strong point of the ROC curve aprroach is that we can summarize the ROC curve into a single number – area under ROC curve (AUC) – which can be used as a criterion of quality of a binary classifier. For a random classifier, AUC=0.5, for an ideal classifier, AUC=1. The higher the AUC, the better discrimination of the classifier. Classifiers with AUC below 0.5 are actually* worse *than a random classifier.*

4. Experiments

To find out whether the methods for assessing confidence measures described in the previous sections can really predict the improvement in the classification quality of a dynamic classifier system, we designed the following experiment. Suppose we have a classifier team $(\mathcal{T}, \mathcal{K})$. Given a dataset, we put apart 20% of the data (this was done only for the datasets which contained more than 500 patterns; for smaller datasets, we used the whole dataset) to measure $\mathcal{F}(\mathcal{T}, \mathcal{K}, \mathcal{M})$ using 5-fold crossvalidation. After that, we use the remaining data to measure the relative improvement of the error rate of a dynamic classifier system (aggregated using DWM) compared to the error rate of a confidence-free classifier system (aggregated using MV), using 10-fold crossvalidation:

$$\mathcal{I}(S_1, S_2) = \frac{Err(S_1) - Err(S_2)}{Err(S_1)}, \tag{14}$$

where $Err(S_1)$ denotes the error rate of the reference classifier system (using MV aggregator), and $Err(S_2)$ denotes the error rate of the dynamic classifier system (using DWM aggregator).

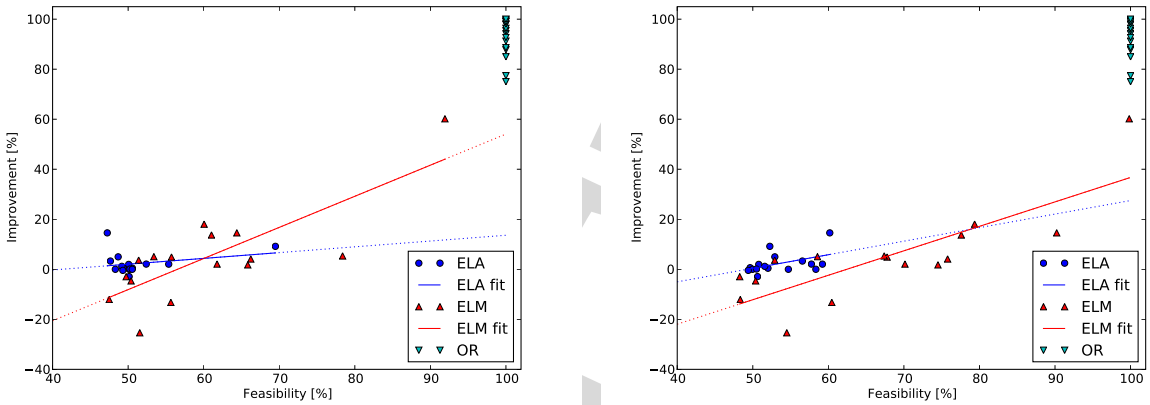
Our goal in this experiment was to study the correlation between $\mathcal{F}$ and $\mathcal{I}$. We performed the experiment on 5 artificial and 11 real-world datasets from the Elena database [27] and from the UCI repository [28]. The classifier teams were created using the Random Forest method [18], and as the classification confidences we used both ELA and ELM. For reference purposes, we also used the Oracle confidence measure (for which $\mathcal{F} = 1$ by definition). For assessing the confidence measures, we used methods descibed in the previous section, i.e., similarity to the Oracle confidence (SOR) and the area under ROC curve of the OK/NOK classifier (AUC), measured on the restricted validation set $U(s)$, for s such that $U(s)$ covers 5% of the data.

For each feasibility measure, we obtained a scatterplot of $(\mathcal{F}, \mathcal{I})$ values, which is shown in Fig. 5. We also computed a least-squares linear approximation of the scatterplot. To test the statistical significance of the results, we used the Spearman's rank correlation test [29], implemented in the Scipy Python package [30]. The Spearman's rank correlation test computes the Spearman's rank correlation coefficient $\rho \in [-1, 1]$, which expresses the degree of correlation of two variables X, Y based on their order in X and Y domains. $\rho = 0$ means there is no correlation between X and Y, $\rho = 1$ means there is a total correlation, and $\rho = -1$ indicates anticorrelation. The value of ρ is then compared to a critical value for a chosen significance level α, under the null hypothesis that there is no correlation between the variables.

For $\mathcal{F}^{(SOR)}$, the scatterplot shows a statistically significant correlation between $\mathcal{F}$ and $\mathcal{I}$ for the ELM confidence measure (at 1% significance level). For the ELA confidence measure, the correlation is not clear, and is not statistically significant. The linear least-squares fit shows that there is an increasing tendency for both confidence measures (however, much smaller for ELA). Regrettably, values of $\mathcal{F}$ for ELA are clustered mainly in the area between 50% and approx. 60%, and thus we cannot study the improvement for higher AUC values.

For $\mathcal{F}^{(AUC)}$, the scatterplot shows a statistically significant correlation between $\mathcal{F}$ and $\mathcal{I}$ for both the ELA (at 5% significance level) and ELM (at 1% significance level) confidence measures. The linear lease-squares fit shows clear increasing tendency for both confidence measures. Again, values of $\mathcal{F}$ for ELA span only the area between 50% and approx. 60%, and thus we cannot study the improvement for higher AUC values.

DRAFT 14. srpna 2009 - 14:32



(a) *SOR*, ELA: $\rho = -0.07, p = 80\%$, ELM: $\rho = 0.64, p = 0.8\%$

(b) *AUC*, ELA: $\rho = 0.53, p = 3.4\%$, ELM: $\rho = 0.76, p = 0.1\%$

Figure 5: Scatterplot of $\mathcal{I}$ versus $\mathcal{F}$ for restricted validation set $U(s)$, covering 5% of the validation data for 16 datasets for the ELA, ELM, and OR confidence measures. The solid/dotted lines represent least-squares linear intrapolations/extrapolations of the data. ρ denotes the Spearman's rank correlation coefficient and p denotes the statistical significance level of the Spearman's test.

These results suggest that the methods for assessing confidence measures could be used for predicting the performance of a dynamic classifier system using classification confidence. As ELM obtains better feasibility values than ELA, the correlation between its feasibility and the improvement is more visible than for ELA. In this experiment, the AUC approach for assessing confidences showed better results than the SOR approach.

5. Summary & Future Work

In this paper, we have introduced a general framework of dynamic classifier systems, built on three main elements – the individual classifiers, their confidence measures, and the aggregator of the system. We have shown examples of one static (Global Accuracy), two dynamic (Euclidean Local Accuracy, Euclidean Local Match), and one reference (Oracle) classification confidence measures, which can be used in the framework.

We have introduced two different heuristics (the similarity to the Oracle confidence measure, and the area under ROC curve of a OK/NOK histogram) how the feasibility of a confidence measure can be assessed for a particular classifier and data. We have also shown that it is useful to compute the feasibility of a confidence measure on a set of patterns for which there is no consensus in the classifier system.

In the experiments, we have shown a correlation between the feasibility of a confidence measure and the improvement of the classification quality of a dynamic classifier system, compared to a confidence-free classifier system (at least for the OK/NOK histogram-based approach).

In our future research, we would like to study methods for assessing classification confidence measures in more detail. We would like to study deeper the way how dynamic classifier systems work and why (and when) the dynamic classification confidence can improve the classification quality.

We would also like to perform experiments with dynamic classifier systems for other classifier types than Quadratic Discriminant Classifiers and Random Forests, mainly Support Vector Machines and k-Nearest Neighbor classifiers. Apart from that, we would like to incorporate dynamic classification confidence into more advanced classifier aggregation methods, for example fuzzy t-conorm integral.

References

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[2] L.I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[3] X. Zhu, X. Wu, and Y. Yang, "Dynamic classifier selection for effective mining from noisy data

streams," in *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)*, (Washington, DC, USA), pp. 305–312, IEEE Computer Society, 2004.

[4] M. Aksela, "Comparison of classifier selection methods for improving committee performance.," in *Multiple Classifier Systems*, pp. 84–93, 2003.

[5] K. Woods, J.W. Philip Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 405–410, 1997.

[6] L.I. Kuncheva, J.C. Bezdek, and R.P.W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison.," *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[7] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[8] M. Robnik-Šikonja, "Improving random forests," in *ECML* (J. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, eds.), vol. 3201 of *Lecture Notes in Computer Science*, pp. 359–370, Springer, 2004.

[9] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Dynamic integration with random forests.," in *ECML* (J. FĂL'rnkranz, T. Scheffer, and M. Spiliopoulou, eds.), vol. 4212 of *Lecture Notes in Computer Science*, pp. 801–808, Springer, 2006.

[10] R. Avnimelech and N. Intrator, "Boosted mixture of experts: An ensemble learning scheme," *Neural Computation*, vol. 11, no. 2, pp. 483–497, 1999.

[11] D.R. Wilson and T.R. Martinez, "Combining cross-validation and confidence to measure fitness," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), paper 163*, 1999.

[12] S.J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating estimates of classification confidence for a case-based spam filter," in *Case-Based Reasoning, Research and Development, 6th Int. Conf., ICCBR 2005, Chicago, USA* (H. Muñoz-Avila and F. Ricci, eds.), vol. 3620 of *LNCS*, pp. 177–190, Springer, 2005.

[13] W. Cheetham, "Case-based reasoning with confidence," in *EWCBR '00: Proceedings of the 5th European Workshop on Advances in Case-Based Reasoning*, (London, UK), pp. 15–25, Springer-Verlag, 2000.

[14] D.J. Hand, *Construction and Assessment of Classification Rules*. Wiley, 1997.

[15] D. Štefka and M. Holeňa, "Classifier aggregation using local classification confidence," in *Proceedings of the ICAART 2009 First International Conference on Agents and Artificial Intelligence, Porto, Portugal*, pp. 173–178, INSTICC Press, 2009.

[16] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[17] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, pp. 148–156, 1996.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] S.D. Bay, "Nearest neighbor classification from multiple feature subsets," *Intelligent Data Analysis*, vol. 3, no. 3, pp. 191–209, 1999.

[20] L.I. Kuncheva and C.J. Whitaker, "Measures of diversity in classifier ensembles," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[21] D. Štefka, "Confidence of classification and its application to classifier aggregation," in *Doktorandské dny KM FJFI ČVUT 2007, Prague, Czech Republic, 16. and 23. 11. 2007* (Z. Ambrož, P. Masáková, ed.), pp. 201–210, Česká technika ČVUT, 2007.

[22] Y.S. Huang and C.Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 1, pp. 90–94, 1995.

[23] L.I. Kuncheva, "Fuzzy versus nonfuzzy in combining classifiers designed by boosting," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 729–741, 2003.

[24] D. Štefka and M. Holeňa, "The use of fuzzy t-conorm integral for combining classifiers," in *Proceedings of the ECSQARU 2007 Ninth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Hammamet, Tunisia, 31.10.-02.11. 2007* (K. Mellouli, ed.), vol. 4724 of *Lecture Notes in Computer Science*, pp. 755–766, Springer, 2007.

[25] D. Štefka and M. Holeňa, "Dynamic classifier systems and their applications to random forest ensembles," in *Proceedings of the ICANNGA 2009 Ninth International Conference on Adaptive and Natural Computing Algorithms, Kuopio, Finland*,

DRAFT 14. 5. 2009 – 16. 5.

vol. 5495 of *Lecture Notes in Computer Science*, p. 458-468, Springer, 2009.

[26] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[27] UCL MLG, "Elena database," 1995, http://www.dice.ucl.ac.be/mlg/?page=Elena.

[28] C.B. D.J. Newman, S. Hettich, and C. Merz, "UCI repository of machine learning databases," 1998, http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[29] C. Spearman, "The proof and measurement of association between two things. By C. Spearman, 1904.," *The American journal of psychology*, vol. 100, no. 3-4, pp. 441–471, 1987.

[30] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," 2001.

DRAFT
14. srpna 2009 - 14:32

COMP – Comparison of Matched Ontologies in Protégé

Post-Graduate Student:
ING. PAVEL TYL

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ

Faculty of Mechatronics, Informatics and Interdisciplinary Studies
Technical University of Liberec
Hálkova 6
461 17 Liberec 1, CZ

pavel.tyl@tul.cz

Supervisor:
ING. JÚLIUS ŠTULLER, CSC.

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ

stuller@cs.cas.cz

Field of Study:
Technical Cybernetics

This project is partly realized under the state subsidy of the Czech Republic within the research and development project “Advanced Remediation Technologies and Processes Center” 1M0554 – Programme of Research Centers PP2-DP01 supported by the Ministry of Education and under the financial support of the ESF and the state budget of the Czech Republic within the research project “Intelligent Multimedia E-Learning Portal”, registration No. CZ.1.07/2.2.00/07.0008 – ESF OP EC.

Ontology integration is important in various areas of ontology engineering in e. g. semantic web services, social networks, etc. While particular ontologies usually cover one specific domain, many applications require data from several domains, in general overlapping. Among promising partial solutions to such semantic heterogeneity surely belongs the ontology matching.
Ontology matching can be supported in various ways: by improving matching strategies, tools and systems, basic techniques and methods or by explaining, representing and further processing and evaluating matching results.
The paper describes a matching plug-in into the well-known ontology editor, Protégé [2]. The plug-in is called COMP and it is a general tool for comparing and evaluating matching techniques and strategies.

Ontology matching is in most cases performed *manually* or *semi-automatically*, in general with support of some *graphical user interface*. Manual specification of ontology parts for matching is *time consuming* and moreover *error prone process*. Therefore there is a strong need for the development of faster and/or less laborious methods, which can process ontologies at least semi-automatically.

COMP (Comparing Ontology Matching Plug-in) is a plug-in to Protégé-OWL 4.0. The Protégé-OWL editor is an extension of Protégé that supports the Web Ontology Language (OWL) [1]. An OWL ontology may include descriptions of classes, properties and instances.

COMP is one of the *tools for comparing ontology matching*: it compares various matching algorithms results. Beside being an *evaluation tool*, it can also help to *find appropriate algorithms, methods or their combinations* for different kinds (formats, size, etc.) or parts (basic root concepts, leave concepts, instances, properties, etc.) of ontologies, in case they have different feature set.

The plug-in for ontology matching was thoughtfully proposed with the view of logical separation of objects creating this plug-in. It was very important to elaborate interface of particular objects for easier implementation of advanced (especially) testing classes. The plug-in development is in permanent progress and there are no doubts it may be further improved.

To our best knowledge it is the first attempt to implement matching system to the latest “pure OWL” version of the powerful system Protégé.

References

[1] OWL – Web Ontology Language / W3C Semantic Web Activity [online]: http://www.w3.org/2004/OWL.

[2] Protégé – Ontology Editor and Knowledge Acquisition System [online]: http://protege.stanford.edu.

DRAFT 14. srpna 2009 - 11:32

Information Extraction from Medical Texts

Post-Graduate Student:
ING. KAREL ZVÁRA
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
zvara@euromise.cz

Supervisor:
DOC. ING. VOJTĚCH SVÁTEK, DR.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
svatek@vse.cz

Field of Study:
Biomedical Informatics

Thanks go to my family which supports me with love, morale and patience.

Abstract

This paper is about information extraction from Czech medical texts (mostly summaries). It discusses specifics of Czech medical summaries in the relation to general texts.

1. Introduction

Text-mining is a current field of science and is being widely applied, especially on English texts. Common approaches to text mining combine preprocessing and using statistical methods. The preprocessing phase (tokenization, stemming, lemmatisation, disambiguation) of text analysis deeply depends on underlying language and its syntax. Therefore applying these methods on different languages may require different approaches to the preprocessing phase.

Medical texts, especially medical summaries are the basis for medical services providers. In the Czech Republic, medical summaries bear the form of free form texts. Such medical summaries must adhere to the ordinance of the Ministry of Health No. 64/2007. This ordinance enumerates contentual, formal and temporal features of any part of any medical record.

Concurrent trends of European integration lead to more intense cross-border healthcare cooperation including medical treatment of foreigners. Technologies to hold the information about patient's health status and past procedures regardless of language exist and are being improved (structured EHRs: ASTM CCR, HL7 CDA; domain wide classification systems: SNOMED CT, LOINC). But the problem of converting free text medical texts to such structured records remains. There isn't (and probably wouldn't be in the future) enough will and/or spare resources to manually convert free form medical summaries to some interoperable structured form. The need of converting "free-text" medical summaries to structured electronic form may be fulfilled by text analysis means. This is the motivation of my research.

2. Classic Approaches to Information Extraction from Free Text

Classic approaches to natural language processing consists of these main phases:

1. tokenization,
2. sentence splitting,
3. grammatical tagging,
4. part-of-speech tagging.

Tokenization is a process of separating basic tokens, e.g. word in the field of text analysis. The most common approach of tokenization consists of specifying rules as regular expressions and using some well established tools like *lex*.

Sentence splitting is usually the second phase of the text analysis. It's task is to split tokenized input into groups to be analysed. The free form text is usually formated using sentences, on higher level using paragraphs.

The task of **grammatical tagging** phase is to convert tokens to some kind of common form. Such form is usually nominative of noun, infinitive of verb and so on. Grammatical tagging usually consists of stemming, lemmatizing and other disambiguation. The stemming subphase cuts prefixes, suffixes and so on. Other disambiguation techniques include The lemmatisation aggregates different words with the same meaning base to the same group (e.g. bad and worst). The information

DRAFT 14. září 2009 - 14:32

"virtually lost" during this phase is stored for the further use in the form of grammatical tags.

The **part of speech tagging** (or *PoS tagging*) phase should classify a token. Previous phase of grammatical tagging helps to group words with the same root. The challenge of the phase of PoS tagging is to grammatically disambiguate words of the same spelling (e.g. "srdce", as it may be grammatically classified as singular nominative, singular genitive, singular vocative, plural nominative or plural vocative noun).

Symbolic analysis and various statistical methods are being used to analyze the tokenized and tagged text. The most commonly used approaches employ hidden Markov models (HMMs).

3. Information Extraction

The task of information extraction is to draw an inference from preprocessed input (which is textual in this case). Typical tasks of information extraction include:

- terminology extraction,
- named entity recognition (NER),
- identification of anaphoras (co-reference),
- identification of relationships.

The task of **terminology extraction** identifies individual tokens and token groups as terminology extraction candidates.

The task of **named items extraction** (NER) is to identify and classify basic elements of input such as proper nouns, numbers and dates. NER systems may be either based on grammar or on statistical methods.

The task of **co-reference** identification is to identify relationship between an anaphora (usually pronoun) and an another entity.

Identification of **relationships** usually depends on terminology extraction and named entity recognition.

4. Medical Summaries Specifics

Czech medical summaries are very specific documents. Such documents usually contain a very compressed information (opposing to other texts, e.g. newspaper articles). Some important information is also hidden in the structure of the document because of adherence to the Ministry of Health ordinance.

Since HMMs are a form of generative model, one must enumerate all possible observation sequences. That's problematic because the meaning of parts of speech often depend on the context and sometimes over large range of sequences.

Therefore discriminative probabilistic models (using conditional probability distribution) may be more feasable. One of such models is a model of conditional random fields.

Other authors show that using just regular expressions may be useful but is limited [3] [4], that HMMs are useful but are limited because such models tend to be unable to handle long range (global) correlations [2].

5. Conclusion

I am at the very beginning of my research project. I have at my disposal some underlying data to study (medical summaries) from different sources. Now it's time for me to start analysing them. I am really willing to undertake the challenge.

References

[1] M. Konchady, "Text Mining Application Programming", *Charles River Media, Boston*, 2006.

[2] M. Labský, "Information Extraction from Websites using Extraction Ontologies", *University of Economics, Prague*, 2009.

[3] J. Semecký, "Multimediální záznam o nemocném v kardiologii", *Charles University, Faculty of Mathematics and Physics, Prague*, 2001.

[4] P. Smatana, "(S)pracovanie lekárských správ pre účely analýzy a dolovania v textoch, *Technická univerzita v Košiciach, Košice*, 2005.

[5] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning", *Introduction to Statistical Relational Learning, MIT Press*, 2006.

[6] H.M. Wallach, "Conditional Random Fields", *Technical Report MS-CIS-04-21, University of Pennsylvania*, 2004.

[7] L. Wasserman, "All of Statistics", *Springer, USA*, 2003.

DRAFT 14. srpna 2009 - 14:32

Základní parametry dokumentů doporučených postupů českých lékařských společností publikovaných prostřednictvím Internetu

doktorand:
MUDR. MIROSLAV ZVOLSKÝ
Oddělení medicínské informatiky
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
zvolsky@euromise.cz

školitel:
DOC. ING. ARNOŠT VESELÝ, CSC.
Oddělení medicínské informatiky
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
vesely@pef.czu.cz

obor studia:
Biomedicínská informatika

Abstrakt

Lékařské doporučené postupy jsou odborné dokumenty publikované Odbornými lékařskými společnostmi v tištěné a v poslední době i elektronické formě. Jako zdroj pro další zpracování informací v nich obsažených, ať již odborným či laickým čtenářem, či pro potřeby aplikací biomedicínské informatiky, jsou tyto dokumenty publikovány v různých formátech a v různé míře dodržují základní identifikační a kvalitativní kritéria. Na dvacet českých lékařských společností uveřejňuje v rámci svých internetových prezentací celkem 426 těchto dokumentů, necelá polovina z nich je ve formátu PDF. Pouze u 63,4 procent dokumentů je ve vlastním textu uveden jeho autor, ve 47,4 procentech dokumentů pak vlastní odborná společnost. Uvádění těchto informací ve vlastnostech dokumentů je zcela ignorováno.

1. Úvod

Lékařské doporučené postupy jsou informativní dokumenty publikované lékařskými odbornými autoritami (národními či nadnárodními odbornými společnostmi, lékařskými sdruženími, státními zdravotními institucemi apod.), jejichž cílem je stanovení nejlepšího postupu a popis rozhodovacího procesu v daném typu klinických případů na základě nejnovějších vědeckých poznatků a uvedení těchto procesů do klinické praxe [1].

Tyto dokumenty jsou publikovány prostřednictvím odborných periodik a věstníků, v rámci monografií i jako samostatné tištěné dokumenty. V posledních letech dochází k publikování lékařských doporučených postupů elektronicky na paměťových médiích nebo prostřednictvím Internetu, přičemž tyto dokumenty jsou buď elektronickou kopií textů uveřejňovaných v tištěné podobě, nebo jsou i primárně šířeny pouze v elektronické formě.

V České republice patří mezi odborné autority publikující lékařské doporučené postupy prostřednictvím Internetu České lékařská společnost Jana Evangelisty Purkyně (ČLS JEP, http://www.cls.cz), odborné lékařské společnosti, které zastřešuje a odborné lékařské společnosti působící samostatně mimo rámec ČLS JEP.

Publikování prostřednictvím Internetu přináší výhody ve formě nízkých ekonomických nákladů, snadné dostupnosti pro širokou odbornou i laickou veřejnost a možnosti rychlé aktualizace a nízkých omezení velikosti uveřejňovaných informací. Mezi nevýhody patří také široká dostupnost a s ní spojená nutnost udržovat informace aktuální, dále opatření a náklady spojené s uspořádaností informací (struktura webové prezentace, kde jsou dokumenty publikovány, registrace a propagace v katalozích a dalších internetových službách, SEO) a důvěryhodností publikovaných informací [2, 3, 4].

Mezi nejzákladnější opatření, která zvyšují uživateli Internetu přístup k publikovaným dokumentům a zvyšují důvěryhodnost a použitelnost informací v nich obsažených je dodržování základních formálních pravidel pro publikování elektronických informací na Internetu - tedy publikování ve standardních formátech, poskytování informací o autorovi, odborném garantovi a časové aktuálnosti dokumentu.

I v případě publikování elektronickou cestou jsou ovšem dokumenty pouze textové bez informací o vnitřní struktuře obsahu. V některých případech obsahují doplňující materiál ve formě tabulek, obrázku,

či schémat. Postup rozhodovacího procesu nebývá formálně popsán a k použití těchto lékařských doporučených postupů v aplikacích biomedicínské informatiky (např. v systémech pro podporu rozhodování) je nutné textové dokumenty dále zpracovávat a vytvářet jejich formální modely [5, 6].

V současnosti jsou vyvíjeny nástroje, které umožňují ruční nebo do různé míry automatizované zpracování textů lékařských doporučených postupů do formy dále použitelné v biomedicínských aplikacích. Existují také nástroje na hodnocení kvalitativních kritérií lékařských informací zveřejnovaných prostřednictvím internetových služeb. Neexistuje ovšem zatím žádné kvalitativní hodnocení ani přehled publikačních autorit ani publikovaných dokumentů lékařských doporučených postupů v České republice [7, 8].

Cílem této práce je zmapovat aktivitu českých odborných lékařských autorit, které publikují lékařské doporučené postupy elektronicky prostřednictvím Internetu a získat přehled o formátech dokumentů a jejich základních identifikačně-kvalitativních parametrech, neboť tyto dokumenty mohou působit jako podklady pro vytváření formálních modelů doporučených postupů s následným použitím v systémech pro podporu rozhodování ve zdravotnictví či jiných biomedicínských aplikacích.

2. Metody

Pro srovnání jsem zvolil z přibližně sedmdesáti českých odborných lékařských společností, které vlastní internetovou prezentaci, dvě desítky těch, které na svých stránkách publikují vždy více než pět dokumentů doporučených postupů a je tedy předpoklad, že se tvorbě a publikování těchto doporučení soustavně věnují. Základní podmínkou bylo, aby posuzované dokumenty byly volně zobrazitelné z internetové prezentace lékařské společnosti jakémukoliv zájemci bez registrace a zdarma. Soubor s textem lékařského doporučeného postupu musel být z internetové prezentace nejen přímo odkazován, musel být také umístěn na stejné doméně (dle adresy Uniform Resource Locator).

Seznam českých odborných lékařských společností a jejich Internetových prezentací splňujících výše zmíněná kritéria je obsahem Tabulky 1.

Název společnosti	URL webové prezentace
Česká angiologická společnost	http://www.angiologie.cz
Česká dermatovenerologická společnost	http://www.lfhk.cuni.cz
Česká diabetologická společnost	http://www.diab.cz
Česká gastroenterologická společnost	http://www.cgs-cls.cz
Česká hematologická společnost	http://www.hematology.cz
Česká hepatologická společnost	http://www.ceska-hepatologie.cz
Česká kardiologická společnost	http://www.kardio-cz.cz
Česká neurologická společnost	http://www.czech-neuro.cz
Česká onkologická společnost	http://www.linkos.cz
Česká pneumologická a ftizeologická společnost	http://www.pneumologie.cz
Česká revmatologická společnost	http://www.revma.cz
Česká společnost anesteziologie, resuscitace a intenzívní medicíny	http://www.csarim.cz
Česká společnost klinické biochemie	http://www.cskb.cz
Česká společnost pro aterosklerózu	http://www.athero.cz
Radiologická společnost	http://www.crs.cz
Společnost českých patologů	http://www.patologie.info
Společnost infekčního lékařství	http://www.infekce.cz
Společnost pro transfuzní lékařství	http://www.transfuznispolecnost.cz
Společnost urgentní medicíny a medicíny katastrof	http://www.urgmed.cz
Společnost všeobecného lékařství	http://www.svl.cz

Tabulka 1: Přehled dvaceti odborných lékařských společností publikujících prostřednictvím Internetu více než 5 dokumentů lékařských doporučených postupů

Jednoduchá kritéria hodnocení, kterým jsem zde vystavené dokumenty podrobil, se dají rozdělit do tří skupin:

a) formát souboru
b) základní identifikační údaje v textu či vlastnostech souboru
c) zabezpečení souboru

DRAFT 14. srpna 2009 – 14:32

Běžně používanými formáty souborů v soudobé elektronické komunikaci jsou HyperText Markup Language (HTML), Portable Document Format (PDF), Rich Text Format (RTF), Microsoft Office Word Document Format (DOC), Office Open XML (OOXML), OpenDocument Format (ODF), Joint Photographic Experts Group File Format (JPEG) a Graphics Interchange Format (GIF).

Prostřednictvím formátu HTML jsou prezentované dokumenty integrovány do vlastní webové prezentace a je nutné je pak považovat za její součást, ať již grafickou a typografickou formou, zařazením do struktury a kontextu stránek, tak i dalším využitím. Takto formátované dokumenty nejsou primárně určeny pro samostatné šíření a další zpracování.

Formát PDF je naproti tomu určen k profesionálnímu šíření samostatných dokumentů a pro publikaci lékařských doporučených postupů v textové podobě se jeví jako optimální. Neumožňuje přímou editaci obsahu bez speciálního komerčního software, naopak poskytuje nástroje na zabezpečení textu - šifrování obsahu, uzamčení souboru, resp. přístup pod heslem, omezení kopírování nebo tisku jednotlivých částí obsahu, ap. Díky uložení všech podrobných informací o formátování textu se formát vyznačuje vysokou kompatibilitou výsledného zobrazení na různých výstupních zařízeních. Pro zobrazení souboru existuje zdarma dostupný software.

Souborový formát RTF je starší formát vyvinutý firmou Microsoft pro sdílení textů mezi jednotlivými textovými editory, je široce kompatibilní napříč technickými platformami, ovšem neumožňuje pokročilejší zabezpečení dokumentu.

Formát DOC je široce rozšířený díky majoritnímu postavení firmy Microsoft na trhu kancelářského aplikačního software, existuje v několika verzích vztažených k verzím aplikačního balíku Microsoft Office. Ve verzi Word 97-2003 umožňuje šifrování a ochranu úprav heslem. K dispozici existuje zdarma distribuovaný prohlížeč.

Grafické formáty JPEG a GIF nejsou určeny k publikaci textových informací, ve výjimečných případech ovšem slouží jako prostředek k publikování vytištěného a následně digitalizovaného textu, když není jiná elektronická forma textu k dispozici.

Nové volně dostupné otevřené formáty OOXML a ODF jsou sice určeny k publikaci a sdílení textových informací, nejsou však v současné době tak běžně rozšířeny.

Základní identifikační údaje, které mohou sloužit k ověření kvality dokumentu jako informačního zdroje, jsou uvedení autora, uvedení názvu lékařské odborné společnosti, či jiné odborné autority, která vznik a autenticitu dokumentu garantuje a uvedení data, od kterého dokument nabývá platnost. Pokud je text elektronicky šířen například vystavením na Internetu, všechny tyto údaje by měly být uvedeny přímo v textu, aby se s nimi čtenář mohl seznámit a použít je jako parametry k posouzení kvality textu a vhodnosti jeho využití v dané konkrétní klinické situaci.

Za uvedení autora v textu bylo pro potřeby této práce považováno uvedení příjmení a křestního jména (nebo zkratka křestního jména) alespoň jednoho autora.

Za uvedení lékařské odborné společnosti v textu bylo považováno uvedení v textu názvu nebo loga té společnosti, jejíž internetové prezentace byl dokument součástí.

Údaje o autorovi a instituci je možné ve všech formátech (kromě grafických) uvést též ve vlastnostech dokumentu. Tyto informace bývají automaticky předvyplněny textovými editory, přičemž jako autor je zmíněn aktuálně přihlášený uživatel, jako jméno instituce je použit řetězec zadaný při registraci software. Vzhledem k časté absenci bezpečnostní politiky při používání výpočetní techniky a nedbalosti při registraci software nemají většinou předvyplněné hodnoty žádnou informační hodnotu. Navíc autor textu většinou používá software, kteý není registrovaný odbornou lékařskou společností.

Pro potřeby této práce bylo za uvedení autora ve vlastnostech dokumentu považováno uvedení příjmení a křestního jména (nebo zkratka křestního jména) alespoň jednoho autora zmíněného v samotném textu dokumentu.

Za uvedení lékařské odborné společnosti ve vlastnostech dokumentu bylo považováno uvedení názvu té společnosti, jejíž internetové prezentace byl dokument součástí.

Za uvedení data zahájení platnosti bylo pro potřeby této práce považováno uvedení data, nebo měsíce a roku, nebo jen roku v hlavičce či zápatí textu dokumentu, datum schválení příslušným orgánem odborné společnosti uvedené v textu, nebo alespoň údaj o roce v názvu dokumentu.

Údaj o datu resp. času vzniku souboru je součástí informací obsažených v souborovém systému a je dohledatelný. Ovšem nemusí být totožný s datem

DRAFT 14. srpna 2009 - 14:12:02

zahájení platnosti dokumentu lékařského doporučeného postupu, proto by bylo jeho užití při posuzování kvalitativních parametrů dokumentu spekulativní.

Zabezpečení proti změně obsahu, kopírování obsahu, tisku obsahu dokumentu umožňují z posuzovaných formáty PDF a DOC. Zjišťováno bylo, zda jsou posuzované dokumenty opatřeny některou z těchto forem zabezpečení.

3. Výsledky

Průzkumu bylo podrobeno celkem 426 dokumentů lékařských doporučených postupů publikovaných v rámci webových prezentací dvaceti českých lékařských odborných společností. Pro porovnání jsou uvedeny i údaje popisující 305 dokumentů doporučených postupů poblikovaných Českou lékařskou společností Jana Evangelisty Purkyně v letech 1999-2001. Uvedená data byla ověřena ke dni 17. 7. 2009.

Výsledky dokumentů dvaceti odborných lékařských společností					
Formát PDF	Formát HTML	Formát DOC	Formát RTF	Formát GIF,JPEG	Celkem
195	155	73	1	2	426
Srovnání - výsledky dokumentů ČLS JEP					
Formát PDF	Formát HTML	Formát DOC	Formát RTF	Formát GIF,JPEG	Celkem
0	0	0	305	0	305

Tabulka 2: Přehled zjištěných formátů souborů publikovaných dokumentů lékařských doporučených postupů

Souhrnné výsledky výskytu jednotlivých formátů souborů ukazuje Tabulka 2, podle které je nejčastějším formátem PDF, ve kterém je publikováno téměř 46 procent všech dokumentů. Podrobný přehled formátů souborů podle jednotlivých odborných lékařských společností je zobrazen v Tabulce 3. Jak z ní vyplývá, pouze dvě odborné společnosti zvolily jednotný formát publikovaných dokumentů - jsou to Společnost všeobecného lékařství a Česká diabetologická společnost. Ostatní společnosti publikují dokumenty v různých nebo paralelně ve více formátech.

Výsledky podle jednotlivých odborných lékařských společností						
Název autority	PDF	HTML	DOC	RTF	GIF/JPEG	Celkem
Společnost všeobecného lékařství	47	0	0	0	0	47
Česká kardiologická společnost	9	15	0	0	0	24
Česká dermatovenerologická společnost	0	36	0	0	0	36
Česká gastroenterologická společnost	0	4	10	0	0	14
Česká neurologická společnost	0	4	14	0	0	18
Česká pneumologická a ftizeologická společnost	13	0	9	0	0	22
Česká angiologická společnost	0	2	6	0	0	8
Česká diabetologická společnost	13	0	0	0	0	13
Česká hepatologická společnost	5	13	0	0	0	18
Česká onkologická společnost	40	39	0	0	0	79
Česká revmatologická společnost	1	8	0	0	0	9
Česká společnost klinické biochemie	18	16	0	0	0	34
Česká společnost pro aterosklerózu	6	4	0	0	0	10
Společnost infekčního lékařství	9	6	4	0	0	19
Společnost urgentní medicíny a medicíny katastrof	12	1	2	0	2	17
Společnost českých patologů	0	1	7	1	0	9
Společnost pro transfuzní lékařství	1	0	7	0	0	8
Radiologická společnost	6	5	3	0	0	14
Česká hematologická společnost	1	1	11	0	0	13
Česká společnost anesteziologie, resuscitace a intenzívní medicíny	14	0	0	0	0	14

Tabulka 3: Přehled zjištěných formátů souborů podle jednotlivých odborných lékařských společností

14. srpna 2009, 4:52

V Tabulce 4 jsou uvedeny počty dokumentů splňujících výše uvedená kritéria informací o autorovi, odborné autoritě a aktuálnosti obsahu dokumentu. Není zobrazeno kritérium uvedení odborné autority/společnosti ve vlastnostech dokumentu, protože toto kritérium nesplnil žádný dokument.

Výsledky podle jednotlivých odborných lékařských společností					
Název autority	A1	A2	S	D	Celkem
Společnost všeobecného lékařství	47	0	47	47	47
Česká kardiologická společnost	24	0	9	7	24
Česká dermatovenerologická společnost	29	0	10	0	36
Česká gastroenterologická společnost	12	4	8	5	14
Česká neurologická společnost	18	2	13	5	18
Česká pneumologická a ftizeologická společnost	13	1	7	4	22
Česká angiologická společnost	8	2	2	3	8
Česká diabetologická společnost	5	1	13	9	13
Česká hepatologická společnost	11	0	7	6	18
Česká onkologická společnost	0	0	0	0	79
Česká revmatologická společnost	4	0	3	1	9
Česká společnost klinické biochemie	29	0	23	30	34
Česká společnost pro aterosklerózu	10	0	9	0	10
Společnost infekčního lékařství	15	0	3	7	19
Společnost urgentní medicíny a medicíny katastrof	16	1	14	14	17
Společnost českých patologů	7	0	5	6	9
Společnost pro transfuzní lékařství	2	0	5	6	8
Radiologická společnost	3	0	4	3	14
Česká hematologická společnost	4	2	7	5	13
Česká společnost anesteziologie, resuscitace a intenzívní medicíny	13	0	13	9	14
Celkem	270	13	202	167	426
Celkem v procentech počtu dokumentů	63,4	3	47,4	39,2	100
Česká lékařská společnost Jana Evangelisty Purkyně	305	0	0	305	305

Tabulka 4: Přehled počtu dokumentů lékařských doporučených postupů splňujících základní kritéria - uvedení autora, odborné společnosti a data zahájení platnosti podle jednotlivých odborných lékařských společností. Vysvětlivky: A1 - autor uveden v textu, A2 - autor uveden ve vlastnostech dokumentu, S - odborná autorita/společnost uvedena v textu, D - Datum uvedeno v textu

Zabezpečení souboru bylo zaznamenáno pouze u jednoho dokumentu publikovaného Českou společností klinické biochemie. Jednalo se o zákaz tisku souboru ve formátu PDF.

4. Diskuse

Průzkum mezi v oblasti doporučených postupů publikačně nejaktivnějšími autoritami v České republice ukázal velkou nejednotnost formátů publikovaných dokumentů a to nejen mezi jednotlivými odbornými společnostmi, ale často i v rámci jedné webové prezentace.

V oblasti uveřejňování základních identifikačních údajů, které jsou mnohdy klíčové pro posuzování kvality informací vyhledaných v Internetu jen 63,4 procent dokumentů lékařských doporučených postupů uvádí v textu autora, 47,4 procenta odbornou autoritu a 39,2 procenta časový údaj vztahující se k období platnosti dokumentu. V naprosté většině jsou ignorovány možnosti umístění těchto kvalitativních informací do vlastností dokumentu u formátů PDF či DOC nebo hlavičky HTML dokumentů.

Obecně jednotnou formu a dodržování obsahových náležitostí splňují dokumenty primárně vytvořené pro publikování v tištěných periodicích, což s sebou ovšem nese jiné komplikace, především otázky autorských práv, formátu určeného pro tisk (například nadbytečných typografických informací), ale i matoucí grafické zpracování a údaje v záhlaví dokumentů.

V publikování dokumentů lékařských doporučených postupů prostřednictvím elektronických formátů a Internetu se projevuje nejednotnost a absence metodiky,

pouze dokumenty ČLS JEP (které však již nejsou aktualizovány a jejich informační hodnota rychle zastarává) a dokumenty Společnosti všeobecného lékařství (pod patronací Centra doporučených postupů SVL) dodržují jednotnou formu. V případě posledně jmenovaných také proto, že se jedná o elektronické verze knižně vydávaných publikací.

Identifikační údaje, stejně jako další někdy i podrobné informace o formátu či velikosti dokumentů, jsou mnohdy uváděny v rámci textu webové prezentace související s odkazem na soubor dokumentu lékařských doporučených postupů, přičemž v jeho vlastním textu poté chybí. Vzhledem k tomu, že samotné soubory jsou ovšem indexovány internetovými katalogy a automatickými webovými službami a také proto, že se mohou šířit i samostatně, jsou takové informace nedostačující a pokud mají být efektivně šířeny, musí být zopakovány v textu dokumentu.

Zabezpečení dokumentů je často podceňováno, především formát DOC lze upravovat v mnoha volně dostupných aplikacích a měnit jeho obsah. Zvláště při dalším šíření z Internetu získaných a tištěných informací by pak snadno mohlo dojít k pozměnění obsahu těchto odborných dokumentů. Jediný dokument, který měl aktivované nadstandardní bezpečnostní funkci (byl formátu PDF), měl pouze omezenou možnost vytištění dokumentu a tím spíše omezoval šíření jistým způsobem garantovaných informací.

Protože všechny posuzované dokumenty lékařských doporučených postupů mohou sloužit jako zdroj informací při vytváření jejich formálních modelů, dodržování jednoduchých pravidel při jejich tvorbě, především uvádění identifikačních a katalogizačních informací a co největší strukturovanost textu, mohou další práci s dokumenty velmi usnadnit.

Literatura

[1] M. Peleg, "Guideline and Workflow Models", *In: Medical Decision-Making: Computational Approaches to Achieving Healthcare Quality and Safety, Robert A. Greenes (ed.)*, Elsevier/Academic Press, 2006.

[2] J.M. Grimshaw and I.T. Russell, "Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations", *Lancet 342 (8883)* 1317-1322, 1993.

[3] G. Eysenbach and D.L. Diepgen, "Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information", BMJ;317, s.1496-1502 http://bmj.com/cgi/content/full/317/7171/1496, 1998.

[4] J. Menoušek, "Medicínské informace na internetu. Klasifikace hodnotících systémů", *Inforum*, 2003.

[5] D. Buchtela, J. Peleška, A. Veselý, J. Zvárová, and M. Zvolský, "Model reprezentace znalostí v doporučeních", EJBI, 2008, http://www.ejbi.cz/articles/200812/34/2.html.

[6] A. ten Teije, M. Marcos, M. Balser, J. van Croonenborg, C. Duelli, F. van Harmelen, et al., "Improving medical protocols by formal methods", *Artif. Intell. Med.* 36 (3) 193-209, 2006.

[7] P. Kasal, A. Janda, J. Feberova, T. Adla, M. Hladikova, J.P. Naidr, and R. Potuckova, "Evaluation of health care related web resources based on web citation analysis and other quality criteria", Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference, Issue, 17-18 Jan. 2006 Page(s):2391 - 2394.

[8] J. Kosek, M. Labsky, J. Nemrava, M. Ruzicka, and V. Svatek, "Projekt MedIEQ: hodnocení zdravotnických webových zdrojů s využitím extrakce informací (in Czech)", In: Datakon 2006, Proceedings of the Annual Database Conference, October 2006, Brno, Czech Republic, 267-270.

DRAFT 14. srpna 2009 11:32

DRAFT
14. srpna 2009 - 14:32

Ústav informatiky AV ČR, v. v. i.
DOKTORANDSKÉ DNY '09

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
University Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou – *not yet* – . publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2009

ISBN – *not yet* –