Doktorandské dny ’08

Hakl, František
2008


September 29.9.–1.10., 2008, Jizerka
Proceedings of the XIII. PhD. Conference
Edited by F. Hakl
Ústav informatiky AV ČR, v. v. i.
Academy of Sciences of the Czech Republic, v. v. i.
Institute of Computer Science
Doktorandské dny '08

Doktorandské dny ’08

Ústav informatiky AV ČR, v. v. i.

Jizerka

29. září – 1. října 2008

vydavatelství Matematicko-fyzikální fakulty
Univerzity Karlovy v Praze

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

ISBN 978-80-7378-054-8

Doktorandské dny Ústavu informatiky AV ČR, v. v. i., se konají již potřinácté, nepřetržitě od roku 1996. Tento seminář poskytuje doktorandům, 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 doktorandů podává průřezovou informaci o odborném rozsahu pedagogických aktivit, které jsou realizovány na pracovištích či za spoluúčasti Ústavu informatiky.

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

Vedení Ústavu informatiky jakožto organizátor doktorandských dnů věří, že toto setkání mladých doktorandů, jejich školitelů 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 kontaktů.

1. září 2008

Obsah

Methods for Identifying Candidate Genes for Cardiovascular Diseases by Using Microarrays

Post-Graduate Student:
MGR. JANA ADÁŠKOVÁ
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
adaskova@euromise.cz

Supervisor:
PROF. RNDR. JANA ZVÁROVÁ, DRSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
zvarova@euromise.cz

Field of Study:
Biomedical Informatics

The work was supported by the grant 1M06014 of the Ministry of Education, Youth and Sport of the Czech Republic.

Abstract

Microarrays present new powerful technique for high-throughput, global transcriptomic profiling of gene expression. It permits to investigate the expression levels of thousands of genes simultaneously. The global snapshots of gene expression, both among different cell types and among different states of a particular cell type can help in identifying candidate genes that may be involved in a variety of normal or disease processes. This promises to provide insight into the pathophysiology of human syndromes such as cardiovascular diseases, whose etiologies are due to multiple genetic factors and their interaction with the environment.

Microarrays also present new statistical and bioinformatical problems because the data are very high dimensional with very little replication. Almost all research employing microarray expression analysis depends heavily on statistical analysis to extract the most useful information from the huge number of data points generated.

The aim of this paper is to present possibilities of use of microarrays for identifying candidate genes for cardiovascular diseases and specially attention is devoted to statistical methods for identifying differentially expressed genes from microarray data.

Keywords: microarray, gene expression, cardiovascular diseases, microarray data, SAM, Bayes T-test, samroc, Zhao-Pan method.

1. Introduction

Identification of genetic determinants that predispose to common diseases such as cardiovascular diseases is a major challenge for current biomedical research. Despite recent advances in molecular and statistical genetics and the availability of complete genome sequences of humans and animal models, however, the underlying molecular pathogenic mechanisms for these disorders are still largely unknown. Nowadays a valuable tool for increasing our understanding of the regulatory and functional complexity of the molecular basis of multifactorially determined diseases is expression profiling.

Gene expression profiling is a logical next step after sequencing a genome: the sequence tells us, what the cell could possibly do, while the expression profile tells us, what it is actually doing now. Genes contain the instructions for making messenger RNA (mRNA), but at any moment each cell makes mRNA from only a fraction of the genes it carries. If a gene is used to produce mRNA, it is considered ”on”, otherwise ”off”. Expression profiling experiments involve measuring the relative amount of mRNA expressed in two or more experimental conditions. This is because altered levels of a specific sequence of mRNA suggest a changed need for the protein coded for by the mRNA, perhaps indicating a homeostatic response or a pathological condition. Therefore gene expression profiling can help in identifying candidate genes that may be involved in a variety of normal or disease processes. Additionally, characterization of genes abnormally expressed in diseased tissues may lead to the discovery of genes that can serve as diagnostic markers, prognostic indicators or targets for therapeutic intervention.

The development of several gene expression profiling methods, such as comparative genomic hybridization (CGH), differential display, serial analysis of gene expression (SAGE) and gene microarray, together with the sequencing of the human genome, has provided an opportunity to monitor and investigate the complex

cascade of molecular events leading to cardiovascular diseases [2]. High-throughput technologies can be used to follow changing patterns of gene expression over time. Among them, gene microarray has become prominent because it is easier to use, does not require large-scale DNA sequencing, and allows for the parallel quantification of thousands of genes from multiple samples. Nowadays gene microarray technology is rapidly spreading worldwide and has the potential to drastically change the therapeutic approach to patients affected with cardiovascular or others complex diseases [3]. Therefore, it is important to know the principles underlying the analysis of the huge amount of data generated with microarray technology.

2. Microarray technology

Microarray technology takes advantage of hybridization properties of nucleic acid (DNA or RNA) and uses complementary molecules attached to a solid surface, referred to as probes, to measure the quantity of specific nucleic acid transcripts (mRNA) of interest that are present in a sample, referred to as the target. The molecules in the target are labelled, and specialized scanner is used to measure the amount of hybridized target at each probe, which is reported as an intensity. The raw or probe-level data are the intensities of each spot on the hybridization array, from which the initial concentrations of the corresponding transcripts are inferred.

Various manufacturers provide a large assortment of different platforms. The different platforms can be divided into two main classes that are differentiated by the data they produce. The high-density oligonucleotide array platforms produce one set of probe-level data per microarray with some probes designed to measure specific binding and others to measure non-specific binding. The two-color spotted platforms produce two sets of probe-level data per microarray (the red and green channels), and local background noise levels are measured from areas in the glass slide not containing probes [4]. Despite the differences among the different platforms, the steps of microarray data analysis are similarly to all microarray technology.

3. Microarray data analysis

Microarray experiments produce a huge amount of data. A single microarray run can produce between 100,000 and a million data points, and a typical experiment may require tens or hundreds of runs [5]. Microarray data analysis consist of three parts: (i) data preparation, in which data are adjusted for the downstream algorithms; (ii) algorithm selection for data analysis; and (iii) interpretation, in which the results from the algorithms are explained in a biological context. In Fig. 1 are shown the major phases of microarray data analysis (colored icons) and their connectivity (arrows) in the microarray workflow process.

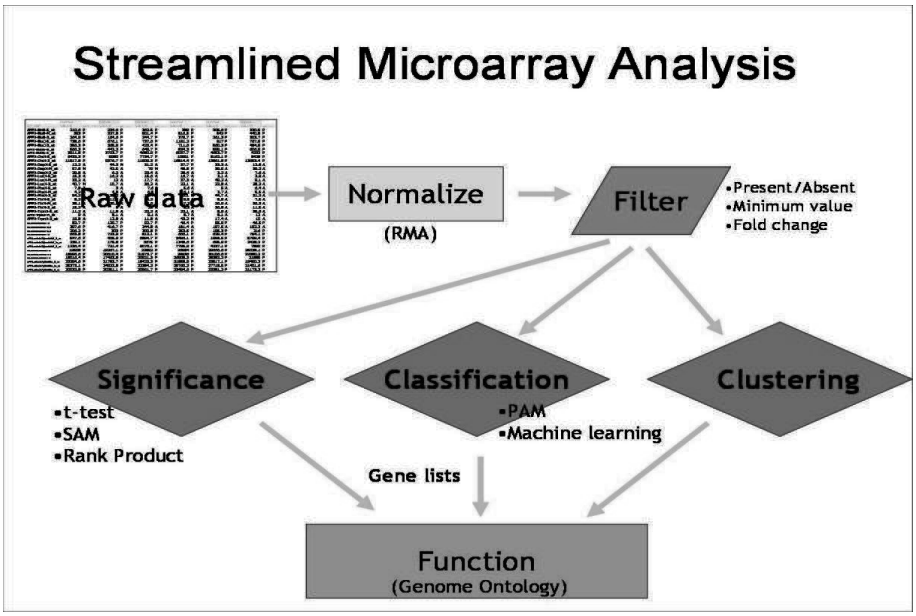


Figure 1: Microarray data analysis.

3.1. Low-Level analysis

Primary image data having been collected from a microarray experiment. The aims of the first level of analysis, so-called low-level analysis or data preprocessing, are image analysis, background elimination, filtration, normalization and data transformation, all of which should contribute to the removal of systematic variation between chips, enabling group comparisons.

Image analysis permits us to convert pixel intensities in the scanned images into probe-level data. Many image-processing approaches have been developed, among which the main differences relate to how spot segmentation, distinguishing foreground from background intensities, is carried out [4]. Another important preprocessing step is normalization. Normalization involves comparing different microarrays relative to some standard intensity value. This could be the overall intensity of the microarray, the overall intensity of all of the genes on the microarray, the intensity of so-called housekeeping genes (the expression of which are supposedly constant), or spiked targets, containing a known and constant amount of a labelled control. Negative normalization controls might be represented by target sequences from a different organism. Several normalization approaches have been introduced, and are discussed elsewhere [4]. Data are often then subjected to log transformation to improve the characteristics of the distribution of the expression values.

3.2. Statistical analysis

Microarrays present new statistical problems because the data are very high dimensional with a very small number of replications. A common task in analyzing microarray data is to determine which genes are differentially expressed across two tissue samples or samples obtained under two experimental conditions.

In early days, the simple method of fold changes was used. Simple and intuitive, this method, involves the calculation of a ratio relating the expression level of a gene under control and experimental conditions. An arbitrary ratio (usually 2-fold) is then selected as being ”significant.” Because this ratio has no biological merit, this approach amounts to nothing more than a blind guess. The selection of an arbitrary threshold results in both low specificity (false positives, particularly with low-abundance transcripts or when a data set is derived from a divergent comparison) and low sensitivity (false negatives, particularly with high-abundance transcripts or when a data set is derived from a closely linked comparison) [6]. It is now accepted that the use of the fold change method should be discontinued.

Since then, many more sophisticated methods have been proposed (e.g. Chen et al 1997, Efron et al 2000, Ideker et al 2000, Newton et al 2001, Tusher et al 2001, Lin et al 2001, Pan et al. 2001) [3]. It has been also noticed that data based on a single array may not reliable and may contain high noises. As the technology advances, microarray experiments are becoming less expensive, which make the use of multiple arrays feasible. Most, if not all, statistical tests can be modified accordingly for a multiple comparison adjustment.

In this section I would like to review more in detail two types of parametric methods (such as T-test and Bayes T-test) and three types of non-parametric methods (such as samroc, SAM, and a modified mixture model proposed by Zhao and Pan) recently used for identifying differentially expressed genes in microarray data. Suppose that the experimental data consist of measurements y_{gi} under two conditions, where i ($i = 1, 2, ..., k$) denotes the i-th array, g ($g = 1, 2, ..., G$) denotes the g-th gene, and k_1 and k_2 are the number of arrays for each condition, that is, $k = k_1 + k_2$. Let the sample means and the sample variances of y_{gi}'s for gene g under two conditions be denoted as $\overline{y}_{g1}$, s^2_{g1} and, $\overline{y}_{g2}$, s^2_{g2} respectively. Here, diff is the difference between $\overline{y}_{g1}$ and $\overline{y}_{g2}$, and s_g and Se_g represent the pooled standard deviation and the standard error of the diff across the replicates for the gene, respectively.

3.2.1 T-statistics: The two sample T-statistics with two independent normal samples without assuming the equal variances between two samples could be written as follows:

$$t_g = \frac{diff}{Se_g}, Se_g = \sqrt{\frac{s^2_{g1}}{k_1} + \frac{s^2_{g2}}{k_2}}$$

A gene with very small variance due to its low expression level contributes to have large absolute t-value regardless of the mean difference under two conditions, and thus this gene can be selected as the differentially expressed gene although it is not truly differentially expressed. To overcome this problem of the traditional T-test, various methods have been proposed. Among these methods, there are SAM and samroc (see below).

3.2.2 Bayes T-test: Baldi and Long [7] developed a Bayesian probabilistic framework for microarray data analysis. Their statistics is used to solve small variance problems in low expression level and uses the parametric Bayesian method to have the parameters (mean, standard deviation and so on.)

for T-statistics. This statistics is well known for its effectiveness in analyzing the samples having small size, but it still heavily depends on the parametric assumption. Bayes T-test uses the estimate of parameters such as population mean (μ) and variance (σ^2) by Bayesian method instead of sample mean and sample variance of the traditional T-statistics. The mean of posterior estimate in each group is given as

$$\mu_j = \mu_{nj}, \sigma_j^2 = \frac{\nu_j \sigma_{nj}^2}{\nu_j - 2},$$

where the mean of the posterior estimate (μ_{nj}) is a convex weighted average of the prior mean (μ_{0j}) and the sample mean $\overline{y}_j$ for group j, $j = 1, 2$, that is,

$$\mu_{nj} = \frac{\lambda_{0j}}{\lambda_{0j} + k_j}\mu_{0j} + \frac{k_j}{\lambda_{0j} + k_j}\overline{y}_j$$

The hyperparameters μ_{0j} and σ_j^2/λ_{0j} can be interpreted as the location and the scale of μ_j, respectively, and k_j is the sample size for each group. σ_{nj}^2 is posterior variance component and posterior sum of squares is

$$\nu_j \sigma_{nj}^2 = \nu_{0j}\sigma_{0j}^2 + (k_j - 1)s_j^2 + \lambda_{0j}k_j/(\lambda_{0j} + k_j)(\overline{y}_j - \mu_{0j})^2,$$

and the posterior degree of freedom is $v_j = v_{0j} + k_j$. In Bayes T-test, the hyperparameters for the prior v_{0j} and σ_{0j}^2 can be interpreted as the degree of freedom and scale of σ_j^2, respectively [7]. Owing to the complicated theoretical background, I will not discuss it here in more detail. This statistics is currently implemented in the Limma software package [8] as part of project Bioconductor accessible at www.bioconductor.org .

3.2.3 Significant analysis of microarrays (SAM): To avoid the small variance problem of T-test, SAM uses a statistics similar to T-statistics and the permutation of repeated measurements to estimate the false discovery rate [9]. At low expression levels, the absolute value of t_{sam} can be high because of small values in Se_g. The shortcoming of the traditional T-test is that genes with small sample variances due to the low expression levels have high chance of being declared as the differentially expressed genes. Thus SAM added a small positive constant a to alleviate this problem. The SAM statistics is

$$t_{sam} = \frac{diff}{Se_g + a}, Se_g = s_g\sqrt{\frac{1}{k_1} + \frac{1}{k_2}},$$

where the value for a is chosen to minimize the coefficient of variation. SAM is similar to the method by Efron et al. [10], which use a to be equal to the 90th percentile of the standard errors of all the genes. SAM assigns a score based on changes that is related to the standard deviation of repeated measurements for that gene. Genes with scores greater than a cutoff value are determined to be significant.

3.2.4 Samroc: Broberg [11] proposed a method for ranking genes in the order of likelihood of being differentially expressed, which is often called as samroc. The main purpose of this method is to estimate the false negative (FN) and false positive (FP) rates. The procedure sets out to minimize these errors. The samroc method is similar to SAM, although an added constant in the denominator of the statistics is different. The proposed statistics is

$$t_{sam} = \frac{diff}{Se_g + b}.$$

Main interest is to find the optimal constant b for given significance level of α. This procedure proposed a criterion, which is the distance of points on the curve to the origin, for choosing a good receiver operating characteristic (ROC) curve. ROC curve allows users to compare the FP error rate and FN error rate of various test statistics without involving P-values. This minimizes the number of genes that are falsely declared positive and falsely declared negative for a given significance level of α and a value b [11].

3.2.5 Zhao-Pan method: Zhao and Pan [12] adopted a modified non-parametric approach to detect the differentially expressed genes in replicated microarray experiments. The basic idea of this non-parametric method lies in estimating the null distribution of test statistics, say Z_g, by directly constructing a null statistics, say z_g, such that the distribution of z_g is the same as the distribution of Z_g under the null hypothesis. This avoids the strong assumptions about the null distribution of the parametric methods. A common problem with these methods is that the numerator and the denominator of z_g and Z_g are assumed to be independent of each other. In practice, this independency is violated by z_g, and z_g and Z_g are used to overcome this problem. For more details refer to the Zhao and Pan [12].

Method	Sample	Distributional	Equal variance assumption between groups
T-statistics	Large	Strong	Unequal
B-statistics	Small	Strong	Unequal
SAM	Small	None	Equal
samroc	Small	None	Equal
Zhao-Pan	Large	Weak	Equal

Table 1: The main features of the statistical methods .

Table 1 summarizes main features of the previous described methods in the context of sample size, distributional assumption, and variance condition between two groups. In general, SAM, samroc and Bayes T-test are known to work well with the small sample size, and T-statistics and Zhao-Pan method are known to perform well with large sample size. This difference may be related to the fact that SAM and samroc do not need any distributional assumption, whereas the others need distributional assumptions for the analysis. Of these five methods, SAM, samroc and Zhao-Pan method require the equal variance assumption between two groups.

3.3. High-Level analysis

High-level microarray analysis is required to identify groups of genes that are similarly regulated across the biological samples under study. A variety of mathematical procedures have been developed that partition genes or samples into groups, or clusters, with maximum similarity, thus enabling the identification of gene signatures or informative gene subsets. Methods for classification are either unsupervised or supervised. Supervised methods use existing biological information about specific genes that are functionally related to ”guide” or ”test” the cluster algorithm. With unsupervised methods, no prior test set is required. The most commonly employed unsupervised classification methods are the clustering techniques [13]. However discussion of these techniques more in detail is beyond the scope of this paper.

Conclusion

Nowadays comprehensive gene expression approaches like microarrays have fundamental role in providing basic information integral to biological and clinical investigation of complex diseases such as cardiovascular diseases. The statistical analysis of microarray data is probably the most difficult problem associated with the use of these technique. We can see, that the selection of the significant genes heavily depends on the choice of the testing methods. We can also see that the performance of the testing methods is affected by sample size, distributional assumption, the variance structure and so on (see Table 1). Therefore, to obtain the reliable testing results for detecting significant genes in microarray data analysis, we first need to explore the characteristic of the data and then apply the most appropriate testing method under the given situation. It is also important to choose the measure of differential expression based on the biological system of interest and particular problem specification. In a situation where the most reliable list of genes is desirable, the best approach may be to examine the intersection of genes identified by more methods.

In our future work we would like to apply the statistical methods described in this paper to the real microarray dataset from project of Centre of Biomedical Informatics *(The goal of this experiment is to identify genes that are differentially expressed in acute myocardial infarction patients and cerebrovascular accident patients)* and compare selected top significant genes by each of testing methods and also compare it with reference selected candidate genes (from well-curated publicly available databases), which are believed to be truly differentially expressed.

References

[1] S. Archacki, Q. K. Wang, “Expression profiling of cardiovascular disease”, *Human Genomics*, vol. 1, pp. 355–370, 2004.

[2] Q.K. Wang, S. Archacki, “Cardiovascular diseases”, *Humana Press*, vol. 129, pp. 1–13, 2007.

[3] J. L. Haines, M. Pericak-Vance, “Genetic analysis of complex diseases”, *John Wiley and Sons Publisher*, 2006.

[4] R. Gentleman, V. J. Carey, W. Huber, R. Irizarry, S. Dudoit, “Bioinformatics and Computational Biology Solutions Using R and Bioconductor”, *Springer Publisher*, 2005.

[5] D. B. Allison, X. Cui, G. P. Page, M. Sabripour, “Microarray data analysis: from disarray to consolidation and consensus”, *Nature Reviews*, vol. 7, pp. 55–65, 2006.

[6] D. Murphy , “Gene expression studies using microarrays: Principles, problems and prospects”, *Advan. Physiol. Edu.*, vol. 26, pp. 256–270, 2002.

[7] P. Baldi, A. D. Long, “A Bayesian framework for the analysis of microarry expression data: regularized t-test and statistical inferences of gene changes”, *Bioinformatics*, vol. 17, pp. 509–19, 2001.

[8] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments", *Statistical Applications in Genetics and Molecular Biology 3*, vol. 1, Article 3, Epub. 2004.

[9] V. Tusher, R. Tibshirani, G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response", *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5116–21, 2001.

[10] B. Efron, R. Tibshirani, J. D. Strey, V. Tusher, "Empirical Bayes analysis of a microarray experiment", *Journal of the American Statistical Association*, vol. 96, pp.: 1151–60, 2001.

[11] P. Broberg, "Ranking genes with respect to differential expression", *Genome Biology*, vol. 3: preprint0007.1-0007.23, from http://genomebiology.com/2002/3/9/preprint/0007 , 2002.

[12] Y. Zhao, W. Pan, "Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments", *Bioinformatics*, vol. 19, pp. 1046–54, 2003.

[13] R. B. Altman, "Whole-genome expression analysis: challenges beyond clustering", *Curr Opp Structural Biol.*, vol. 11, pp. 340-347, 2001.

Modeling Costs of Program Runs in Fuzzified Propositional Dynamic Logic

Post-Graduate Student:
MGR. LIBOR BĚHOUNEK

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic

behounek@cs.cas.cz

Supervisor:
DOC. PHDR. PETR JIRKŮ, CSC.

Faculty of Arts
Charles University in Prague
Celetná 20
116 42 Prague, Czech Republic

petr.jirku@ff.cuni.cz

Field of Study:
Logic

The work was supported by grant No. IAA900090703 *Dynamic Formal Systems* of the Grant Agency of the Academy of Sciences of the Czech Republic and Institutional Research Plan No. AV0Z10300504. The advisor for my research in the area of fuzzy logic is Prof. RNDr. Petr Hájek, DrSc. I have profitted from discussions with Marta Bílková and Petr Cintula.

Abstract

The paper introduces a logical framework for representing costs of program runs in fuzzified propositional dynamic logic. The costs are represented as truth values governed by the rules of a suitable t-norm fuzzy logic. A translation between program constructions in dynamic logic and fuzzy set-theoretical operations is given, and the adequacy of the logical model to the informal motivation is demonstrated. The role of tests of conditions in programs is discussed from the point of view of their costs, which hints at the necessity of distinguishing between the fuzzy modalities of admissibility and feasibility of program runs.

1. Introduction

It has been argued in [1] that t-norm fuzzy logics can be interpreted as logics of resources or costs, besides their usual interpretation as logics of partial truth. Particular instances of costs are the costs of program runs: typically, a run of a program needs various kinds of resources like machine time for performing instructions, operational memory or disk space for data, access to peripheries or special computation units, etc. Depending on the amount of the resources needed, some runs of programs can be more costly than others. The most usual logical model of programs and program runs is presented by propositional dynamic logic, which will be used as a basis for the present generalization. The aim of this paper is to sketch a logical framework for handling the costs of program runs by means of fuzzy logic, with programs modeled abstractly in propositional dynamic logic, and present some basic observations on the proposed model.

The paper has the following structure: A brief description of t-norm fuzzy logics and their cost-based interpretation is given in Sections 2 and 3. The apparatus of propositional dynamic logic is recalled in Section 4. A combination of these approaches, leading to a model of costs of program runs in fuzzified propositional dynamic logic, is given in Section 5. The role of tests of conditions in programs, which necessitates distinguishing the feasibility and admissibility of program runs in fuzzified propositional dynamic logic, is discussed in Section 6.

It should be noted that the paper only presents an initial sketch of the proposed approach to logical modeling of program costs. The work on this approach is currently in progress and a more comprehensive elaboration is being prepared, with Marta Bílková and Petr Cintula as co-authors.

2. T-norm fuzzy logic

In this section we give a short overview of the most important t-norm fuzzy logics that will be needed later on. Only the standard semantics of t-norm fuzzy logics is presented here, as it suffices for the needs of this paper. For more details on t-norm logics, including their axiomatic systems and general semantics, see [2, 3].

In the standard semantics, formulae of t-norm fuzzy logics are evaluated truth-functionally in the real unit interval $[0, 1]$; i.e., propositional connectives

are semantically realized by operations on $[0,1]$. In particular, the connective called *strong conjunction* $\&$ is in t-norm fuzzy logics realized by a *left-continuous t-norm*, i.e., a left-continuous binary operation on $[0,1]$ which is commutative, associative, monotone, and has 1 as its neutral element. The most important (left-) continuous t-norms are

$$\begin{aligned} x *_{\mathrm{G}} y &= \min(x,y) && \text{Gödel t-norm} \\ x *_{\Pi} y &= x \cdot y && \text{product t-norm} \\ x *_{Ł} y &= \max(0, x+y-1) && \text{Łukasiewicz t-norm} \end{aligned}$$

Every left-continuous t-norm $*$ has a unique *residuum* $\Rightarrow_*$, defined as

$$x \Rightarrow_* y = \sup\{z \mid z * x \leq y\},$$

which interprets *implication* $\rightarrow$ in the logic $\mathrm{L}(*)$ of the left-continuous t-norm $*$. If $x \leq y$, then $x \Rightarrow_* y = 1$; for $x > y$ the residua of the above three t-norms evaluate as follows:

$$\begin{aligned} x \Rightarrow_{\mathrm{G}} y &= y \\ x \Rightarrow_{\Pi} y &= y/x \\ x \Rightarrow_{Ł} y &= \min(1, 1-x+y) \end{aligned}$$

Further propositional connectives of $\mathrm{L}(*)$ are interpreted in the following way:

- *Negation* $\neg$ as $\neg_* x = x \Rightarrow_* 0$
- *Equivalence* $\leftrightarrow$ as

$$x \Leftrightarrow_* y = \min(x \Rightarrow_* y, y \Rightarrow_* x)$$

- *Disjunction* $\vee$ as the maximum, and
- *Weak conjunction* $\wedge$ as the minimum

Optionally, the *delta connective* Δ is added to $\mathrm{L}(*)$ with standard interpretation $\Delta x = 1$ if $x = 1$, and $\Delta x = 0$ otherwise. (We shall always use t-norm logics with Δ in this paper.) The algebra

$$[0,1]_* = \langle [0,1], *, \Rightarrow, \vee, \wedge, 0, \Delta \rangle$$

defining an interpretation of propositional t-norm logic is called the *t-algebra* of $*$ (with Δ).

Formulae that always evaluate to 1 are called *tautologies* of the logic $\mathrm{L}(*)$. The formulae that are tautologies of $\mathrm{L}(*)$ for all $*$ from some class $\mathcal{K}$ of left-continuous t-norms form the t-norm logic of the class $\mathcal{K}$. In particular, Hájek's [2] logic BL is the logic of all *continuous* t-norms and the logic MTL of [3] is the logic of all *left-*continuous t-norms: these general logics capture rules valid independently of a particular t-norm realization of $\&$. The proofs in this paper will be carried out in the logic MTL, thus sound for all left-continuous t-norms.

Propositional t-norm logics can be extended to their first-order and higher-order variants. These are needed for mathematical reasoning about fuzzy properties and will be employed later in this paper. For the formal apparatus of first-order fuzzy logic I refer the reader to [2]; Higher-order fuzzy logic has been introduced in [4] and described in a primer [5] freely available online. Here we shall only recall that the quantifiers $\forall, \exists$ are respectively realized as the infimum and supremum of the truth values, and that higher-order logic is a theory of fuzzy sets and relations with terms $\{x \mid \varphi(x)\}$, each of which represents the fuzzy set to which any element x belongs to the degree given by the truth value of the formula $\varphi(x)$.

3. Fuzzy logics as logics of costs

In fuzzy logic, truth values $x \in [0,1]$ are usually interpreted as degrees of truth, with 1 representing full truth and 0 full falsity of a proposition. As argued in [1], the truth values can also be interpreted as measuring *costs,* with propositional connectives representing natural operations on costs. Under this interpretation, we abstract from the nature of costs (be they time, money, space, or any other kind of resources) and only assume that they are linearly ordered and normalized into the interval $[0,1]$.

(The assumption of linear ordering can actually be relaxed to more general *prelinear* orderings, which cover most usual kinds of resources. In particular, direct products of linear orderings fall within the class, which allows *vectors* of costs, e.g., pairs of disk space and computation time, to be represented within this framework. In general, the cost-interpretation of fuzzy logic is based on the fact that most common resources show the structure of a prelinear residuated lattice. However, for simplicity we shall only consider linearly ordered costs that can be embedded in the real unit interval here.)

Under the cost-based interpretation, the truth value 1 represents the zero cost ("for free") and the truth value 0 represents a maximal or unaffordable cost. Intermediary truth values represent various degrees of costliness, with the usual ordering of $[0,1]$ inverse to that of costs (the truth values can thus be understood as expressing degrees of truth of the fuzzy predicate "is cheap"). Strong conjunction $\&$ represents the *fusion* of resources, or the "sum" of costs. Various left-continuous t-norms

represent various ways by which costs may sum, and particular t-norm logics thus capture the rules that govern particular ways of cost addition. For example, the Łukasiewicz t-norm $*_{\text{Ł}}$ corresponds to the *bounded sum* of costs: assume that costs sum up to a bound $b > 0$; if we normalize the interval $[0, b]$ to $[0, 1]$ with the cost $c \in [0, b]$ represented by $1 - c/b \in [0, 1]$, then the bounded sum on $[0, b]$ corresponds to the Łukasiewicz t-norm on $[0, 1]$, since

$$(1 - x) *_{\text{Ł}} (1 - y) = 1 - (x + y)$$

unless the bound 0 (representing b) is achieved. Similarly the product t-norm corresponds to the *unbounded sum* of costs (via the negative logarithm), with 0 representing the infinite cost. The Gödel t-norm corresponds to taking the *maximum* cost as the "sum", which is also natural for some kinds of costs (e.g., the disk space for temporary results of calculation, which can be erased before the program proceeds). Other left-continuous t-norms correspond to variously distorted addition of costs, possibly suitable under some rare circumstances.

Obviously, disjunction and weak conjunction correspond, respectively, to the minimum and maximum of the two costs. The meaning of implication is that of surcharge: the cost expressed by $A \to B$ is the cost needed for B, provided we have already got the cost of A. (Observe that if the cost of B is less than or equal to that of A, then indeed $A \to B$ evaluates to 1, as we have already got the cost of B if we have the cost of A; i.e., the "upgrade" from A to B is "for free", which is represented by the value 1.) The equivalence connective represents the "difference" (in terms of $\&$) between two costs, and negation the remainder to the maximal cost.

Tautologies of a given t-norm logic represent combinations of costs that are always "for free". More importantly, tautologies of the form $A_1 \& \ldots \& A_n \to B$ express the rules of preservation of "cheapness", as their cost-based interpretation reads: if we have the costs of all A_i together, then we also have the cost of B. Particular t-norm fuzzy logics thus express the rules of reasoning *salvis expensis,* in a similar manner as classical Boolean tautologies of the above form express the rules of reasoning *salva veritate.*

In the following sections we shall apply this interpretation of fuzzy logic to a particular kind of costs, namely the costs of program runs as modeled in propositional dynamic logic.

4. Propositional dynamic logic

Propositional dynamic logic (PDL) provides an abstract apparatus for logical modeling of behavior of programs. For details on PDL see [6, 7].

PDL models programs as (non-deterministic) transitions in an abstract space of states. (As such, PDL programs can represent any kind of actions over an arbitrary set of states, not only programs operating on the states of a computer; the applicability of both PDL and the present approach is thus much broader than just to computer programs.) Programs can in PDL be composed of simpler programs by means of a fixed set of constructions (the usual choice is that of regular expressions with tests, by which all common programming constructions are expressible), applied recursively on a fixed countable set of atomic programs (representing, e.g., the instructions of a processor). Propositional formulae of PDL express Boolean propositions about the states of the state space, and include, besides usual connectives of Boolean logic, modalities corresponding to programs, by means of which it is possible to reason about programs and their preconditions and postconditions.

Formally, the sets $\mathbf{Form}$ of formulae and $\mathbf{Prog}$ of programs of PDL are defined by simultaneous recursion from fixed countable sets of atomic formulae and atomic programs as follows:

- Every atomic formula is a formula; every atomic program is a program.
- If φ and ψ are formulae, then $\neg\varphi$ and $(\varphi \wedge \psi)$ are formulae (meaning *not* φ resp. φ *and* ψ). The abbreviations $\top$, $\bot$, $(\varphi \vee \psi)$, $(\varphi \to \psi)$, and $(\varphi \leftrightarrow \psi)$ are defined as usual in Boolean logic, with usual conventions on omitting parentheses.
- If α and β are programs, then α^*, $(\alpha \cup \beta)$, and $(\alpha; \beta)$ are programs (meaning *repeat* α *finitely many times, do* α *or* β, and *do* α *and then* β, respectively, where *or* and *finitely many* means a non-deterministic choice).
- If φ is a formula and α is a program, then $[\alpha]\varphi$ is a formula (meaning φ *holds after any run of* α). The expression $\langle\alpha\rangle\varphi$ abbreviates $\neg[\alpha]\neg\varphi$.
- If φ is a formula, then $\varphi?$ is a program (meaning *continue iff* φ).

The semantic models of PDL are multimodal Kripke structures $\langle W, R, V\rangle$ with W a non-empty set (of states),

$R\colon \mathbf{Prog} \to 2^{W^2}$ an evaluation of programs by binary relations on W (representing possible transitions between states by the program), and $V\colon \mathbf{Form} \to 2^W$ an evaluation of formulae by subsets of W (namely, the sets of verifying states), such that

$$V_{\neg\varphi} = W \setminus V_\varphi \tag{1}$$
$$V_{\varphi\wedge\psi} = V_\varphi \cap V_\psi \tag{2}$$
$$V_{\langle\alpha\rangle\varphi} = R_\alpha {}^{\leftarrow} V_\varphi \tag{3}$$
$$R_{\alpha;\beta} = R_\alpha \circ R_\beta \tag{4}$$
$$R_{\alpha\cup\beta} = R_\alpha \cup R_\beta \tag{5}$$
$$R_{\alpha^*} = R_\alpha^* \tag{6}$$
$$R_{\varphi?} = \mathrm{Id} \cap V_\varphi \tag{7}$$

where $\circ$ denotes the composition of relations, ${}^{\leftarrow}$ the preimage under a relation, R^* the reflexive and transitive closure of R, and Id the identity of relations. A formula φ is valid in the model iff $V_\varphi = W$, and is a tautology iff is valid in all models.

PDL is sound and complete w.r.t. the axiomatic system consisting of all propositional tautologies, the axioms

$$[\alpha;\beta]\varphi \leftrightarrow [\alpha][\beta]\varphi \tag{8}$$
$$[\alpha\cup\beta]\varphi \leftrightarrow [\alpha]\varphi \wedge [\beta]\varphi \tag{9}$$
$$[\alpha^*]\varphi \leftrightarrow \varphi \wedge [\alpha][\alpha^*]\varphi \tag{10}$$
$$[\varphi?]\psi \leftrightarrow (\varphi \to \psi) \tag{11}$$
$$[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi) \tag{12}$$

and the rules of modus ponens (from φ and $\varphi \to \psi$ infer ψ), necessitation (from φ infer $[\alpha]\varphi$), and induction (from $\varphi \to [\alpha]\varphi$ infer $\varphi \to [\alpha^*]\varphi$).

For simplicity, we shall not consider expansions of PDL by further program constructions like intersection, converse, etc.

5. Modeling the costs of program runs

PDL does not take costs of program runs into consideration. In PDL, possible runs of a program α are modeled as transitions from a state w to a state w' such that $R_\alpha ww'$. The relation R_α representing the program α is binary (crisp): thus the states w' are either accessible or unaccessible from w by a run of α. In practice, however, it often occurs that although a state w' is theoretically achievable from w by α, the run of α from w to w' is not *feasible*—e.g., is too long (for example, needs to perform 10^{100} instructions, a frequent case in exponentially complex problems), requires too much memory, etc. Obviously, such unfeasible runs should not play a role in the practical assessment whether some condition φ can or cannot hold after the possible runs of α. Nevertheless, classical PDL cannot exclude such unfeasible runs, as there is no sharp boundary between feasible and unfeasible runs (i.e., the feasibility of runs is a fuzzy property).

A more realistic model can be obtained by considering costs of program runs, by means of which we can measure their feasibility. A simple model, which nevertheless covers many common situations, would assign the triple α, w, w' such that $R_\alpha ww'$ in a model of PDL a real number $C_\alpha ww'$ representing the cost of the run of α from w to w'. The cost thus would be represented by a function

$$C\colon \mathbf{Prog} \times W^2 \to [0, +\infty],$$

i.e., we are weighting the arrows in the co-graph of R_α by their costs; we assign the cost $+\infty$ to impossible runs with $\neg R_\alpha ww'$. The cost of a run of $\alpha_1;\alpha_2;\ldots;\alpha_n$ going from w_0 through $w_1, w_2, \ldots$ to w_n would be a function f (most often, the sum) of the costs of the runs of α_i from w_{i-1} to w_1. If there are different paths between w_0 and w_n through which $\alpha_1;\alpha_2;\ldots;\alpha_n$ can run, we are interested in the cheapest path, i.e., the run of $\alpha;\beta$ from w to w' will be understood as costing

$$C_{\alpha;\beta}ww' = \inf_{w''} f(C_\alpha ww'', C_\beta w''w'). \tag{13}$$

This model would allow us to work with the costs of program runs in the expanded models of PDL and define and investigate many useful notions related to costs by means of classical mathematics and logic. Nevertheless, since the important property of *feasibility* of a program run is essentially a fuzzy predicate, we shall recast this model in terms of the cost-based interpretation of fuzzy logic. This will allow us to employ fuzzy logic for a convenient definition of feasible runs and use the apparatus of fuzzy logic for reasoning about the costs on the propositional level, by replacing classical rules of reasoning with those of fuzzy logic. For a methodological discussion of this approach see [4, 5, 8, 9].

Thus we shall assume that the structure of costs is that of some t-norm algebra (see Section 3 for possible extension to more general algebras). Then, instead of weighting the arrows in the co-graph of R_α with costs, we can directly replace R_α with a *fuzzy relation* $\tilde{R}_\alpha \in [0,1]^{W^2}$, with the truth values of $\tilde{R}_\alpha ww'$ representing the cost of the run of α from w to w'.

Since the sum of costs now translates to conjunction in a suitable t-norm logic and since we are interested in the cheapest runs if more paths are possible, (13) now

translates to

$$\tilde{R}_{\alpha;\beta}ww' \equiv (\exists w'')(\tilde{R}_\alpha ww'' \,\&\, \tilde{R}_\beta w''w') \qquad (14)$$

with logical symbols interpreted in a t-norm fuzzy logic, i.e., in semantics,

$$\tilde{R}_{\alpha;\beta}ww' = \sup_{w''}(\tilde{R}_\alpha ww'' * \tilde{R}_\beta w''w')$$

It can be observed that the formula (14) has exactly the same form as in classical PDL where $R_{\alpha;\beta} = R_\alpha \circ R_\beta$, since by definition

$$(R_\alpha \circ R_\beta)ww' \equiv (\exists w'')(R_\alpha ww'' \,\&\, R_\beta w''w') \quad (15)$$

The only difference between (14) and (15) is that the relations in (14) are fuzzy, and that the logical operations are (therefore) interpreted in a t-norm fuzzy logic instead of Boolean logic. This is in fact a general feature of using the framework of formal fuzzy logic that natural definitions usually take the same form as in the crisp case, only with the logical symbols reinterpreted in fuzzy logic (cf. [4, 5, 8, 9]): we shall see that further definitions will follow this pattern, too. Indeed, analogously to (15) it is usual [10] in fuzzy logic to define the composition of fuzzy relations $\tilde{R}$ and $\tilde{S}$ as

$$\begin{aligned}(\tilde{R} \circ \tilde{S})ww' &\equiv (\exists w'')(\tilde{R}ww'' \,\&\, \tilde{S}w''w'), \text{ i.e.,}\\ &\equiv \sup_{w''}(\tilde{R}ww'' * \tilde{S}w''w')\end{aligned}$$

Consequently, we can write

$$\tilde{R}_{\alpha;\beta} = \tilde{R}_\alpha \circ \tilde{R}_\beta$$

in our setting, in full analogy with the definition (4) of $R_{\alpha;\beta}$ in classical PDL.

Similarly it is natural to assume $\tilde{R}_{\alpha\cup\beta} = \tilde{R}_\alpha \cup \tilde{R}_\beta$ as in (5), where $(\tilde{R} \cup \tilde{S})ww'$ is defined for any fuzzy relations $\tilde{R}, \tilde{S}$ as $\tilde{R}ww' \vee \tilde{S}ww'$, since the cost of a run of $\alpha \cup \beta$ between w and w' should be the smaller of the two costs of the runs of α and β between the same states (which in $[0,1]_*$ is represented by the larger of the two truth values). Analogously one verifies that the cost of α^* is represented by the transitive and reflexive closure $\tilde{R}^*_\alpha$ of the fuzzy relation $\tilde{R}_\alpha$ defined as usual in the theory of fuzzy relations [10], in full analogy to (6).

The reinterpretation in fuzzy logic of (3), which expands to

$$V_{\langle\alpha\rangle\varphi}w \equiv (\exists w')(R_\alpha ww' \,\&\, V_\varphi w') \qquad (16)$$

yields a very natural modality expressing that after a *feasible* run of α the condition φ can hold. (Notice that this definition reflects the motivation for taking the costs of program runs into account, described in the beginning of this section.)

It can be observed in (16) that even if V_φ is crisp, a fuzzy R_α will yield a fuzzy $V_{\langle\alpha\rangle\varphi}$. Thus, because of the interplay of programs and formulae in PDL, our fuzzification of programs necessitates a fuzzification of formulae as well. A model of our fuzzified PDL is thus a triple $\langle W, \tilde{R}, \tilde{V}\rangle$, where W is a non-empty crisp set of states, $\tilde{R}$ maps programs α to fuzzy relations $\tilde{R}_\alpha \in [0,1]^{W^2}$, and $\tilde{V}$ gives fuzzy sets $\tilde{V}_\varphi \in [0,1]^W$ of states which fuzzily validate φ (i.e., $\tilde{V}_\varphi w$ is the truth value of φ in w).

Thus in the fuzzified (16), which reads

$$\tilde{V}_{\langle\alpha\rangle\varphi}w \equiv (\exists w')(\tilde{R}_\alpha ww' \,\&\, \tilde{V}_\varphi w'), \qquad (17)$$

the subformula $\tilde{R}_\alpha ww'$ can be understood as expressing the fuzzy proposition "w' is cheaply accessible from w by a run of α" (which is a fuzzy-propositional reading of the cost represented by $\tilde{R}_\alpha ww'$) and $\tilde{V}_\varphi w'$ as the fuzzy proposition "φ holds in w'" (viz, to the degree expressed by $\tilde{V}_\varphi w'$). Both $\tilde{R}_\alpha ww'$ and $\tilde{V}_\varphi w'$ can thus be understood as fuzzy propositions, and their combination in a single formula thus does not present a type mismatch: we only assume that the cost is represented by a truth value of the fuzzy proposition "the run is cheap", and that the mapping of costs to $[0,1]_*$ is such that the conjunction $*$ of truth values coincides with summation of costs. (This assumption is more natural if $\tilde{V}_\varphi$ for non-modal φ are assumed to be crisp, since then the fuzziness of $\tilde{V}_\psi$ for modal ψ arise exactly from considering the costs $\tilde{R}_\alpha ww'$ in (16). However, in many real-world applications of fuzzified PDL it may be desirable to have non-modal formulae fuzzy as well: then, if different algebras of degrees are needed for $\tilde{V}$ and $\tilde{R}$ in a particular model, one can use suitable direct products of t-norm algebras; I omit details here.) Particular interpretations $*$ of $\&$ and particular mappings of actual costs under consideration to $[0,1]_*$ will then yield concrete ways of calculating the truth values of this expression in particular models; importantly, however, the rules of general fuzzy logics like BL or MTL allow deriving theorems on program costs that are valid independently of a concrete representation in $[0,1]_*$.

Returning to (16), one can observe that again it coincides with the usual definition of preimage of a fuzzy set in a fuzzy relation (see, e.g., [11]). Thus we can write

$$\tilde{V}_{\langle\alpha\rangle\varphi} = \tilde{R}_\alpha \leftarrow \tilde{V}_\varphi,$$

again in full analogy with (3).

The derived semantical clause for $[\alpha]\varphi$, which in the classical case reads

$$V_{[\alpha]\varphi}w \equiv (\forall w')(R_\alpha ww' \to V_\varphi w'), \qquad (18)$$

yields in the fuzzy reinterpretation

$$\tilde{V}_{[\alpha]\varphi}w \equiv (\forall w')(\tilde{R}_\alpha ww' \to \tilde{V}_\varphi w'), \qquad (19)$$

a useful fuzzy modality expressing that after all feasible (or cheap enough) runs of α the fuzzy condition φ will hold. (Similar comments as in the case of $\langle\alpha\rangle\varphi$ are applicable.) The operation defined by (18) for crisp R_α and V_α and by (19) for fuzzy $\tilde{R}_\alpha$ and $\tilde{V}_\alpha$ is denoted by $\leftarrowtriangle$ and called the *subproduct preimage* in [11], where it is studied as a particular case of BK-subproduct $\triangleleft$. (These notions were introduced by Bandler and Kohout in [12] for crisp relations and generalized for fuzzy relations in [13]. Further references to the literature on $\leftarrowtriangle$ and its properties in fuzzy logic are given in [11].) Thus we can write

$$V_{[\alpha]\varphi} = R_\alpha \leftarrowtriangle V_\varphi$$
$$\tilde{V}_{[\alpha]\varphi} = \tilde{R}_\alpha \leftarrowtriangle \tilde{V}_\varphi$$

respectively for crisp and fuzzy PDL. Notice that unlike in classical PDL, $[\alpha]\varphi$ and $\langle\alpha\rangle\varphi$ are no longer interdefinable in fuzzified PDL, as the clauses (17) and (19) do not generally satisfy $\tilde{V}_{\neg\langle\alpha\rangle\varphi} = \tilde{V}_{[\alpha]\neg\varphi}$ in fuzzy logic, unless the negation $\neg$ is involutive. Both $[\alpha]$ and $\langle\alpha\rangle$ therefore need to be present in the language of fuzzified PDL as primitive symbols.

As an example of theorems that can be proved in our framework, we shall check the soundness of the axioms (8)–(12) and the three inference rules of classical PDL in our fuzzified PDL semantics. The validity of the axiom (8) in any model $M = \langle W, \tilde{R}, \tilde{V}\rangle$ is proved as follows:

$$\begin{aligned}
&M \models [\alpha;\beta]\varphi \leftrightarrow [\alpha][\beta]\varphi\\
&\text{iff}\quad \tilde{V}_{[\alpha;\beta]\varphi} = \tilde{V}_{[\alpha][\beta]\varphi}\\
&\text{iff}\quad \tilde{R}_{\alpha;\beta} \leftarrowtriangle \tilde{V}_\varphi = \tilde{R}_\alpha \leftarrowtriangle \tilde{V}_{[\beta]\varphi}\\
&\text{iff}\quad (\tilde{R}_\alpha \circ \tilde{R}_\beta) \leftarrowtriangle \tilde{V}_\varphi = \tilde{R}_\alpha \leftarrowtriangle (\tilde{R}_\beta \leftarrowtriangle \tilde{V}_\varphi),
\end{aligned}$$

where the last identity is an easy property of $\leftarrowtriangle$, see [11, Cor. 5.17].

Similarly, the validity of the axiom (9) is proved by

$$\begin{aligned}
&M \models [\alpha \cup \beta]\varphi \leftrightarrow [\alpha]\varphi \wedge [\beta]\varphi\\
&\text{iff}\quad \tilde{V}_{[\alpha\cup\beta]\varphi} = \tilde{V}_{[\alpha]\varphi\wedge[\beta]\varphi}\\
&\text{iff}\quad \tilde{R}_{\alpha\cup\beta} \leftarrowtriangle \tilde{V}_\varphi = \tilde{V}_{[\alpha]\varphi} \cap_\wedge \tilde{V}_{[\alpha]\varphi}\\
&\text{iff}\quad (\tilde{R}_\alpha \cup \tilde{R}_\beta) \leftarrowtriangle \tilde{V}_\varphi = (\tilde{R}_\alpha \leftarrowtriangle \tilde{V}_\varphi) \cap_\wedge (\tilde{R}_\beta \leftarrowtriangle \tilde{V}_\varphi),
\end{aligned}$$

where the last identity is again an easy property of $\leftarrowtriangle$, see [11, Cor. 5.16]. Notice that weak conjunction $\wedge$ is in order in the fuzzy version of (9), corresponding in the proof to *min-intersection* defined for any fuzzy sets $\tilde{U}, \tilde{V}$ as $(\tilde{U} \cap_\wedge \tilde{V})w \equiv \tilde{U}w \wedge \tilde{V}w$.

In order to verify the axiom (10), we need a few definitions and lemmata. First, define for any fuzzy relation $\tilde{R}$ its iterations

$$\tilde{R}^0 = \mathrm{Id} \qquad (20)$$
$$\tilde{R}^{n+1} = \tilde{R} \circ \tilde{R}^n \qquad (21)$$

for all $n \in \mathbb{N}$. Furthermore, the union $\bigcup\mathcal{A}$ of a crisp or fuzzy set $\mathcal{A}$ of fuzzy relations is in higher-order fuzzy logic defined as

$$\big(\bigcup\mathcal{A}\big)ww' \equiv (\exists\tilde{R})(\mathcal{A}\tilde{R} \,\&\, \tilde{R}ww').$$

Obviously, for any fuzzy relation $\tilde{R}$,

$$\bigcup_{n=0}^{\infty}\tilde{R}^n = \tilde{R}^0 \cup \bigcup_{n=1}^{\infty}\tilde{R}^n = \mathrm{Id} \cup \bigcup_{n=1}^{\infty}\tilde{R}^n$$

by (20). It can trivially be verified that by definitions, $\mathrm{Id} \leftarrowtriangle \tilde{V} = \tilde{V}$, thus also $\tilde{R}^0 \leftarrowtriangle \tilde{V} = \tilde{V}$, for any fuzzy relation $\tilde{R}$ and any fuzzy set $\tilde{V}$. Finally, it can be proved (cf. [10]) that the transitive and reflexive closure $\tilde{R}^*$ of a fuzzy relation $\tilde{R}$ is in fuzzy logic characterized in the same way as in classical mathematics, viz

$$\tilde{R}^* = \bigcup_{n=0}^{\infty}\tilde{R}^n = \mathrm{Id} \cup \bigcup_{n=1}^{\infty}\tilde{R}^n$$

Now we can show the soundness of (10), which amounts to the general validity of $\tilde{V}_{[\alpha^*]} = \tilde{V}_{\varphi\wedge[\alpha][\alpha^*]\varphi}$. We have the following chain of identities, justified by definitions and previous lemmata:

$$\begin{aligned}
\tilde{V}_{[\alpha^*]\varphi} &= \tilde{R}_{\alpha^*} \leftarrowtriangle \tilde{V}_\varphi =\\
&= \Big(\bigcup_{n=0}^{\infty}\tilde{R}_\alpha^n\Big) \leftarrowtriangle \tilde{V}_\varphi\\
&= \Big(\mathrm{Id} \cup \bigcup_{n=1}^{\infty}\tilde{R}_\alpha^n\Big) \leftarrowtriangle \tilde{V}_\varphi\\
&= (\mathrm{Id} \leftarrowtriangle \tilde{V}_\varphi) \cap_\wedge \left(\Big(\bigcup_{n=1}^{\infty}\tilde{R}_\alpha^n\Big) \leftarrowtriangle \tilde{V}_\varphi\right)\\
&= \tilde{V}_\varphi \cap_\wedge \left(\Big(\tilde{R}_\alpha \circ \bigcup_{n=0}^{\infty}\tilde{R}_\alpha^n\Big) \leftarrowtriangle \tilde{V}_\varphi\right)\\
&= \tilde{V}_\varphi \cap_\wedge \tilde{V}_{[\alpha;\alpha^*]\varphi} = \tilde{V}_{\varphi\wedge[\alpha][\alpha^*]\varphi}.
\end{aligned}$$

Notice again that weak conjunction is in order in fuzzified (10).

The soundness of the rule of induction amounts to the validity of inferring

$$\tilde{V}_\varphi \subseteq \tilde{R}_\alpha^* \leftarrowtriangle \tilde{V}_\varphi \quad \text{from} \quad \tilde{V}_\varphi \subseteq \tilde{R}_\alpha \leftarrowtriangle \tilde{V}_\varphi.$$

By induction, we shall prove that from $\tilde{V}_\varphi \subseteq \tilde{R}_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi$ we can infer $\tilde{V}_\varphi \subseteq \tilde{R}^n_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi$ for all $n \in \mathbb{N}$, i.e., by [14, Lemma B.8(L5)],

$$\tilde{V}_\varphi \subseteq \bigcap_{n\in\mathbb{N}} (\tilde{R}^n_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi),$$

which is by [11, Cor. 5.16] equivalent to the required

$$\tilde{V}_\varphi \subseteq \Big(\bigcup_{n\in\mathbb{N}} \tilde{R}^n_\alpha\Big) \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi.$$

The first step $\tilde{V}_\varphi \subseteq \tilde{R}^0_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi$ of the induction is trivially valid by $\tilde{R}^0_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi = \mathrm{Id} \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi = \tilde{V}_\varphi$. For the induction step, we need to infer

$$\tilde{V}_\varphi \subseteq \tilde{R}^{n+1}_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi \quad \text{from} \quad \tilde{V}_\varphi \subseteq \tilde{R}^n_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi,$$

i.e., by [14, Th. 4.3(I14)],

$$(\tilde{R}^n_\alpha \circ \tilde{R}_\alpha) \mathbin{\rightarrow} \tilde{V}_\varphi \subseteq \tilde{V}_\varphi, \quad \text{from} \quad \tilde{R}_\alpha \mathbin{\rightarrow} \tilde{V}_\varphi \subseteq \tilde{V}_\varphi.$$

By [11, Cor. 4.14], the former is equivalent to

$$\tilde{R}_\alpha \mathbin{\rightarrow} (\tilde{R}^n_\alpha \mathbin{\rightarrow} \tilde{V}_\varphi) \subseteq \tilde{V}_\varphi,$$

which follows from $\tilde{R}_\alpha \mathbin{\rightarrow} \tilde{V}_\varphi \subseteq \tilde{V}_\varphi$ by monotony of $\mathbin{\rightarrow}$ w.r.t. $\subseteq$ [11, Cor. 4.7].

A discussion of the test construction is postponed to Section 6; therefore we shall skip checking the soundness of the the axiom (11). The soundness of the rule of modus ponens and the axioms of propositional logic is demonstrated in [15], as $\langle W, \tilde{V}\rangle$ forms the usual intensional semantics for fuzzy logic. The soundness of the rule of necessitation amounts to the validity of inferring $W \subseteq \tilde{R}_\alpha \mathbin{\leftarrow\!\!\triangleleft} \tilde{V}_\varphi$, i.e., $\tilde{R}_\alpha \mathbin{\rightarrow} W \subseteq \tilde{V}_\varphi$, from $W \subseteq \tilde{V}_\varphi$; but since $\tilde{R}_\alpha$ only operates on W, it is immediate that $\tilde{R}_\alpha \mathbin{\rightarrow} W \subseteq W \subseteq \tilde{V}_\varphi$.

On the other hand, the axiom (12) fails in fuzzy PDL, as it is well known (already from [2]) that fuzzified Kripke frames do not in general validate the modal axiom K. Since also the interdefinability of $\langle\alpha\rangle$ and $[\alpha]$ fails for non-involutive negation, dual axioms and rules for $\langle\alpha\rangle$ need to be added to a prospective axiomatic system of fuzzified PDL. I omit the discussion of these axioms here; it can nevertheless be hinted that since the relationship between the semantic clauses for $\langle\alpha\rangle$ and $[\alpha]$ is that of Morsi's duality [16] (combined with the duality between fuzzy relations and their converses), the formulation and soundness of the dual axioms and rules for $\langle\alpha\rangle$ can be obtained from the axioms and rules for $[\alpha]$ automatically by the same duality.

6. The role of tests

In classical PDL, tests $\varphi?$ have the role in branching complex programs: they are employed in definitions of such programming constructions as if–then–else, while–do, or repeat–until. They do not themselves affect the state in which a program run is, but bar a further execution of the program if their condition is not met. A straightforward fuzzification of the semantic condition (7), $\tilde{R}_{\varphi?} = \mathrm{Id} \cap \tilde{V}_\varphi$, would interpret tests in fuzzy PDL as programs which do not change the state, but can decrease the "passability" of the run through the current state according to the truth value of the condition φ. This, however, does not correspond to the primary motivation of $\tilde{R}_\alpha ww'$ as the *cost* of the run of α from w to w': the condition φ may be cheap to test, but can have a low truth degree in w, or vice versa. The two roles of the truth value yielded by the test $\varphi?$ do not match in fuzzy PDL: the *truth degree* of φ should affect the possibility of further execution, while the *cost* of performing the test of φ should contribute to the overall cost of the run of a complex program. Neither of the two roles can be sacrificed, since the former is necessary for branching the program (by the fuzzy if–then–else and cycle constructions), while without the latter we would be unable to distinguish between feasible and unfeasible runs (which was our primary motivation for the fuzzification of PDL).

Unless we want to stipulate that the conventional complexity (or cost) of a test be identified with the truth value it yields, thus equating the accessibility of paths of program execution with their costs (by which the actual cost of performing the computation is replaced by a different conventional measure), we may have to admit that the identification of the feasibility (or cost) value with the value of accessibility was too bold and that these two fuzzy relations on W have to be distinguished. If we denote the fuzzy accessibility relation by $\tilde{R}_\alpha$ and the feasibility relation by $\tilde{C}_\alpha$, then the test $\varphi?$ would contribute to $\tilde{R}_\alpha$ by the truth value of φ, while to $\tilde{C}_\alpha$ by the cost of performing the test. For instance, performing a test of a difficult tautology may contribute a lot to the cost of the run, while not decreasing the "correctness" degree of the run at all. We may then distinguish the modality $\langle\alpha\rangle^{\tilde{R}}\varphi$ expressing that there is a "correct" run to a state where φ holds from $\langle\alpha\rangle^{\tilde{R}\cap\tilde{C}}\varphi$ expressing that there is a "correct feasible" run validating φ (all conditions understood fuzzily). Their semantic clauses are, respectively:

$$\tilde{V}_{\langle\alpha\rangle^{\tilde{R}}\varphi} w \equiv (\exists w')(\tilde{R}_\alpha ww' \,\&\, \tilde{V}_\varphi w')$$
$$\tilde{V}_{\langle\alpha\rangle^{\tilde{R}\cap\tilde{C}}\varphi} w \equiv (\exists w')(\tilde{R}_\alpha ww' \,\&\, \tilde{C}_\alpha ww' \,\&\, \tilde{V}_\varphi w')$$

The apparatus of costs of program runs thus appears

to operate best on PDL with fuzzified accessibility relations of programs, whose truth degrees do not express the degrees of feasibility (or costs) of program runs, but the degrees of their admissibility (or "correctness", in the sense of the satisfaction of conditions passed through). The fuzzification of admissibility can be developed independently, without regarding costs of runs at all, thus making the same idealization as regards costs as classical PDF, i.e., with equating feasibility and admissibility of runs. Such fuzzification only generalizes the framework of PDL to permit fuzzy conditions like "if the temperature is high, do α" (which may be quite useful in real-world applications) and a measure of "correctness" of some transitions between states by programs (capturing for instance such phenomena as rounding numerical results etc.).

Adding moreover the apparatus for costs then makes the (already fuzzified) model more realistic by the possibility of distinguishing not only (the degree of) correctness, but also (the degree of) feasibility of (more or less correct) runs of programs. The double nature of tests regarding the truth and cost degrees, however, seems to exclude the possibility of adding the apparatus of costs directly to crisp rather than already fuzzified PDL, unless we forbid tests on feasibility (e.g., of the form $(\langle\alpha\rangle^{\tilde{R}\cap\tilde{C}}\varphi)?$), which automatically fuzzify the admissibility of runs.

Various kinds of restrictions on tests (e.g., allowing only tests of atomic formulae, non-modal formulae, formulae not referring to feasibility, etc.) would, however, strongly affect the requirements on the models and their properties. An elaboration of these considerations is left for future work, as are the problems of axiomatizability of such systems of fuzzy PDL and a detailed investigation of their properties.

References

[1] L. Běhounek, "Fuzzy logics interpreted as logics of resources," in *XXII Logica Volume of Abstracts*, (Prague), Institute of Philosophy, Academy of Sciences of the Czech Republic, 2008. XXII International Conference Logica, held on June 16–19, 2008 in Hejnice, Czech Republic.

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

[3] 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.

[4] L. Běhounek and P. Cintula, "Fuzzy class theory," *Fuzzy Sets and Systems*, vol. 154, no. 1, pp. 34–55, 2005.

[5] L. Běhounek and P. Cintula, "Fuzzy Class Theory: A primer v1.0," Tech. Rep. V-939, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2006. Available at www.cs.cas.cz/research/library/reports_900.shtml.

[6] D. Harel, "Dynamic logic," in *Handbook of Philosophical Logic* (D. M. Gabbay and F. Guenthner, eds.), vol. II: Extensions of Classical Logic, pp. 497–604, Dordrecht: D. Reidel, 1st ed., 1984.

[7] D. Harel, D. Kozen, and J. Tiurin, *Dynamic Logic*. Cambridge MA: MIT Press, 2000.

[8] L. Běhounek and P. Cintula, "From fuzzy logic to fuzzy mathematics: A methodological manifesto," *Fuzzy Sets and Systems*, vol. 157, no. 5, pp. 642–646, 2006.

[9] L. Běhounek and P. Cintula, "Fuzzy class theory as foundations for fuzzy mathematics," in *Fuzzy Logic, Soft Computing and Computational Intelligence: 11th IFSA World Congress*, vol. 2, (Beijing), pp. 1233–1238, Tsinghua University Press/Springer, 2005.

[10] L. A. Zadeh, "Similarity relations and fuzzy orderings," *Information Sciences*, vol. 3, pp. 177–200, 1971.

[11] L. Běhounek and M. Daňková, "Relational compositions in Fuzzy Class Theory." To appear in Fuzzy Sets and Systems (doi:10.1016/j.fss.2008.06.013), 2008.

[12] W. Bandler and L. J. Kohout, "Mathematical relations, their products and generalized morphisms," Tech. Rep. EES-MMS-REL 77-3, Man–Machine Systems Laboratory, Department of Electrical Engineering, University of Essex, Essex, Colchester, 1977.

[13] W. Bandler and L. J. Kohout, "Fuzzy relational products and fuzzy implication operators," in *International Workshop of Fuzzy Reasoning Theory and Applications*, (London), Queen Mary College, University of London, 1978.

[14] L. Běhounek, U. Bodenhofer, and P. Cintula, "Relations in Fuzzy Class Theory: Initial steps," *Fuzzy Sets and Systems*, vol. 159, no. 14, pp. 1729–1772, 2008.

[15] L. Běhounek, "Fuzzification of Groenendijk–Stokhof propositional erotetic logic," *Logique et Analyse*, vol. 47, no. 185–188, pp. 167–188, 2004.

[16] N. N. Morsi, W. Lotfallah, and M. El-Zekey, "The logic of tied implications, part 2: Syntax," *Fuzzy Sets and Systems*, vol. 157, pp. 2030–2057, 2006.

Agent-based Simulation of Processes in Medicine

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, Czech Republic

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, Czech Republic

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

Process modelling has proven itself as a useful technique for capturing the work practice in companies. In this paper, we focus on its usage in the domain of medical care. We analyze the problem of the simulation of processes and present an approach based on agent-based simulations. We formally define an enhanced process language, the algorithm transforming these enhanced processes into the definition of agents' behavior, and the architecture of the target multi-agent system simulating the modeled processes in some environment. The example of usage is given in the form of a critiquing expert system proposal that uses formalized medical guidelines as the knowledge base.

1. Introduction

Process modelling is a widely used technique offering a simple and understandable view on the work practice within a team or a company, and it is mainly utilized by managers and executives in various fields of industry. The area of medical care also offers an opportunity for process modelling, and its usage in computer systems, such as hospital information system (HIS) or workflow management systems (WfMS). However, there are many problems with applying this proved technique into the medical care [1], hence it is not as spread as it could be. In our work, we focus on the simulation of processes in general, we study the possibility of using agents and multi-agent system for this purpose, but we also want to apply these state-of-the-art methods into a development of an expert system for physicians that would use processes as a knowledge base.

In the area of processes in medical care, most of the authors distinguish two main categories [1]. Firstly, there are processes that directly relate to treatment of a patient (e.g. describing treatment of a patient with a chronic ailment), and secondly there are processes that relate to organizational duties (e.g. the process of a reservation of a clinical bed for a patient within different hospital facilities). In our approach, however, we do not differentiate between these two types of processes and we try to work with them in the same way as a set of process diagrams and use them together. The reason for combining different sources of knowledge is to enable validation of applied procedures, on a general level, as well as with a local practice in a hospital, that can differ in each facility and that is captured as clinical processes.

The main goal of this paper is to summarize all aspects necessary for agent-based process simulation in a medical environment leading to an critiquing expert system. Therefore we firstly discuss more exhaustively processes and process simulation and its specific characteristics in the medical domain in Secion 2, where we also reason about the advantages that utilization of agents can bring into the field of process simulations. In Section 3 we present formal definitions of our enhanced processes. We describe the architecture of a multi-agent system that can simulate these processes in an environment in Section 4. Then in Section 5 we propose the vision of the whole expert critiquing systems, that can use this approach, and conclude in Section 6.

2. Process Modelling in Medicine

The work practice (i.e. duties of employees and organizational procedures – such as a specification of an activities' order, an assignment of employees as well as necessary resources to these activities, etc.) is

usually captured using a set of processes describing the functioning of a work team or the whole company. These processes can be stored as a document in a textual form, and often these documents also contains their models, made using some of process modelling languages, as visual diagrams, which improve understanding and lucidity of the information.

There are several studies [1, 2, 3] 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 is the usage of processes rather low and insufficient. The main reasons were identified as more complex processes than in other fields of industry, or problems with interoperability resulting from inconsistencies of databases and used ontology or protocols. Finally, real processes in medical care are very variable hence the system that uses them has to be prepared for such a dynamic environment and multiple variations of similar processes. This factor prohibits us from applying standart workflow systems, that can not handle exceptions nor irregular situations.

As we have already stated, processes in medical care can be seen in several levels. Using terminology from [1] we can differentiate the *organizational processes* and the *medical treatment processes*. The latter type can be seen as medical guidelines that represent the recommended diagnosis and treatment procedures for a patient in a specific area of healthcare. They are approved by medical experts in related field based on the newest studies, literature reviews, and expert knowledge. There have been several surveys aimed at the importance of guidelines and generally they are considered to be a useful method for standardizing the medical practice, improving quality of treatment [4], or lowering the patient's medical expenses [5]. Currently, the guidelines are being approved as a document (i.e. in a textual form), which prohibits one from using them directly in a computer-based system – such as hospital information system, or an expert system helping a physician or a patient. Hence there has been a significant focus on the formalization of medical guidelines into a formal language [6]. Many formal languages have been developed, such as ASBRU [7], EON [8], GLIF [9], or PRO*forma* [10]. They are all quite different and based on different foundations, but they are all trying to capture the same thing – the recommended process of treatment of a patient in the specific area of healthcare.

In order to achieve corresponding simulation of processes, we need to simulate both of these types, as they affect each other – an organizational process is strictly limited by patient's health conditions, on the other hand, a physician has to take specific clinical processes and hospital organization into consideration when treating a patient. Furthermote we use both types in the same way – i.e. formalized using the same theoretical foundations, but in possibly different languages. In spite of several other proposed approaches (e.g. like in [11]), we do not try to convert one formalism into other one (e.g. formalize medical guideline using a business process modelling language), but modify existing formal languages in order to capture all necessary information for the agent-based simulation.

2.1. Using Agents in Simulations of Processes

Generally, we are interested in a simulation of processes in a certain environment by means of agents and we want to create a multi-agent system that would be coordinated and organized by a set of processes. Let us therefore discuss the advantages and disadvantages of this approach.

Using agents to simulate processes in companies is a promising alternative to standard process simulation methods based on statistical calculations [12]. There are several studies addressing this issue [13, 14, 15], and they all highlight the advantages, that agents are more accordant with people, they can be autonomous, they can plan their assignments and they can distribute and coordinate their activities. However, in practice, there are not many existing applications that would interpret a process language in a multi-agent system and let agents be guided directly by modeled processes. In some cases, even though the agents are supposed to simulate processes, their behavior is hand-coded depending on processes (e.g. in [14]) using some standard decision mechanism (e.g. rules, FSM, etc.). Several approaches in the area of WfMS were discussed in [16] or even processes modelling in [17], however no existing implementation or transforming algorithm for agents' behavioral definition was presented. In both these cases authors try to cover much wider concepts (e.g. different views on a single process by different agents, concept of trust etc.) which prohibits them from proposing the universal MAS architecture and algorithm interpreting the processes into behavior of agents. Therefore we introduced a new approach to a process simulation in [18] that defines a universal multi-agent system and transforming algorithm that enables process simulation by means of reactive autonomous agents.

When we closely focus on using agents in process simulation in medical care, we can see that several problems, mentioned in previous section, can be

overcome. When agents represent the hospital staff or patinets, they do not have only to follow the modeled processes, but also their own pre-defined goals, hence the exceptions or interruptions of process execution can be handled much easier. Furthermore, using enhanced processes described in [18], the variability of processes can be assured.

3. Formal Definition of Agent-based Process Simulation

In order to correctly define multi-agent system that simulates modeled processes we firstly need to properly define processes modeled in process diagrams.

Definition 1: We call a seven-tuple $D = (P, S, E, C, O, A, R)$ a process diagram, when:

- P is a non-empty set of processes (activities).
- S is a set of passive states that describes current state of environment.
- C is a set of connectors that can split or join the control flow.
- $E \subseteq (\{P, S, C\} \times \{P, S, C\})$ is a non-empty set of control edges that connect processes and define a control flow of a diagram.
- O is a set of objects from the environment. Each object has a set of parameters that can be modified by processes.
- A is a non-empty set of roles of agents that participate in activities.
- $R \subseteq (\{P, O, A\} \times \{P, O, A\})$ is a set of auxiliary edges (relations) connecting agents and objects with processes.
- Process diagram is a directed graph $G = (V, X)$, where $V = (P \cup S \cup C \cup O \cup A)$ and $X = (E \cup R)$

Furthermore, when D is a process diagram, and

- p_v is a set of vertexes preceding to the vertex v; $p_v = \{n \in V; (n, v) \in E\}$
- s_v is a set of vertexes succeeding to the vertex v; $s_v = \{n \in V; (v, n) \in E\}$

following conditions have to hold:

- The sets of vertexes P, S, C, O, A are pairwise disjoint. The same condition holds for the sets of edges E and R.
- There is at most one edge outgoing of and incoming to each node except connectors; $\forall v \in \{V \setminus C\} : (|p_v| \leq 1) \wedge (|s_v| \leq 1)$
- Logical connectors have at least one incoming and at least one outgoing edge. We distinguish exactly two types of connectors – splitters and joiners. We thus define two disjoint subsets of the set of connectors as: $C = T \cup J$, where $T \cap J = \emptyset$. Now the following corollaries hold:
  - splitters – connectors that have exactly one incoming edge and at least two outgoing edges; $\forall t \in T : (|p_t| = 1) \wedge (|s_t| > 1)$
  - joiners – connectors that have at least two incoming arc-edges and exactly one outgoing arc-edge; $\forall j \in J : (|p_j| > 1) \wedge (|s_j| = 1)$

This definition of process is quite universal. It is based on EPC language definition [19] that is widely used in business process modelling. However, it is extended in order to cover other specific languages as well, specifically GLIF, that we use to formalize medical guidelines. We use three control entities (processes, states and connectors) that forms the control flow, and two auxiliary entities (agents, objects) that describe processes in more detail. Note, that in definition of relations (the set R), we allow the connections between different roles as well. This corresponds to definition of organizational hierarchy in a team using roles (e.g. Jane is a nurse and she also is a general employee).

Now, let us define the enhancements for a general process language identified in [18], in order to be able to properly simulate them using a multi-agent system.

Definition 2: We say, that $D' = (P, S, C, E, O, A, R)$ is an enhanced process diagram, when for each $p \in P$ hold:

- $O_{p_i} = \{o \in O; (o, p) \in R\}$ is a set of input objects of the process. Following properties of each input object have to be specified:

  optional – relation that represents whether this object is necessary for executing the process or not

  utilization – float number representing the amount of usage of the input object in order to use it in several processes at the same time

- $O_{p_o} = \{o \in O; (p, o) \in R\}$ is a set of output objects of the process. If an output object is not also an input object, the process creates a new object in the environment.
- $A_p = \{a \in A; (a, p) \in R\}$ is a set of roles of executing agents. Following properties have to be specified for each role:

  optional – a relation that represents whether this agent is necessary for executing the process or not

  utilization – a float number representing the amount of agent's utilization in order to enable the possibility of multi-tasking of agents

  replace – a relation that represents whether agent should be replaced by another agent possesing this role when it interrupts the execution of this process, or not
- *location* – an optional characteristic represented by one of the input objects. As we are running the simulation in a certain environment, there can be a need for executing each process at precise location (e.g. an examination should be executed in the appropriate room of hospital that can be modeled as a virtual world for the visualization of the whole simulation).
- *priority* – an integer number representing the priority of the process
- *transition function* – a description of the course of the activity as such. Let X_i be the domains of changing parameters of output objects of the process. Then we say, that

  $$f^p_{O_{p_i}} : \mathbb{N} \longmapsto (X_1, X_2, \ldots, X_m)$$

  is a transition function of the process that for each timestep (a natural number) returns the actual values for each changing parameters of the output objects.

These enhancements are mostly natural and correctly specify input and output objects with their characteristics, or participating agents. Note, that we allow cooperation of several agents on a single process ($\mid A_p \mid \geq 1$), and we introduce multi-tasking of agents as well.

We explain the definition of the transition function, that represents the course of the process. We use a concept of mathematical functions that can be defined for each output effect of a single process separately, meaning we are modeling several courses of changes in time – one for each output parameter (e.g. the state of a patient examination request can change during an execution of a single process from "new" through "verified" to "prepared"). Thanks to using general mathematical functions we can determine the precise state of all output effects at any time and we are able to apply partial results into a virtual world when an interruption of the process occurs. Because all of the described functions have only one input variable – discrete time – we can transform the set of functions as a single multidimensional transition function of the process. Finally, according to a real-life practice, we expect the real course of the function during the simulation to depend on the actual state of environment and input objects (e.g. if we have some of the optional input objects we need less time to accomplish a tas). Hence the transition function is parametrized by these aspects.

4. Multi-agent System for a Process Simulation

In previous section we defined the enhanced processes nad now let us define a multi-agent system (MAS), that simulates them. Firstly we present an existing MAS architecture, which was proved usefull as prototype implementation in [18], following by several enhancements that we want to implement in our current approach in order to improve the course of the simulation as such.

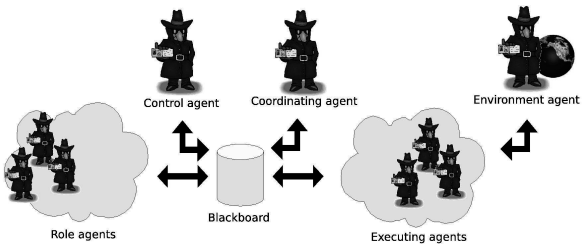


Figure 1: The architecture of the MAS simulating enhanced processes.

The organizational scheme, shown in Figure 1, represent a multi-agent system that can simulate processes captured in a formalism for enhanced processes. We differentiate several types of agents, but there are three main groups. The first one is an agent representing the environment in which the simulation proceeds. Secondly, there are *executing agents* that correspond with the modeled hospital staff (e.g. physicians) that act within the environment. Finally, we identify three types of auxiliary agents (*control*, *coordinating* and *role agents*) which help to organize *executing agents* in case of more complicated scenarios. Communication of agents uses the blackboard architecture, where every

agent is able to read and write facts (e.g. activation of specific processes for an execution agent, etc.) at the common blackboard. As the decision mechanism for the execution agents, the hierarchical reactive plans were used, as they are easy to automatically generate from process diagrams and they can define reasonably complex behavior of an agent.

Let us now describe the functioning of the system. For auxiliary agents, we present their behavior in the pseudocode, that for brevity handles with only one instance of process diagrams in the system. In the implementation, however, several instances of the same process diagram can be active. Firstly, we focus on a simple scenario – simulation of a single process. An *executing agent* reads from the blackboard a set of currently allowed actions (they are allowed entirely based on progress in process diagrams), it autonomously chooses one of them on the basis of its internal rules, priority of the processes, the ability to satisfy the input conditions, and commits itself to execute it. It asks the *transition function* of the process, what is the expected finish time of this instance of the activity (as it can depend on the actual values of input objects parameters), and after the specified time it applies the target values of the effects of the activity as provided by the transition function and marks the activity as finished at the blackboard. However, during the execution of the activity, the agent can suspend its work (e.g. because it needs to accomplish a task with higher priority). At the time of the occurrence of this suspension, the agent asks the transition function for actual values of all effects and reflects the partial changes in the environment.

Algorithm 1 Rules for the *control agent*

1: **if** $\exists p \in P : finished(CoordAgent, p)$ **then**
2: choose processes $P' \subseteq P$ subsequent to p according to the process rules
3: **if** $P' \neq \emptyset$ **then**
4: $remove(finished(CoordAgent, p))$
5: **for all** $p' \in P'$ **do**
6: $store(active(CoordAgent, p'))$
7: **end for**
8: **end if**
9: **end if**

Described scenario was the simplest one, however in more advanced cases, the three auxiliary agent types are used. The *control agent* is the one who controls the correct order of the process execution according to the process diagrams and sets the set of currently allowed activities. We can demonstrate its behavior using pseudocode shown in Algorithm 1. Note, that movement in the process chain in line 2 can contain several steps or possibly splitting or joining the flow using a connector.

In the case of cooperation of several agents in a process execution, the *coordinating agent* takes responsibility for notifying the correct subordinate agents (lines 1–4), it selects which agent is so called *master agent* (i.e. the one, that actually modifies objects used in the process; lines 5–6), and monitors the progress of the execution (lines 8–27). Coordination agent is also necessary in the case of an interruption, when it chooses one of the other participating agents to be the master agent (lines 16–25):

Algorithm 2 Rules for the *coordinating agent*

1: **if** $\exists p \in P : (active(CoordAgent, p) \wedge (\neg\exists a \in A_p, \exists p' \in P : \neg optional(a, p) \wedge active(a, p') \wedge (priority(p') > priority(p))))$ **then**
2: **for all** $a \in A_p$ **do**
3: $store(active(a, p))$
4: **end for**
5: choose one $a \in A_p$
6: $store(master(a, p))$
7: **end if**
8: **for all** $p \in P$ **do**
9: **if** $(\exists a \in A_p) : (active(a, p) \wedge master(a, p) \wedge \neg working(a, p))$ **then**
10: **if** $finished(a, p)$ **then**
11: $remove(finished(a, p))$
12: **for all** $a' \in A_p$ **do**
13: $remove(active(a', p))$
14: **end for**
15: $store(finished(CoordAgent, p))$
16: **else if** $interrupted(a, p)$ **then**
17: **if** $\neg optional(a, p)$ **then**
18: **for all** $a' \in A_p$ **do**
19: $remove(active(a', p))$
20: **end for**
21: **else**
22: choose one $a' \in \{e \in \{A_p \setminus a\} : working(e, p)\}$
23: $store(master(a', p))$
24: **end if**
25: **end if**
26: **end if**
27: **end for**

Finally, we describe the *role agents*. We are using the concept of roles, hence the role agent reads the set of currently active processes for the given role (set by the coordinating agent, line 1) and activates them for selected executing agent (lines 2–4). When

an interruption occurs and the suspended agent should be replaced, a role agent is responsible for notifying another executing agent possessing the same role (lines 11–14).

Algorithm 3 Rules for a *role agent*

Input: a is this role agent; Ex_a is a set of *executing agents* that posses this role

1: **if** $(\exists p \in P) : active(a,p) \wedge \neg working(a,p)$ **then**
2: choose one $e \in \{c; c \in Ex_a \wedge (\neg\exists p' \in P : working(c,p') \wedge (priority(p') > priority(p)))\}$
3: $remove(interrupted(a,p))$
4: $store(\{active(e,p)), deleg(a,e,p), working(a,p)\})$
5: **end if**
6: **for all** $p \in P$ **do**
7: **if** $\exists e \in Ex_a : deleg(a,e,p) \wedge \neg working(e,p)$ **then**
8: **if** $finished(e,p)$ **then**
9: $remove(\{active(a,p), deleg(a,e,p)\})$
10: $store(finished(a,p))$
11: **else if** $\neg finished(e,p) \wedge replace(a,p)$ **then**
12: $remove(\{active(e,p), deleg(a,e,p)\})$
13: choose one $e' \in \{c; c \in Ex_a \setminus \{e\} \wedge (\neg\exists p' \in P : working(c,p') \wedge (priority(p') > priority(p)))\}$
14: $store(\{active(e',p), deleg(a,e',p)\})$
15: **else**
16: $remove(working(a,p))$
17: $store(interrupted(a,p))$
18: **end if**
19: **else if** $working(a,p) \wedge \neg active(a,p)$ **then**
20: **for all** $e \in Ex_a : deleg(a,e,p) \wedge active(e,p)$ **do**
21: $remove(\{active(e,p), deleg(a,e,p)\})$
22: **end for**
23: **end if**
24: **end for**

4.1. Transforming Algorithm

Let us now describe how the set of rules for an executing agent is automatically generated and how its action-selection mechanism works.

As we have already stated, we are using the reactive architecture for *execution agents*, hence each goal of the plan is represented by a fuzzy if-then rule. For each process the executing agent can participate in, one rule is automatically generated. These rules are for each agent ordered by the descending priority of the activities and they create the first level of hierarchical architecture of the agent. The second layer is created by several sets of rules, where each set is related to one first-level rule. This second-level set of rules represents several partial activities that are necessary to execute according to the conventions in the environment (e.g. transporting movable objects to the location of the execution of the process), and one rule for executing the simulation of the activity as such (modeled by a *transition function* as described in Section 3). Except the last one, the nature of these rules depends on the conventions that hold in the virtual world and therefore cannot be generalized.

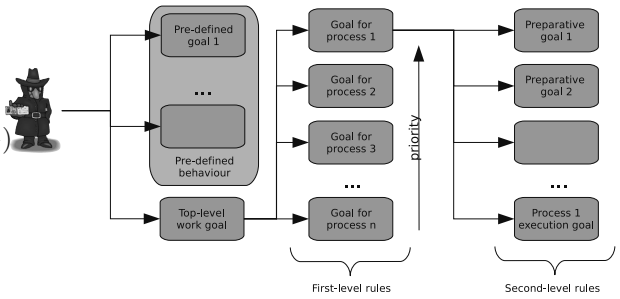


Figure 2: A hierarchy of reactive plans of each *executing agent*

The condition of a first-level rule is created as a conjunction of all constrains related to properties of input objects and agents (i.e. correct values of their utilization (whether they can execute this activity) and possibly other attributes, such as the state of an patient etc.), and activation of an appropriate process. Moreover, if an input object or a participant is not mandatory, related conditions do not need to hold in order to fire the rule.

4.2. Improved Architecture of the MAS

The architecture presented in previous section can successfully simulate modeled processes and as such can suit our intention to create an expert critiquing system based on the simulation of clinical processes and formalized medical guidelines. However, several issues can be improved in previous approach. First of all, the control and coordination of *execution agents* using specialized auxiliary agents together with a blackboard architecture is quite stiff and it partially limits the autonomy of *executing agents*. Moreover, in order to enhance *executing agents* with planning or advanced architectures, much more organization-related communication would be needed.

Therefore we propose a new architecture that, according to our experiences gained during implementation and testing the previous one, should emphasize more the positive concepts of agents paradigm and enable implementation of further improvements and functionality, such as planning and better *executing agents* coordination. Currently, we do not change the

reactive architecture of the *executing agents* as there is not known correct interpretation of common knowledge in form of processes for more deliberative agents. We argue that processes have stronger conceptual meaning than a plan library for an agents, as not only an agent knows what actions it needs to execute, but also what actions other agents should execute and how their actions would affect the state of the environment. This remains an open problem which we want to address in further research.

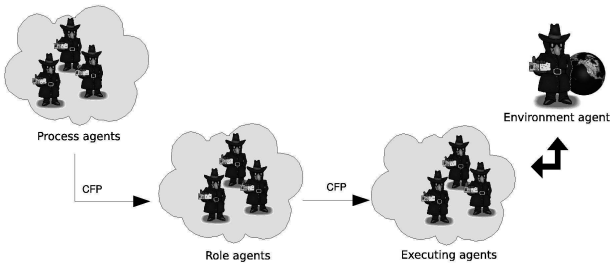


Figure 3: The improved architecture of the MAS simulating enhanced processes.

The schema of the new architecture is shown in Figure 3. Both control and coordinating agent were replaced by a set of agents – for each of modeled process one *process agent* is automatically generated. Each of them is responsible for executing one type of activity (possibly several instances of one process) and the duties of removed auxiliary agents are distributed within this set. Note that in the new architectural schema, the blackboard is no longer used. The simplified organizational concept enables the possibility of usage the direct messaging as well as a standard concept of the Contract-Net Protocol (CNP) [20].

The pseudocode of a *process agent* is shown in Algorithm 4 and it presents how a it acts in the simulation. Note that the pseudocode is reduced (several lines regarding the responses to rejections are omitted). We can see, that agent keeps to the CNP and each *process agent*, when notified (lines 2–6), finds appropriate role agents (lines 25–26 and 11–15), monitors the progress of the master agent in case of cooperation, and passes the information of the success to the next process agent(lines 16–24). Other agents, *role* and *executing*, are acting in the same way, except the changes in the communication.

Let us point out the advantages, that these modifications can bring. The key change is the shift from the blackboard architecture to the direct messaging within agent community together with using standard protocols. At the cost of increasing the overall number of agents we simplify the communication within agents (compare the organizational communication issues in *control* and *coordinating agent* with *process agents*). Moreover, we expect easier integration of planning that can be added as further communication within *process* and *role agents* (e.g. one *process agent* knows, what the subsequent processes are, hence it can notify appropriate agents in advance and negotiate executing some of the auxiliary actions (see the second-level rules in Section 4.1) to save time).

Algorithm 4 Rules for a *process agent*

Input: p is a process assigned to this agent; I_p is the set of currently active instances of p; m_i master agent of the process for $i \in I_p$; X_i is a set of returned proposals for $i \in I_p$ asdasd

1: **for all** $msg \in IncMsgQueue$ **do**
2: **if** msg is activation of i **then**
3: $I_p = I_p \cup i$
4: $m_i = \emptyset$
5: i is *new*
6: $X_i = \emptyset$
7: **else if** msg is proposal for i **then**
8: $X_i = X_i \cup msg$
9: **end if**
10: **end for**
11: **if** $CFPTimeOut \wedge X_p \neq \emptyset$ **then**
12: choose one agent, m_i, from X_i
13: $sendAcceptProposal(i, m_i)$
14: i is *started*
15: **end if**
16: **for all** $i \in I_p$ **do**
17: **if** i is *finished* **then**
18: $I_p = I_p \setminus i$
19: **for all** $p \in (s_p \cap P)$ **do**
20: $sendActivation(success(i), a)$
21: **end for**
22: **else if** i is *interrupted* $\wedge \{A_p \setminus m_i\} \neq \emptyset$ **then**
23: $X_i = \emptyset$
24: $sendProposal(i, A_p)$
25: **else if** i is *new* **then**
26: $sendProposal(i, A_p)$
27: **end if**
28: **end for**

5. Future Work

So far we discussed processes, problems related to their simulation, and proposed a solution based on a multi-agent system. Now we present the vision of the critiquing expert system which can profit from these methods.

The critiquing system runs in the background of the standard applications of HIS and controls the inserted data about a patient. From these data values it tries to recognize a medical guideline that physician is following and furthermore recognize the state of the patient. After a successfull matching, it further predicts the future progress of possible patient's treatment with respect to the guidelines and database of existing cases in the facility. This prediction follows the next steps in guidelines (note, that patient can have several diseases hence we need to take all of them into consideration) and tries to simulate the future actions of the physician and in case of missing current data value (e.g. a result from an examination that patient have not undergone yet) the approximation using similar patients from the database is made. Also, this prediction would be probabilistic, hence multiple branches of the guidelines would be evaluated. Therefore, in case of for example omitting an optional examination, physician can be alerted by the system that similar patients had results that negatively affect their further progress. Finally, the simultaneous work with several guidelines for different diseases can bring attention of the physician that treatment of a disease she/he is focused on can conflict with another treatment that this patient is going through.

In this system, we want to combine several existing techniques. For a guideline recognition we want to use ideas from existing plan recognition techniques (such as using Bayesian network), and for guideline simulation we want to apply the approach described in this paper. However, the advantage of usage of agents for a guideline simulation purpose (and the whole system as such) is not so evident. We argue that focusing on distributed artificial intelligence can simplify the implementation and also the adaptivity of the system (e.g. learning of the specialized *process agents*). Finally, in the future a system designed on such general principles could also be integrated into more advanced HIS based on processes, which could help to plan and organize work in a hospital facility with a close relation to specific patients' treatment.

6. Conclusions

In this paper we presented an approach to an agent-based simulation of processes in an environment and described its possible utilization in medical care – specifically in the development of an critiquing expert system that would use formalized medical guidelines as a knowledge base. We formally defined processes and their enhancement which helped us to closely describe the functioning of the multi-agent system that simulates the processes and finally, we presented our vision of application of this approach in medicine.

Because such a direct usage of processes to control a multi-agent system has not been till now a not very explored area, there are several open issues: further improvement of the architecture of the MAS, implementation of planning and learning, or using more deliberative decision mechanisms for *executing agents*. In the following work we want to address some of them and prove the usefulness of this method by implementation of the working critiquing system that would help the physician with their work.

References

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

[2] 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.

[3] 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.

[4] A. G. Ellrodt, L. Conner, M. Riedinger, and S. Weingarten, "Measuring and Improving Physician Compliance with Clinical Practice Guidelines: A Controlled Interventional Trial," *Ann Intern Med*, vol. 122, no. 4, pp. 277–282, 1995.

[5] J. Cartwright, S. de Sylva, M. Glasgow, R. Rivard, and J. Whiting, "Inaccessible information is useless information: addressing the knowledge gap," *J Med Pract Management*, vol. 18, pp. 36–41, 2002.

[6] P. A. de Clercq, J. A. Blom, H. H. M. Korsten, and A. Hasman, "Approaches for creating computer-interpretable guidelines that facilitate decision support.," *Artificial Intelligence in Medicine*, vol. 31, pp. 1–27, 2004.

[7] Y. Shahar, S. Miksch, and P. Johnson, "The asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines.," *Artificial Intelligence in Medicine*, vol. 14, pp. 29–51, 1998.

[8] S. Tu and M. Musen, "A flexible approach to guideline modeling," *Proc AMIA Symp.*, pp. 420–424, 1999.

[9] M. Peleg, A. Boxwala, and O. Ogunyemi, "Glif3: The evolution of a guideline representation format.," *Proc AMIA Annu Fall Symp.*, pp. 645–649, 2000.

[10] J. Fox, N. Johns, A. Rahmanzadeh, and R. Thomson, "Proforma: A method and language for specifying clinical guidelines and protocols," in *Amsterdam*, 1996.

[11] L. Dazzi, C. Fassino, R. Saracco, S. Quaglini, and M. Stefanelli, "A patient workflow management system built on guidelines.," *Proc AMIA Annu Fall Symp*, pp. 146–150, 1997.

[12] A. W. Scheer and M. Nüttgens, "ARIS architecture and reference models for business process management," in *Bus. Proc. Management, Models, Techniques, and Empirical Studies*, (London, UK), pp. 376–389, Springer-Verlag, 2000.

[13] M. Sierhuis, *Modeling and Simulating Work Practice*. PhD thesis, University of Amsterdam, 2001.

[14] N. R. Jennings, P. Faratin, T. J. Norman, P. O'Brien, and B. Odgers, "Autonomous agents for business process management," *Int. Journal of Applied Artificial Intelligence*, vol. 14, no. 2, pp. 145–189, 2000.

[15] A. Moreno, A. Valls, and M. Marín, "Multi-agent simulation of work teams," in *Multi-Agent Systems and Applications III: 3rd Int. CEEMAS*, (Prague, Czech Republic), June 16-18 2003.

[16] M. P. Singh and M. N. Huhns, "Multiagent systems for workflow," *Int. Journal of Intelligent Syst. in Accounting, Finance and Management*, vol. 8, pp. 105–117, 1999.

[17] C. de Snoo, "Modelling planning processes with talmod," Master's thesis, University of Groningen, 2005.

[18] B. Bosansky, "A virtual company simulation by means of autonomous agents," Master's thesis, Charles University in Prague, 2007.

[19] A. Finkelstein, J. Kramer, B. Nuseibeh, L. Finkelstein, and M. Goedicke, "Viewpoints: A framework for integrating multiple perspectives in system development," *Int. Journal of Software Eng. and Knowledge Engineering*, vol. 2, no. 1, pp. 31–57, 1992.

[20] R. G. Smith, "The contract net protocol: high-level communication and control in a distributed problem solver," pp. 357–366, 1988.

On the Independence of Axioms in BL and MTL

Post-Graduate Student:
MGR. KAREL CHVALOVSKÝ
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
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, Czech Republic
bilkova@cs.cas.cz

Field of Study:
Logic

This work was supported by GA ČR EUROCORES project ICC/08/E018.

Abstract

We show by standard automated theorem proving methods and freely available automated theorem prover software that axiom (A2), stating that multiplicative conjunction implies its first member, is provable from other axioms in fuzzy logics BL and MTL without using axiom (A3), which is known to be provable from other axioms [1]. We also use freely available automated model generation software to show that all other axioms in BL and MTL are independent.

1. Introduction

Among propositional fuzzy logics Hájek's basic logic BL [3] and Esteva and Godo's monoidal t-norm based logic MTL [2] play prominent roles. BL, which was introduced as a common fragment of Łukasiewicz, Gödel and product logics, is the logic of continuous t-norms[1] and their residua[2]. However, in [2] was shown that the minimal condition for a t-norm to have a residuum is left-continuity and authors proposed logic MTL, which was later proved to be the logic of left-continuous t-norms and their residua.

Standard Hilbert style calculus for BL comes from Hájek. Esteva and Godo slightly addapted this system for MTL by replacing one axiom by three other axioms. Generally, both systems are almost identical. In a short note by Cintula [1], it was shown that axiom (A3), stating commutativity of multiplicative conjunction, is provable from other axioms and thus redundant. Lehmke proved that also axiom (A2), stating that multiplicative conjunction implies its first member, is provable from other axioms by using his own Hilbert style proof generation software [4]. However, the proof used axiom (A3) and thus was not a proof of independence of both axioms (A2) and (A3).

We use a well known technique of automated theorem proving to encode the Hilbert style calculus of a fuzzy propositional logic into classical first order logic, and standard automated theorem proving software to prove axiom (A2), without using axiom (A3), in BL and MTL. Moreover, by an easy application of similar technique and standard automated model generation software we show that none of the other axioms is redundant in BL and MTL, independently of presence of axioms (A2) and (A3).

The interest of this paper is solely in above stated properties of Hilbert style calculus of BL and MTL. The technique used to obtain them can be in our case used completely naive.

The paper is organised as follows. In Section 2 we set up notation and terminology. In Section 3 we give a brief exposition of techniques used to obtain presented results. Section 4.1 contains the proof of derivability of axiom (A2) for MTL and Section 4.2 for BL. In Section 5 the semantic proofs of independence of other axioms are presented.

2. Preliminaries

We will touch only a few aspects of the theory. For simplicity of notation, we use fuzzy logic for fuzzy propositional logic and first order logic (FOL) for classical first order logic. First order fuzzy logics and classical propositional logic are not discussed in this paper.

We define standard Hilbert style calculus for the *Basic*

[1] A t-norm is a binary function $\star$ on linearly ordered real interval $[0, 1]$ which satisfies commutativity, monotonicity, associativity and 1 acts as identity element.

[2] The operation $x \Rightarrow y$ is the residuum of the t-norm $\star$ if $x \Rightarrow y = \max\{z \mid x \star z \leq y\}$.

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

Definition 2.1 *We define the basic logic BL 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) $(\varphi \,\&\, \psi) \rightarrow (\psi \,\&\, \varphi)$,

(A4) $(\varphi \,\&\, (\varphi \rightarrow \psi)) \rightarrow (\psi \,\&\, (\psi \rightarrow \varphi))$,

(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 BL 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 [3]. Axiom (A1) is transitivity of implication. Axiom (A2) states that multiplicative conjunction implies its first member. Axiom (A3) is commutativity of multiplicative conjunction. In BL, additive conjunction $\varphi \wedge \psi$ is definable as $\varphi \,\&\, (\varphi \rightarrow \psi)$. The equivalence of these two formulae is the divisibility axiom. Axiom (A4) is commutativity of additive conjunction. 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.

Definition 2.2 *Hilbert style calculus BL$^-$ is obtained by dropping axioms* (A2) *and* (A3) *from BL.*

We obtain a Hilbert style calculus of the monoidal t-norm based logic MTL by weakening properties on additive conjunction. In BL, we define $\varphi \wedge \psi$ as an abbreviation for $\varphi\&(\varphi\rightarrow\psi)$. In MTL, we define additive conjunction directly by three new axioms which state that additive conjunction is commutative, implies its first member and one implication of divisibility property.

Definition 2.3 *We obtain the monoidal t-norm based logic MTL by replacing axiom* (A4) *in BL by following three axioms*

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

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

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

Definition 2.4 *Hilbert style calculus MTL$^-$ is obtained by dropping axioms* (A2) *and* (A3) *from MTL.*

2.1. First order logic and automated theorem proving

A FOL *model* is a pair $\langle D, I\rangle$, where domain D is a set of elements and I is an interpretation of symbols of a language.

In FOL, *terms* are defined inductively as the smallest set of all variables and constants closed under function symbols in given language. We will have only one predicate symbol Pr and thus all our *atomic formulae* have a form $Pr(t)$, where t is a term. A *literal* l is an atomic formula (*positive literal*) or a negative atomic formula (*negative literal*). A *clause* C is a finite disjunction of literals. Specifically, a *Horn clause* is a clause with at most one positive literal. All clauses will be for our purposes implicitly universally quantified. *Unification of literals l and l'* is a substitution σ which gives $l\sigma = l'\sigma$. So called *most general unifier of l and l'*, denoted $\mathrm{mgu}(l, l')$, is a unification σ such that for every unification θ of l and l' exists a unification η satisfying $\theta = (\sigma)\eta$.

The standard FOL automated theorem proving strategy is resolution [5]. We can transform a problem of $\Gamma \vdash \varphi$ to the problem of deciding whether set $\{\Gamma, \neg\varphi\}$ is contradictionary. Let $\sigma = \mathrm{mgu}(l, l')$, then resolution calculus with (binary) resolution rule

$$\frac{C \vee l \qquad D \vee \neg l'}{(C \vee D)\sigma}$$

and factoring rule

$$\frac{C \vee l \vee l'}{(C \vee l)\sigma}$$

is refutational complete [5], which means that for every contradictionary set eventually find a derivation of empty clause which represents a contradiction.

3. Usage of ATP methods

There is a well known technique for encoding propositional Hilbert style calculus into classical FOL through terms. The key idea is that formula variables in axioms and rules are encoded as universally quantified first order variables and propositional connectives as first order function symbols. Moreover, we use one unary predicate which says which terms are provable (encoding of axioms) and how another provable term can be obtained from provable terms (encoding of rules). It is evident that our axioms and modus ponens rule can be encoded easily. However, for more complicated axioms and rules problems may arise.

For simplicity of notation, we write Fle_L instead of the set of all formulae in language L.

Definition 3.1 *Let L be BL or MTL or their fragment. We define term encoding $f : Fle_L \to Fle_{FOL}$.*

First, a function $f' : Fle_L \to Fle_{FOL}$ is defined recursively as follows

$$f'(\varphi) = \begin{cases} 0_f & \varphi \text{ is } \overline{0}, \\ f'(\psi) \to_f f'(\chi) & \varphi \text{ is } \psi \to \chi, \\ f'(\psi) \,\&_f\, f'(\chi) & \varphi \text{ is } \psi \,\&\, \chi, \\ f'(\psi) \wedge_f f'(\chi) & \varphi \text{ is } \psi \wedge \chi, \\ X_\psi & \varphi \text{ is a formula variable } \psi, \end{cases}$$

where $\&_f$, $\wedge_f$ and $\to_f$ are new binary function symbols, written for better readability in infix notation, 0_f is a new FOL constant and X_ψ is a new FOL variable for every formula variable ψ, but the same for every occurrence of ψ in the encoded formula.

Second, formula $f(\varphi)$ is the universal closure of formula $Pr(f'(\varphi))$, where Pr is a common new unary predicate saying which terms are provable.

Finally, let $\varphi_1, \ldots, \varphi_n \vdash \psi$ be a propositional rule (in our case just (MP)*), we define term encoding f into classical FOL as the universal closure of formula $(f'(\varphi_1) \wedge \ldots \wedge f'(\varphi_n)) \Rightarrow f'(\psi)$, where $\wedge$ and $\Rightarrow$ are standard logical connectives for conjunction and implication in classical FOL and function f' is defined as above.*

Example Let us have a system with axioms (A2), (A3) and the only rule (MP). This propositional system will be formalised, for better readability with X and Y instead of X_φ and X_ψ, in FOL as follows

$$(\text{A1}_f)\ \ (\forall X, Y) Pr((X \,\&_f\, Y) \to_f X),$$

$$(\text{A2}_f)\ \ (\forall X, Y) Pr((X \,\&_f\, Y) \to_f (Y \,\&_f\, X)),$$

$$(\text{MP}_f)\ \ (\forall X, Y)(Pr(X) \wedge Pr(X \to_f Y) \Rightarrow Pr(Y)).$$

Before stating a crucial lemma we make some remarks. For a set of formulae Γ, we define $f(\Gamma)$ as a set of all f-translated formulae from Γ. We write $f(MP)$ for the term encoding f of modus ponens rule.

By an easy observation we realize that all translated axioms and modus ponens translation, written in form of disjunction, are Horn clauses.

Lemma 3.2 *Let L be BL or MTL or their fragment with the set of axioms Δ, Γ arbitrary set of formulae, and φ arbitrary formula, both in language of L. Then $\Gamma \vdash_L \varphi$, if and only if $f(\Delta), f(\Gamma), f(MP) \vdash_{FOL} f(\varphi)$.*

Proof: A Hilbert style proof of φ from Γ can be easily translated into a Hilbert style proof of $f(\varphi)$ from $f(\Delta), f(\Gamma)$ and $f(MP)$ in classical FOL using generalisation rule, if $\vdash_{FOL} \psi$ then $\vdash_{FOL} \forall x \psi$, and $\vdash_{FOL} \forall x \psi \to \psi$.

The opposite direction can be shown by using a resolution refutation. It is an easy observation that only Horn clauses occur in such a resolution refutation. And this fragment has a property that given resolution refutation can be reordered in such a way that a backward translation gives a proof of φ in Γ. ∎

Demonstrating the independence of some axiom, we are also interested in unprovability. There is a standard model theoretical technique for proving that some formula φ is unprovable from a set of formulae Γ. From soundness theorem in FOL it is enough to show a FOL model in which all formulae from Γ are true and formula φ is false. By previous lemma we can easily transform a problem of unprovability φ from Γ in a Hilbert style calculus to a problem of finding classical FOL model in which $f(\Delta), f(\Gamma), f(MP)$ and $\neg f(\varphi)$ are true.

We have thus transformed the problem of provability of formula in propositional fuzzy logic Hilbert style calculus into FOL and we can try to solve it by standard automated theorem proving software. We can use a theorem prover for showing that some formula (in an encoded form) is provable from other formulae using given rules, or a model generator software to find a model which demonstrates its unprovability. Traditionally, both computations are executed in parallel.

Generally speaking, because of undecidability of FOL, this technique cannot be fully satisfiable. Moreover, abilities of automated theorem provers and automated model generators are very limited and highly dependent on software configuration. However, several results were obtained by this or similar techniques, which proved its usability, see for instance Wos's papers [6].

We are not going to describe technique used by automated theorem provers and model generators, because these systems are rather complicated. For our experiments we used freely available E prover in version 0.999-001[3], which is based on superposition (restricted paramodulation) calculus. For building models we used freely available Paradox 2.3[4] finite model finder which iteratively tries to find finite models by transforming a given problem into SAT problems.

Tuning software for obtaining results can be highly complicated. Nevertheless, for all our results standard configuration is sufficient as well as almost any state of the art prover or model generator. However, the presented form of results was obtained by experimenting with software configuration and some configurations are better suited for direct extraction of proofs in Hilbert style calculus.

4. Provability of axiom (A2)

We present a proof of axiom (A2) separetely for MTL^- and BL^-. Both proofs are obtained by proving weakening formula $\varphi \to (\psi \to \varphi)$ which immediately gives a proof of axiom (A2). We note that the original prover proofs were slightly adapted.

4.1. MTL^-

First, we present proof for MTL^-which is shorter. It may look surprising, because MTL^- is weaker than BL^-. However, for the proof of axiom (A2), axioms (A4a)–(A4c) are evidently better suited than axiom (A4).

Lemma 4.1 *The following formulae are provable in MTL^-:*

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

(b) $(((\varphi \to \psi) \to \varphi) \to \chi) \to (\varphi \to \chi)$,

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

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

Proof:

a)

1: $((\varphi \,\&\, (\varphi \to \psi)) \to (\varphi \wedge \psi)) \to (((\varphi \wedge \psi) \to \chi) \to ((\varphi \,\&\, (\varphi \to \psi)) \to \chi))$	(A1)
2: $((\varphi \wedge \psi) \to \chi) \to ((\varphi \,\&\, (\varphi \to \psi)) \to \chi)$	by (A4a), 1
3: $(\varphi \,\&\, (\varphi \to \psi)) \to \varphi$	by (A4b), 2

b)

4: $\varphi \to ((\varphi \to \psi) \to \varphi)$	by (a), (A5b)
5: $(((\varphi \to \psi) \to \varphi) \to \chi) \to (\varphi \to \chi)$	by 4, (A1)

c)

6: $((\varphi \to ((\varphi \to \psi) \to \varphi)) \to \varphi) \to (\varphi \to ((\varphi \to \psi) \to \varphi))$	by 4, 4
7: $\varphi \to (\varphi \to ((\varphi \to \psi) \to \varphi))$	by 6, (b)
8: $((\varphi \to ((\varphi \to \psi) \to \varphi)) \to \chi) \to (\varphi \to \chi)$	by (7), (A1)
9: $\varphi \to (\varphi \to \varphi)$	by 8, (b)

[3] `http://www.eprover.org`
[4] `http://www.cs.chalmers.se/~koen/folkung/`

d)

10:	$((\varphi \to \varphi) \to \psi) \to (\varphi \to \psi)$	by (c), (A1)
11:	$((\varphi \to (\psi \to (\varphi \,\&\, \psi))) \to \varphi) \to (((\varphi \,\&\, \psi) \to (\varphi \,\&\, \psi)) \to \varphi)$	by (A5b), (A1)
12:	$\varphi \to (((\varphi \,\&\, \psi) \to (\varphi \,\&\, \psi)) \to \varphi)$	by 11, (b)
13:	$(\varphi \to (\varphi \to \varphi)) \to ((((\varphi \to (\varphi \to \varphi)) \,\&\, \psi) \to ((\varphi \to (\varphi \to \varphi)) \,\&\, \psi)) \to (\varphi \to (\varphi \to \varphi)))$	12
14:	$(((\varphi \to (\varphi \to \varphi)) \,\&\, \psi) \to ((\varphi \to (\varphi \to \varphi)) \,\&\, \psi)) \to (\varphi \to (\varphi \to \varphi))$	by (c),13
15:	$((\varphi \to (\varphi \to \varphi)) \,\&\, \psi) \to (\varphi \to (\varphi \to \varphi))$	by 14, 10
16:	$(\varphi \to (\varphi \to \varphi)) \to (\psi \to (\varphi \to (\varphi \to \varphi)))$	by 15, (A5b)
17:	$\psi \to (\varphi \to (\varphi \to \varphi))$	by (c), 16
18:	$((\varphi \to (\varphi \to \varphi)) \to \varphi) \to (\psi \to \varphi)$	by 17, (A1)
19:	$\varphi \to (\psi \to \varphi)$	by 18, (b)

■

Now by application of (A5a) we immediately obtain

Corollary 4.2 *Axiom* (A2) *is derivable in MTL*$^-$.

Let us note that we do not use axioms (A4c), (A6) and (A7). On the contrary, all other axioms are necessary, which can be demonstrated by Section 5 methods.

Corollary 4.3 (see Cintula [1]) *Axiom* (A3) *is derivable in MTL*$^-$.

It is worth pointing out that axiom (A3) can be proved by similar technique used to prove axiom (A2).

Proof:

a)

1:	$((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (((\varphi \to \varphi) \,\&\, \varphi) \to \varphi)$	(A5a)
2:	$(((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (((\varphi \to \varphi) \,\&\, \varphi) \to \varphi)) \to ((((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (((\varphi \to \varphi) \,\&\, \varphi) \to \varphi)) \to (((\varphi \to \varphi) \,\&\, \varphi) \to \varphi))$	(A6)
3:	$(((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (((\varphi \to \varphi) \,\&\, \varphi) \to \varphi)) \to (((\varphi \to \varphi) \,\&\, \varphi) \to \varphi)$	by 1, 2
4:	$((\varphi \to \varphi) \,\&\, \varphi) \to \varphi$	by 1, 3
5:	$(((\varphi \to \varphi) \,\&\, \varphi) \to \varphi) \to ((\varphi \to \varphi) \to (\varphi \to \varphi))$	by (A5b)
6:	$(\varphi \to \varphi) \to (\varphi \to \varphi)$	by 4, 5

4.2. BL$^-$

We are going to prove a similar lemma for logic BL$^-$. Let us note that we will use axiom (A6) and axiom (A7), which are not necessary, but shorten the proof, whereas all other axioms are necessary.

Lemma 4.4 *The following formulae are provable in BL*$^-$*:*

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

(b) $(\varphi \,\&\, (\varphi \to \overline{0})) \to \psi$,

(c) $(\varphi \,\&\, \psi) \to \psi$,

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

7: $((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (\varphi \to \varphi))$ (A6)

8: $((\varphi \to \varphi) \to (\varphi \to \varphi)) \to (\varphi \to \varphi)$ by 6, 7

9: $\varphi \to \varphi$ by 6, 8

b)

10: $((\varphi \to \psi) \to (\varphi \to \psi)) \to (((\varphi \to \psi) \,\&\, \varphi) \to \psi)$ (A5a)

11: $((\varphi \to \psi) \,\&\, \varphi) \to \psi$ by (a), 10

12: $(((\varphi \to \psi) \,\&\, \varphi) \to \psi) \to ((\psi \to \chi) \to (((\varphi \to \psi) \,\&\, \varphi) \to \chi))$ (A1)

13: $(\psi \to \chi) \to (((\varphi \to \psi) \,\&\, \varphi) \to \chi)$ by 11, 12

14: $(\overline{0} \to (\psi \to \varphi)) \to ((\overline{0} \,\&\, \psi) \to \varphi)$ (A5a)

15: $(\overline{0} \,\&\, \psi) \to \varphi$ by (A7), 14

16: $((\overline{0} \,\&\, \psi) \to \varphi) \to (((\chi \to (\overline{0} \,\&\, \psi)) \,\&\, \chi) \to \varphi)$ 13

17: $((\chi \to (\overline{0} \,\&\, \psi)) \,\&\, \chi) \to \varphi$ by 15, 16

18: $(((\varphi \to (\overline{0} \,\&\, \psi)) \,\&\, \varphi) \to \chi) \to ((\varphi \to (\overline{0} \,\&\, \psi)) \to (\varphi \to \chi))$ (A5b)

19: $(\varphi \to (\overline{0} \,\&\, \psi)) \to (\varphi \to \chi)$ by 17, 18

20: $(\varphi \,\&\, (\varphi \to \overline{0})) \to (\overline{0} \,\&\, (\overline{0} \to \varphi))$ (A4)

21: $((\varphi \,\&\, (\varphi \to \overline{0})) \to (\overline{0} \,\&\, (\overline{0} \to \varphi))) \to ((\varphi \,\&\, (\varphi \to \overline{0})) \to \psi)$ 19

22: $(\varphi \,\&\, (\varphi \to \overline{0}) \to \psi$ by 20, 21

c)

23: $\varphi \to ((\varphi \to \overline{0}) \to \psi)$ by (b), (A5b)

24: $(\varphi \to ((\varphi \to \overline{0}) \to \psi)) \to ((((\varphi \to \overline{0}) \to \psi) \to \chi) \to (\varphi \to \chi))$ (A1)

25: $(((\varphi \to \overline{0}) \to \psi) \to \chi) \to (\varphi \to \chi)$ by 23, 24

26: $(\varphi \to \psi) \to (\overline{0} \to \psi)$ by (A7), (A1)

27: $(((\varphi \to \overline{0}) \to \psi) \to (\overline{0} \to \psi)) \to (\varphi \to (\overline{0} \to \psi))$ 25

28: $\varphi \to (\overline{0} \to \psi)$ by 26, 27

29: $\varphi \to (\psi \to (\varphi \,\&\, \psi))$ by (a), (A5b)

30: $(\varphi \to \varphi) \to (\psi \to ((\varphi \to \varphi) \,\&\, \psi))$ 29

31: $\psi \to ((\varphi \to \varphi) \,\&\, \psi)$ by (a), 30

32: $(\varphi \to (\overline{0} \to \psi)) \to ((\chi \to \chi) \,\&\, (\varphi \to (\overline{0} \to \psi)))$ 31

33: $(\chi \to \chi) \,\&\, (\varphi \to (\overline{0} \to \psi))$ by 28, 32

34: $((\varphi \to \varphi) \,\&\, ((\varphi \to \varphi) \to (\overline{0} \to \psi))) \to ((\overline{0} \to \psi) \,\&\, ((\overline{0} \to \psi) \to (\varphi \to \varphi)))$ (A4)

35: $(\overline{0} \to \psi) \,\&\, ((\overline{0} \to \psi) \to (\varphi \to \varphi))$ by 33, 34

36: $((\varphi \,\&\, \psi) \to (\overline{0} \to \chi)) \to (\varphi \to (\psi \to (\overline{0} \to \chi)))$ (A5b)

37: $\varphi \to (\psi \to (\overline{0} \to \chi))$ by 28, 36

38: $(\varphi \to (\psi \to (\overline{0} \to \chi))) \to (((\psi \to (\overline{0} \to \chi)) \to \xi) \to (\varphi \to \xi))$ (A1)

39: $((\psi \to (\overline{0} \to \chi)) \to \xi) \to (\varphi \to \xi)$ by 37, 38

40: $(\varphi \to (\overline{0} \to \psi)) \to (((\overline{0} \to \psi) \to \chi) \to (\varphi \to \chi))$ (A1)

41: $((\varphi \to (\overline{0} \to \psi)) \to (((\overline{0} \to \psi) \to \chi) \to (\varphi \to \chi))) \to (\xi \to (((\overline{0} \to \psi) \to \chi) \to (\varphi \to \chi)))$ 39

42: $\xi \to (((\overline{0} \to \psi) \to \chi) \to (\varphi \to \chi))$ by 40, 41

43: $(\varphi \to (((\overline{0} \to \psi) \to \chi) \to (\xi \to \chi))) \to ((\varphi \,\&\, ((\overline{0} \to \psi) \to \chi)) \to (\xi \to \chi))$ (A5a)

44: $(\varphi \,\&\, ((\overline{0} \to \psi) \to \chi)) \to (\xi \to \chi)$ by 42, 43

45: $((\overline{0} \to \varphi) \,\&\, ((\overline{0} \to \varphi) \to (\psi \to \psi))) \to (\varphi \to (\psi \to \psi))$ 44

46: $\varphi \to (\psi \to \psi)$ by 35, 45

47: $(\varphi \,\&\, \psi) \to \psi$ by 46, (A5a)

d)

48: $((\psi \,\&\, \chi) \to \chi) \to (((\varphi \to (\psi \,\&\, \chi)) \,\&\, \varphi) \to \chi)$ 13

49: $((\varphi \to (\psi \,\&\, \chi)) \,\&\, \varphi) \to \chi$ by (c), 48

50: $(\varphi \to (\psi \,\&\, \chi)) \to (\varphi \to \chi)$ by 49, (A5b)

51: $((\varphi \,\&\, (\varphi \to \psi)) \to (\psi \,\&\, (\psi \to \varphi))) \to ((\varphi \,\&\, (\varphi \to \psi)) \to (\psi \to \varphi))$ 50

52: $(\varphi \,\&\, (\varphi \to \psi)) \to (\psi \to \varphi)$ by (A4), 51

53: $\varphi \to ((\varphi \to \psi) \to (\psi \to \varphi))$ by 52, (A5b)

54: $(((\varphi \to \psi) \to (\psi \to \varphi)) \to \chi) \to (\varphi \to \chi)$ by 53, (A1)

55: $((\psi \to \varphi) \to (\psi \to \varphi)) \to (((\varphi \to \psi) \to (\psi \to \varphi)) \to (\psi \to \varphi))$ (A6)

56: $((\varphi \to \psi) \to (\psi \to \varphi)) \to (\psi \to \varphi)$ by (a), 55

57: $\varphi \to (\psi \to \varphi)$ by 56, 54

■

Now again by application of (A5a) we immediately obtain

Corollary 4.5 *Axiom* (A2) *is derivable in BL*$^-$.

Corollary 4.6 (see Cintula [1]) *Axiom* (A3) *is derivable in BL*$^-$.

It is worth pointing out that axiom (A3) can be again proved by similar technique used to prove axiom (A2).

5. The independence of axioms

We know that axioms (A2) and (A3) are redundant in BL and MTL. Is any other axiom redundant in BL or MTL? We answer this question negatively for every remaining axiom by presenting a model and a valuation which make the axiom false, but all other axioms including (A2) and (A3) and modus ponens rule are true in the model. It means that none of the axioms but (A2) and (A3) is redundant in original systems BL and MTL. We obtain immediately that all axioms in BL$^-$ and MTL$^-$ are independent.

All models are finitely valued structures with elements labeled by natural numbers, presented in form of truth tables. Let us note that in all models except for (A7) we interpret constant $\overline{0}$ as the minimal element 0 and truth as the maximal value in a model, e.g. in a four member model it has value 3.

The important point to note is that checking falsity of axiom in a given model under a given valuation is an easy task. On the other hand, to show that all other axioms are true in the model, exhausting checking is sometimes needed. Fortunately, for computer it is an easy task. We naturally do not present these proofs.

For shortening the presentation we present models for BL and MTL at once. Only models for logic specific axioms (A4) and (A4a)–(A4c) are presented separately. Moreover, we prefer the same definition for multiplicative and additive conjunction.

We start by a group of axioms common to BL and MTL.

5.1. Axiom (A1)

For showing the independence of axiom (A1) we need a model in which implication is not transitive. We present such a model which falsifies axiom (A1) for valuation $\varphi = 1, \psi = 0$ and $\chi = 2$.

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

$\rightarrow$	0	1	2	3
0	3	3	3	3
1	3	3	1	3
2	3	3	3	3
3	1	1	1	3

Table 1: Truth tables for (A1)

5.2. Axiom (A5a)

First of the residuation axioms (A5a) fails evidently for $\varphi = 2, \psi = 1$ and $\chi = 0$. Both conjunctions are defined separately.

$\&$	0	1	2	3
0	0	0	0	0
1	0	0	2	2
2	0	2	0	2
3	0	2	2	3

$\wedge$	0	1	2	3
0	0	0	0	0
1	0	1	1	1
2	0	1	1	1
3	0	1	1	3

$\rightarrow$	0	1	2	3
0	3	3	3	3
1	1	3	3	3
2	2	3	3	3
3	0	2	1	3

Table 2: Truth tables for (A5a)

5.3. Axiom (A5b)

To demonstrate the independence of axiom (A5b), much easier model than for axiom (A5a) is needed. A two valued model with classical implication and both conjunctions false for all values is sufficient. Axiom (A5b) fails for $\varphi = 1, \psi = 1$ and $\chi = 0$.

$\&, \wedge$	0	1
0	0	0
1	0	0

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

Table 3: Truth tables for (A5b)

5.4. Axiom (A6)

The independence of axiom (A6) can be easily shown by an algebraic arguments. It represents prelinearity and logics without prelinearity have been already studied. Moreover, MTL without axiom (A6) represents Höhle Monoidal Logic ML. Nevertheless, we present our standard semantic argument. Axiom (A6) fails for φ, ψ and χ represented by $1, 2$ and 3.

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

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

Table 4: Truth tables for (A6)

5.5. Axiom (A7)

It is evident that axiom (A7) is independent of other axioms, because of new symbol $\overline{0}$. For demonstration it is enough to interpret $\overline{0}$ as truth and all connectives classically. In such model, axiom (A7) easily fails and all other axioms are evidently true.

$\&, \wedge$	0	1
0	0	0
1	0	1

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

Table 5: Truth tables for (A7)

Now we present BL and MTL specific cases.

5.6. Axiom (A4)

If we take $\varphi \wedge \psi$ as an abbreviation for $\varphi \,\&\, (\varphi \to \psi)$, axiom (A4) represents commutativity of additive conjunction in BL. For $\varphi = 1$ and $\psi = 2$, additive conjunction is not commutative.

$\&$	0	1	2	3
0	0	0	0	0
1	0	0	0	1
2	0	0	0	2
3	0	1	2	3

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

Table 6: Truth tables for (A4)

We show the independence of axioms (A4a)–(A4c) by small models, in which axioms (A1)–(A3) and (A5a)–(A7) are evidently true, because of $\&$ and $\to$ definition. Therefore to complete the proof it is sufficient to show the (in)validity of axioms (A4a)–(A4c) in the corresponding truth tables only.

5.7. Axiom (A4a)

Axiom (A4a) fails for $\varphi = 1$ and $\psi = 1$, but axioms (A4b) and (A4c) are evidently true.

$\&$	0	1
0	0	0
1	0	1

$\wedge$	0	1
0	0	0
1	0	0

$\to$	0	1
0	1	1
1	0	1

Table 7: Truth tables for (A4a)

5.8. Axiom (A4b)

Axiom (A4b) fails for $\varphi = 0$ and $\psi = 1$, but axioms (A4a) and (A4c) are evidently true.

$\&$	0	1
0	0	0
1	0	1

$\wedge$	0	1
0	0	1
1	1	1

$\to$	0	1
0	1	1
1	0	1

Table 8: Truth tables for (A4b)

5.9. Axiom (A4c)

Axiom (A4c) fails for $\varphi = 1$ and $\psi = 0$, but axioms (A4a) and (A4b) are evidently true.

$\&$	0	1
0	0	0
1	0	1

$\wedge$	0	1
0	0	0
1	1	1

$\to$	0	1
0	1	1
1	0	1

Table 9: Truth tables for (A4c)

Corollary 5.1 *All axioms but* (A2) *and* (A3) *are independent of each other in BL.*

Corollary 5.2 *All axioms but* (A2) *and* (A3) *are independent of each other in MTL.*

It is worth pointing out that the independence of axioms could be presented also by studying some known algebraic structures, which has several indisputable theoretical advantages. On the other hand, our approach seems to be easier for presentation.

6. Summary and conclusion

We presented the complete solution of dependence and independence of axioms in prominent fuzzy propositional logics BL and MTL by using simple technique from automated theorem proving. Also other similar problems can be solved using these methods and state of the art theorem provers and model generators.

Nevertheless, our approach has several drawbacks. First, abilities of current theorem provers are limited and in some situations even short proofs are inaccessible for them without special settings. Second, abilities of automated model generators are also very limited, e.g. infinite models are highly problematic.

References

[1] P. Cintula, “Short note: on the redundancy of axiom (A3) in BL and MTL,” *Soft Computing*, vol. 9, no. 12, pp. 942–942, 2005.

[2] 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.

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

[4] S. Lehmke, “Fun with automated proof search in basic propositional fuzzy logic,” in *Abstracts of the Seventh International Conference FSTA 2004* (P. E. Klement, R. Mesiar, E. Drobná, and F. Chovanec, eds.), (Liptovský Mikuláš), pp. 78–80, 2004.

[5] J. A. Robinson, “A machine-oriented logic based on the resolution principle,” *Journal of the ACM*, vol. 12, no. 1, pp. 23–41, 1965.

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

Změkčování rozhodovacích stromů maximalizací plochy pod částí ROC křivky

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

V návaznosti na plochu pod ROC křivkou jakožto obvyklou míru kvality klasifikátoru zavádíme plochu pod počáteční částí ROC křivky, která je mírou kvality klasifikátoru zaměřeného na dosažení nízké chybovosti na negativních (background) případech. Tato míra je použita jako cílová funkce při změkčování rozhodovacích stromů pomocí optimalizace. Pro optimalizaci je použit algoritmus Nelder-Mead. Experimenty na datech „Magic Telescope" ukazují účinnost této metody.

1. Úvod

Změkčování hran v rozhodovacích stromech umožňuje zlepšení klasifikátoru při zachování většiny dobrých vlastností rozhodovacích stromů. Změkčené stromy oproti klasickým mohou dosahovat lepšího poměru správné / chybné klasifikace a dalším přínosem je spojitost výstupu klasifikátoru. Zachována zůstává snadná interpretovatelnost modelu a přímočará převoditelnost na systém pravidel (v případě změkčeného stromu půjde o fuzzy-pravidla). Nevýhodou je zvětšení paměťové náročnosti modelu a hlavně časové složitosti jak učení, tak klasifikace.

Zde se budeme zabývat změkčováním jakožto postprocessingem stromů získaných standardní metodou CART [2]. Základní tvar změkčení je stejný, jako je v metodě C4.5 [6], ale liší se způsob určení (učení) parametrů, tj. hranic intervalů změkčení. Zatímco C4.5 určuje parametry změkčení pomocí směrodatné odchylky klasifikační chyby nezměkčeného stromu bez ohledu na to, jaký efekt má změkčení na chování klasifikátoru, my budeme hledat změkčení pomocí optimalizace výsledků změkčeného stromu.

Při změkčování pomocí optimalizace je pro kvalitu výsledného klasifikátoru i pro rychlost optimalizace zásadní volba cílové funkce. Použití relativního počtu chybných klasifikací se ukázalo jako nevhodné, protože to je funkce po částech konstantní a má velké množství lokálních minim. Varianty založené na sumaci transformované diference spojitého výstupu klasifikátoru a očekávané klasifikace pomohou získat spojitou funkci, ale stále trpí problémem lokálních minim a pro jejich optimalizaci byla používána metoda založená na simulovaném žíhání, jak bylo popsáno v [3], tento algoritmus je však časově velmi náročný.

V tomto příspěvku ukážeme využití plochy pod počáteční částí ROC křivky jakožto cílové funkce pro optimalizaci změkčení rozhodovacího stromu. Ukazuje se, že pro takovouto optimalizaci je možné použít simplexový algoritmus (Nelder-Mead) [5], což vede k podstatně rychlejšímu učení, než předchozí přístup se simulovaným žíháním.

2. ROC křivka a plocha pod křivkou

ROC křivka (Receiver Operating Characteristic curve) je standardním nástrojem pro analýzu chování klasifikátoru. V této sekci uvádíme především informace podstatné pro další vysvětlení změkčování rozhodovacích stromů. Čerpáme zejména z [4] a další literatury.

Pro klasifikátor, který rozděluje data do dvou tříd (nazývejme je pozitivní a negativní, někdy též signal resp. background), ROC křivka ukazuje vztah relativního počtu správně klasifikovaných pozitivních vzorů a relativního počtu chybně klasifikovaných negativních vzorů (signal acceptance vs. background acceptance) při různě nastavené „citlivosti".

Pokud výstupem klasifikátoru je pro každý datový vzor $\mathbf{x}$ reálné číslo „response" $\mathrm{R}(\mathbf{x})$, přičemž jeho vyšší hodnota reprezentuje vyšší pravděpodobnost, že předložený případ je pozitivní, potom různé nastavení citlivosti odpovídá různým volbám hodnoty prahu, kterým oddělujeme případy, jež podle response považujeme za pozitivní od případů, které zařadíme k negativním.

Plocha pod ROC křivkou (Area Under Curve, AUC) je skalárním vyjádřením kvality klasifikátoru. AUC klasifikátoru, který zařadí všechny vzory správně, je rovna jedné. Čím je hodnota nižší, tím je klasifikátor horší. AUC pro náhodný klasifikátor je $1/2$. Hodnoty v intervalu $\langle 0, 1/2)$ by charakterizovaly klasifikátor horší než náhodný.

Máme-li množinu, jež obsahuje P pozitivních vzorů $\mathbf{x}_1^+, \ldots, \mathbf{x}_P^+$ a Q negativních vzorů $\mathbf{x}_1^-, \ldots, \mathbf{x}_Q^-$ a definujeme-li funkci

$$g(u,v) = \begin{cases} 1 & \text{když } u > v \\ 1/2 & \text{když } u = v \\ 0 & \text{když } u < v \end{cases}$$

potom z této množiny vypočteme

$$\mathrm{AUC} = \frac{1}{PQ} \sum_{i=1}^{P} \sum_{j=1}^{Q} g\left(\mathrm{R}(\mathbf{x}_i^+), \mathrm{R}(\mathbf{x}_j^-)\right)$$

3. Metoda změkčování

Mějme nezměkčený rozhodovací strom, který pro vstupní vzor $\mathbf{x} = (x_1, \ldots, x_m)$ testuje ve vnitřních uzlech $v_j, j = 1, \ldots, s$ podmínky tvaru

$$x_{k_j} \leq c_j \tag{1}$$

V listech jsou uloženy hodnoty response z intervalu $\langle 0, 1\rangle$. Klasifikace tímto stromem probíhá tak, že pro předložený vzor se počínaje kořenem stromu testuje nerovnost (1), je-li splněna, pokračuje se v levém podstromu, jinak v pravém podstromu, dokud není dosaženo listu, který určí výslednou response.

Odpovídající změkčený strom bude mít stejnou strukturu, hodnoty response v listech zůstanou stejné, ale každý vnitřní uzel bude kromě hodnot k_j, c_j z podmínky (1) určovat reálné parametry změkčení $a_j, b_j \geq 0$. Potom definujeme změkčující funci f_j jež lineárně interpoluje body uvedené v tabulce:

t	$-\infty$	$-a_j$	0	b_j	∞
$f_j(t)$	1	1	1/2	0	0

Response změkčeného stromu je definována rekurzívně: v listu stromu je pro libovolný vstupní vzor response daná hodnotou uloženou v tomto listu. Jinak pro strom s kořenem v_j a vzor $\mathbf{x}$ je výsledkem průměr response levého a pravého podstromu vážený hodnotami $r_{j,\mathbf{x}}$ a $(1 - r_{j,\mathbf{x}})$, kde $r_{j,\mathbf{x}} = f_j(x_{k_j} - c_j)$.

Úlohou změkčování je pak určení parametrů $a_j, b_j, j = 1, \ldots, s$, k čemuž používáme optimalizaci funkce založené na tom, jak změkčený strom s danými parametry klasifikuje vzory z trénovací množiny.

V mnohých skutečných klasifikačních úlohách (včetně klasifikace dat „Magic Telescope" použitých v našich experimentech), je podstatné dosažení nízké úrovně background acceptance. Protože background acceptance tvoří horizontální osu ROC křivky, charakterizuje chování klasifikátoru při nízkých hodnotách background acceptance počáteční část ROC křivky. Naše metoda proto používá jako cílovou funkci pro optimalizaci plochu pod nejmenší částí ROC křivky, jež pokrývá celou oblast, kde background acceptance není větší, než zvolená hodnota $0 \leq \Theta \leq 1$. Tuto částečnou AUC označujme AUC_Θ.

Předpokládejme dále bez újmy na obecnosti, že vzory v množině, z níž počítáme AUC_Θ, jsou očíslovány tak, aby

$$\mathrm{R}(\mathbf{x}_1^+) \geq \mathrm{R}(\mathbf{x}_2^+) \geq \cdots \geq \mathrm{R}(\mathbf{x}_P^+)$$
$$\mathrm{R}(\mathbf{x}_1^-) \geq \mathrm{R}(\mathbf{x}_2^-) \geq \cdots \geq \mathrm{R}(\mathbf{x}_Q^-)$$

Označme ϑ nejvyšší hodnotu prahu, při níž je hodnota background acceptance alespoň Θ:

$$\vartheta = \mathrm{R}(\mathbf{x}^-_{\lceil \Theta Q \rceil})$$

Dále počty pozitivních a negativních případů, jejichž response je alespoň ϑ označme:

$$P_\vartheta = \max\left\{i; \quad \mathrm{R}(\mathbf{x}_i^+) \geq \vartheta\right\}$$

$$Q_\vartheta = \max\left\{j; \quad \mathrm{R}(\mathbf{x}_j^-) \geq \vartheta\right\}$$

Potom

$$\mathrm{AUC}_\Theta = \frac{1}{PQ} \sum_{i=1}^{P_\vartheta} \sum_{j=1}^{Q_\vartheta} g\left(\mathrm{R}(\mathbf{x}_i^+), \mathrm{R}(\mathbf{x}_j^-)\right)$$

Tato hodnota je vypočtena lehce modifikovaným algoritmem pro výpočet standardní AUC uvedeným v [4].

Pro optimalizaci cílové funkce je použit simplexový algoritmus pro minimalizaci (Nelder-Mead) [5]. Minimalizuje se $-\mathrm{AUC}_\Theta$ vypočtená z trénovacích dat. Algoritmus vyžaduje, aby ve vstupním prostoru měly všechny dimenze stejnou škálu, tedy aby jednotkový

krok v libovolném směru měl vždy přibližně stejný význam. Použitá škála byla definována následovně: Nejprve celý prostor ve všech směrech omezíme nejzazšími trénovacími vzory, tak získáme základní hyperkvádr. Když v uzlu v_j podmínka (1) rozděluje hyperkvádr vyšší úrovně, který je v proměnné x_{k_j} omezen hodnotami $z_{j,1}, z_{j,2}$, kde $z_{j,1} < c_j < z_{j,2}$, potom za jednotkový krok v parametru a_j resp. b_j považujeme $c_j - z_{j,1}$ resp. $z_{j,2} - c_j$. Zároveň jako iniciální hodnoty parametrů pro změkčování se použijí:

$$a_j^0 = \frac{1}{4}(c_j - z_{j,1}); \quad b_j^0 = \frac{1}{4}(z_{j,2} - c_j)$$

4. Výsledky experimentů

Pro experimenty byla použita data „Magic Telescope"[1], která jsou zkoumána také v [1] a [3]. Trénovací množina obsahovala 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. Strom byl vytvořen metodou CART, velikost stromu je možno řídit nastavením parametrů prořezávání (viz [2]).

Pro změkčení byla použita výše popsaná metoda s parametrem $\Theta = 1/10$, jakožto data pro výpočet částečné AUC byla použita celá trénovací množina. Pro hodnocení získaného klasifikátoru byla použita testovací množina o velikosti 6340 vzorů.

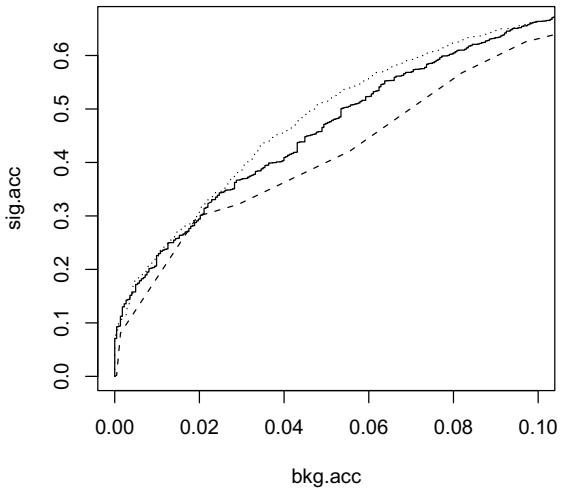
Obrázky 1 a 2 ukazují získané části ROC křivek pro vybrané stromy. Na obrázcích je čárkovaně vyznačena ROC křivka nezměkčeného stromu na testovacích datech; tečkovaná je ROC křivka změkčeného stromu na trénovacích datech, tzn. jedná se o křivku, která figurovala v cílové funkci; plnou čarou je ROC křivka změkčeného stromu na testovacích datech.

Z obrázků je patrné, že změkčený strom je v oblasti nízké úrovné background acceptance lepší klasifikátor, než nezměkčený strom.Takové chování se ukázalo jako typické i na dalších stromech.

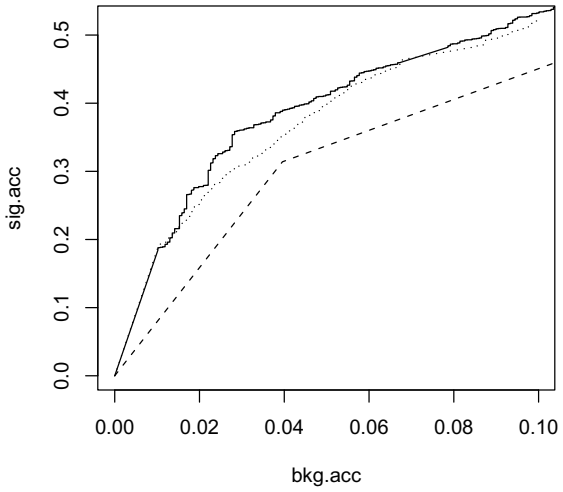
5. Závěr

Plocha pod částí ROC křivky se ukazuje jako vhodná cílová funkce pro změkčování rozhodovacích stromů pomocí optimalizace. Tuto cílovou funkci lze optimalizovat metodou Nelder-Mead, což proti doposud zkoumaným cílovým funkcím optimalizovaným pomocí simulovaného žíhání vede k významnému snížení časové náročnosti změkčování. Dalším přínosem je možnost preferovat nízkou background acceptance klasifikátoru.

V dalším výzkumu se zaměříme na ladění parametrů optimalizačního algoritmu a budeme ještě zkoumat modifikace cílové funkce. Pozornost bude také věnována skutečnosti, že v provedených experimentech byla na menších stromech zkoumaná část ROC křivky vypočtené z testovacích dat lepší, než ROC z trénovacích dat. Tento aspekt je viditelný na obrázku 2 a byl pozorován i na dalších stromech.



Obrázek 1: Části ROC křivek pro strom se 45 vnitřními uzly



Obrázek 2: Části ROC křivek pro strom s 10 vnitřními uzly

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

Literatura

[1] R.K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, 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, C.J. Stone, *Classification and Regression Trees*, Belmont CA: Wadsworth, 1993

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

[6] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo — California, 1993

Verification of Hybrid Systems

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

Supervisor:
ING. STEFAN RATSCHAN, PH.D.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
stefan.ratschan@cs.cas.cz

Field of Study:
Theoretical Computer Science

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 some continuous environment. Usually in such applications there are some states that will be dangerous for the system or its user. Hence, for a hybrid system we define some states as unsafe. Verification is an algorithm that, for a safe hybrid system, proves that no unsafe state will be reached. In our work we improve the method for verification of hybrid systems by constraint propagation based abstraction refinement proposed by Stefan Ratschan and Zhikun She. Hybrid systems often contain variables with linear time evolution, which we call clocks. We introduce hyperplane barriers into the abstraction of hybrid system, which we can compute from linear clock constraints. This will give us more precise information about the hybrid system and saves us computation steps. Later we will extend the method also to non-linear constraints using interval arithmetics.

Induction of User Preferences in Semantic Web

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

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

Field of Study:
Software Engineering

This work was supported by Czech projects 1ET 100300517 and MSM 0021620838.

Abstract

Uncertainty querying of large data can be solved by providing top-k answers according to a user fuzzy ranking/scoring function. Usually different users have different fuzzy scoring function – a user preference model. Main goal of this paper is to assign a user a preference model automatically. To achieve this we decompose user's fuzzy ranking function to ordering of particular attributes and to a combination function. To solve the problem of automatic assignment of user model we design two algorithms, one for learning user preference on particular attribute and second for learning the combination function. Methods were integrated into a Fagin-like top-k querying system with some new heuristics and tested.
These user preference models can be used by an artifical agent, which automatically selects objects that are most suitable for the user and present them to the user. The agent's proposal can be modified by the user, making a feedback for the agent in this way. This feedback is crucial for better representation of user preferences.

The abstract was originally published in paper [5]. Due to the copyright issues, only the abstract is presented here, extended with second paragraph containing newer issues.

References

[1] A. Eckhardt "Návrh agenta řízeného uživatelskými preferencemi", *To appear in Proc. of ITAT 2008*, Hotel Hrebienok, Vysoké Tatry, Slovakia

[2] J. Dědek, A. Eckhardt, P. Vojtáš " Experiments with Czech Linguistic Data and ILP", *To appear in Proc. of 18th conference on Inductive Logic Programming 2008*, Prague, Czech Republic.

[3] A. Eckhardt, T. Horváth, D. Maruščák, R. Novotný, P. Vojtáš " Uncertainty Issues in Automating Process Connecting Web and User", *Proceedings of the Third ISWC Workshop on Uncertainty Reasoning for the Semantic Web 2007*, pp.104-115, Busan, Korea.

[4] A. Eckhardt, T. Horváth and P. Vojtáš " PHASES: A User Profile Learning Approach for Web Search", *In Proc. of Web Intelligence 2007*, pp.780-783, Silicon Valley, USA.

[5] A. Eckhardt, T. Horváth and P. Vojtáš " Learning different user profile annotated rules for fuzzy preference top-k querying", *In Proc. of SUM 2007*, pp. 116-130, Washington DC, USA.

[6] A. Eckhardt, J. Pokorný and P. Vojtáš " Integrating user and group preferences for top-k search from distributed web resources", *In Proc. of DEXA 2007*, pp. 317-322 Regensburg, Germany.

[7] A. Eckhardt, J. Pokorný and P. Vojtáš " A system recommending top-k objects for multiple users preferences", *In Proc. of Fuzz-IEEE 2007*, pp. 1101-1106, London, England.

[8] A. Eckhardt " Inductive models of user preferences for semantic web", *In Proc. of Dateso 2007*, pp. 103-114, Desná, Czech Republic.

[9] A. Eckhardt, P. Vojtáš " Uživatelské preference při hledání ve webovských zdrojích", *In Proc. of Znalosti 2007*, pp. 179-190, Ostrava, Czech Republic.

Logistic Regression and Classification and Regression Trees (CART) in Acute Myocardial Infarction Data Modeling

Post-Graduate Student:
MGR. VÁCLAV FALTUS, MSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
faltus@euromise.cz

Supervisor:
PROF. RNDR. JANA ZVÁROVÁ, DRSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
zvarova@euromise.cz

Field of Study:
Biomedical Informatics

The work was suported by the grant 1M06014 of the Ministry of Education of the Czech Republic.

Within the last 15 years there has been increasing interest in the use of the classification and regression tree (CART) analysis as competitive means to the logistic regression. Especially when modeling biomedical data, a common goal is to develop a reliable clinical decision rule, which can be used later to classify patients into clinically relevant categories. In these situations, the logistic regression does not always prove to be the best choice. Instead we use CART, which is the binary recursive partitioning method used to construct classification and regression trees. In such trees the classification of each patient is simpler and more evident to clinicians and medical doctors. Furthermore, the advantage of tree-base methods is that it does not require that one parametrically specify the nature of the relationship between the predictor variables and the outcome. The assumption of linearity made in generalized linear models is also relaxed.

Logistic regression works well in modeling categorical data in various fields. However, the interpretation of its results is not always straightforward and logistic regression equations are sometimes difficult to use in clinical practice, especially in situations in which the outcome variable has more than two levels or when there is too many predictor variables with unknown interactions etc. In these situations one usually uses stepwise procedures such as forward or backward selection or its combinations to obtain a feasible model. However, not depending on the choice of testing criteria (F-test, AIC, BIC, Mallows' Cp) this is not sometimes a good choice because it leaves or omits all interactions in the model in situations where we expect some significant onces. Another problem is with interpretation of sequentially used p-values and biased tests.

We use both methods, CART and logistic regression, in acute myocardial infarction (AMI) in-hospital mortality modelling. Our data were available on a sample of patients with acute myocardial infarction consecutively admitted to six municipal hospitals in the Czech Republic during the years 2003–2007. Data were obtained by yearly retrospective chart reviews. The registry hospitals were: Čáslav, Kutná Hora and Znojmo in years 2003–2007, Jindřichův Hradec and Písek in 2004, Chrudim in years 2005–2007. All of them are non-PCI hospitals from geographically different rural regions of the Czech Republic and collaborate with different PCI centers.

There was 3185 cases of patients who in the time period 2003–2007 presented with AMI to one of the registry hospitals, but since it was not possible to identify patients who were present more than once during the five years period (with more than one AMI), we omit all the cases in which it is not possible to uniquely discriminate one patient from another. For this task we use the categorical variables such as date of birth, date of MI, in-hospital mortality, previous MI, gender and local hospital. This process yielded 312 AMI cases which is 9.8 % of all the data. This leaves 2873 observed patients with AMI. In our data there is also more than one hundred cathegorical and continuous predictor variables with various mechanisms and amounts of missing data. We discuss the impacts of missing data on models obtained from both conventional logistic regression and cart data modelling.

The purpose of this study is to compare the predictive ability of logistic regression with that of regression tree methods in our sample and to discuss the impact of missing data on models obtained from logistic regression and CART. Great deal of effort is also dedicated to the interpretation of the results in clinical practice. We use repeated split sample validation using our dataset of patients hospitalized with acute myocardial infarction.

Metainformace ke zdrojovému kódu jazyka Python

doktorand:
ING. FRANTIŠEK JAHODA
Katedra matematiky
FJFI ČVUT
Trojanova 13
120 00 Praha 2
jahoda@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:
Struktura zdrojového kódu

Děkuji vedoucímu své diplomové práce Ing. Zdeňku Čulíkovi za odborné vedení a svému školiteli Ing. Júliu Štullerovi CSc. za trpělivou pomoc s formální stránkou tohoto článku.

Abstrakt

Běžně bývá zdrojový kód programovacích jazyků uložen v textových souborech, které jsou snadno editovatelné, ale nereprezentují přímo strukturu zdrojového kódu. V tomto článku zkoumám koncept ukládání a práce se zdrojovým kódem (jazyka Python) ve formě stromu příkazů hierarchicky uspořádaných podle syntaxe, což přináší výhody v reprezentaci dodatečných informací popisující zdrojový kód a při práci s nimi.

Dále v článku popisuji knihovnu pro práci se zdrojovým kódem v tomto formátu, kterou jsem vytvořil. Tato knihovna umožňuje efektivně zpracovávat zdrojový kód v navržené reprezentaci (importovat existující zdrojový kód do stromu příkazů, načítat, ukládat, upravovat a vizualizovat strom příkazů).

1. Úvod

Nástroje pro práci se zdrojovým kódem jako: *automatické doplňování*, *navigace v kódu*, různé *generátory kódu* nebo *nástroje pro refaktorizaci*[1] potřebují znát syntaktickou strukturu zdrojového kódu pro svoji práci. Jsou situace, kdy je potřeba některé informace vztažené ke zdrojovému kódu uchovávat i mezi jednotlivými kroky editace.

Příkladem může být informace o tom, kdo editoval naposled danou část kódu, která by se měla zachovat při editaci jiné části kódu.

Dalším příkladem mohou být informace získané od uživatele pro provedení refaktorizace, které by bylo vhodné zachovat, pokud se nemění jejich platnost.

Je tedy vhodné takovéto informace uchovávat a upravovat společně se zdrojovým kódem. Zdrojový kód by také měl být vhodně strukturován, aby se změny v něm daly izolovat na konkrétní místo a nebyla tak ovlivněna platnost výše uvedených informací v jiných částech.

K uchování těchto dodatečných informací lze využít komentáře ve zdrojovém kódu. Komentáře ale nejsou navrženy k uchování většího množství strukturovaných dat. Nezachycují strukturu dat a při editaci zdrojového kódu mohou uživateli překážet.

Ve své diplomové práci jsem proto navrhl strukturu, která uchovává zdrojový kód v hierarchické formě stromu příkazů uspořádaných podle syntaxe jazyka. Tato struktura zachycuje základ syntaktického stromu, ale bez jeho detailnější struktury. Umožňuje ke zdrojovému kódu uchovávat dodatečné informace a zároveň umožňuje dotazovat a editovat kód přímo v této struktuře.

Praktickou realizací výše uvedeného návrhu je knihovna, která umí zpracovávat zdrojový kód přímo v navržené struktuře. Existující kód umí do dané struktury převést a v ní ho upravovat a uchovávat. Knihovna také umí kód v dané struktuře zobrazit. Postupy použité při návrhu struktury a při implementaci pro jazyk Python se dají aplikovat na jakýkoliv jiný strukturovaný programovací jazyk.

2. Strom příkazů

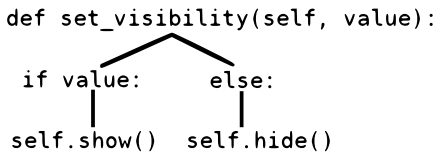
Zdrojový kód je možné reprezentovat za pomoci stromu složeného z příkazů (viz. obr. 1). **Příkazem** chápu *terminály* vzniklé přepisem *neterminálu* `simple_stmt` (např. *přiřazení*, *volání funkce*, *metody*, *komentář*).

[1] Refaktorizace představuje malé změny programu neměnící jeho funkčnost, které mají za cíl zlepšit jeho přehlednost a rozšiřitelnost. Tyto změny jako *přejmenování proměnné*, *přesun metody* do jiné třídy, atd. lze mnohdy provádět automaticky. V dynamicky typovaných jazycích (jako je Python) je to však obtížnější a některé informace nutné pro provedení refaktorizace lze získat pouze od uživatele.

Blokovým příkazem rozumím řádku s *definicí metody*, *třídy*, *podmínky*, *cyklu*. Blokové příkazy mohou obsahovat další příkazy a blokové příkazy.

Příkazy jsou složené z logických tokenů. **Logický token** je posloupnost stejného typu tokenů[2] jazyka Python. Rozlišuji tyto základní typy logických tokenů: *komentář*, *identifikátor*, *klíčové slovo* a *nerozlišené tokeny*. Rozdělení příkazů na logické tokeny umožňuje snadněji dotazovat těla příkazů.

```
1: def set_visibility(self, value):
2:     if value:
3:         self.show()
4:     else:
5:         self.hide()
```



Obrázek 1: Příklad na strom příkazů

3. Scénáře použití stromu příkazů

3.1. Dotazování kódu

Zdrojový kód ve stromu příkazů lze lépe dotazovat než textový soubor a zároveň je snadněji editovatelný než syntaktický strom (který však zachycuje více podrobností):

- Z komentářů a dokumentačních řetězců[3] obsažených ve stromu příkazů lze *generovat dokumentaci* ke zdrojovému kódu.
- Strom příkazů lze rozšířit a umožnit v něm uchování informací, jaké *identifikátory* jsou v konkrétních částech kódu používány, jakého jsou *typu*, zda jsou *lokální* nebo *globální*, atd.
- Pokud jsme schopni určit a do kódu poznamenat vazby mezi identifikátory, lze pak implementovat vlastnosti jako *automatické doplňování* zdrojového kódu (nabídku kódu k doplnění na základě místa, kam kód vkládáme) nebo nástroje pro zrychlení *navigace v kódu* (skok na další použití nebo definici identifikátoru).

3.2. Transformace kódu

- Kód (se všemi rozšiřujícími informacemi) lze měnit *transformacemi* ve stromu příkazů, a proto lze např. provádět *refaktorizace* přímo v navržené struktuře.
- Pokud převádíme kód ze stromu příkazů do textové podoby, lze ho *filtrovat*, a to i na základě syntaxe (např. lze jednoduše smazat dokumentační řetězce). Při výstupu lze kód doplňovat automaticky vygenerovaným kódem.
- Pomocí transformací lze vynutit *jednotné formátování* zdrojového kódu. Takže např. identifikátory mohou mít specifický tvar. Do jazyka je možné zavést rozšiřující *vlastní syntaktické konstrukce* a pomocí transformací je při vytváření výstupu převést na příkazy jazyka Python.
- Lze registrovat *změny mezi jednotlivými verzemi* souboru na základě pozice v syntaktickém stromu (a ne pouze na základě čísla řádky). Tímto způsobem lze vytvořit nástroje *diff* a *patch*.
  Program *diff* vytvoří na základě dvou verzí souboru tzv. *patch soubor*, který obsahuje rozdíly mezi verzemi.
  Program *patch* umí aplikovat tento soubor na starší verzi původního souboru a získat tak novější verzi. Pokud *patch soubor* reprezentuje místa změn na základě syntaxe, je možné jej použít i na odlišnou verzi původního souboru, pokud zůstala zachována přibližně stejná syntaktická struktura tohoto souboru.

3.3. Editace kódu

- Zdrojový kód reprezentovaný stromem příkazů lze spolu s rozšiřujícími informacemi *vytvářet*, *editovat* a *přímo dotazovat* (bez nutnosti vytváření nezávislých pomocných struktur, které by pak bylo třeba synchronizovat s měněným zdrojovým kódem).
- Ke zdrojovému kódu lze připojovat libovolné další *rozšiřující informace*. Například informace z programů pro práci se zdrojovým kódem, jako je *debugger*, *profiller* a *kontrola chyb*. Tyto informace pak lze spolu s kódem *vizualizovat*.
- Při vytváření kódu v této struktuře je možné sledovat *editační změny* (tedy kdo a kdy změnil konkrétní příkaz) a získat tak konkrétnější

[2] nejmenší jednotka syntaktické analýzy, např. klíčové slovo, operátor nebo číslo
[3] řetězec nacházející se hned za definicí třídy nebo metody, který slouží k jejímu popisu

informace, než jsou dostupné ze systému správy verzí.

4. Struktura vytvořené knihovny

Knihovna pro zpracování kódu ve formě stromu příkazů je rozdělena do několika vrstev, přičemž spodní vrstvy jsou nezávislé na vyšších a dají se tedy bez nich použít.

4.1. Objektová struktura

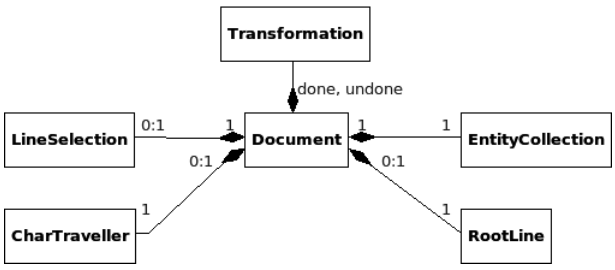
Základní vrstvou knihovny jsou objekty reprezentující *příkazy jazyka* (Python) a *rozšiřující informace*.
Z nich vytvořenou stromovou strukturu lze uložit do *souboru typu XML* a také z něj nahrát. *Strom příkazů* se skládá z objektů pro reprezentaci *blokového příkazu*, *příkazu*, *logických tokenů* a *rozšiřujících vlastností*. Logika pro procházení stromu příkazů je oddělena od objektů tvořících tento stromu.

4.2. Syntaktický analyzátor

Jednou z nejdůležitějších částí knihovny je syntaktický analyzátor (parser) [3] zdrojového kódu jazyka Python. Jedná se o parser vybudovaný nad standardním pythonským tokenizérem.
Výhodou našeho parseru je, že narozdíl od standardního pythonského parseru *zachovává formátování* zdrojového kódu spolu s *komentáři* a pamatuje si *pozice příkazů* v souboru.
Analyzátor rozpoznává pouze základní strukturu syntaktického stromu (od kořene po *nonterminál* `simple_stmt`). Pokud někde vznikne chyba v důsledku špatné syntaxe vstupního textu, pokusí se z ní analyzátor sám zotavit. Ve většině případů pomocí ignorování nečekaných tokenů. Analyzátor lze tedy použít i pro vkládání nového textu do stromu příkazů.

4.3. Vrstva transformací

Nad vrstvou základních objektů se nachází objekt `Document` (viz. obr. 2), který umožňuje na stromu příkazů provádět různé transformace.



Obrázek 2: Objekt Document

Mezi základní transformace patří *odstranění příkazu*, *přidání příkazu* a *přesun příkazu*. Ostatní transformace jsou tvořeny pomocí těchto základních. Provedení základní transformace může způsobit vznik *události*, která se šíří ve stromu příkazů směrem k listům a umožňuje tak aktualizovat údaje, které se provedením transformace změnily.
Rozlišuji události `on_set` a `on_clear`. První se volá při změně nebo vytvoření nějakého příkazu a druhá naopak při odstranění příkazu. Všechny transformace jsou implementovány pomocí návrhového vzoru příkaz (Command pattern [2]). Transformace lze vracet a skládat dohromady. Složené transformace pak lze vrátit najednou. Ve stromu příkazů lze také vyznačit oblast mezi dvěma příkazy se společným rodičem a na všechny příkazy v této oblasti aplikovat transformace zároveň. Mimo to objekt `Document` umožňuje využívat vlastnosti, které se ve stromu příkazů uchovávají jen tehdy, pokud se jejich nastavení liší od základního. Tyto vlastnosti umožňují snížit paměťové nároky knihovny.

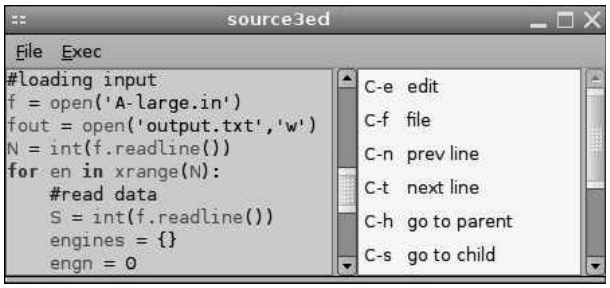
4.4. Vizualizace

Nejvyšší vrstva vizualizace je rozdělena na dvě části:

- **Obecnou vizualizační komponentu**, která poskytuje rozhraní pro zobrazení stromu příkazů a provádění operací na něm. Obecná vizualizace obsahuje také systém hierarchických menu, která umožňují implementovat ovládací rozhraní nezávisle na použité zobrazovací knihovně.
- **Specifickou implementaci** vizualizační komponenty s využitím multiplatformní knihovny `wxWidgets`. Knihovnu `WxWidgets`, která se stará o vlastní zobrazení, jsem si vybral, protože je zdarma dostupná pro platformy Windows, Linux i MacOS.

  Vizualizační komponenta umí zobrazit strom příkazů do lineárního formátu komponenty pro vkládání textů knihovny `wxWidgets` (viz. obr. 3). Objekt `Document` při změnách stromu příkazů, výběru příkazů nebo kurzoru umí volat nadřazený objekt, který se postará o vizualizaci těchto změn. Tímto objektem je většinou právě vizualizační komponenta.

Pokud ve stromu příkazů provedeme změny, umí vizualizace efektivně zobrazit pouze tyto změny. Komponenta pro vkládání textů, kterou zobrazení realizuji, umožňuje vymazat část textu a vložit text na určenou pozici. Určení pozice v této komponentě provádím pomocí vlastnosti, která je nastavena ke všem příkazům a určuje úroveň odsazení příkazu (hloubku ve stromu), délku příkazu v textu a vzdálenost příkazu od svého rodiče v textu.



Obrázek 3: Ukázka vizualizace stromu řádek

5. Příklady použití vytvořené knihovny

5.1. Syntaktický analyzátor

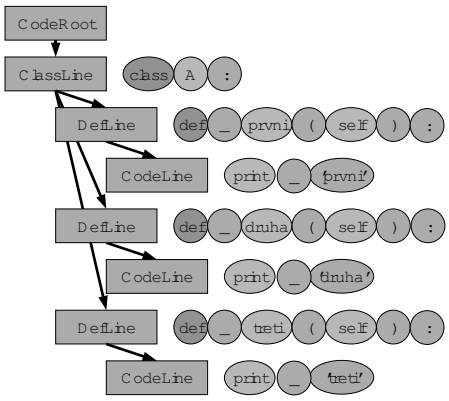
Následující příkazy na obrázku 4 převedou řetězec `test_code` (obsahující kód jazyka Python) na strom příkazů.

```
test_code = \
'''class A:
    def prvni(self):
        print 'prvni'

    def druha(self):
        print 'druha'

    def treti(self):
        print 'treti'
'''
from document import Document
from plugins import SplittingParser
root = SplittingParser().parse_string(test_code)
document = Document()
document.insert_source_tree(root)
```

Obrázek 4: Volání parseru



Obrázek 5: Objekty reprezentující kód

Samotný převod realizuje objekt `SplittingParser`. Vrchol stromu příkazů pak přiřadíme objektu `Document`, který se stará o manipulaci s tímto stromem. Na obrázku 5 jsou v obdélnících vyznačeny typy objektů, které tvoří strom příkazů. V oválech nalevo od každého příkazu jsou pak vypsány tokeny jazyka Python, které ho tvoří. Barva oválu označuje typ tokenu. Tokeny stejného typu vyskytující se za sebou jsou reprezentovány jedním objektem.

5.2. Uložení do souboru

Příkazy na obrázku 6 provedené po příkazech na obrázku 4 uloží strom příkazů do XML souboru (obr. 7) a pak ho z něj znovu načtou.

```
document.save('vystup.pyt')
document = Document()
document.load('vystup.pyt')
```

Obrázek 6: Příkazy pro uložení a nahrání dokumentu

Odpovídající XML soubor je rozdělen na hlavičku a tělo. V hlavičce je umístěno mapování jmen v souboru na verze a názvy v programu. V těle je pak uložen strom příkazů. Každý příkaz a token je uložen pomocí html tagu `entity`. Rozdílné názvy v souboru a programu používám z důvodu délky těchto názvů.

Při nahrání souboru se pomocí DOM parseru[4] načte celý soubor a pro každé jméno entity se za pomoci hlavičky vybere objekt *"importer"*, který převede *serializovaná* data na aktuální verzi objektu. Tímto způsobem je zajištěna kompatibilita se staršími verzemi uložených dat.

```
<source ver="0.1">
  <head>
    <mapping>
      <map name="classline"
           eid="plugins.code.ClassLine" ver="1" />
      <map name="defline"
           eid="plugins.code.DefLine" ver="1" />
      <map name="code"
           eid="base.CodeToken" ver="1" />
      <map name="croot"
           eid="plugins.code.CodeRootLine" ver="1" />
      <map name="keyword"
           eid="base.KeywordToken" ver="1" />
      <map name="name"
           eid="base.NameToken" ver="1" />
    </mapping>
  </head><entity name="croot" uid="int_2">
    <list name="lines">
      <entity name="classline" uid="int_3">
        <list name="tokens">
          <entity name="keyword" text="str_class" />
          <entity name="code" text="str_ " />
          <entity name="name" text="str_A" />
...
```

Obrázek 7: Obsah XML souboru

[4] Document Object Model parser - syntaktický analyzátor, který převede XML soubor na jeho objektovou reprezentaci v paměti

5.3. Editace

Objekt `Document` umí strom příkazů transformovat. První odstavec kódu (obr. 8) vytvoří v dokumentu výběr mezi ukazateli `start` a `end`. Další příkaz pak tento výběr příkazů vymaže. Poslední příkaz pak vše vrátí do původního stavu.

```
root = document.get_root()
start = document.get_traveller(root[0][0])
end = start.copy()
end.go_next_sibling()
document.set_selection(start, end)

document.delete()
document.undo()
```

Obrázek 8: Příkazy pro editaci dokumentu

5.4. Rozšiřující informace

Strom příkazů lze rozšířit o libovolné další informace. V následujícím příkladu (obr. 9) zavádím skupinu informací, která bude reprezentovat, kdo a kdy změnil naposledy konkrétní příkaz. Tato skupina je reprezentována objektem `EditMark` a musí být odvozena od třídy `Entity`. Ke každému příkazu ve stromu příkazů se naváže jedna instance této skupiny (příkaz, ke kterému je instance skupiny navázána nazvu *kontrolovaný příkaz*). Nově vytvořené třídě je potřeba nastavit identifikátor v XML souboru, jednoznačný identifikátor v programu a verzi ukládaných dat.

Metoda `serialize` ukládá perzistentní data vlastnosti `EditMark` do XML souboru a metoda `deserialize` se stará o nahrání těchto dat z XML souboru. Při změně *kontrolovaného příkazu* se zavolá metoda `on_set`, která vlastnosti nastaví aktuální údaje.

Důležitý je poslední řádek příkladu, který knihovně říká, jakým způsobem má vytvořit objekt této skupiny vlastností, pokud na jeho data narazí v XML souboru. Proto je potřeba tento řádek provést před jakýmkoliv nahráváním souboru.

```
class EditMark(Entity):
    name = 'editmark'
    entity_id = 'plugins.editmarks.EditMark'
    version = Version('0.1.0')

    def on_set(self, root, source):
        #aktualizuj udaje
        ...

    def serialize(self, doc):
        e = Entity.serialize(self, doc)
        doc.serialize_type(e, 'editor',
                           self.editor, int)
        doc.serialize_type(e, 'time',
                           self.time, datetime)
        return e

    def deserialize(self, e, doc):
        Entity.deserialize(self, e, doc)
        self.editor = doc.deserialize_type(e,
                           'editor', int)
        self.time = doc.deserialize_type(e,
                           'time', datetime)

register_importer(EditMark.entity_id,
        DefaultImporter(EditMark))
```

Obrázek 9: Příklad rozšiřující vlastnosti

6. Závěr

Vytvořená knihovna pro práci se zdrojovým kódem ve stromu příkazů je volně k použití [4] pod licencí BSD pro vytváření nástrojů pro práci se zdrojovým kódem.

Literatura

[1] František Jahoda, *"Metainformace pro zdrojový kód jazyka Python"*, Diplomová práce, KM FJFI ČVUT

[2] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley Professional, ISBN 0201485672

[3] Dick Grune, Ceriel J.H. Jacobs, *Parsing Techniques - A Practical Guide*, Ellis Horwood, Chichester, England, ISBN 0136524316

[4] Domovská stránka vytvořené knihovny, http://sourceforge.net/projects/source3ed

[5] Roedy Green, *esej o SCID IDE* http://mindprod.com/projects/scid.html

Evolutionary Algorithms for Constrained Optimization Problems

Post-Graduate Student:
ING. DAVID KOZUB

Department of Mathematics
Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Trojanova 13

120 00 Prague, Czech Republic

zub@linux.fjfi.cvut.cz

Supervisor:
ING. RNDR. MARTIN HOLEŇA, CSC.

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

182 07 Prague, Czech Republic

martin@cs.cas.cz

Field of Study:
Mathematical Engineering

Abstract

This paper presents an overview of the techniques used to solve constrained optimization problems using evolutionary algorithms. The construction of the fitness function together with the handling of feasible and infeasible individuals is discussed. Approaches using penalty functions, special representations, repair algorithms, methods based on separation of objective and constraints and multiobjective techniques are mentioned.

1. Introduction

Evolutionary algorithms have been successfully used in a range of applications. [1] Majority of the papers presented pertain to unconstrained optimization problems. As [2] argues, virtually all real problems are constrained. Thus, the study of constraint-handling methods that can be used with evolutionary algorithms is an important subject.

Evolutionary algorithms are based on a analogy with the evolution process occurring in nature: The individuals have genes that encode the solution. The individuals are compared with others and those that perform better (have higher fitness) get higher probability of propagating their genes into the next generation. The genes of the offspring population are the product of applying genetic operators to the genes of their parent individuals.

For an evolutionary algorithm, the following is needed:

- A representation of the potential solution (an individual).
- A way of initializing the population of the individuals.
- Genetic operators that act on the (parent) population – typically recombination and mutation.
- Selection operator that chooses which individuals propagate to the next generation.

Evolutionary algorithm can be formally defined as follows (based on [1]):

Definition 1 *(Evolutionary algorithm) The following algorithm is called an Evolutionary Algorithm:*

1. $t \leftarrow 0$
2. *initialize:*
$$P_0 = \left\{a_0, \ldots, a_{\mu^{(0)}}\right\} \subseteq \mathcal{I}$$
3. *while (* $\iota\left((P_0, \ldots, P_t)\right) \neq 1$ *) do*
   (a) *recombine:*
   $$P'_t \leftarrow r^{(t)}_{\phi_r^{(t)}}(P_t)$$
   (b) *mutate:*
   $$P''_t \leftarrow m^{(t)}_{\phi_m^{(t)}}(P'_t)$$
   (c) *select: if* $\chi = 1$*:*
   $$P_{t+1} \leftarrow s^{(t)}_{(\phi_s^{(t)})}(P''_t)$$
   else:
   $$P_{t+1} \leftarrow s^{(t)}_{(\phi_s^{(t)})}(P''_t \cup P_t)$$
   (d) $t \leftarrow t + 1$

where:

- $\mathcal{I} \neq \varnothing$ *is the individual space*
- $a_0, \ldots, a_{\mu^{(0)}}$ *is the initial population*
- $\left(\mu^{(i)}\right)_{i \in \mathbb{N}_0}$ *is a sequence of the parent population sizes*
- $\left(\mu'^{(i)}\right)_{i \in \mathbb{N}_0}$ *is a sequence of the offspring population sizes*
- $\iota : \left\{ \left(\mathcal{I}^{\mu^{(i)}} \right)_{i=0}^{t} \middle| t \in \mathbb{N}_0 \right\} \rightarrow \{0,1\}$ *is the terminating criterion*
- $\chi \in \{0,1\}$ *chooses between* (μ, λ) *and* $(\mu + \lambda)$ *selection method*
- $\left(r^{(i)}\right)_{i \in \mathbb{N}_0}$ *is a sequence of recombination operators:*

  $$r^{(i)} : \Xi_r^{(i)} \rightarrow \left[\mathcal{I}^{\mu^{(i)}} \rightarrow \mathcal{I}^{\mu'^{(i)}} \right]$$

  where $\Xi_r^{(i)}$ *is the set of recombination parameters and* $\theta_r^{(i)} \in \Xi_r^{(i)}$
- $\left(m^{(i)}\right)_{i \in \mathbb{N}_0}$ *is a sequence of mutation operators:*

  $$m^{(i)} : \Xi_m^{(i)} \rightarrow \left[\mathcal{I}^{\mu'^{(i)}} \rightarrow \mathcal{I}^{\mu'^{(i)}} \right]$$

  where $\Xi_m^{(i)}$ *is the set of mutation parameters and* $\theta_m^{(i)} \in \Xi_m^{(i)}$
- $\left(s^{(i)}\right)_{i \in \mathbb{N}_0}$ *is a sequence of selection operators:*

  $$s^{(i)} : \Xi_s^{(i)} \rightarrow \left[\mathcal{I}^{\mu'^{(i)} + \chi\mu'^{(i)}} \rightarrow \mathcal{I}^{\mu^{(i+1)}} \right]$$

  where $\Xi_s^{(i)}$ *is the set of mutation parameters and* $\theta_s^{(i)} \in \Xi_s^{(i)}$

In this paper we focus on applying evolutionary algorithms to constrained optimization problems. By this we mean the following:

$$\min_{x \in \Omega} f(x) \tag{1}$$

subject to:

$$g_i(x) \leq 0 \quad \forall i \in \{1, \ldots, n_g\} \tag{2}$$
$$h_j(x) = 0 \quad \forall j \in \{1, \ldots, n_h\} \tag{3}$$

where the set Ω is the *search space*. Let n denote the total number of constraints:

$$n = n_g + n_h$$

The constraints (3) and (2) implicitly define the feasible set Φ:

$$\Phi = \{ x \in \Omega | g_i(x) \leq 0 \wedge h_j(x) = 0 \\ \forall i \in \{1, \ldots, n_g\}, \forall j \in \{1, \ldots, n_h\} \}$$

We make no additional assumptions about the feasible set. In general it can be a non-convex, even a disconnected set.

Defining $\Upsilon = \Omega - \Phi$, it can be stated that the search space Ω is partitioned into two disjoint sets: the feasible set Φ and the infeasible set Υ.

The level of violation of the constraints (2) and (3) by a point $x \in \Omega$ can be measured as follows:

$$G_i(x) = \max\{0, g_i(x)\} \tag{4}$$
$$H_j(x) = |h_j(x)| \tag{5}$$

Note that for all $x \in \Phi$

$$G_i(x) = 0$$
$$H_j(x) = 0$$

for all $i \in \{1, \ldots, n_g\}$, $j \in \{1, \ldots, n_h\}$.

An equality constraint $h_j(x) = 0$ can be transformed into inequality constraints in the following way:

$$|h_j(x)| \leq \varepsilon$$

where ε is a small constant specifying the tolerance.

This approach allows the equality constraints to be treated as inequalities, which can be useful for methods that do not treat equality constraints separately.

2. Fitness function

The fitness function is a function $F : \mathcal{I} \rightarrow \mathbb{R}$ that evaluates the individuals according to how well they solve given problem.

The design of the fitness function can be a non-trivial task even for an unconstrained problem. In case of constrained problems, the design of a good fitness function is even more difficult. In [2] the following points guiding the design of the fitness function are listed:

1. How should two feasible points be compared?
2. How should two infeasible points be compared?

3. How are the functions for feasible and infeasible points related? Should feasible points be always ”better” than infeasible ones?
4. Should infeasible points be considered harmful and removed from the population?
5. Should infeasible points be ”repaired”?
6. If individuals are repaired, should this repaired individual be used only for evaluating its fitness (*Baldwin effect*) or should the individual be replaced (*Lamarckian evolution*)?
7. Should infeasible individuals be penalized?
8. Should the algorithm start with a feasible population and keep the feasibility throughout the run of the algorithm?

During the run of the algorithm, the population can generally contain both feasible and infeasible individuals. In the end though, the answer must be a feasible solution, as the infeasible individual, no matter its fitness from the point of view of the evolutionary algorithm, is not a solution to the original problem.

An obvious method of ensuring this works by removing all the infeasible solutions, so that the population never contains an infeasible individual. While this method has been used, in many problems it does not work. (See section 3 for more information on this approach.)

This leads to the conclusion that the evolutionary algorithm should allow the infeasible individuals in the population. Because of this, a decision has to be made on how to compare the feasible and the infeasible individuals.

One way to tackle this task is to define the fitness function as follows:

$$F(x) = \begin{cases} F_\Phi(x) & x \in \Phi \\ F_\Upsilon(x) & x \in \Upsilon \end{cases} \tag{6}$$

When evaluating F_Φ, the actual value of the constraints should not be important, as the point is in the feasible set. When evaluating F_Υ, the question is if the value of the objective function f should be taken into account. F_Υ should react to the fact that the solution is not feasible and direct the search into the feasible set. Yet, should it be based on the amount of the violation, or should it only reflect the number of violated constraints?

While the inclusion of the objective f in F_Υ might help guide the search, sometimes (in case the objective is not defined outside of the feasible region Φ) this is not possible.

It should be noted that in some evolutionary algorithms the fitness function is not explicitly needed. For example, if the evolutionary algorithm uses the tournament selection, all that is needed is an ordering relation defined over the individual space $\mathcal{I}$. Still, this does not relieve us of the burden of satisfactorily answering the aforementioned questions.

An overview of some of the methods that were used to solve constrained optimization problems follows. The methods differ by how they answer the aforementioned questions.

3. Penalty functions

The oldest and most common approach to solving constrained optimization problems using evolutionary algorithms is the use of a penalty function. The method is based in the idea of adding to the objective function f a function that penalizes solutions laying in the infeasible set, thus decreasing their fitness.

There are two basic options: *interior penalty functions* – this approach starts from a feasible solution and the penalty function is defined so that its value approaches to infinity as the solution moves towards the boundary of the feasible set, and *exterior penalty functions* – this approach starts from any (generally infeasible) point in the search space and the penalty is used to guide the search into the feasible set.

An advantage of the exterior approach is that it does not require an initial feasible population.

The generic formula for the fitness function with an exterior penalty is:

$$F(x) = f(x) + P^{(t)}(x) \tag{7}$$

where $P^{(t)} : \mathcal{I} \to \langle 0, +\infty)$ is the penalty function satisfying for all $x \in \Phi$ and for all $t \in \mathbb{N}_0$:

$$P^{(t)}(x) = 0$$

A problem with this approach is the choice of the value of the penalty: Too small penalty value does not discourage the algorithm from the infeasible set, possibly resulting in an infeasible optimum. On the other hand, too high penalty value might prohibit the algorithm from crossing the feasible set boundary (which might be useful or even necessary in case the feasible set is non-convex or disconnected) and from exploring the boundary of the feasible set.

In [3] author suggests the relation between an infeasible individual and the feasible set plays an important role in the penalization. There are several ways how this relationship could be reflected in the penalty function:

1. the penalty is constant – the individual is being penalized for being infeasible
2. the penalty reflects the amount of constraint violation
3. the penalty reflects the effort needed to make the individual feasible

This method was advanced in several directions in order to tackle this issue:

static penalties In this approach, the value of the penalties is independent of the generation number. Typical choice for $P^{(t)}$ is:

$$P^{(t)}(x) = \sum_{i=1}^{n_g} a_i G_i(x)^\beta + \sum_{j=1}^{n_h} b_j H_j(x)^\gamma$$

with $\beta, \gamma \in \{1, 2\}$, a_i, b_i positive constants called *penalty factors* and G_i, H_j as defined in (4) and (5).

dynamic penalties In this approach, the value of the penalties is dependent on the generation number. Typically, the penalties rise over time. This enables the population to explore the search space (low penalties) and eventually move into the feasible set. An example of this approach is:

$$P^{(t)}(x) = (ct)^\alpha \left(\sum_{i=1}^{n_g} a_i G_i(x)^\beta + \sum_{j=1}^{n_h} b_i H_j(x)^\gamma \right)$$

annealing penalties This method was inspired by simulated annealing: The penalties change when the algorithm gets stuck in a local optimum. The penalty rises over time to penalize infeasible solutions in the end of the run of the algorithm.

adaptive penalties Within this approach, the penalty uses the previous states of the algorithm: The penalty with respect to a constraint is increased if all the individuals in the previous generation were infeasible. The penalty is decreased if all the individuals in the previous generation were feasible.

co-evolutionary penalties In this approach, there are more populations, for example a population for the evolution of solutions and a population for the evolution of the penalty factors. A co-evolution scheme is then used.

death penalty This is a simple method that works by eliminating all the non-feasible individuals form the population. While it can be easily implemented, it tends to work only if the feasible set is a reasonably large subset of the search space and when the feasible set is convex. [2]

Another approach in this category works by focusing the search on the boundary of the feasible set Φ. According to [1], many real-world tasks have optimum for which at least some constraints are active, so the focus on the boundary of the feasible set seems reasonable. The way the border is explored is by varying a penalty and thus forcing the individuals to cross between the feasible and the infeasible set.

The main disadvantage of the penalty methods is their dependency on multiple parameters. While some guidance has been provided, often the parameters have to be empirically determined. [1] Also, penalty methods often do not perform well when the problem is highly-constrained or when the feasible set is disconnected. [2]

4. Special representations

This approach tackles the optimization problem by designing a special, problem-dependent, representation of the individuals. This in turn calls for special operators to be used on those individuals. The operators used typically preserve the feasibility of the population. The motivation behind this approach is to simplify the feasible set Ω.

The representation is problem-specific. While the approach was successfully used on specific problems, it is difficult to generalize this approach.

5. Repair algorithms

This approach works by repairing infeasible individuals. Two ways are possible: The repaired individual is used only to evaluate the fitness of the original, or the infeasible individual is replaced with the repaired one.

The resulting individual is not necessarily feasible, but the amount of constraint violation is reduced.

This method was generalized into the area of constrained multiobjective evolutionary optimization in [4] and [5].

The repair approach often has problems with keeping the diversity of the population. Also, the repair operator can sometimes introduce a strong bias into the search process. [3]

6. Separation of constraints and objectives

The following approaches do not mix the objective and the constraints together. There are several different methods reported in [2] and [3].

6.1. Superiority of feasible points

In this approach feasible individuals are always considered superior to infeasible ones.

One way to ensure this is to map the objective function onto a bounded-above interval, e. g. $(-\infty, 1)$ and specify the fitness function like:

$$F(x) = \begin{cases} f(x) & x \in \Phi \\ L(x) & x \in \Upsilon \end{cases} \tag{8}$$

where $L : \Upsilon \to (1, +\infty)$ is a function measuring the level of constraint violation.

An interesting adaptation that does not require the objective to be bounded-above is:

$$F(x) = \begin{cases} f(x) & x \in \Phi \\ f_{max}^{(t)} + L(x) & x \in \Upsilon \end{cases} \tag{9}$$

where $f_{max}^{(t)} = \max_{x \in P_{(t)} \cap \Phi} f(x)$ and $L : \Upsilon \to \mathbb{R}^+$ is a function measuring the level of constraint violation.

A different way to ensure the feasible points are always superior is to use tournament selection with the rules (x and y denotes the individuals being compared) from table 1.

Table 1: Tournament selection for the superiority of feasible points method

$x \in \Phi$	$y \in \Upsilon$	x is preferred over y
$x \in \Upsilon$	$y \in \Phi$	y is preferred over x
$x \in \Phi$	$y \in \Phi$	decide based on $f(x)$ and $f(y)$
$x \in \Upsilon$	$y \in \Upsilon$	decide based on constraint violation

6.2. Behavioral memory

This method requires a linear ordering of the constraints. Then it proceeds as follows:

1. initialize the population randomly
2. evolve the individuals to minimize the violation of the first constraint; stop when the percentage of individuals feasible with respect to the first constraint surpasses given percentage
3. $j \to 2$
4. while $j \leq n$ do:
   (a) evolve the individuals to minimize the violation of the j-th constraint while removing individuals which do not satisfy any of the constraints $1 \ldots j$; stop when the percentage of individuals feasible with respect to the j-th constraint surpasses given percentage
   (b) $j \to j + 1$
5. evolve the individuals to minimize the objective f while removing infeasible individuals from the population (*death penalty* – see section 3)

This approach is similar to the lexicographic ordering approach mentioned in subsection 7. A drawback is that the initial ordering of the constraints influences the results obtained.

Those methods do not work well when the size of the feasible set is relatively small (when the constraints are difficult to satisfy). Another problem mentioned in [3] is the difficulty of maintaining the diversity of the population.

An interesting point to make is that those approaches never evaluate the objective on infeasible points, making it interesting for problems with hard constraints.

7. Multiobjective techniques

The technique works by transforming the original constrained optimization problem into an unconstrained multiobjective problem, turning the original constraints into additional objectives. The problem (1) – (3) turns into:

$$\min_{x \in \Omega} (f, G_1(x), \ldots, G_{n_g}(x), H_1(x), \ldots, H_{n_h}(x)) \tag{10}$$

The ideal solution of (10) is an $x^{ideal} \in \Phi$ such that:

$$\begin{aligned} f(x^{ideal}) &= min_{x \in \Phi} f(x) \\ G_i(x^{ideal}) &= 0 \ \forall i \in \{1, \ldots, n_g\} \\ H_j(x^{ideal}) &= 0 \ \forall j \in \{1, \ldots, n_h\} \end{aligned}$$

Unlike in actual multiobjective optimization, here we are not interested in finding good trade-offs between

the objectives (the original objective (1) and the constraints): Any feasible point might be acceptable, no matter the actual value of the constraint violation values. On the other hand, a global minimum that lies in the infeasible set is no solution to the original problem, even if it means a good trade-off in the multiobjective problem.

In [6] a min-max-like approach was described: The evolutionary algorithm uses the tournament selection with the rules (x and y denotes the individuals that are compared) according to table 2.

Table 2: Tournament selection for the min-max approach in [6]

$x \in \Phi$	$y \in \Upsilon$	x is preferred over y
$x \in \Upsilon$	$y \in \Phi$	y is preferred over x
$x \in \Phi$	$y \in \Phi$	decide based on $f(x)$ and $f(y)$
$x \in \Upsilon$	$y \in \Upsilon$	select the individual having the smallest maximal constraint violation.

8. Conclusion

This paper presents several ways of handling constrains together with evolutionary optimization. Majority of the approaches does need to evaluate the objective outside the feasible set, which renders the methods unusable for constraints that cannot be relaxed. Handling such problems with evolutionary algorithms seems therefore like an interesting option for further research.

References

[1] C. A. Coello Coello, D. A. Van Veldhuisen, G. B. Lamont "Evolutionary Algorithms for Solving Multi-Objective Problems", *Kluwer Academic Publishers*, 2002.

[2] Z. Michalewicz, M. Schmidt "Evolutionary Algorithms and Constrained Optimization", *Evolutionary Optimization*, New York, Kluwer Academic Publishers, pp. 57–86, 2003.

[3] C. A. Coello Coello, "Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art", *Computer Methods in Applied Mechanics and Engineering*, vol. 191, pp. 1245–1287, 2002.

[4] K. Harada, J. Sakuma, K. Ikeda, I. Ono, S. Kobayashi, "Local search for multi-objective function optimization: Pareto descent method", in: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, New York, NY, ACM Press, pp. 659–666, 2007.

[5] K. Harada, J. Sakuma, I. Ono, S. Kobayashi "Constraint-Handling Method for Multi-objective Function Optimization: Pareto Descent Repair Operator", in: *Proceedings of the Evolutionary Multi-Criterion Optimization (EMO 2007)*, Springer, Berlin, 156–170, 2007.

[6] F. Jiménez, J. L. Verdegay "Evolutionary Techniques for Constrained Optimization Problems", in: *Seventh European Congress on Intelligent Techniques and Soft Computing*, Springer, Aachen, 1999.

A Note on Steady Flows of an Incompressible Fluid with Pressure- and Shear Rate-dependent Viscosity

Post-Graduate Student:
MGR. MARTIN LANZENDÖRFER

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic ,
Mathematical Institute
Charles University
Sokolovská 83

186 75 Prague, Czech Republic

lanz@cs.cas.cz

Supervisor:
DOC. RNDR. JOSEF MÁLEK, CSC.

Mathematical Institute
Charles University
Sokolovská 83

186 75 Prague, Czech Republic

malek@karlin.mff.cuni.cz

Field of Study:
Mathematical Modeling

This work was supported by GAČR 201/06/0352.

Abstract

A class of incompressible fluids whose viscosities depend on the pressure and the shear rate is considered. The existence of weak solutions for flows of such fluids under different settings was studied lately. In this short note, two recent existence results are adverted and their direct generalization into different setting is indicated; in this setting the corresponding energy estimates are derived showing the existence of a solution to an approximate system. A minor correction to one of the referred papers is also stated.

1. Introduction

The Newtonian homogeneous incompressible fluid is described by Navier-Stokes equations, where a linear relation between the stress tensor and the symmetric part of the velocity gradient is assumed, with a given constant called viscosity. However, in many important applications a non-Newtonian model is required. In this short note, the existence of a weak solution for steady flows of fluids with the viscosity increasing with the pressure and decreasing with the shear rate is addressed.

1.1. Fluid model

The theoretical analysis of the following problem is considered: Find the pressure and the velocity $(p, \boldsymbol{v}) = (p, v_1, \dots, v_d) : \Omega \to \mathbb{R}^{d+1}$ ($\Omega \subset \mathbb{R}^d$ being an open bounded domain, $d \geq 2$) solving the equations:

$$\operatorname{div} \boldsymbol{v} = 0 \quad \text{in } \Omega\,, \tag{1}$$

$$\operatorname{div}(\boldsymbol{v} \otimes \boldsymbol{v}) - \operatorname{div}[\nu(p, |\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v})] = -\nabla p + \boldsymbol{b} \quad \text{in } \Omega\,, \tag{2}$$

(∇ denotes the Eulerian spatial gradient, $\mathbf{D}(\boldsymbol{v}) = \frac{1}{2}(\nabla \boldsymbol{v} + (\nabla \boldsymbol{v})^T)$ the symmetric part of the velocity gradient) completed by:

$$\int_\Omega p \,\mathrm{d}\boldsymbol{x} = 0 \tag{3}$$

and by the Dirichlet boundary condition

$$\boldsymbol{v} = \boldsymbol{\varphi} \qquad \text{on } \partial\Omega\,, \tag{4}$$

where $\boldsymbol{\varphi} : \partial\Omega \to \mathbb{R}^d$ and $\boldsymbol{b} : \Omega \to \mathbb{R}^d$ are given. We shall denote the system (1)-(4) by Problem (P). Standard notation[1] concerning function spaces is used.

For the viscosity $\nu(p, |\mathbf{D}|^2)$ the following assumptions are considered:

A1 For a given $r \in (1, 2)$, there are positive constants C_1 and C_2 such that for all symmetric linear transformations $\mathbf{B}$, $\mathbf{D}$ and all $p \in \mathbb{R}$

$$C_1(1 + |\mathbf{D}|^2)^{\frac{r-2}{2}}|\mathbf{B}|^2 \leq \frac{\partial[\nu(p, |\mathbf{D}|^2)\mathbf{D}]}{\partial \mathbf{D}} \cdot (\mathbf{B} \otimes \mathbf{B}) \leq C_2(1 + |\mathbf{D}|^2)^{\frac{r-2}{2}}|\mathbf{B}|^2\,,$$

where $(\mathbf{B} \otimes \mathbf{B})_{ijkl} = \mathbf{B}_{ij}\mathbf{B}_{kl}$.

[1] For $1 \leq r \leq \infty$, the symbols $(\mathrm{L}^r(\Omega), ||\cdot||_r)$ and $(\mathrm{W}^{1,r}_{(0)}(\Omega), ||\cdot||_{1,r})$ denote the standard Lebesgue and Sobolev spaces (with zero trace on $\partial\Omega$). If $\mathrm{X}(\Omega)$ is a Banach space of functions defined on Ω then $(\mathrm{X}(\Omega))^*$ denotes its dual space. Also, $\mathbf{X}(\Omega) := \mathrm{X}(\Omega)^d = \{\boldsymbol{u} : \Omega \to \mathbb{R}^d; u_i \in \mathrm{X}(\Omega), i = 1, \dots, d\}$. Further, $(\mathbf{W}^{-1,r'}(\Omega), ||\cdot||_{-1,r'}) := (\mathbf{W}^{1,r}_0)^*$, where $r' = \frac{r}{r-1}$. We use the Einstein summation convention in the text.

A2 For all symmetric linear transformations $\mathbf{D}$ and for all $p \in \mathbb{R}$

$$\left|\frac{\partial[\nu(p,|\mathbf{D}|^2)\mathbf{D}]}{\partial p}\right| \leq \gamma_0(1+|\mathbf{D}|^2)^{\frac{r-2}{4}} \leq \gamma_0\,,$$

with

$$\gamma_0 < \frac{1}{C_{\mathrm{div},2}}\frac{C_1}{C_1+C_2} \leq \frac{1}{2C_{\mathrm{div},2}}\,.$$

The constant $C_{\mathrm{div},q}$ originates in the following problem, which is instrumental in the proof of the existence: For $g \in \mathrm{L}^q(\Omega)$ given, $\int_\Omega g\,\mathrm{d}\boldsymbol{x} = 0$, find $\boldsymbol{z}$ solving

$$\operatorname{div}\boldsymbol{z} = g \quad \text{in } \Omega, \quad \boldsymbol{z} = \mathbf{0} \quad \text{on } \partial\Omega\,. \tag{5}$$

For $q \in (1,\infty)$, the bounded linear Bogovskii operator $\mathcal{B} : \mathrm{L}^q(\Omega) \to \mathbf{W}^{1,q}_0(\Omega)$, assigning $\boldsymbol{z} := \mathcal{B}(g)$ the solution of (5), fulfills

$$||\boldsymbol{z}||_{1,q} = ||\mathcal{B}(g)||_{1,q} \leq C_{\mathrm{div},q}||g||_q\,. \tag{6}$$

Moreover, if $g = \operatorname{div}\boldsymbol{f}$, with $\boldsymbol{f} \in \mathbf{W}^{1,q}(\Omega)$ and $\boldsymbol{f}\cdot\boldsymbol{n} = 0$ on $\partial\Omega$, then

$$||\boldsymbol{z}||_q = ||\mathcal{B}(\operatorname{div}\boldsymbol{f})||_q \leq D_{\mathrm{div},q}||\boldsymbol{f}||_q\,. \tag{7}$$

Note that the assumptions **(A1)** and **(A2)** determine the fluid model to be shear-thinning and allow it to be pressure-thickening. Examples and more details can be found e.g. in [1]. Note also that the following inequalities result from **(A1)** and **(A2)**, see [1, 2] for their proofs. First,

$$\nu(p,|\mathbf{D}|^2)\mathbf{D} : \mathbf{D} \geq \frac{C_1}{2r}(|\mathbf{D}|^r - 1)\,, \tag{8}$$

$$|\nu(p,|\mathbf{D}|^2)\mathbf{D}| \leq \frac{C_2}{r-1}(1+|\mathbf{D}|)^{r-1} \tag{9}$$

holds for all symmetric $\mathbf{D}$ and all $p \in \mathbb{R}$. Then, defining

$$I^{1,2} := \int_0^1 \left(1+|\mathbf{D}^1 + s(\mathbf{D}^2-\mathbf{D}^1)|^2\right)^{\frac{r-2}{2}} |\mathbf{D}^1-\mathbf{D}^2|^2\,\mathrm{d}s\,, \tag{10}$$

there hold

$$\frac{C_1}{2}I^{1,2} \leq \left(\nu(p^1,|\mathbf{D}^1|^2)\mathbf{D}^1 - \nu(p^2,|\mathbf{D}^2|^2)\mathbf{D}^2\right) : (\mathbf{D}^1-\mathbf{D}^2) + \frac{\gamma_0^2}{2C_1}|p^1-p^2|^2\,, \tag{11}$$

$$\left|\nu(p^1,|\mathbf{D}^1|^2)\mathbf{D}^1 - \nu(p^2,|\mathbf{D}^2|^2)\mathbf{D}^2\right| \leq C_2\left(I^{1,2}\right)^{\frac{1}{2}} + \gamma_0\,|p^1-p^2|\,. \tag{12}$$

1.2. Results

The model described above has been systematically studied in last decade or more; the reader is kindly asked to find references given in [1] and [2].

In [1], the existence of a weak solution to Problem (P) including the non-homogeneous Dirichlet boundary condition (4) was proved, either for small data or assuming the inner flows:

$$\boldsymbol{\varphi}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega\,. \tag{13}$$

The proof is given for $d = 2$ or 3 and for

$$\frac{3d}{d+2} \leq r < 2\,.$$

The lower bound relates to the fact, that with $r \geq \frac{3d}{d+2}$ the solution is a possible test function in the weak formulation and a standard monotone operator theory is applicable, supplied by proper estimates on the pressure. Within the proof, the following ε-approximate system is utilized, replacing equation (1) by

$$-\varepsilon\Delta p + \operatorname{div}\boldsymbol{v} = 0 \ \text{ in } \Omega, \quad \frac{\partial p}{\partial \boldsymbol{n}} = 0 \ \text{ on } \partial\Omega \tag{14}$$

for $\varepsilon > 0$. The solution to Problem (P) is obtained by the limit $\varepsilon \to 0$.

Recently in [2], the theory was extended to the case

$$\frac{2d}{d+2} < r \leq \frac{3d}{d+2}\,,$$

considering the homogeneous Dirichlet boundary condition

$$\boldsymbol{\varphi} = \mathbf{0} \quad \text{on } \partial\Omega\,.$$

The starting point is the following η,ε-approximate system, replacing (1) by (14) and replacing (2) by

$$\left.\begin{aligned}&\eta|\boldsymbol{v}|^{2r'-2}\boldsymbol{v} + \operatorname{div}(\boldsymbol{v}\otimes\mathcal{P}\boldsymbol{v})\\ &\quad - \operatorname{div}\left[\nu(p,|\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v})\right] = -\nabla p + \boldsymbol{b}\end{aligned}\right\} \tag{15}$$

for $\eta > 0$, where $\mathcal{P}$ is a projection to divergence-free functions.

The goal of the presented paper is to follow these two results and to study the existence of a weak solution to Problem (P) with

$$r < \frac{3d}{d+2}$$

and subject to non-homogeneous Dirichlet boundary condition. Section 2 derives the energy estimates for the corresponding η,ε-approximate system, thereby showing the existence of its weak solution. In Section 3, the main existence theorem is merely stated, the remaining parts of the proof–the limit procedures $\varepsilon \to 0$ and $\eta \to 0$–being left to the reader, referring to [2]. The theorem assumes non-homogeneous Dirichlet b.c. with small data, its corollary then treats inner flows with large data. In the last section, some minor correction to [1] is mentioned.

2. Energy estimates

The main result of this paper is the following variation of Lemma 4.1, which is the starting point of the result established in [2].

Lemma 1 *Let $\varepsilon, \eta > 0$ be arbitrary. Let $\Omega \in \mathcal{C}^{0,1}$, $d \geq 2$ and $\boldsymbol{b} \in \mathbf{W}^{-1,r'}(\Omega)$ be given. Let*

$$\frac{2d}{d+1} < r < \min\left\{2, \frac{3d}{d+2}\right\} \tag{16}$$

and the assumptions **A1** *and* **A2** *be satisfied. There are certain positive constants H_1, H_2 which depend on r, Ω, C_1, C_2 and $\boldsymbol{b}$ and which are small enough such that they meet the inequality* (23)*. Let there exist $\lambda \geq 1$ and $\boldsymbol{\Phi} \in \mathbf{W}^{1,r}(\Omega)$ such that, with $q := \frac{rd}{r(d+1)-2d}$,*

$$\operatorname{div}\boldsymbol{\Phi} = 0 \ \text{ in } \Omega, \quad \operatorname{tr}\boldsymbol{\Phi} = \boldsymbol{\varphi} \qquad \text{and} \qquad ||\boldsymbol{\Phi}||_q \leq H_1\lambda^{r-2} \qquad \text{and} \qquad ||\nabla\boldsymbol{\Phi}||_r \leq ||\boldsymbol{\Phi}||_{1,r} \leq H_2\lambda\,. \tag{17}$$

Then there exists a couple $(p, \boldsymbol{v})$ satisfying

$$\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{\Phi}, \qquad \boldsymbol{u} \in \mathbf{W}^{1,r}_0(\Omega) \cap \mathbf{L}^{2r'}(\Omega) \qquad \text{and} \qquad p \in \mathrm{W}^{1,2}(\Omega) \cap \mathrm{L}^2_0(\Omega)\,, \tag{18}$$

$$\varepsilon \int_\Omega \nabla p \cdot \nabla\xi \,\mathrm{d}\boldsymbol{x} + \int_\Omega \xi \operatorname{div}\boldsymbol{v} \,\mathrm{d}\boldsymbol{x} = 0 \qquad \text{for all } \xi \in \mathrm{W}^{1,2}(\Omega)\,, \tag{19}$$

$$\left.\begin{aligned} &\eta \int_\Omega |\boldsymbol{u}|^{2r'-2}\boldsymbol{u}\cdot\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} + \int_\Omega \nu(p,|\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v}) : \mathbf{D}(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{x} - \int_\Omega (\boldsymbol{v}\otimes\boldsymbol{v}) : \nabla\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} \\ &\quad - \frac{1}{2}\int_\Omega (\operatorname{div}\boldsymbol{u})\,\boldsymbol{u}\cdot\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} = \int_\Omega p\operatorname{div}\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} + \langle\boldsymbol{b},\boldsymbol{\psi}\rangle \qquad \text{for all } \boldsymbol{\psi} \in \mathbf{W}^{1,r}_0(\Omega)\cap\mathbf{L}^{2r'}(\Omega)\,. \end{aligned}\right\} \tag{20}$$

Moreover, the following estimates hold:

$$\varepsilon\,||p||^2_{1,2} + \eta\,||\boldsymbol{v}||^{2r'}_{2r'} + ||\mathbf{D}(\boldsymbol{v})||^r_r \leq C < +\infty\,, \tag{21}$$

$$||\nu(p,|\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v})||_{r'} \leq C < +\infty \qquad \text{and} \qquad ||p||_{\frac{2dr}{r(d-2)+d}} \leq C(\eta) < +\infty\,. \tag{22}$$

Proof: Note that all integrals make sense:

$$\begin{aligned} &\boldsymbol{v} \in \mathbf{W}^{1,r}(\Omega)\cap\mathbf{L}^{2r'}(\Omega) \quad \Leftarrow \quad \boldsymbol{\Phi} \in \mathbf{W}^{1,r}(\Omega)\cap\mathbf{L}^q(\Omega), \ \text{ where } q > 2r' \text{ since } r < \frac{3d}{d+2}\,, \\ &\xi\operatorname{div}\boldsymbol{v} \in \mathrm{L}^1(\Omega) \quad \Leftarrow \quad \xi \in \mathrm{W}^{1,2}(\Omega) \hookrightarrow \mathrm{L}^{r'}(\Omega) \ \text{ since } r > \frac{2d}{d+2}\,, \\ &\nu(p,|\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v}) : \mathbf{D}(\boldsymbol{\psi}) \in \mathrm{L}^1(\Omega) \quad \Leftarrow \quad \boldsymbol{v},\boldsymbol{\psi} \in \mathbf{W}^{1,r}(\Omega) \ \text{ and since (9).} \end{aligned}$$

The pair $(p, \boldsymbol{v})$ fulfilling (18)-(20) can be found as a limit of Galerkin approximations. The proof uses Brouwer's fixed point theorem, the compact embedding argument, the monotonicity conditions (11), (12) and Vitali's theorem. Here the first steps are provided in detail and, in time, the remainings are referred to [1].

Take $\{\alpha^k\}_{k=1}^\infty$ and $\{\boldsymbol{a}^k\}_{k=1}^\infty$ any bases of $\mathrm{W}^{1,2}(\Omega)$ and $\mathbf{W}^{1,2}_0(\Omega)$, respectively. Define the Galerkin approximations as follows:

$$\left.\begin{aligned} p^N &:= \textstyle\sum_{k=1}^N c_k^N(\alpha^k - \frac{1}{|\Omega|}\int_\Omega \alpha^k\,\mathrm{d}\boldsymbol{x}) \\ \boldsymbol{v}^N &:= \boldsymbol{\Phi} + \textstyle\sum_{k=1}^N d_k^N\boldsymbol{a}^k =: \boldsymbol{\Phi} + \boldsymbol{u}^N \end{aligned}\right\} \quad \text{for } N = 1,2,\ldots\,,$$

where $\boldsymbol{c}^N = (c_1^N,\ldots,c_N^N)$ and $\boldsymbol{d}^N = (d_1^N,\ldots,d_N^N)$ solve the algebraic system

$$\mathcal{M}([\boldsymbol{c}^N,\boldsymbol{d}^N]) = \mathbf{0}\,,$$

with $\mathcal{M} : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ being a continuous mapping:

$$\begin{aligned} \mathcal{M}_k([\boldsymbol{c}^N,\boldsymbol{d}^N]) &:= \varepsilon\int_\Omega \nabla p^N\cdot\nabla\alpha^k\,\mathrm{d}\boldsymbol{x} + \int_\Omega \alpha^k\operatorname{div}\boldsymbol{v}^N\,\mathrm{d}\boldsymbol{x}\,, \quad k = 1,2,\ldots,N \\ \mathcal{M}_{N+l}([\boldsymbol{c}^N,\boldsymbol{d}^N]) &:= \eta\int_\Omega |\boldsymbol{u}^N|^{2r'-2}\boldsymbol{u}^N\cdot\boldsymbol{a}^l\,\mathrm{d}\boldsymbol{x} - \int_\Omega (\boldsymbol{v}^N\otimes\boldsymbol{v}^N) : \nabla\boldsymbol{a}^l\,\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_\Omega (\operatorname{div}\boldsymbol{u}^N)\boldsymbol{u}^N\cdot\boldsymbol{a}^l\,\mathrm{d}\boldsymbol{x} \\ &\quad + \int_\Omega \nu(p^N,|\mathbf{D}(\boldsymbol{v}^N)|^2)\mathbf{D}(\boldsymbol{v}^N) : \mathbf{D}(\boldsymbol{a}^l)\,\mathrm{d}\boldsymbol{x} - \int_\Omega p^N\operatorname{div}\boldsymbol{a}^l\,\mathrm{d}\boldsymbol{x} - \langle\boldsymbol{b},\boldsymbol{a}^l\rangle\,, \quad l = 1,2,\ldots,N\,. \end{aligned}$$

The basic estimate is obtained by testing the equation by $(p^N, \boldsymbol{u}^N)$ as follows. First, realize that (recall $\operatorname{div} \boldsymbol{v}^N = \operatorname{div} \boldsymbol{u}^N$)

$$\begin{aligned}\mathcal{M}([\boldsymbol{c}^N, \boldsymbol{d}^N]) \cdot ([\boldsymbol{c}^N, \boldsymbol{d}^N]) &= \varepsilon \, ||\nabla p^N||_2^2 + \eta \, ||\boldsymbol{u}^N||_{2r'}^{2r'} \overbrace{- \int_\Omega (\boldsymbol{v}^N \otimes \boldsymbol{v}^N) : \nabla \boldsymbol{u}^N \, \mathrm{d}\boldsymbol{x} - \frac{1}{2} \int_\Omega (\operatorname{div} \boldsymbol{u}^N) \, |\boldsymbol{u}^N|^2 \, \mathrm{d}\boldsymbol{x}}^{=: I_{\mathrm{conv}}} \\ &\quad + \int_\Omega \nu(p^N, |\mathbf{D}(\boldsymbol{v}^N)|^2) \mathbf{D}(\boldsymbol{v}^N) : \mathbf{D}(\boldsymbol{u}^N) \, \mathrm{d}\boldsymbol{x} - \langle \boldsymbol{b}, \boldsymbol{u}^N \rangle \, .\end{aligned}$$

Since $\frac{1}{2} \int_\Omega (\operatorname{div} \boldsymbol{u}^N) \, |\boldsymbol{u}^N|^2 \, \mathrm{d}\boldsymbol{x} = - \int_\Omega (\boldsymbol{u}^N \otimes \boldsymbol{u}^N) : \nabla \boldsymbol{u}^N \, \mathrm{d}\boldsymbol{x}$, it follows that

$$I_{\mathrm{conv}} = - \int_\Omega (\boldsymbol{\Phi} \otimes \boldsymbol{\Phi} + \boldsymbol{\Phi} \otimes \boldsymbol{u}^N + \boldsymbol{u}^N \otimes \boldsymbol{\Phi}) : \nabla \boldsymbol{u}^N \, \mathrm{d}\boldsymbol{x} \, ,$$

which implies (using Hölder's, Korn's and embeddings inequalities and using $r > \frac{2d}{d+1}$)

$$|I_{\mathrm{conv}}| \leq ||\nabla \boldsymbol{u}^N||_r \left(2 \, ||\boldsymbol{u}^N||_{\frac{rd}{d-r}} \, ||\boldsymbol{\Phi}||_q + ||\boldsymbol{\Phi}||_{2r'}^2 \right) \leq C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r^2 \, ||\boldsymbol{\Phi}||_q + C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r \, ||\boldsymbol{\Phi}||_q^2 \, ,$$

where $q = \frac{dr}{r(d+1)-2d} > 2r'$. Throughout this text, the symbols C denote positive, generally different constants. Further,

$$\begin{aligned}\int_\Omega \nu(p^N, |\mathbf{D}(\boldsymbol{v}^N)|^2) \mathbf{D}(\boldsymbol{v}^N) : \mathbf{D}(\boldsymbol{u}^N) \, \mathrm{d}\boldsymbol{x} &= \int_\Omega \nu(p^N, |\mathbf{D}(\boldsymbol{v}^N)|^2) \mathbf{D}(\boldsymbol{v}^N) : (\mathbf{D}(\boldsymbol{v}^N) - \mathbf{D}(\boldsymbol{\Phi})) \, \mathrm{d}\boldsymbol{x} \\ &\geq \frac{C_1}{2r} \int_\Omega |\mathbf{D}(\boldsymbol{v}^N)|^r \, \mathrm{d}\boldsymbol{x} - \frac{C_1}{2r} |\Omega| - \frac{C_2}{r-1} \int_\Omega (1 + |\mathbf{D}(\boldsymbol{v}^N)|)^{r-1} |\mathbf{D}(\boldsymbol{\Phi})| \, \mathrm{d}\boldsymbol{x} \\ &\geq C \, ||\mathbf{D}(\boldsymbol{u}^N) + \mathbf{D}(\boldsymbol{\Phi})||_r^r - C - C \, ||\mathbf{D}(\boldsymbol{\Phi})||_r \, ||1 + |\mathbf{D}(\boldsymbol{u}^N) + \mathbf{D}(\boldsymbol{\Phi})| \, ||_r^{r-1} \, .\end{aligned}$$

Using $|a+b|^{r-1} \leq |a|^{r-1} + |b|^{r-1}$ due to $r - 1 < 1$, it follows

$$\begin{aligned}\int_\Omega \nu(p^N, |\mathbf{D}(\boldsymbol{v}^N)|^2) \mathbf{D}(\boldsymbol{v}^N) : \mathbf{D}(\boldsymbol{u}^N) \, \mathrm{d}\boldsymbol{x} &\geq C \, ||\mathbf{D}(\boldsymbol{u}^N) + \mathbf{D}(\boldsymbol{\Phi})||_r \left(||\mathbf{D}(\boldsymbol{u}^N)||_r^{r-1} - ||\mathbf{D}(\boldsymbol{\Phi})||_r^{r-1} \right) \\ &\quad - C - C \, ||\mathbf{D}(\boldsymbol{\Phi})||_r \left(1 + ||\mathbf{D}(\boldsymbol{u}^N)||_r^{r-1} + ||\mathbf{D}(\boldsymbol{\Phi})||_r^{r-1} \right) \\ &\geq D \, ||\mathbf{D}(\boldsymbol{u}^N)||_r^r - C \, ||\mathbf{D}(\boldsymbol{\Phi})||_r \, ||\mathbf{D}(\boldsymbol{u}^N)||_r^{r-1} - C \, ||\mathbf{D}(\boldsymbol{\Phi})||_r^{r-1} \, ||\mathbf{D}(\boldsymbol{u}^N)||_r - C \, ||\mathbf{D}(\boldsymbol{\Phi})||_r^r - C \, .\end{aligned}$$

Finally, since $|\langle \boldsymbol{b}, \boldsymbol{u}^N \rangle| \leq C \, ||\boldsymbol{b}||_{-1,r'} \, ||\mathbf{D}(\boldsymbol{u}^N)||_r$ and noticing that there holds $||\nabla p^N||_2 \geq C \, ||p^N||_{1,2}$, we arrive at

$$\begin{aligned}\mathcal{M}([\boldsymbol{c}^N, \boldsymbol{d}^N]) : ([\boldsymbol{c}^N, \boldsymbol{d}^N]) &\geq \varepsilon \, C \, ||p^N||_{1,2}^2 + \eta \, ||\boldsymbol{u}^N||_{2r'}^{2r'} + D \, ||\mathbf{D}(\boldsymbol{u}^N)||_r^r \\ &\quad - C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r^2 \, ||\boldsymbol{\Phi}||_q - C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r \, ||\boldsymbol{\Phi}||_q^2 - C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r^{r-1} \, ||\nabla \boldsymbol{\Phi}||_r \\ &\quad - C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r \, ||\nabla \boldsymbol{\Phi}||_r^{r-1} - C \, ||\nabla \boldsymbol{\Phi}||_r^r - C - C \, ||\mathbf{D}(\boldsymbol{u}^N)||_r \, .\end{aligned}$$

At this point the assumption (17) is recalled and, denoting $\rho := ||\mathbf{D}(\boldsymbol{u}^N)||_r / \lambda$, the following is observed:

$$\begin{aligned}\mathcal{M}([\boldsymbol{c}^N, \boldsymbol{d}^N]) : ([\boldsymbol{c}^N, \boldsymbol{d}^N]) &\geq \varepsilon \, C \, ||p^N||_{1,2}^2 + \eta \, ||\boldsymbol{u}^N||_{2r'}^{2r'} + D \rho^r \lambda^r \\ &\quad - C \rho^2 \lambda^2 H_1 \lambda^{2r-2} - C \rho \lambda H_1 \lambda^{2r-4} - C \rho^{r-1} \lambda^{r-1} H_2 \lambda - C \rho \lambda H_2^{r-1} \lambda^{r-1} - C H_2^r \lambda^r - C \rho \lambda - C \\ &\geq \varepsilon \, C \, ||p^N||_{1,2}^2 + \eta \, ||\boldsymbol{u}^N||_{2r'}^{2r'} + D \rho^r \lambda^r \\ &\quad - C H_1 \rho^2 \lambda^r - C H_1 \rho \lambda^{2r-3} - C H_2 \rho^{r-1} \lambda^r - C H_2^{r-1} \rho \lambda^r - C H_2^r \lambda^r - C \rho \lambda - C \, .\end{aligned}$$

Since $1 \leq \lambda \leq \lambda^r$ and $\lambda^{2r-3} \leq \lambda^r$, this can be rewritten as

$$\begin{aligned}\mathcal{M}([\boldsymbol{c}^N, \boldsymbol{d}^N]) : ([\boldsymbol{c}^N, \boldsymbol{d}^N]) &\geq \varepsilon \, C \, ||p^N||_{1,2}^2 + \eta \, ||\boldsymbol{u}^N||_{2r'}^{2r'} \\ &\quad + \lambda^r \left[\left(\frac{D}{2} \rho^r - C \rho - C \right) + \left(\frac{D}{2} \rho^r - C H_1 \rho^2 - C H_1 \rho - C H_2 \rho^{r-1} - C H_2^{r-1} \rho - C H_2^r \right) \right] \, .\end{aligned}$$

Define $E > 0$ such that $\frac{D}{2}E^r - CE - C \geq 0$. The values of C, D and E define the following constraint, which is assumed to be fulfilled by the constants H_1 and H_2:

$$\frac{D}{2}E^r - (CE^2 + CE)H_1 - CE^{r-1}H_2 - CEH_2^{r-1} - CH_2^r \geq 0\,. \tag{23}$$

Note that, since $\frac{D}{2}E^r > 0$, some H_1, H_2 small enough to meet (23) can be found. Note that the values of C, D, E and consequently H_1 and H_2 depend only on C_1, C_2, r, Ω and $\boldsymbol{b}$.

It follows that the inequality

$$\mathcal{M}([\boldsymbol{c}^N, \boldsymbol{d}^N]) : ([\boldsymbol{c}^N, \boldsymbol{d}^N]) \geq \mathbf{0} \tag{24}$$

holds for any $[\boldsymbol{c}^N, \boldsymbol{d}^N]$, provided that $||\mathbf{D}(\boldsymbol{u}^N)||_r = E$. Moreover, there exists some $C > 0$ independent of ε and η, such that (24) holds also for any $[\boldsymbol{c}^N, \boldsymbol{d}^N]$, provided that $\varepsilon\,||p^N||_{1,2}^2 \geq C$ or provided that $\eta\,||\boldsymbol{u}^N||_{2r'}^{2r'} \geq C$. Applying Brouwer's fixed point theorem, a solution $(p^N, \boldsymbol{v}^N)$ of the Galerkin approximate system is obtained, fulfilling the estimate (21)

$$\varepsilon\,||p^N||_{1,2}^2 + \eta\,||\boldsymbol{v}^N||_{2r'}^{2r'} + ||\mathbf{D}(\boldsymbol{v}^N)||_r \leq C < \infty\,, \tag{25}$$

where C does not depend on ε neither on η. The estimate $(22)_1$

$$||\nu(p^N, |\mathbf{D}(\boldsymbol{v}^N)|^2)\mathbf{D}(\boldsymbol{v}^N)||_{r'} \leq C < \infty \tag{26}$$

then follows from (9).

With the estimates (25)-(26) in hand, the limit passage $N \to \infty$ follows exactly the steps given e.g. in [1]; the compact embedding, the monotonicity (11) and Vitali's theorem are used and a couple $(p, \boldsymbol{v})$ is found, which solves (18)-(20) and fulfills the estimates (21), $(22)_1$.

In order to obtain an estimate for pressure uniform with respect to ε, test the equation (20) with $\boldsymbol{\psi} := \mathcal{B}(|p|^{s-2}p - \frac{1}{|\Omega|}\int_\Omega |p|^{s-2}p\,\mathrm{d}\boldsymbol{x})$, denoting $s := \frac{2rd}{r(d-2)+d}$. Note that

$$\begin{aligned} &||\boldsymbol{\psi}||_{1,s'} \leq 2C_{\mathrm{div},s'}\,||p||_s^{s-1} \\ &||\boldsymbol{\psi}||_{2r'} = ||\boldsymbol{\psi}||_{\frac{ds'}{d-s'}} \leq C\,||\boldsymbol{\psi}||_{1,s'}\,, \qquad r \leq s' \quad \text{and} \quad s \leq r'\,. \end{aligned}$$

Since $\int_\Omega p\,\mathrm{div}\,\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} = ||p||_s^s$, this yields

$$\begin{aligned} ||p||_s^s &= \eta\int_\Omega |\boldsymbol{u}|^{2r'-2}\boldsymbol{u}\cdot\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} - \int_\Omega (\boldsymbol{v}\otimes\boldsymbol{v}) : \nabla\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_\Omega (\mathrm{div}\,\boldsymbol{u})\boldsymbol{u}\cdot\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} + \int_\Omega \nu(p, |\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v}) : \mathbf{D}(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{x} - \langle \boldsymbol{b}, \boldsymbol{\psi}\rangle \\ &\leq \eta\,||\boldsymbol{\psi}||_{2r'}||\boldsymbol{u}||_{2r'}^{2r'-1} + C\,||\boldsymbol{\psi}||_{1,s'}\,||\boldsymbol{v}\otimes\boldsymbol{v}||_s + C\,||\mathbf{D}(\boldsymbol{u})||_r\,||\boldsymbol{\psi}||_{2r'}\,||\boldsymbol{u}||_{2r'} + C\,||\boldsymbol{\psi}||_{1,r}(1 + ||\mathrm{D}(\boldsymbol{v})||_r)^{r-1} \\ &\quad + ||\boldsymbol{b}||_{-1,r'}\,||\boldsymbol{\psi}||_{1,r} \leq C(\eta)\,||\boldsymbol{\psi}||_{1,s'} \quad \leq \quad C(\eta)\,||p||_s^{s-1}\,, \end{aligned}$$

which finally implies $(22)_2$

$$||p||_{\frac{2dr}{r(d-2)+d}} \leq C(\eta) < \infty\,. \tag{27}$$

□

3. Existence theorem

Lemma 1 allows to establish the following results. First, the generalization of Theorem 1 stated in [1] and of Theorem 2.1 stated in [2] can be formulated:

Theorem 2 *Let $\Omega \in \mathcal{C}^{0,1}$, $d \geq 2$ and $\boldsymbol{b} \in \mathbf{W}^{-1,r'}(\Omega)$ be given. Let*

$$\frac{2d}{d+1} < r < \min\left\{2, \frac{3d}{d+2}\right\}$$

and the assumptions **A1** *and* **A2** *be satisfied. Let there exist $\lambda \geq 1$ and $\boldsymbol{\Phi} \in \mathbf{W}^{1,r}(\Omega)$ fulfilling* (17)*, with H_1 and H_2 meeting the inequality* (23)*.*

Then there exists at least one weak solution $(p, \boldsymbol{v})$ *to Problem (P) such that*

$$\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{\Phi}, \qquad (p, \boldsymbol{u}) \in \mathrm{L}_0^{\frac{dr}{2(d-r)}}(\Omega) \times \mathbf{W}^{1,r}_{\mathrm{div},0}(\Omega)\,,$$

and such that, for all $\boldsymbol{\psi} \in \mathcal{C}_0^\infty(\Omega)^d$,

$$\int_\Omega \nu(p, |\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v}) : \mathbf{D}(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{x} - \int_\Omega (\boldsymbol{v} \otimes \boldsymbol{v}) : \nabla\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} = \int_\Omega p\,\mathrm{div}\,\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{x} + \langle \boldsymbol{b}, \boldsymbol{\psi}\rangle\,.$$

For the proof, the reader is asked to follow the complete procedure given in [2], starting with the above established Lemma 1 and using the method of Lipschitz approximations of Sobolev functions, developed in [3, 4].

The assumptions (17) on the non-homogeneous Dirichlet boundary condition contains, deliberately, the "free" parameter $\lambda \geq 1$. This allows, due to Lemma 3 in [1], to proceed to the following analogy of Corollary 4 in [1] concerned with the inner flows:

Corollary 3 *Let* Ω *and* $\boldsymbol{b}$ *be the same as in Theorem 2. Let the assumptions* **(A1)** *and* **(A2)** *be satisfied with*

$$d = 3$$

and with

$$2 - \frac{1}{d} = \frac{5}{3} < r < \frac{9}{5} = \frac{3d}{d+2}\,. \tag{28}$$

Let $\boldsymbol{\varphi} = \mathrm{tr}\,\boldsymbol{\Phi}$ *for some* $\boldsymbol{\Phi} \in \mathbf{W}^{1,q}(\Omega) \cap \mathbf{L}^\infty(\Omega)$, $q = \frac{rd}{r(d+1)-2d}$, *where* $\boldsymbol{\varphi}$ *satisfies* (13)

$$\boldsymbol{\varphi} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega\,.$$

Then there is at least one weak solution to Problem (P).

A short proof given in [1] is reproduced here. The goal is to find $\boldsymbol{\Phi}^\eta$, $\eta \in (0, 1\rangle$ and $\lambda \geq 1$ such that the condition (17) is fulfilled, i. e.

$$||\boldsymbol{\Phi}^\eta||_{\frac{rd}{r(d+1)-2d}} \leq H_1\lambda^{r-2}\,, \tag{29}$$

$$||\boldsymbol{\Phi}^\eta||_{1,r} \leq H_2\lambda\,. \tag{30}$$

Then the assertion follows from Theorem 2.

For any $\eta \in (0, 1\rangle$, Lemma 3 in [1] gives a suitable extension $\boldsymbol{\Phi}^\eta$ of the boundary data $\boldsymbol{\varphi}$ and the estimate

$$||\boldsymbol{\Phi}^\eta||_q < H\eta^{\frac{1}{q}}\,, \tag{31}$$

$$||\boldsymbol{\Phi}^\eta||_{1,q} < H\eta^{\frac{1}{q}-1}\,, \tag{32}$$

where $q \in (0, \infty)$ and where H depends only on Ω and $\boldsymbol{\Phi}$. Since $r > 2 - \frac{1}{d}$, an s can be found such that

$$\frac{r-1}{r} < s < \frac{r(d+1)-2d}{rd(2-r)}\,.$$

Setting $\lambda := \eta^{-s}$ this means that for any positive constants H, H_1 and H_2, suitable $\eta \in (0, 1\rangle$ can be found such that

$$H_1\lambda^{r-2} = H_1\eta^{s(2-r)} > H\eta^{\frac{r(d+1)-2d}{rd}}\,,$$
$$H_2\lambda = H_2\eta^{-s} > H\eta^{\frac{1-r}{r}}\,.$$

For such η, the assertions (29)-(30) follow from (31) and (32).
□

4. Further notes

Note that in comparison to Theorem 1 in [1], its assumption (15) is not of any use here and is simply missing in Lemma 1 and Theorem 2. This is, however, not a generalization of the previous result but merely a correction of a mistake. The energy estimates procedure provided in [1] is formulated in terms of $\boldsymbol{v}^N$ instead of $\boldsymbol{u}^N$, which is (in the context of applying Brouwer's fixed point theorem) not correct. The author apologizes for this inconvenience.

Note that the constraint $r > 2 - \frac{1}{d}$ does not allow to extend the result for inner flows in case of two dimensions, because $2 - \frac{1}{d} = \frac{3}{2} = \frac{3d}{d+2}$. In three dimensions, while the "homogeneous Dirichlet" Theorem 2.1 in [2] holds for r down to $\frac{2d}{d+2} = \frac{6}{5}$, the "small data" Theorem 2 requires $\frac{2d}{d+1} = \frac{3}{2} < r$ and the "inner flows" Corollary 3 assumes $2 - \frac{1}{d} = \frac{5}{3} < r$.

References

[1] M. Lanzendörfer, "On steady inner flows of an incompressible fluid with the viscosity depending on the pressure and the shear rate ", *Nonlinear Analysis: Real World Applications*, in press, 2008

[2] M. Bulíček, V. Fišerová,"Existence Theory for Steady Flows of Fluids with Pressure and Shear Rate Dependent viscosity, for low values of the power-law index", *Zeitschrift f. Analysis und Anwendungten*, accepted, 2008

[3] J. Frehse, J. Málek and M. Steinhauer,"On analysis of steady flows of fluids with shear-dependent viscosity based on the Lipschitz truncation method", *SIAM J. Math. Anal.*, 34(5):1064-1083, 2003

[4] L. Diening, J. Málek and M. Steinhauer,"On Lipschitz truncations of Sobolev functions (with variable exponent) and their selected applications." *ESAIM: Control, Optimisation and Calculus of Variations*, to appear, 2008

Integrace dat na sémantickém webu

doktorand:
ING. ZDEŇKA LINKOVÁ

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

Katedra matematiky
FJFI ČVUT
Trojanova 13
120 00 Praha 2

linkova@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:
Matematické inženýrství

Práce byla podpořena projektem 1ET100300419 programu Informační společnost (Tématického programu II Národního programu výzkumu v ČR: "Inteligentní modely, algoritmy, metody a nástroje pro vytváření sémantického webu") a výzkumným záměrem AV0Z10300504 "Informatika pro informační společnost: Modely, algoritmy, aplikace".

Abstrakt

V tomto příspěvku je popsán přístup k virtuální integraci dat využívající současných principů, metod a nástrojů sémantického webu. Přístup pracuje s daty ve formátu RDF a předpokládá dostupnost ontologií, které je popisují. Ontologie jsou základem pro všechny kroky prezentovaného integračního procesu. Jsou využity jak k určení vztahů mezi daty a poskytovaným integrovaným pohledem, tak i k zápisu nalezených korespondencí. Ty jsou dále použity při zpracování dotazů kladených na integrovaná data.

1. Úvod

Úloha zpracování dat z různých (i distribuovaných) datových zdrojů je známa více než 40 let. Tato úloha je označována jako integrace dat a je předmětem mnoha výzkumných prací a projektů zabývajících se celou škálou typů dat - od dat relačních databází přes obecná (heterogenní) data. Současným velmi rozšířeným tématem je integrace dat pocházejících z webu, případně dat sémantického webu.

V případě webových dat je obvykle používána tzv. virtuální integrace dat [18]. Tento přístup je někdy také označován jako integrace pomocí pohledů či pomocí mediátorů. Je založený na tom, že se na data poskytne globální integrovaný pohled (který je ovšem virtuální), místo aby byla úloha řešena vytvořením nového materializovaného zdroje. Definovaný pohled zprostředkovává přístup k datům, která zůstávají fyzicky uložena v původních zdrojích, nicméně díky němu je možné původní data zpracovávat takovým způsobem, jako kdyby byla uložena na jednom místě, v jednom zdroji, v jednom prostředí, se stejným schématem atd.

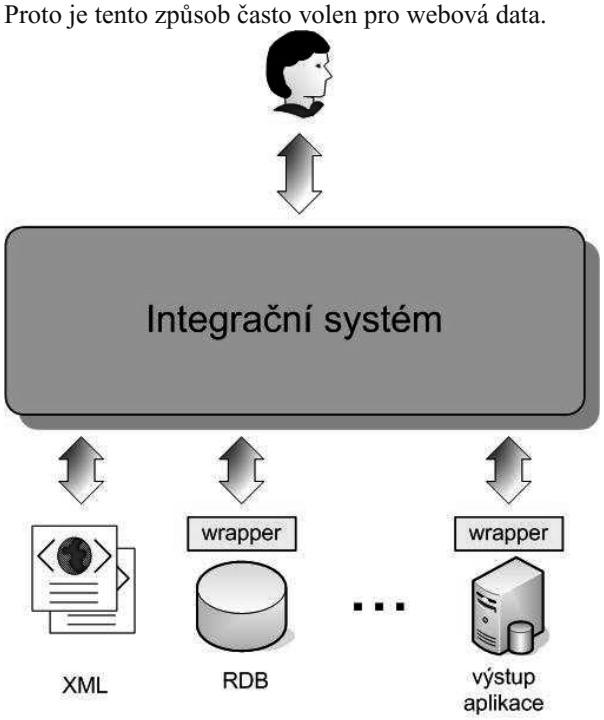
Abychom více omezili obecný typ dat, která chceme integrovat, zaměříme se na data sémantického webu. Integrace takovýchto dat může vycházet z toho, že na sémantickém webu by měla být počítačově zpracovávatelná data. Současnými prostředky a technikami, které jsou využívány k podpoře této myšlenky je jazyk XML, model RDF a OWL ontologie. Na základě hlavní motivace sémantického webu - umožnit zpracování dat bez nutnosti lidského zásahu, mohou tedy přístupy řešení integrace založené na těchto principech očekávat lepší zautomatizování řešené úlohy.

Současné projekty v této oblasti se zaměřují hlavně na využití ontologií. Ontologie mohou být použity v mnoha krocích integračního procesu. Nejčastěji jsou ovšem využity ve fázi hledání korespondencí mezi integrovanými daty. Tento článek popisuje přístup, ve kterém jsou ontologie kromě výše uvedeného použity také k definování nalezených korespondencí. Součástí popisu přístupu je nejen jak získat potřebné korespondence a jakým způsobem je v ontologii zapsat, ale také jak je poté využít při zpracování dotazů.

Článek je organizován následovně: Část 2 poskytuje základní popis obecného přístupu virtuální integrace dat, v podrobnostech se pak dále orientuje na přístup založený na ontologiích a prezentuje ideu využití ontologie jako prostředku k popisu vztahů mezi jednotlivými elementy zdrojů. Část 3 pak popsaného přístupu využívá při zpracování dotazů. Srovnání s jinými ontologicky zaměřenými přístupy je předmětem části 4. Celý článek shrnuje část 5.

2. Integrace dat s využitím ontologií

Běžným způsobem jak kombinovat data pocházející z velkého množství zdrojů nebo ze zdrojů s relativně často se měnícím obsahem je virtuální integrace dat. V takovém přístupu řešení úlohy integrace zůstávají data uložena v původních zdrojích a přístup k nim je umožněn prostřednictvím integrovaného pohledu nebo pomocí rozhraní integračního systému, který takový pohled poskytuje. Z této myšlenky vyplývá hlavní výhoda přístupu: nevytváří se kopie dat v novém materializovaném zdroji - není třeba se zabývat aktuálností dat a nemusí být řešeny paměťové nároky. Proto je tento způsob často volen pro webová data.



Obrázek 1: Virtuální integrace dat

Základem přístupu na Obr. 1. jsou datové zdroje. Vyšší vrstva je reprezentována komponentami označovanými jako *wrappery* - ty přísluší k lokálním zdrojům. Každý wrapper poskytuje přístup ke zdroji a plní funkci rozhranní mezi lokálním prostředím zdroje a prostředím integračního systému.

Vlastní jádro integrace představuje integrační systém, který použije uživatel, chce-li přistupovat k integrovaným datům. Uživatel formuluje své dotazy v prostředí globálního pohledu prezentovaného systémem. Protože však dotaz musí být vyhodnocen nad daty ve zdrojích, jejichž prostředí může být naprosto odlišné, musí systém dotaz nějakým způsobem zpracovat, než jej může vyhodnotit nad zdroji, aby mohl vrátit odpověď uživateli. K umožnění požadované funkcionality jsou definovány korespondence mezi globálním a jednotlivými lokálními prostředími.

Integrační proces je možné vidět jako kolekci úloh, které spolu zajistí žádaný výsledek. Základními kroky ve virtuálně řešené integraci jsou:

- *matching* - úloha hledání korespondencí mezi daty
- *mapování* - způsob, jak zaznamenat nalezené korespondence
- *dotazování* - úloha vyhodnocení dotazů za pomoci informací uložených v mapování

V prezentovaném přístupu jsou uvažována data pocházející ze sémantického webu. Proto jsou předpokládány zdroje obsahující RDF data vyjádřená pomocí syntaxe XML. Dalším důležitým předpokladem jsou OWL ontologie popisující integrované zdroje. Presentovaný přístup těží z dostupných informací obsažených v ontologiích, proto je jejich dostupnost klíčovým předpokladem tohoto způsobu řešení integrace dat.

2.1. Korespondence mezi daty

Při hledání vztahů mezi daty obsaženými v různých datových zdrojích lze nalézt různé typy vzájemných korespondencí. V obecném případě může jeden element jednoho zdroje korespondovat s jedním nebo více jinými elementy (i jiných zdrojů), může korespondovat s kombinací elementů, nebo nemusí korespondovat s žádným jiným elementem. V této souvislosti se obvykle při hledání korespondencí používá pojem *kardinalita*, která pro určitou korespondenci vyjadřuje, kolik elementů mapovaných schémat do vztahu vstupuje. Kardinalita korespondence může být 1:1, 1:N, N:1, N:M. Většina existujících přístupů využívá kardinalit 1:1 nebo 1:N.

Prezentovaný přístup uvažuje vztahy následujících kardinalit:

- **1:1** - při vzájemném porovnávání dvou schémat. Tento případ vyjadřuje, že element jednoho schématu je ve vztahu s jedním elementem druhého schématu.
- **1:N** - při porovnávání jednoho schématu s více dalšími schématy. Tento případ je možné vidět jako množinu korespondencí kardinalit 1:1.

Uvažovaným vztahem mezi daty jsou následující *druhy korespondencí*:

- **Is-a** hierarchický vztah (tj. jeden element je obecnější než druhý, nebo naopak) - tento druh je označen jako $\subseteq$, resp. $\supseteq$.
- **Ekvivalence** mezi elementy - tento druh je označen jako =.
- **Disjunktnost** - tj. mezi elementy není žádná souvislost.

Výsledek úlohy hledání vzájemných vztahů mezi schématy, tedy nalezené korespondence, se často označuje jako *mapování*. Obecně může mapování představovat libovolná struktura. Kromě například používání mapovacích pravidel jako tvrzení pro elementy globálních a lokálních schémat (ať už ve formě 1-1 pravidel či pohledů), které jsou orientovány na konkrétní řešenou úlohu, je možné využít složitější a dokonce standardizovanou strukturu, jenž by pokrývala všechna mapování. K popisu mapování mezi elementy schématu globálního pohledu a schémat lokálních zdrojů bude použita *ontologie OWL*.

K popisu mapování bude v závislosti na typu vztahu využit odpovídající konstrukt. Abstraktním mechanismem pro seskupování popisovaných zdrojů v OWL je třída (class). Zdrojem na webu je jakákoli identifikovatelná entita. Proto bude pojetí `owl:Class` použito pro korespondenci elementů:

- **Is-a** hierarchický vztah, tj. element1 $\subseteq$ element2, lze vyjádřit pomocí podtříd. Příslušným rysem OWL je `rdfs:subClassOf`, který umožňuje vyjádřit, že extenze jedné třídy je podmnožinou extenze jiné třídy.
- Vztah **ekvivalence**, tj. element1 = element2, lze v OWL vyjádřit s `owl:equivalentClass`. `owl:equivalentClass` umožňuje vyjádřit, že dvě třídy mají stejnou extenzi. V tomto případě může být také použit `rdfs:subClassOf` tak, že definujeme element1 jako podtřídu třídy element2 a současně element2 jako podtřídu třídy element1.
- **Disjunktnost** (neboli tvrzení, že extenze jedné třídy nemá žádné společné prvky s extenzí jiné třídy) lze vyjádřit pomocí `owl:disjointWith`.

2.2. Hledání korespondencí v případě sdílené ontologie

Důležitým předpokladem prezentovaného přístupu je dostupnost ontologií, které popisují integrovaná data. Ke každému uvažovanému zdroji je tedy předpokládána existence nějaké popisující ontologie. Situace přitom nemusí být taková, že jeden zdroj je popsán právě jednou ontologií, ale zdroj může být popsán více ontologiemi, přičemž každá z nich jej popisuje pouze částečně, nebo naopak jediná ontologie může popisovat data více zdrojů současně.

V nejjednodušším případě je popis všech zdrojů dostupný v jediné ontologii. Tato ontologie je lokálními zdroji sdílena a pokrývá popis všech lokálních dat. Vztahy mezi elementy není třeba hledat - mohou být nalezeny přímo v této ontologii.

Uvažujeme-li dříve zmíněné typy korespondencí, je možné přístup založit na is-a hierarchii definované sdílenou ontologií. Některé vztahy nemusí být v ontologii vyjádřeny přímo, ale je možné je z ontologie získat využitím tranzitivity is-a vztahu. Je-li například použit přístup k ontologii jako grafu s třídami popisujícími jednotlivé pojmy jako uzly a s orientovanými hranami vyjadřujícími existenci is-a vztahu, korespondenci nepopisuje pouze existující hrana, ale také ohodnocená cesta v grafu.

V případě, že jsou elementy disjunktní, znamená to, že v is-a hierarchii neexistuje žádná cesta a není tedy nutné nějaký vztah hledat. V praxi vede tato situace ke stejnému efektu, jako když je vztah hledán, ale žádný není nalezen. Ovšem je vhodné tuto informaci o disjunktnosti dále uchovávat, protože může být dále využita při rozšiřování přístupu například o další usuzování apod.

2.3. Obecný případ hledání korespondencí založený na ontologiích

Obecně nemusí být ontologie, která by popisovala všechna zpracovávaná data, dostupná. Některé zdroje mohou sdílet některé pojmy, avšak sdílení všech pojmů všemi zdroji nelze předpokládat. Je třeba pracovat obecně s více ontologiemi. Sloučením všech ontologií, které popisují integrované datové zdroje, získáme ”novou” sdílenou ontologii, a tak je tento obecný případ převeden na předchozí.

Slučováním ontologií se zabývá řada výzkumů v oblastech ontology alignment a ontology merging [5] a je tedy možné využít některou ze známých metod. V souvislosti s ontologemi, pojmy alignment a merging spolu úzce souvisí. Pro oba jsou také relevantní úlohy hledání korespondencí (matching) a mapování (mapping). *Ontology alignment* obvykle označuje stanovení binárních vztahů mezi dvěma ontologiemi. To umožňuje definovat způsob, jak tyto

ontologie sloučit. Výsledkem *ontology merging* je nová integrovaná ontologie.

Metodami pro ontology merging, jež je možné při hledání sdílené ontologie použít, se zabývá mnoho výzkumných projektů, například Chimaera [7], PROMPT [12], FCA-MERGE [16], HCONE [6]. V této fázi integračního procesu je možné využít některý z již vytvořených nástrojů. To je výhoda, která vyplývá z faktu, že k zachycení potřebných vztahů využíváme standardizovaný nástroj.

3. Dotazování nad integrovanými daty

Vytvoření mapování uvedené v předchozí kapitole je stěžejní úloha, jejíž výsledek hraje důležitou roli při přístupu k datům pomocí dotazů. Dotazy jsou tvořené nad poskytovaným pohledem (využívají jeho jazyk, schéma apod.). Pro vyhodnocení dotazu nad daty uloženými v lokálních datových zdrojích je třeba původní dotaz nějakým způsobem zpracovat.

Zpracováním dotazu [13] se zabývají dva základní přístupy. Prvním je *přepisování dotazů* (query rewriting) - dotaz je dekomponován na části odpovídající lokálním zdrojům. Ty jsou dále přepsány tak, aby byly vyjádřeny v prostředí příslušného lokálního zdroje. Nad zdrojem jsou pak vzniklé lokální dotazy vyhodnoceny a ze získaných lokálních odpovědí je následně sestavena globální odpověď, která je vrácena jako odpověď na původní (uživatelův) dotaz.

Druhou možností je *odpovídání dotazů* (query answering), která nijak nespecifikuje, jak má být daný dotaz zpracován. Jejím cílem je využít všechny dostupné informace k získání odpovědi na dotaz. Příkladem může být hledání takových dat, u nichž lze dle dostupných znalostí usuzovat, že jsou hledaným výsledkem.

V konkrétní situaci, kterou se zabývá tento článek, jsou uvažována RDF/XML data. RDF/XML data jsou obsažena v původních zdrojích a jsou také prezentována jako data integrovaného pohledu. V obou případech - na lokální i globální úrovni - je tedy jako dotazovací prostředek využíván jazyk SPARQL. Úlohou je globálně vyjádřený kladený dotaz vyjádřit v takové formě, aby bylo možné dotaz vyhodnotit nad zdroji.

K přepsání globálního dotazu do příslušných lokálních subdotazů je využito mapování zachycené v ontologii. Z této ontologie jsou patrné uvažované vztahy mezi pojmy použitými v dotaze a pojmy, které používají lokální zdroje. Přirovnáme-li ontologii ke grafu, ve kterém jsou pojmy zobrazeny jako uzly a vztahy mezi nimi jako ohodnocené hrany, lze přepsání pojmu, který byl v dotaze použit, získat z ontologie následujícím způsobem: všechny pojmy, do kterých vede z daného pojmu cesta ohodnocená uvažovanými vztahy korespondence (např. ekvivalence nebo hierarchie) jsou relevantní a použitelné při přepsání dotazu. Každého kandidáta na přepsání tedy získáme průchodem grafu ontologie od daného pojmu přes hrany korespondencí.

Není nutné využívat pouze hrany vyjadřující ekvivalenci. Například při uvažování hierarchie pojmů lze využít také is-a vztah. Jde přitom o pravidlo, jehož princip je dobře znám například v objektově orientovaném programování: potomek může zastoupit svého předka. Chceme-li uvažovat bohatší škálu korespondencí, je třeba přepisovací mechanismus doplnit o adekvátní mechanismy, aby bylo možné vztahů v přepisování využít.

Zvolený způsob zpracování dotazů v prezentovaném přístupu je popsán následujícími přepisovacími algoritmy. Základní situací je tzv. *jednoduchý dotaz*, tj. dotaz obsahující pouze jednoduchou podmínku na požadovaná data trojice RDF, RDF trojice v dotaze nijak nekombinujeme. Dotaz tedy není třeba rozkládat a získané odpovědi není třeba kombinovat. Globální odpověď získáme přepsáním lokálních odpovědí do globálního prostředí.

Algoritmus 1 *Přepsání jednoduchého dotazu I*

vstupy: globální dotaz, mapovací ontologie
výstupy: lokální dotazy, lokální odpovědi, globální odpověď

- pro každý pojem t generuj ***množinu všech možných přepsání pojmu r(t)***
- použitím všech r(t) generuj ***množinu všech možných přepsání dotazu****, tj. množinu všech lokálních dotazů*
- všechny ***lokální dotazy*** *vyhodnoť nad všemi lokálními zdroji a získej* ***lokální odpovědi***
- využitím reversního přepsání vrať odpovědi v globálním prostředí, tj. ***globální odpověď***

Základní případ nemusí nutně vést k situaci, že by odpovědí musela být jediná trojice RDF. Hledaná data mohou být obsažena ve více zdrojích. Dá-li každý takový zdroj odpověď, jsou všechny tyto získané RDF trojice součástí výsledku, který získáme jejich sjednocením. Může následovat další zpracování výsledku, například odstranění duplicit. V této fázi je též možné, že odhalíme nekonzistenci v datech zdrojů.

Uvedený algoritmus je možné (a je to dokonce žádoucí) dále zefektivňovat. Ptáme-li se všech zdrojů s využitím všech možných přepsání, je jednak

u některých kombinací zdrojů a dotazů předem očekávána prázdná množina s odpovědí a jednak narůstá počet všech možných přepsání dotazu. V případě jednoduchého dotazu s podmínkou na jedinou trojici jde o zanedbatelný fakt, ovšem ve složitějších případech při kombinování trojic či kladení složitějších podmínek objem lokálních dotazů neúnosně narůstá.

V optimalizované formě postupu přepisování je proto zohledněn fakt, zda je daný pojem zdrojem podporován či nikoliv, tedy dotaz je přepisován přímo do formy pro konkrétní datový zdroj. Využity jsou tedy pouze podporované pojmy neboli relevantní k danému zdroji. Takovou informaci je možné získat přímo z ontologie zdroje, schémata zdroje, nebo také předzpracováním zdroje, pokud je požadováno tuto množinu co nejvíce omezit. To je velmi efektivní v případech, kdy je podporovaná ontologie mnohem rozsáhlejší vzhledem ke zdroji, schéma obsahuje velké množství nepovinných prvků a podobně.

Algoritmus 2 *Přepsání jednoduchého dotazu II*

vstupy: globální dotaz, mapovací ontologie, množiny podporovaných pojmů pro každý zdroj
výstupy: lokální dotazy, lokální odpovědi, globální odpověď

- pro každý pojem t generuj ***množinu všech relevantních přepsání pojmu r(t)***
- použitím všech r(t) generuj ***množinu všech relevantních přepsání dotazu****, tj. množinu všech lokálních dotazů*
- všechny ***lokální dotazy*** *vyhodnoť nad všemi lokálními zdroji a získej* ***lokální odpovědi***
- využitím reversního přepsání vrať odpovědi v globálním prostředí, tj. ***globální odpověď***

V případě, že globální dotaz obsahuje složenou podmínku, například při kombinaci více RDF trojic, je nutné složený dotaz nejprve rozdělit do více jednoduchých dotazů s jednoduchými podmínkami. Získané jednoduché odpovědi je nutné před vrácením odpovědi adekvátním způsobem opět složit. Rozklad dotazu na jednoduché dotazy je určen strukturou podmínek na data RDF. Obecně jde například o kombinaci sjednocením či průnikem, adekvátní složení je tedy průnik odpovědí, či jejich sjednocení.

Při rozkladu složeného dotazu však nejde pouze o podmínku specifikovanou v dotaze. Ovlivněn bude také požadovaný výstup - jde-li v dotaze o kombinaci trojic, je nutné, aby v jednoduché odpovědi byly obsaženy prvky, přes které je pak skládána globální složená odpověď. Před vlastním rozkladem dotazu je proto nutné tyto výstupy (pokud nejsou uvedeny) doplnit. Při rozkladu dotazu pak není rozdělena jen vlastní podmínka, ale také výstupy tak, aby každý jednoduchý dotaz obsahoval pouze vzájemně relevantní části.

Algoritmus 3 *Přepsání složeného dotazu*

vstupy: globální dotaz, mapovací ontologie, množiny podporovaných pojmů pro každý zdroj
výstupy: globální jednoduché dotazy, lokální jednoduché dotazy, lokální jednoduché odpovědi, globální odpověď

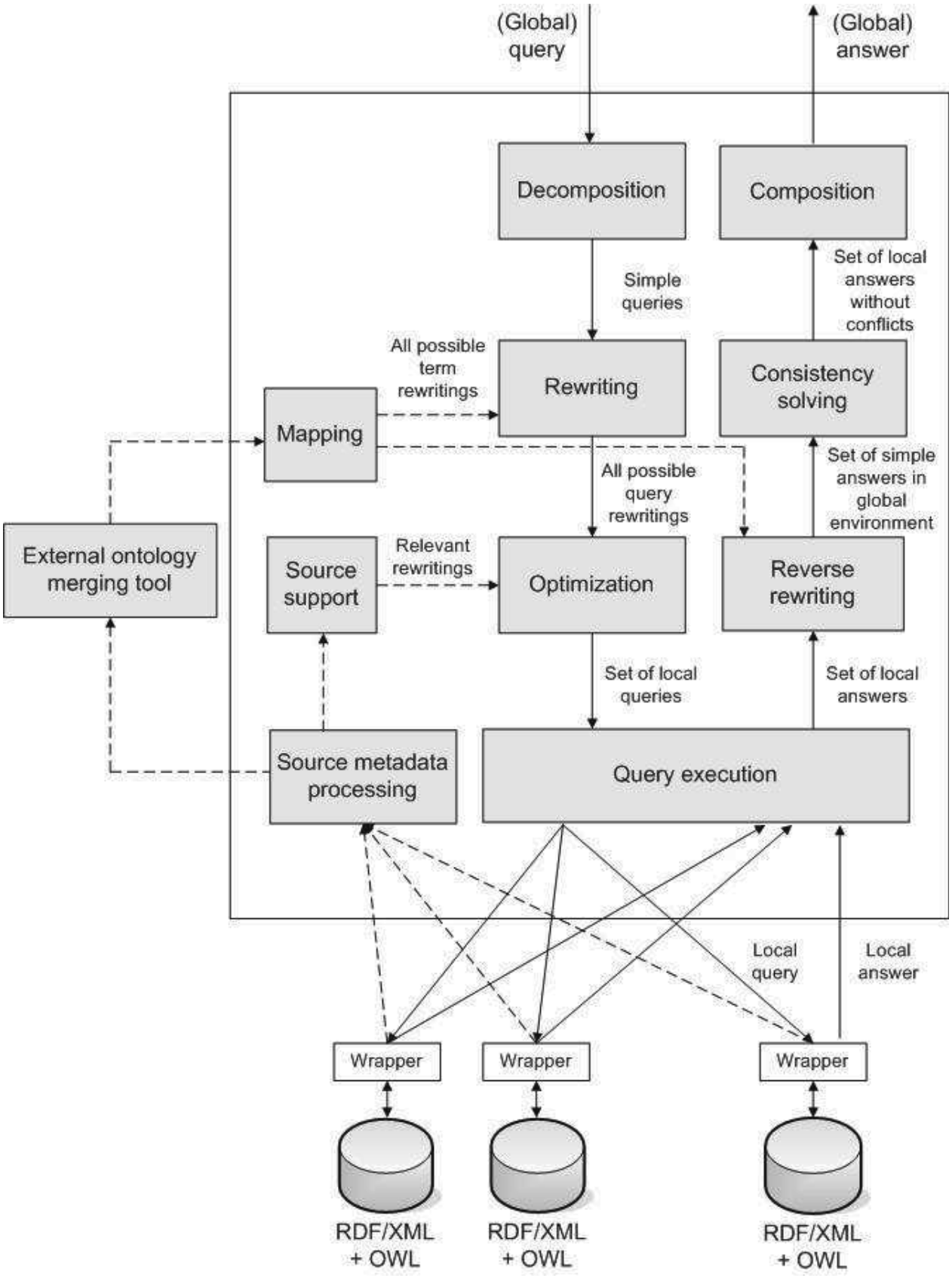
- rozložením složených podmínek na jednoduché rozlož dotaz na ***jednoduché dotazy***
- pro každý jednoduchý dotaz přepisovacím algoritmem získej ***jednoduché odpovědi***
- jednoduchých odpovědí sestav ***globální složenou odpověď***

Celý proces zpracování dotazu pomocí uvedených přepisovacích algoritmů, včetně zpracovávaných dat v jednotlivých fázích je znázorněn na Obr. 2.

4. Srovnání přístupů

Integrace dat je složitá úloha, která zahrnuje celou sadu podúloh, které je třeba řešit, abychom v konečné fázi získali požadovaný výsledek. I jednotlivé fáze procesu integrace jsou značně obsáhlé a speciálně se jimi zabývá řada výzkumných článků.

Přístupy, které se věnují hledání korespondencí [10], [14], [15], se dají klasifikovat dle úrovně informací, kterou o datech využívají. Jedná se o metody pracující na úrovni instancí (korespondence mezi schématy zdrojů), na úrovni používaných pojmů (lingvisticky založené metody, zpracování slov jako řetězců znaků) nebo na úrovni struktury (grafové metody). Velmi častá je ovšem kombinace těchto přístupů a uplatňují se i funkce, vyjadřující podobnosti srovnávaných dat [11], [17], [19].



Obrázek 2: Zpracování dotazu

V tomto pohledu by se mohlo zdát, že přístup popsaný v tomto článku je značně odlišný. Podobné metody jako při hledání korespondencí se však uplatňují při slučování ontologií, jichž prezentovaný přístup využívá. Podobnosti lze tedy nalézt, jsou pouze řešeny na jiné úrovni. Toto převedení úlohy integrace dat na úlohu slučování ontologií [9] mimo jiné umožní využít výsledků jiných projektů (např. vytvořených nástrojů) a více zautomatizovat operace probíhající v procesu.

Na rozdíl od úlohy hledání korespondencí řešenou "tradičním" způsobem, kde je často nutná lidská interakce v konečné fázi při určení skutečně korespondujících dat, jsou všechny korespondence získané z ontologie s určitostí přijaty. Není na ně nahlíženo nejprve jako na kandidáty, neboť zde není žádný odhad korespondencí - všechny z nich jsou v dané ontologii definovány. Je však nutné poznamenat, že i v tomto případě je možné, že je určení korespondencí řešeno lidským zásahem, a to v případě využití externího nástroje při sloučení ontologií. Ačkoliv při odvozování vztahů schémat ze sdílené ontologie žádní kandidáti nevznikají a korespondence jsou přímo určeny, v obecném případě mohou vznikat právě při řešení podúlohy hledání sdílené ontologie pomocí existující metody, která s kandidáty pracuje.

K vyjádření mapování lze použít od jednoduchých 1-1 mapovacích pravidel vyjadřujících přímou korespondenci mezi elementy, přes mapování konceptu na dotaz nebo pohled [2], až po pomocné mapovací struktury. Různé projekty obvykle používají vlastní pojetí mapování, často je následován přístup definice mapování LAV (Local As View), GAV (Global As View), či jejich kombinace GLAV [8].

Zpracování dotazů je pak přímo ovlivněno volbou mapování. Podle složitosti jak uvažovaných dotazů, tak i mapování se odvíjí velmi individuálně konkrétní podoba přístupu k dotazům, například Inverse rule algorithm [3], Bucket algorithm a jeho vylepšení v systému MiniCon [13], či Styx [1].

Podobnost prezentovaného přístupu lze nalézt v případě algoritmu Styx, který také využívá vztahů předek - potomek při zpracování dotazů. Inspirován algoritmem Styx byl algoritmus použitý v systému VirGIS [4] integrujícím geografická data. V něm je udržováno mapování separátně pro každý zdroj a tak je dosaženo dotazování na relevantní pojmy. Na rozdíl od toho přístup prezentovaný v tomto článku pracuje s mapováním jako celkem a separátně udržuje pouze informace o podpoře částí pro každý zdroj. To, že celé mapování je obsaženo v jediné struktuře, umožňuje efektivní obohacování mapování při zjištění dalších korespondencí, při přidání nového zdroje do systému či při reakci na změnu některého ze zdrojů. Vše bez nutnosti přepracovat již zjištěné mapování nebo dokonce mapovat každý zdroj znovu.

5. Závěr

Článek popisuje přístup k řešení úlohy virtuální integrace dat pomocí ontologií. Ontologie je využita nejen ke získání informací při hledání souvislostí mezi daty, ale slouží i jako prostředek k zachycení nalezených korespondencí.

Užití ontologie pro mapování umožňuje řešit změny a obohacování systému doplněním ontologie mapování bez nutnosti zasahovat do již existujících částí. Přináší také možnost znovupoužití i v jiných úlohách či situacích. Navíc, bude-li v budoucnu třeba zachytit i další typy vztahů mezi elementy, může být ontologie dále využita, neboť je schopna zachytit různé typy vztahů.

Mapování popsané v ontologii slouží dále jako klíčový zdroj ve fázi zpracování dotazů. Pro zodpovězení dotazů kladených na integrovaná data je v článku prezentován mechanismus, s nímž je daný dotaz z globální úrovně rozložen a přepsán tak, aby mohl být vyhodnocen nad fyzickými daty. Využitím představeného přístupu integrace je tak možné pracovat s daty na globální úrovni bez toho, aby uživatel musel řešit, ve kterém zdroji a v jaké podobě se dotazovaná data nachází.

Literatura

[1] Bernd A., Beeri C., Fundulaki I. a Scholl M., "Querying XML Sources Using an Ontology-Based Mediator", *On the Move to Meaningful Internet Systems, Confederated International Conferences DOA, CoopIS and ODBASE*, Springer-Verlag, pp. 429–448, 2002.

[2] Calvanese D., De Giacomo G. a Lenzerini M., "Ontology of integration and integration of ontologies", *Proceedings of DL 2001 - Description Logic Workshop*, 2001.

[3] Duschka O. M. a Genesereth M. R., "Answering recursive queries using views", *Proceedings of ACM PODS, ACM Press*, pp. 109–116, 1997.

[4] Essid M., Boucelma O., Lassoued Y. a Colonna, F.-M., "Query Processing in a Geographic Mediation System", *Proceedings of The 12th International Symposium of ACM GIS* Washingtin D.C., 2004.

[5] Kalfoglou Y. a Schorlemmer M., "Ontology

mapping: the state of the art", *The Knowledge Engineering Review 18*, 1, pp. 1–31, 2003.

[6] Kotis K. a Vouros G. A., "The HCONE Approach to Ontology Merging", *ESWS*, LNCS 3053, Springer, pp. 137–151, 2004.

[7] McGuinness D. L., Fikes R., Rice J. a Wilder S., "An Environment for Merging and Testing Large Ontologies", *Proceedings of the Seventh International Conference*, 2000.

[8] Lenzerini M., "Data Integration: A Theoretical Perspective", *Proceedings of the 21st ACM SIGMOD - SIGACT - SIGART symposium on Principles of database systems*, ACM Press, pp. 233–246, 2002.

[9] Linková Z., "Ontology-Based Schema Integration", *SOFSEM 2007. Theory and Practice of Computer Science*, Vol.: 2, Institute of Computer Science AS CR, Prague, pp. 71–80,2007.

[10] Mitra P., Wiederhold G. a Jannink J., "Semi-automatic integration of knowledge sources", *Proceeding of the 2nd Int. Conf. On Information FUSION'99*, 1999.

[11] Nottelmann H. a Straccia U., "Information retrieval and machine learning for probabilistic schema matching", *Inf. Process. Manage. 43*, 3, pp. 552–576, 2007.

[12] Noy F. N. a Musen M. A., "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment", *AAAI/IAAI*, pp. 450–455, 2000.

[13] Pottinger R. a Levy A., "A Scalable Algorithm for Answering Queries Using Views", *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, 2000.

[14] Rahm E. a Bernstein P. A., "A survey of approaches to automatic schema matching", *VLDB Journal: Very Large Data Bases 10*, 4, pp. 334–350, 2001.

[15] Shvaiko P. a Euzenat J., "A survey of schema-based matching approaches", *3730*, pp. 146–171, 2005.

[16] Stumme G. a Maedche A., "FCA-MERGE: Bottom-Up Merging of Ontologies", *IJCAI*, pp. 225-234, 2001.

[17] Su X. a Gulla J. A., "An information retrieval approach to ontology mapping", *Data & Knowledge Engineering 58*, 1, pp. 47–69, 2006.

[18] Ullman J. D., "Information integration using logical views", *Theoretical Computer Science 239*, pp. 189–210, 2000.

[19] Yi S., Huang B. a Chan W. T., "Xml application schema matching using similarity measure and relaxation labeling", *Inf. Sci. 169*, 1-2, pp. 27–46, 2005.

Fitness Landscape in Genetic Algorithms

Post-Graduate Student:
MGR. JAROSLAV MORAVEC
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25
118 00 Prague, Czech Republic
jaroslav.moravec@gmail.com

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

Field of Study:
Theoretical Informatics

The author thanks his suppervisor Martin Holena for support of works on this paper. The present work was supported by the Czech Science Foundation under the contract no. 201/05/H014.

Abstract

This paper provide introduction to genetic algorithms and to fitness landscape. It also gives a survey of fitness landscape approximation techniques. Principles of genetic algorithm are described followed by characterization of fitness landscape including its basic features. Summary of improving genetic algorithms performance by approximation of fitness landscape is given including survey of often used approximation models.

1. Introduction

Genetic algorithm can be seen as a tool for solving optimization problems. It is very robust and can be applied to many complicated problems. Robustness of genetic algorithm is paid by computational complexity. This can be partially reduced in some cases with approximation techniques. In second part of this paper genetic algorithm itself is explained. Third part is introduction to fitness landscape generally and in the fourth part some exact approaches to fitness landscape approximation are introduced. At the end of this paper, future work is discussed.

2. Genetic Algorithm

The main idea of genetic algorithms is based on Darwin's Evolution theory. According this theory all animals and humans are developed from primitive organisms. The basic principle of this theory is based on natural selection. The natural selection says that individuals who are better adapted to surrounding conditions have grater chance of survival and therefore grater chance to reproduce themselves. So after many generations there will be a population where these better individuals will predominate. By crossing and mutating new individuals arise. The nature actually solves an optimization problem by this. It tries find out the optimal solution of the fitness function or in other words an organism which is well adapt to surrounding conditions.

In the nature, there is another very important mechanism, it is the genetics. The information about parent is passed to the offspring coded in molecule of deoxyribonucleic acid - DNA. In genetic algorithms there are many possibilities how to code solutions. Often used coding is real coding where solution is represented as a sequence of real numbers. Second possibility is to code solution as a sequence of values which are taken from finite sets, the binary coding belongs to this class. Binary coding means that all components of the sequence are taken from the set of size two.

Before we describe the genetic algorithm itself, let us clear some terminology. Every solution is called phenotype. Coded phenotype we call genotype. For coding it is usually used binary string of fixed length. This mapping should be explicit at least from genotype to phenotype. Every single genotype we call individual and set of individuals which we will work with we call population. An offspring rise from the population by applying the selection and genetic operators (mutation and recombination) and that offspring became a new generation by replacing the old population.

2.1. Basic algorithm

Following algorithm represents basic genetic algorithm.

1. initialization
2. evaluation
3. while(stopping criterion)

(a) selection
(b) recombination
(c) mutation
(d) evaluation

4. end while

We describe all steps in detail now.

2.1.1 Initialization: In initialization the first generation is created. Every single genotype in population is randomly generated. When there is a chance of receiving an inadmissible solution one can randomly generate phenotypes and then transform them to genotypes. Population size is usually constant throw the computation and it is important parameter of the algorithm.

2.1.2 Stopping criterion: Most commonly used stopping criterion is reaching certain number of generations. Sometimes it is convenient to stop the algorithm after some predefined time and get the best solution found till that point. In some cases, the needed level of fitness is known and the algorithm is stopped after such solution is found.

2.1.3 Evaluation: Evaluation is a simple computing of the fitness function value for each individual in the population. Fitness functions often represent complex problems and therefore this step of the genetic algorithm takes usually the most of computing time.

2.1.4 Selection: Selection should ensure that better individuals will survive to the next generation and worse individuals do not. Here again, there are a lot of strategies to choose from, we will describe the famous one named roulette wheel. In this approach, each new individual is chosen from the old population randomly. Chances of individuals in the old population to be chosen are in the same ratio as their fitness values.

2.1.5 Mutation: A realization of mutation depends on used coding. In the case of binary coding, one position in the string is randomly chosen and its value is changed from 0 to 1 or from 1 to 0. In the case of coding by finite sets new value is chosen randomly. In the case of real coding, new value is often chosen randomly with Gaussian probability distribution.

2.1.6 Recombination: Recombination operator represents principle when the genetic code of childe is a combination of genetic information of its parents. We hope, by doing this that we get a better solution then its parents are. The most widely used recombination is so called one point crossover. One point crossover randomly chooses number $i \in \{1, L-1\}$ and from patents $(x_1, ..., x_L)$ and $(y_1, ..., y_L)$ makes children $(x_1, ..., x_i, y_{i+1}, ...y_L)$ and $(y_1, ..., y_i, x_{i+1}, ...x_L)$ as you can see in figure 1.

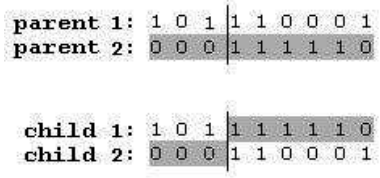


Figure 1: Example of one point crossover for binary string of length 9.

3. Fitness landscape

In this section, we will describe fitness landscape and some of its features. Before we define fitness landscape, we have to introduce configuration space first. More details about configuration spaces and fitness landscapes can be found in [11].

3.1. Configuration space

In fitness landscape, mutual distance relations between data points, and therefore between their fitness values, are very important. For formalizing this, we now define configuration space.

Definition 1 *Configuration space C is pair (X, d). Here d stands for a distance measure and X denotes the set of all coded solutions.*

When fitness function is a function of real variables, then X is a set of vectors of real numbers and d could be Euclidean distance measure. Corresponding configuration space is then Euclidean space. When inputs for fitness function are values from finite sets, Euclidean distance measure can not be used. For dealing with this, we first define Neighborhood structure.

Definition 2 *Neighborhood structure for individual s and operator m is the set of individuals $N_m(s) \subseteq X$ that can be reached from s by single application of a genetic operator m.*

When there is no need of distinguishing between operators is not necessary, we omit the index and write simply $N(s)$. Now we can define the distance measure on X induced by operator m.

Definition 3 *Function* $d_m : X \times X \rightarrow \Re \cup \infty$ *is a distance measure on* X *induced by operator* m *when* $\forall s, t, u \in X$ *following conditions hold:*

- $d_m(s,t) \geq 0$
- $d_m(s,t) = 0 \Leftrightarrow s = t$
- $d_m(s,t) \leq d_m(s,u) + d_m(u,t)$
- $d_m(s,t) = 1 \Leftrightarrow t \in N_m(s)$

First three conditions (non negativity, definiteness and triangular inequality) hold for every distance measure, the forth condition represent connection with genetic operator. X (set of vectors of values from finite sets) together with distance measure on X form configuration space called combinatorial space. Combinatorial spaces can be represent with graphs, where vertices represent individuals and edge (a, b) means that individual b can be reached from a by single application of genetic operator.

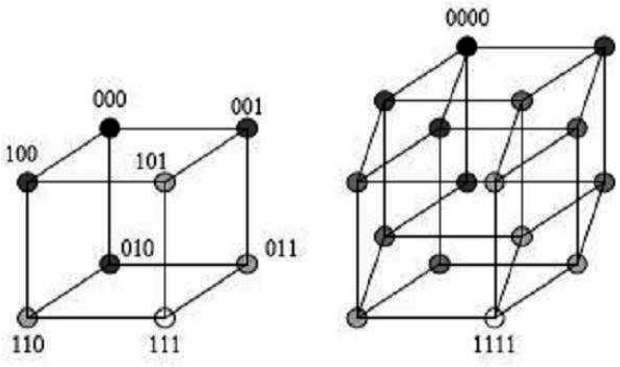


Figure 2: Configuration space for binary string of length 3 (left) and for binary string of length 4 (right).

In figure 2, two examples of combinatorial configuration spaces are shown. On the left, there is a graph of space for binary string of length 3 and traditional mutation which change one bit. Such a graph forms a three dimensional cube, on the right site of the picture is case for coding by binary strings of length 4 which form a four dimensional cube. Besides binary strings, permutations also form and important group of combinatorial spaces. For more information about combinatorial spaces and permutation problems see [7].

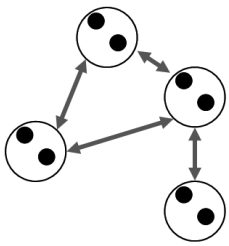


Figure 3: Illustration of a graph of recombination configuration space with complex vertices.

In case of recombination as a genetic operator, situation is more complicated because then we have more then one parent and more then one child, usually two parents and two children. For dealing with this, extended vertices can be used, one vertex in a graph then represent set of individuals (figure 3).

Second possibility is make edges more complex. Edge then became hyperedge which is subset of set of vertices and represents all individuals which can be results of recombination of certain parents (figure 4).

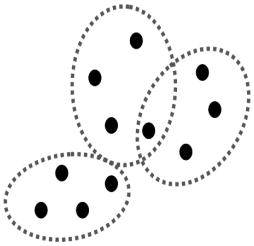


Figure 4: Illustration of a graph of recombination configuration space with complex edges.

Disadvantage of this approach is that it can not be recognized which parents belongs to which children. By adding mapping from sets of parents to hyperedges, which solves the problem, one get structures that are in literature known as P-structures (figure 5).

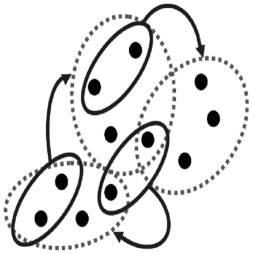


Figure 5: Illustration of a graph of recombination configuration space using P-structures.

3.2. Fitness landscape

A configuration space together with a fitness function form Fitness landscape.

Definition 4 *Fitness landscape is a triple* (X, f, d) *where* X *denotes set of all coded solutions,* f *is a fitness function and* d *stands for a distance measure.*

From this point, it can be seen that changing the operator has a big influence on the fitness landscape. Changing the operator means changing neighborhood structures and therefore the position and mutual distances of local optimums.

Fitness landscape for a two dimensional Euclidean configuration space can be seen as a surface with local optimums in peaks/bottoms of hills/valleys, see figure 6. Individuals then can be seen as points on such a landscape and genetic algorithm computing is then a movement of points on the surface.

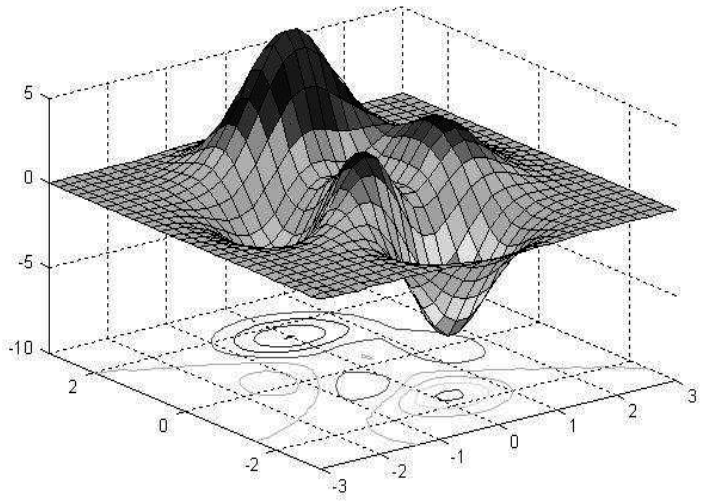


Figure 6: Example of a fitness landscape for a two dimensional Euclidean configuration space.

3.3. Basic features of fitness landscape

Now we describe some basic features of fitness landscape which help us describe different landscapes.

3.3.1 Local optimum: When the configuration space is Euclidean space, local optimum is defined as usual.

Definition 5 *Vector* s *is a local minimum of function* f *if* $\exists\epsilon\forall t\,|t-s|<\epsilon:\ f(s)\leq f(t)$.

The local maximum is defined correspondingly. For combinatorial spaces, the following form of definition is more common.

Definition 6 *For landscape* $L(X, f, d)$ *vector* s *is a local minimum if* $f(s) \leq f(t)t \in N(s)$*, where* $N(s)$ *denotes neighborhood structure for vector* s*.*

Number of local optima in fitness landscape could be used as a measure of fitness rudeness. Generally higher number of local optima indicates a difficult optimization problem. Clearly just one local optimum which is therefore global optimum means usually an easy problem for a genetic algorithm. Actually not just for a genetic algorithm but such problems are easy for other optimization techniques which probably will be able to find optimum in shorter time then a genetic algorithm in that case. However, even among problems with one local optimum, there are some difficult problems.

3.3.2 Basin of attraction: When we talk about local optima, it is not just the number of

them what is matter. Another important feature is the size of the surface area from where optimization algorithm tends to reach certain local optimum - basin of attraction. For a formal definition of a basin of attraction, we need understand to term adaptive walk first. Adaptive walk for minimization is a sequence of points from the configuration space $(z_1, \ldots, z_n)$ defined by steepest descent algorithm where z_1 is starting point and in each step the neighbor $z_{k+1} \in N^k$ is chosen that $f(z_{k+1}) \leq f(z_i)\ \forall z_i \in N^k$ where $N^k = \{z_j : z_j \in N(z_k) \wedge f(z_j) < f(z_k)\}$. Algorithm terminates when z_k is a local minimum. Adaptation of an algorithm for maximization problems is clear. In a Euclidean space, one can use gradient to help choose the direction of the next step of walk, then it is called gradient walk.

Definition 7 *Basin of attraction $B(s)$ for a local optimum s is set of $x \in X$ such that exist adaptation walk $(z_1, \ldots, z_n)$ where $z_1 = x$ and $z_n = s$.*

The size of a basin of attraction corresponds to value of local optimum. Larger basin usually means higher local maximum, respectively a deeper local minimum. For the estimation of basin size one can use average length of the adaptation walks that ends in corresponding local optimum. The length of adaptation walk is the number of elements in the sequence. Small number of large basins indicate an easy problem, on the other hand, lot of small basins indicate a rude fitness landscape.

3.3.3 Examination techniques: Besides already mentioned techniques of examination fitness landscape like estimation of number of local optima or estimation of sizes of basins of attraction, there are other methods. One of them is based on the examination of a random walk, for details see [10]. Another method is spectral analysis. In the case of a Euclidean configuration space can be used traditional Fourier transform for the decomposition of a fitness landscape to a linear combination of trigonometric functions. Similarly in the case of binary coding, Walsh transform can be used for decomposition to a linear combination of Walsh functions. For more details about spectral analyses of fitness landscape, see [8, 9].

4. Fitness landscape approximation

Facing some problem, three levels of approximation can be used. The higher level of approximation is the problem approximation when the original problem is replaced with a problem approximately same but easier to solve. Fitness function approximation is using an approximation model of fitness function instead of the original. We will discuss this approach in details later in this paper. The last approximation level used in evolutionary algorithms is evolutionary approximation. Fitness inheritance and fitness imitation methods belong to this class. In fitness inheritance, individuals from offspring inherit fitness values from their parents. Fitness imitation method divides individuals to clusters. One individual in the center of each cluster is evaluated; fitness values of other individuals in the same cluster are estimated based on that evaluated value. In following, we assume Euclidean configuration space. Another introduction to fitness landscape approximation is given in [1].

4.1. Goals of fitness landscape approximation

The most computationally expensive part of genetic algorithm is usually a population evaluation by computing fitness function. The main idea of fitness landscape approximation is to build a model of fitness landscape and use it instead of the original fitness function. The goal of computation using a fitness landscape model is speed up convergence of genetic algorithm. Fulfilling that leads to either reaching a better solution in the same computational time or reaching a solution of the same quality level in shorter computational time.

The other often mentioned motivations for using a fitness landscape model are absence of explicit model for fitness computation (e.g. evaluation depends on human user) and noisy fitness function. Approximation should smooth out the original noisy fitness function and therefore such a model represents an easier fitness landscape for genetic algorithm.

4.2. Evolution control

One of the main questions is how many individuals should be evaluated by the original fitness function. Here, we want to satisfy two contradictory goals. On one hand, we want to evaluate as few individuals as possible to save computational time, on the other hand we want to evaluate as many individuals as possible to make the model more precise so it do not lead the algorithm to a false optimum. Techniques for solving that, we call evolution control and we will present some of the most widely used principles. More information about evolution control can be found in [4].

4.2.1 Individual based: In individual based evolution control, some individuals are evaluated in each generation. Here, the problem of which individuals should be chosen rises. Again, we want to satisfy two

contradictory goals, exploration and exploitation. One can choose the best individuals for local search in a promising area or individuals from an area where just few original fitness values are known to improve the model in that part and search for new promising areas.

4.2.2 Generation based: In generation based approach all individuals in the population are evaluated in the same time and then the model is used for several generations. This approach can be used with an advantage when a parallel computation of fitness function is possible.

4.2.3 Fixed: Fixed evolution control means that the frequency of evaluation and the number of evaluated individuals are set by parameters of algorithm and they do not change during the computation. This approach is simple and easy to implement comparing to adaptive methods.

4.2.4 Adaptive: In contrast to the fixed evolution control in adaptive evolution control, the frequency of evaluation and the number of evaluated individuals are changing during computation and they are trying to adapt for actual situation. Widely used adaptive generation based evolution control approach is surrogate approach. In surrogate approach, the model is build at the beginning and used till the convergence criteria is reached. Then the whole generation is evaluated by the original fitness function and the model is updated.

4.3. Approximation models

Quality of approximation depends a lot on used model. Survey of often used models is given in this section. Another survey of fitness landscape approximation methods can be found in [5].

4.3.1 Polynomials: The simplest approximation model is polynomial model where the fitness function is approximated by polynomial of certain order based on data set with known fitness values. The most widely used polynomial is third or second order polynomial.

4.3.2 Neural Networks: Neural network is a set of simple computational units (neurons) which are linked to each other. The most widely used type of neural network is multilayered perceptron. In multilayered perceptron, all neurons are organized in layers, where just neurons from neighboring layers are connected so output of one layer is input for the next one. The number of layers and the number of neurons in each layer have to be chosen. Then weights of all connections are set up by process called training, where set of known function values is used. Neural network can be then used for estimating fitness values.

4.3.3 Gaussian processes: This approach builds probabilistic model over data set with known fitness values. Then the model is used for prediction of mean and standard deviation of fitness values of new data. The vector of known function values $\vec{t}_N$ is one sample of multivariate Gaussian distribution with joint probability density $p(\vec{t}_N|\vec{X}_N)$ where $\vec{t}_N = (t_1, t_2, \ldots, t_N)$ and $\vec{X}_N = (x_1, x_2, \ldots, x_N)$ is a vector of inputs. Similarly for $N+1$ data points it is $p(\vec{t}_N, t_{N+1}|\vec{X}_N, x_{N+1})$. By applying the rule $p(A|B) = p(A,B)/p(B)$, we get probability density for t_{N+1} as

$$p(t_{N+1}|\vec{X}_{N+1}, \vec{t}_N) = \frac{p(\vec{t}_{N+1}|\vec{X}_{N+1})}{p(\vec{t}_N|\vec{X}_N)}. \quad (1)$$

From this equation one can get mean as

$$\hat{t}_{N+1} = \vec{k}^T C_N^{-1} \vec{t}_N \quad (2)$$

where correlation matrix C and vector $\vec{k}$ are defined by correlation function $c : X \times X \to \Re$. Example of correlation function follows:

$$c(x_i, x_j) = \alpha \cdot exp\left(-\frac{1}{2}\sum_{k=1}^{n}\frac{(x_{i,k} - x_{j,k})^2}{r_k^2}\right) + \beta \quad (3)$$

where $x_{i,k}$ denotes k-th element of input x_i, r_k is length scale in k-th dimension, α and β are parameters. Elements of C and $\vec{k}$ are $C_{ij} = c(x_i, x_j)$ and $k_i = c(x_i, x_{N+1})$. Variance of $p(t_{N+1}|\vec{X}_{N+1}, \vec{t}_N)$ is given by

$$\sigma^2_{t_{N+1}} = \kappa - \vec{k}^T C_N^{-1} \vec{k} \quad (4)$$

where $\kappa = c(x_{N+1}, x_{N+1})$. More details can be found in [2].

4.3.4 Kriging models: The idea of Kriging model is combining global and local model.

$$U = a(x) + b(x). \quad (5)$$

In this equation U is model of original fitness function, $a(x)$ represents an average behavior along all configuration space and $b(x)$ represents a short distance influence. For global part of the model, the polynomial of low order is often used. Other possibilities are to use trigonometric series or a constant function. The $b(x)$ is defined as follows:

$$b(x) = \sum_{n=1}^{N}[b_n \cdot K(h(x, x_n))] \quad (6)$$

where $h(x, y)$ stands for distance measure of normalized vectors:

$$h(x,y) = \sqrt{\sum_{i=1}^{L} \left(\frac{x_i - y_i}{x_i^{max} - x_i^{min}} \right)^2} \qquad (7)$$

where x_i^{max} respective x_i^{min} are maximum, respective minimum, value in ith dimension. Many functions can be used as K function. The simplest model based on linear function is defined as follow:

$$K(h) = \begin{cases} 1 - \left(\frac{h}{d}\right) & \text{if } h < d, \\ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

Where d is parameter controlling the distance of influence of $b(x)$. When smooth model is required, function K has to satisfy following conditions:

- $K(0) = 1$,
- $K(d) = 0$,
- $\left(\frac{\partial K}{\partial h}\right)_{h=0} = \left(\frac{\partial K}{\partial h}\right)_{h=d} = 0$.

Another possibility is use Gaussian process as $b(x)$. Details about Kriging model can be found in [6, 3].

5. Conclusion

Approximating fitness landscape is approach which speeds up a convergence of genetic algorithm for problems with a markedly time consuming fitness function evaluation. This area is studied in many works yet still there are a lot of questions to answer, for example how to set up parameters or appropriate size of population. Approximation of combinatorial spaces deserve a deeper study as well as spaces with both types of variables, real variables and variables with values from finite sets. Very promising approach for these spaces appears to be Gaussian process model.

References

[1] J. Branke, "Faster Convergence by Means of Fitness Estimation", *Soft Computing*, vol. 9, pp. 13–20, 2005.

[2] D. Buche, N. Schraudolph, P. Koumoutsakos, "Accelerating Evolutionary Algorithms with Gaussian Process Fitness Function Models", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. XX, no. Y, 2004.

[3] M. Emmerich, A. Giotis, M. Ozdemir, T. Back, K. Giannakoglou, "Metamodel-Assisted Evolution Strategies", *Parallel Problem Solving from Nature VII*, LNCS 2439, pp. 361–370, 2002.

[4] Y. Jin, M. Olhofer, B. Sendhoff, "A Framework for Evolutionary Optimization with Approximate Fitness Functions", *IEEE Trans. Evol. Comput.*, vol. 6, pp. 481–494, 2002.

[5] Y. Jin, "A Comprehensive Survey of Fitness Approximation in Evolutionary Computation", *Soft Computing*, vol. 9, pp. 3–12, 2005.

[6] A. Ratle, "Accelerating the Convergence of Evolutionary Algorithms by Fitness Landscape Approximation", *Parallel Problem Solving from Nature V*, vol. 1498/1998, pp. 87, 1998.

[7] Ch. Reidys, P. Stadler, "Copmbinatorial Landscapes", *SIREV*, vol. 44, issue 1, pp. 3–54, 2002.

[8] D. Rockmore, P. Kostelec, W. Hordijk, P. Stadler, "Fast Fourier Transform for Fitness Landscapes", *Appl. Comput. Harmonic Anal.*, 2001.

[9] P. Stadler, "Linear Operators on Correlated Landscapes", *J.Physique*, 4:681–696, 1994.

[10] P. Stadler, "Towards a Theory of Landscapes", *Complex Systems and Binary Networks*, 1995.

[11] P. Stadler, "Fitness Landscapes", *Biological Evolution and Statistical Physics*, pp. 187–207, 2002.

HL7-based Data Exchange in EHR Systems

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, Czech Republic
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, Czech Republic
riha@euromise.cz

Field of Study:
Biomedical Informatics

This work was supported by the project number 1ET200300413 of the Academy of Sciences of the Czech Republic.

Abstract

This paper describes procedures of development of an electronic health record for shared healthcare which include implementation of communication standard HL7 v.3, its application in the environment of existing hospital information systems (HIS) and modeling the semantics of the transferred data. The main part of the solution is so called HL7 broker that serves as a mediator in the communication between the two incorporated systems and implements procedures defined in the HL7 v.3 standard. Data models which describe the systems communicating wit communicating with7 broker are based on the original data models implemented in HISes and are in the proper form, demanded by the HL7 standard. In order to achieve the semantic interoperability of incorporated system the creation of mapping of existing data models to international nomenclatures was necessary. Finally the possibilities of usage of international standards and nomenclatures in comparison to the national ones are discussed.

1. Introduction

My contribution describes the results of the project called ”Information technologies for development of continuous shared health care” supported by Czech national programme ”Information Society” that are covered by the thema of my doctoral thesis – semantic interoperability among systems of electronic health record (EHR). One of goals of the project was to design and implement environment of communicating systems, which would create a base for lifelong EHR of the patient. There are participating two different EHR systems - MUDR EHR [1] and hospital information system (HIS) WinMedicalc 2000 [2]. International standards and nomenclatures were utilised in order to achieve semantic interoperability of concerned systems.

2. Incorporated medical systems

In order to fullfil the main goal of the project an analyzis of the semantics of both participating EHR systems had to be done. The MUDR EHR focuses on efficient, reliable and modular way of data storage and is intented to be part of a more complex system as it does not contain modules engaging in catering services, human resources, drug supply etc. WinMedicalc 2000 is a full featured HIS and for the purpose of the project the interest was limited on its EHR part.

The abbreviation MUDR stands for MUltimedia Distributed Record, which is a pilot solution of structured electronic health record, developed in the Department of Medical Informatics ICS AS CR. MUDR EHR uses a special graph structure called knowledge base and data files to represent the stored information [3]. The WinMedicalc 2000 stores its data in a relational database and thus uses Entity-Relationship model [4] to represent its information model. Preparation of the semantic content of both EHRs in the field of cardiology started from the same modeling basis - the set of important medical attributes for the diagnosis of cardiological patients named the Minimal Data Model for Cardiology [5]. In the MUDR EHR, the modeling process resulted in creating of a part of the knowledge base - the knowledge domain called PATIENT, consisting of basic administrative data, allergy information, family history, social history, subjective information, physical examination, laboratory examination, personal history, treatment information and history of cardio-vascular diseases.

The model of WinMedicalc 2000 system consists of basic administrative information, cardiological

examinations (e.g. ECG examination, Holter monitor, stress test ECG etc.), laboratory examination, physical examination and family history. Each of these data (except administrative information) are connected to a clinical event, that binds together the object and subject of the event, i.e. the patient and the physician. Clinical event contains further information such as place where the event took place (e.g. ward, emergency room). Moreover, WinMedicalc 2000 system covers a broader scope than just clinical data (e.g. catering services, bed management), but these are out of the concern so they are left out.

3. Solving communications

After an initial survey in the field of international communication standards the HL7 v.3 [6] was chosen to enable the data exchange among EHRs. Due to the complexity of the HL7 standard it would be in real life too exhausting to comprehend the whole standard. Therefore the implementation is divided in several parts. The communication was based on:

- creating *local information models* (LIMs) describing the semantic structure of EHR (for this purpose a modelling application named MODELAR was developed)
- establishment of *HL7 brokers* for each information system
- implementation of supporting modules (we call them *LIM fillers*) as parts of the participating systems

A sample communication scenario (see Fig. 1) is based on situation when HIS2 enquires particular data from HIS1 and it is already known which data are going to be transferred. The first step is to retrieve data from the database of HIS1. The LIM filler on the side of HIS1 transforms this data into a LIM message described by a relevant LIM template. LIM templates are described by XML-Schema language and are embedded in the LIM filler. Data proceed in a secure way to the HL7 broker via the SOAP protocol bound to HTTPS protocol using web-services technology. The HL7 broker transforms the data into HL7 message instance according to mapping definitions between the LIM model of HIS1 and HL7 balloted messages. The instance of HL7 message is sent to the receiver of the data which is stated in the header of the LIM message. The accepting HL7 broker transforms the incoming HL7 message into a LIM message according to HIS2. The HIS2 gradually polls (by using web-services) the broker for new data, which in this case will be successful, otherways it gets a message that says that there are no messages. The LIM filler on the side of HIS2 transforms the received LIM message into internal form suitable for storage into its database. The last step of the communication is the storage of received data into HIS2's database.

Hospital information systems contain sensitive data thus the access control and security is one of key issues. In proposed solution a secured HTTP connections are used. The access control is managed by HIS themselves as it was this way before the HL7 communication extensions. All extensions of the HIS developed in the frame of this project are transparent to the hosting HIS as much as possible.

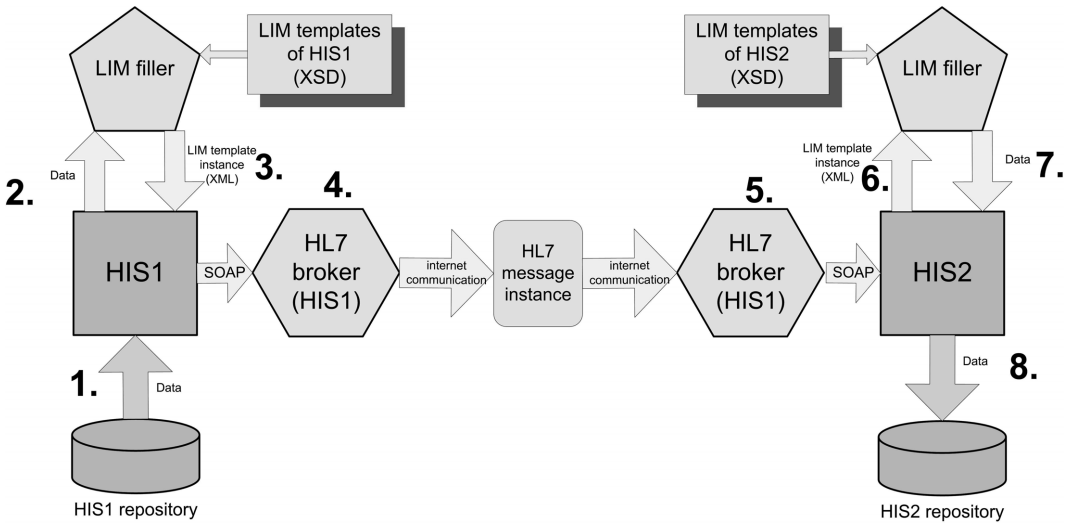


Figure 1: Communication scenario between HIS1 and HIS2.

Connecting a HIS to HL7 communication environment brings a need of dealing with data originating outside of the system. On the other side the system must deal with a new user type or role – the HL7 user. There was a need to store the foreign data separately and mark it clearly that it originates in HL7 communication. In case of querying data from other HIS over HL7 there had to be done an access control exception for testing purposes since the main goal of the project was to design and test the communication possibilities of the HL7 standard. The access control and overall manipulation with data originating from different hospital is gowerned by law, which is still not in the suitable from. Incoming queries have right to read all data about particular patient that are intended to be shared, which are for testing purposes almost all of them. In real life usage a sophisticated access control policy would be needed.

3.1. Describing EHR semantics

The HL7 v.3 standard methodology introduces a reference model (Fig. 2) that should serve as a basis for all semantic objects modelled during the whole process of implementation of communication among EHR systems. Both MUDR and WinMedicalc 2000 systems have been described on a semantic level by classes derived from those in the reference information model. This produced so called local information models (LIMs) of each EHR which are conceptually very close to HL7 D-MIMs (domain message information model). Classes from these models represent collected variables. Moreover, beside the similar concepts both LIMs use also the references to established code systems (LOINC [7], NCLP [8]), giving the possibility of the precise specification of semantics.

3.2. HL7 broker

The main motivation for creating the HL7 broker was to disengage vendors of EHR systems from comprehending and implementing all parts of the HL7 standard, thus saving financial resources. The HL7 broker serves as a configurable communication interface for the EHR system. The configuration is made by a XML file containing the LIM model of a particular EHR.

After creating LIM models for both EHR systems involved in the project, the next step was to produce so called *LIM templates*. These templates consist of classes defined in LIM model which are arranged in a tree structure. Each LIM template represents one integrated part of the EHR system the LIM model describes, e.g. physical examination, medication, ECG data. Having LIM templates the configuration of HL7 broker by mapping classes from LIM models to fragments of balloted HL7 messages could be completed.

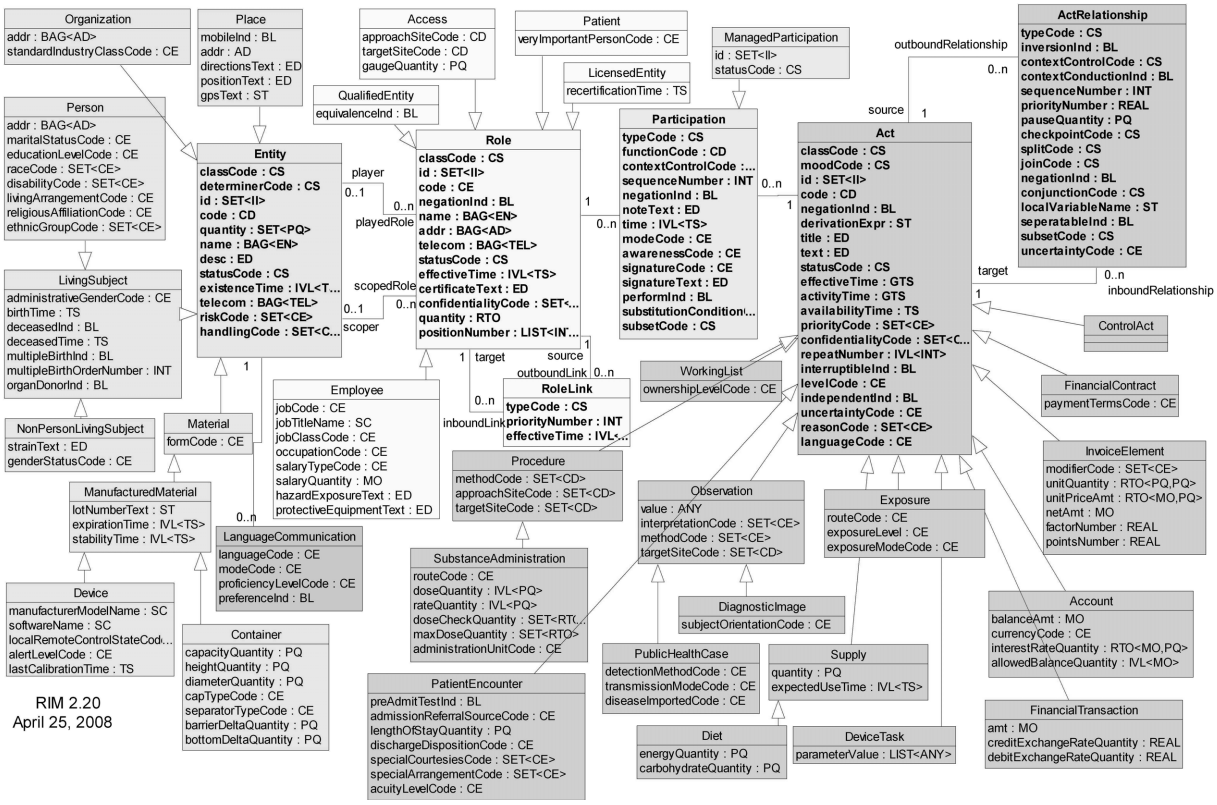


Figure 2: Reference Information Model defined in the HL7 v.3 standard.

The HL7 broker plays a role of an entry point to a HL7 network or some sort of a gateway. The HL7 network consists of HL7 brokers which communicate with each other. The HL7 brokers support peer to peer communitaction and broadcasts are possible as well. New HIS is added into the HL7 newtwork after implementing all three steps mentioned at the beginning of this section. The HL7 broker connected to the new HIS is added to the set of existing HL7 brokers.

3.3. EHR communication modules

Both EHR systems had to be extended by programatic parts supporting the communication with a particular HL7 broker. We call these parts LIM fillers as their main task is to fill in LIM templates with actual data, thus creating LIM message instance.

Secondary task of a LIM filler is communication with the HL7 broker via SOAP protocol. Therefore, a SOAP client had to be implemented on both EHRs. Each filler was created independently but on a similar basis as a pluggable module.

3.4. Classifications and code lists

Uniqueness of term definitions and their precise denomination are necessary for semantic interoperability. We have found that current classification of medical terms is not optimal. Insufficient standardization in medical terminology represents one of the prevailing problems in processing of any kind of medical-related data.

Various classification systems, nomenclatures, thesauri and ontologies have been developed to solve this problem, but the process is complicated by the existence of more than one hundred incompatible systems. The most extensive current project that supports conversions between major classification systems and records relations among terms in heterogeneous sources is the Unified Medical Language System (UMLS) [9].

During the development of MUDR EHR and MDMC, the UMLS Knowledge Source Server was used to 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. More than 50 % are included in SNOMED Clinical Terms [10].

Each information system uses its internal code-lists. There are plenty of them in the information systems and standards. To avoid necessity to implement them in the HL7 broker, a web application enabling each information system developer to define and maintain their own code-lists has been developed. The same mechanism can be used to import the HL7 code-lists. Each code-list is characterised by its name, technical name, version, administrator and user of the code-list (HL7, WinMedicalc, MUDR).

The web application allows the user to define relations between values of individual code-lists describing the possibility to convert value from one code-list to value from the another one. The allowed relation types are equivalence, generalization or specialization.

The entered data about code-lists can be utilized by SOAP method `Translate(val, A, B)`, where `val` is the value from code-list `A` and `B` is the destination code-list. The method returns the value from `B` which is equivalent or generalization of the value `val` from `A`. This method can be used by core of HL7 broker to convert values from messages according to required code-lists.

3.5. Communication interface between MUDR EHR and HL7 broker

Communication between electronic health record and HL7 broker is similar in both participating systems, therefore in the following text the MUDR EHR part will be described.

Communication between MUDR EHR and HL7 broker is based on SSL secured SOAP protocol. The HL7 broker provides several methods (`sendLimMsg()`, `ackLimMsg()`, `getLimMsg()`) for transfer of the data between MUDR EHR and HL7 broker. These methods are exposed by the web-service of the HL7 broker as operations and are described in web-services definiton language (WSDL) file in the following form:

The data are transported in the form of a message described by the LIM template – *LIM message instance*. Several LIM templates are defined, e.g. administrative data, ECG or laboratory results. There are two communication modes - *querying mode* and *passive*.

In the query mode the MUDR EHR receives a special LIM template with a query from the HL7 broker. This LIM template contains only several entered values serving as an identifier of the demanded information – query parameters. After information retrieval from the local database of MUDR EHR, the information 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.

```
...
<portType name='svc-porttype'>
  <operation name='ackLimMsg'>
    <input message='tns:ackLimMsgRequest'/>
    <output message='tns:ackLimMsgResponse'/>
  </operation>

  <operation name='getLimMsg'>
    <input message='tns:getLimMsgRequest'/>
    <output message='tns:getLimMsgResponse'/>
  </operation>

  <operation name='sendLimMsg'>
    <input message='tns:sendLimMsgRequest'/>
    <output message='tns:sendLimMsgResponse'/>
  </operation>
</portType>
...
```

The combination of both modes enables the EHR application triggered by the user request to ask for the data from the other EHR via HL7 broker, wait for the incoming data and store them into its own database structure. Such data should be flagged as externally received.

The result of a query in the EHR initiated by the received LIM template could consist of a 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. Results

Communication between participating EHR systems is realized via the HL7 v.3 communication standard. Local information models describing semantical structure of both EHRs were created in order to support semantic interoperability. Each LIM is derived from the HL7 RIM. The message exchange is realized via HL7 brokers which communicate with corresponding EHRs by using web-services technology based on SOAP protocol.

5. Discussion

The majority of healthcare information systems in the Czech Republic uses so called Data standard of the Ministry of health (DASTA) [8] and National code-list of laboratory items (NCLP)[8], as a communication platform. These standards are developed by the producers from many companies, faculties, research institutions in the Czech Republic. DASTA is based on a predefined limited set of structured data, especially from the field of laboratory examinations, which is possible to transfer by the standard messages. The benefit of the DASTA is its simplicity, allowing an easy implementation of the interface and realization of a communication among information systems. This simplicity however limits its use, in case the information not covered by the actual standard version is to be transferred. The communication of structured general clinical information is not covered satisfactorily by the DASTA and it is usually limited to transfer of the free text messages. On the other side, the possible extension of the standard by another data is much easier on the national level than on the international one.

HL7 v.3 offers the general methodology and tools for the realization of communication between information systems in healthcare and covers this area with a large scale of generality. The large extent of the standard and existing relations to other standards and classifications are a bit demotivating for the developers with the minimal experience with standards of this scale. On the other side, this extensiveness and universality allows to represent the majority of the situations and entities appearing in the data exchange process in healthcare. Thanks to references to external classifications and nomenclatures the HL7 standard provides the method to accurately specify the semantics of the communicated data without the need for ad hoc agreement of communicating parties about the exact meaning of individual elements in the transferred messages.

NCLP and DASTA have only a minimal relations to international classifications and standards. The communication of Czech hospital information systems with other healthcare information systems on the european or international level is not possible without the adherence to the international standards and classifications. Unfortunately, their use in the national environment is very limited without existing Czech localization of a high quality. Such translation would be very expensive and time consuming, but on the other hand it would significantly extend the integration possibilities of Czech eHealth activities into the international context.

6. Conclusion

The structured form of information stored in EHR is an inevitable prerequisite for semantic interoperability establishment among various EHR systems. The research work in the scope of the project ”Information technologies for development of continuous shared health care” demonstrated one possible concept of solving the problem of distributed medical environment. The developed concept is based on international standards and nomenclatures which can be applied as a system for shared lifelong electronic patient’s health documentation.

References

[1] Hanzlicek, P., Spidlen, J., Nagy, M.: Universal Electronic Health Record MUDR. In: Duplaga, M., Zielinski, K., Ingram, D. (eds.) Transformation of Healthcare with Information Technologies. pp. 190–201. IOS Press, Amsterdam (2004)

[2] Medicalc Software s.r.o.: WinMedicalc 2000, http://medicalc.cz/winmedicalc

[3] Spidlen, J.: Databazova reprezentace medicinskych informaci a lekarskych doporuceni (In Czech). Master Thesis at Faculty of Mathematics and Physics, Charles University in Prague, pp. 32–34, (2002)

[4] Chen, P.: The Entity-Relationship Model – Toward a Unified View of Data. ACM Transactions on Database Systems, vol. 1, no. 1, pp. 9–36. (1976)

[5] Tomeckova, M.: Minimalni datovy model kardiologickeho pacienta (In Czech). Cor et Vasa, vol.44, No. 4, pp. 123. (2002)

[6] HL7 Int. Ballot HL7 v3 – May 2008, http://www.hl7.org

[7] Regenstrief Institute Inc.: Logical Observation Identifiers Names and Codes, http://loinc.org

[8] Ministry of Health of the Czech Republic: Datovy standard MZ CR v.4, http://ciselniky.dasta.mzcr.cz

[9] National Library of Medicine – National Institutes of Health: Unified Medical Language System, http://www.nlm.nih.gov/research/umls

[10] International Health Terminology Standards Development Organisation: Snomed CT, http://www.snomed.org

User Preference and Optimization of Relational Queries

Post-Graduate Student:
RADIM NEDBAL
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic ,

Department of Mathematics
Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Trojanova 13

120 00 Prague, Czech Republic

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, Czech Republic

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 notion of preference poses a new prospect of personalization of database queries. In addition, it can be exploited to optimize query execution. Indeed, a novel optimization technique involving preference is developed, and its algorithm presented.

1. Introduction

Preference provides a modular and declarative means for relaxing and optimizing database queries. It is a concept that needs a special framework for embedding in the relational data model: on the one hand, the framework should be rich enough to capture **various kinds of preference** to provide database users with an expressive language to formulate their wishes, and, on the other hand, robust enough to allow for possibly **conflicting preferences** as the assumption of consistency of complex preferences is hard to fulfill in practical applications.

To reach the above goal we consider sixteen kinds of preferences, some of them allowing for expressing uncertainty. Also, basic preference combiners (Pareto or lexicographic composition) are taken into account.

To embed the notion of preference into relational query languages, a **preference operator**, parameterized by user preference, is defined: it filters out not all the bad results, but only worse results than the best matching alternatives and returns the perfect match if present in the database, otherwise, it delivers best-matching alternatives, but nothing worse!

Optimization strategy of **pushing the preference specification** down the query execution tree is governed by both algebraic properties of the preference operator and logical properties of user preference that always is expressed over a set of possible states of the world. This strategy is based on the assumption that early application of the preference operator reduces intermediate results and thus minimizes the data flow during the query execution.

2. Embedding Preference in Relational Query Languages

2.1. Preference Operator

A new, preference operator is added to the relational algebra. Its expressive power depends on the expressivity of the language for expressing user preference – its single parameter.

Definition 1 (Preference operator) *Let U denote a universe and $W^{\mathcal{P}} \subseteq W$ a set of the most preferred worlds with regard to a preference specification $\mathcal{P}$ over a set W of possible worlds. The preference operator $\omega_{\mathcal{P}}$ is a mapping $\omega_{\mathcal{P}} \colon V \to 2^V$ from a set V of discourse into the powerset 2^V of V:*

$$\omega_{\mathcal{P}}(v) = \{v' \subseteq v | \exists u \in U \exists w \in W^{\mathcal{P}} \colon u \models w \wedge v'\} \ . \quad (1)$$

It is important to point out that the preference specification parameter $\mathcal{P}$ allows for complex preference compound from elementary preferences of various kinds. We take into account *locally optimistic*, *locally pessimistic*, *opportunistic*, and *careful* preferences, whose terminology and motivation has been introduced in [1]. Moreover, we consider another two binary choices: a preference can be strict or non-strict and can be evaluated without or with a ceteris paribus proviso, a concept introduced by von Wright [2]. Altogether, we get sixteen various kinds of preference.

On the one hand, this complex preference specification parameter yields a large expressivity, however, on the other hand, it makes the preference operator absent from algebraic properties fundamental for realizing the algebraic optimization strategy that is based on early application of the most selective operators of relational algebra. Thus a more general technique has to be developed.

2.2. Optimization

Algebraic optimization strategy involving the preference operator must provide a transformation (of a given database query) under which the preference operator, which is the last operator to be applied, is invariant.

Example 1 *Let $\mathcal{R}$ be a database schema and $\mathcal{I}$ its instance consisting of two relation instances I_1, I_2. Suppose a user expresses their requirements through a database query*

$$q(\mathcal{I}) = \pi_X(I_1 \cup I_2) \tag{2}$$

over $\mathcal{I}$ and their preferences (wishes) through a preference specification $\mathcal{P}$ over the set

$$W = 2^{q(\mathcal{I})} \tag{3}$$

of possible worlds. Then, the preference operator $\omega_{\mathcal{P}}\big(q(\mathcal{I})\big)$ evaluated over (2) *returns the best matching alternatives with regard to the user preferences.*

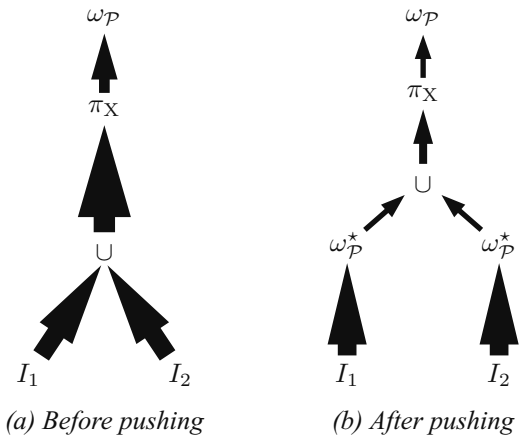
Suppose the preference operator is invariant under the following transformation of $q(\mathcal{I})$ to $q'(\mathcal{I})$:

$$q'(\mathcal{I}) = \pi_X\big(\omega^\star_{\mathcal{P}}(I_1) \cup \omega^\star_{\mathcal{P}}(I_2)\big) \ , \tag{4}$$

where $\omega^\star_{\mathcal{P}}(I_i)$ is a preference operator derivative filtering out "bad" tuples. Then, the preference operator $\omega_{\mathcal{P}}\big(q'(\mathcal{I})\big)$ evaluated over (4) *returns the best matching alternatives with regard to the user preferences:*

$$\omega_{\mathcal{P}}\big(q'(\mathcal{I})\big) = \omega_{\mathcal{P}}\big(q(\mathcal{I})\big) \ .$$

The query execution trees are depicted in Fig. 1, where data flow between the computer's main memory and secondary storage is represented by the drawing width.



(a) Before pushing *(b) After pushing*

Figure 1: *Improving the query plan by pushing the preference operator down the query execution tree*

Supposing that relation instances I_1 and I_2 are too big to fit into the main memory and using the number of the secondary storage I/O's as our measure of cost for an operation, it can be seen that the strategy of pushing the preference operator can improve the performance significantly.

Note that to push the preference specification $\mathcal{P}$ down the expression tree, a special derivative $\omega^\star_{\mathcal{P}}$ of the preference operator $\omega_{\mathcal{P}}$ realizing its filtering potential has been introduced. Unlike the preference operator (cf. 1), it is a mapping $\omega^\star_{\mathcal{P}} \colon V \to V$ from a set V of discourse – a set of all possible tuples over a given relation scheme – into itself. Most importantly, it fulfills the following property:

$$\omega_{\mathcal{P}}\big(\omega^\star_{\mathcal{P}}(I)\big) = \omega_{\mathcal{P}}(I) \ ,$$

i.e., it filters out bad tuples of a given relational instance I without affecting the value of the preference operator.

Furthermore, observe that $\omega_{\mathcal{P}}$ and $\omega^\star_{\mathcal{P}}$ have an identical value of the preference parameter. This value – a user preference $\mathcal{P}$ over W – however, is usually expressed over the result of a query (3). **Does it mean that we need to have computed** (3)**, and thus also** (2)**, before we are able to evaluate** (4)**?** The answer has to be searched for in the definition of the semantics of preference specification [3] and is provided by the following proposition 1.

In brief, a preference specification has the constructive semantics defined by means of a disjunctive logic program (DLP). In the following, $\mathfrak{W}$ stands for the

Herbrand universe for the DLP assigned to a preference specification $\mathcal{P}$, and $g_\mathcal{P}$ for a mapping that can be computed from models of the DLP. Note that models of the DLP can be computed using single exponential time on the cardinality of $\mathfrak{W}$, which , in turn, depends exponentially on the number of elementary preferences composing the preference specification $\mathcal{P}$. This number, however, is supposed to be small, usually between five and ten. Finally, $f_\mathcal{P}$ stands for a mapping that can be expressed as a first order query.

Lemma 1 *Let q denote a database query – a mapping $q\colon \mathrm{inst}(\mathcal{R}) \to \mathrm{inst}(S)$ from a set of database instances over a database schema $\mathcal{R}$ to a set of relation instances over a relation schema S. Given a preference specification $\mathcal{P}$ over a set W of possible worlds, there exist a finite set $\mathfrak{W}$ and a mapping $g_\mathcal{P}\colon 2^\mathfrak{W} \to 2^\mathfrak{W}$ such that the following properties hold for all subsets $\mathfrak{W}'$ of $\mathfrak{W}$ if $W = 2^{q(\mathcal{I})}$:*

$$g_\mathcal{P}(\mathfrak{W}') \subseteq \mathfrak{W}' \ , \tag{5}$$

$$g_\mathcal{P}(\mathfrak{W}') \subseteq f_\mathcal{P}^{-1}(\mathrm{supp}) \subseteq \mathfrak{W}' \Rightarrow W^\mathcal{P} = \langle {\mathfrak{W}'}^\mathcal{P} \rangle_W \ , \tag{6}$$

where $f_\mathcal{P}\colon \mathfrak{W} \to \{\mathrm{unsupp}, \mathrm{supp}\}$ is a function returning supp *for every $\mathfrak{w} \in \mathfrak{W}$ that, loosely speaking, is "supported" by $\mathcal{P}$ over W, and*

$$\langle {\mathfrak{W}'}^\mathcal{P} \rangle_W = \{w \in W |\ \exists \mathfrak{w} \in {\mathfrak{W}'}^\mathcal{P} \colon w \Rightarrow \mathfrak{w}\} \ . \tag{7}$$

Proposition 1 *Suppose $\mathcal{I}$, $\mathfrak{W}$, $f_\mathcal{P}$, and $g_\mathcal{P}$ are as in Lemma 1. Then, the mapping $h_\mathcal{P}\colon 2^\mathfrak{W} \to 2^\mathfrak{W}$:*

$$h_\mathcal{P}(\mathfrak{W}') = \big(\mathfrak{W}' - g_\mathcal{P}(\mathfrak{W}')\big) \cup \big(g_\mathcal{P}(\mathfrak{W}') \cap f_\mathcal{P}^{-1}(\mathrm{supp})\big) \tag{8}$$

has a fixpoint $\mathfrak{W}_\mathrm{fix}$ such that $\mathfrak{W}_\mathrm{fix} \supseteq f_\mathcal{P}^{-1}(\mathrm{supp})$.

Proof: It follows readily from (5) that $\forall \mathfrak{W}' \subseteq \mathfrak{W}\colon h_\mathcal{P}(\mathfrak{W}') \subseteq \mathfrak{W}'$. As $\mathfrak{W}$ is finite, it is clear that $\forall \mathfrak{W}' \subseteq \mathfrak{W}\ \exists n \in \mathbb{N}\ [i \geq n \Rightarrow h^i_\mathcal{P}(\mathfrak{W}') = h^n_\mathcal{P}(\mathfrak{W}')]$, i.e., $h^n_\mathcal{P}(\mathfrak{W}')$ is a fixpoint of $h_\mathcal{P}$. Now the observation: $\forall \mathfrak{W}' \subseteq \mathfrak{W}\ [\mathfrak{W}' \supseteq f_\mathcal{P}^{-1}(\mathrm{supp}) \Rightarrow h_\mathcal{P}(\mathfrak{W}') \supseteq f_\mathcal{P}^{-1}(\mathrm{supp})]$ completes the proof. ∎

The following corollary follows readily from (6) and from the observation

$$h_\mathcal{P}(\mathfrak{W}') = \mathfrak{W}' \iff g_\mathcal{P}(\mathfrak{W}') \subseteq f_\mathcal{P}^{-1}(\mathrm{supp}) \ .$$

Corollary 1 *Suppose q and $\mathcal{P}$ over W are as in Lemma 1. Then, $\mathfrak{W}_\mathrm{fix}$ being the fixpoint from Proposition 1 and $W = 2^{q(\mathcal{I})}$, the following equality holds:*

$$W^\mathcal{P} = \langle \mathfrak{W}^\mathcal{P}_\mathrm{fix} \rangle_W \ .$$

So the answer is: partially. To evaluate the preference operator derivative $\omega^\star_\mathcal{P}$, it suffices to find a relevant part of the query result. Intuitively speaking, this relevant part is subsumed by the fixpoint of (8) (Corollary 1) and computed by stepwise pruning the special set $\mathfrak{W}$ (Proposition 1).

3. An Algorithm

The above corollary is the key to effective computation of (4) in the above example:

Algorithm 5 Preference operator filtering tuples

Input: $q\colon \mathrm{inst}(\mathcal{R}) \to \mathrm{inst}(S), \mathcal{P}, \mathcal{I}$
Output: $\omega^\star_\mathcal{P}(I_1)$

```
1:  W_fix := W
2:  while change do
3:      compute g_P(W_fix)
4:      if ∃w ∈ g_P(W_fix): f_P(w) = unsupp then
5:          remove such w from W_fix
6:      end if
7:  end while
8:  compute W^P_fix
9:  ω*_P(I_1) := I_1
10: for all t ∈ I_1 do
11:     if ∀w ∈ W^P_fix: w ⇒ ¬t then
12:         remove t from ω*_P(I_1)
13:     end if
14: end for
```

On line 1, $\mathfrak{W}$ depends solely on preference specification $\mathcal{P}$. It is independent of the set W over which $\mathcal{P}$ is expressed, and thus it also is independent of the input database instance $\mathcal{I}$. The while block computes a fixpoint of (8): the function $g_\mathcal{P}$ can be computed in exponential time on input $\mathfrak{W}$, and the function $f_\mathcal{P}$ can be expressed as a first order query over $\mathcal{I}$. On line 8, $\mathfrak{W}^\mathcal{P}_\mathrm{fix}$ can be computed in exponential time on input $\mathfrak{W}$. In the for block, the input relation instance I_1 is filtered: on line 11, the logical condition follows from Corollary 1 and analysis of (1) and (7).

4. Related Work

The study of preference in the context of database queries has been originated by Lacroix and Lavency [4]. They, however, don't deal with algebraic optimization. Following their work, *preference datalog* was introduced in [5], where it was shown that 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 attracted broader interest again. Kießling et al. [6, 7, 8, 9, 10, 11] and Chomicki et al. [12, 13, 14, 15] pursued independently a similar, *qualitative* approach within which preference between tuples is specified directly, using binary *preference relations*. They have laid the foundation for preference query optimization that extends established query optimization techniques: preference queries can be evaluated by extended – preference relational algebra. While some transformation laws for queries with preferences were presented in [11, 6], the results presented in [12] are mostly more general.

In brief, Chomicki et al. and Kießling et al. have embedded the concept of preference into relational query languages identically: they have defined an operator parameterized by user preference and returning only the best preference matches. This embedding is similar to ours. However, their operator differs from our preference operator by the parameter: Chomicki et al. and Kießling et al. consider such preference that the operator is partially antimonotonic with respect to its relational argument. By contrast, the preference parameter we consider is more complex and consequently, this property is not fulfilled by the preference operator. As a result, most algebraic properties presented by the above authors don't apply to the preference operator. Specifically, the commutativity and distributivity properties do not hold, and thus the optimization strategy presented in this paper has to rely on different techniques.

Moreover, Chomicki et al. and Kießling et al. are concerned only with one type of preference and don't consider preferences between sets of elements. In terms of logic of preference, they only take into account preferences between singleton worlds[1]. In this sense, their approach is subsumed by the approach presented in this paper, and, in particular, the introduced optimization technique can be applied to the their preference relational algebra.

A special case of the same embedding represents *skyline operator* introduced by Börzsönyi et al. [16]. Some examples of possible rewritings for skyline queries are given but no general rewriting rules are formulated.

[3] is preliminary contribution building on recent advances in logic of preference. Employing non-monotonic reasoning mechanisms, it takes into account various kinds of preferences. The embedding of preference in relational query languages is based on a single preference operator parameterized by a user preference. By contrast to the presented approach, it is assumed that user preference always is expressed over a fixed "universal" domain – a powerset of a universal relation[2]. Consequently, the preference operator has "nice" algebraic properties including conditional commutativity and distributivity. As a result, an optimization strategy of pushing the preference operator down the query expression tree could been developed [17].

A slightly different goal is pursued in [18], where the relational data model is extended to incorporate partial orderings into data domains. The *partially ordered relational algebra* (PORA) is defined by allowing the ordering predicate to be used in formulae of the selection operator. PORA provides users with the capability of capturing the semantics of ordered data. A similar approach to preference modelling in the context of web repositories is presented in [19]: a special algebra is developed for expressing complex *web queries*. The queries employ application-specific ranking and ordering relationships over pages and links to filter out and retrieve only the "best" query results. In addition, cost-based optimization is addressed. Also in [20], actual values of an arbitrary attribute are allowed to be partially ordered according to user preference. Accordingly, relational algebra operations, aggregation functions and arithmetic are redefined. However, some of their properties are lost, and the query optimization issues are not discussed.

A comprehensive work on partial order in databases, presenting the partially ordered sets as the basic construct for modelling data and proposing the embedding of the notion of partial order in relational data model by means of realizer, is [21]. Aiming at an effective representation of information representable by a partial order and proposing a suitable data structure, [22] builds on this framework. 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 [23], SRQL [24], Aquery [25, 26]. Various kinds of ordering on power-domains have also been considered in context of modelling incomplete information: a very extensive and general study is provided in [27].

By contrast to the above qualitative approach, in the *quantitative* approach [28, 29, 30, 31, 32, 33, 34], preference is specified indirectly using *scoring functions* that associate a numeric score with every tuple. On the one hand, this approach enables expressing quantitative

[1] A singleton world is a world containing a single element.

[2] Here, the term *universal relation* denotes that unique relation instance over a relation schema that contains all possible tuples over that schema

aspects of preference, e.g., its strength, however, on the other hand, expressivity of the qualitative aspect of preference is restricted to the weak order – a special case of the partial order.

5. Conclusions

The paper deals with the optimization of relational queries using the concept of preference. It builds on the recent leading ideas that have contributed to remarkable advances in the field:

- 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. By considering the preference operator on its own, we can, on the one hand, focus on the abstract properties of user preference and, on the other hand, study special evaluation and optimization techniques for the preference operator itself.
- 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. Pushing the preference operator, based on its algebraic properties, is a well known technique realizing this strategy.

Furthermore, to express a user preference, we employ the language introduced by Kaci and van der Tore [35], who have extended propositional language with sixteen kinds of preference. In their non-monotonic logic framework, we can capture complex preference, including preference between sets, yet the preference operator parameterized by such complex preference doesn't fulfil the commutativity and distributivity properties. For this reason, the optimization strategy needs to employ different technique: computing preference models over a stepwise pruned special set $\mathfrak{W}$ until the fixpoint is reached and then using a special preference operator derivative to filter out "bad" tuples.

In conclusion, the main contribution of the paper consists in presenting the optimization strategy of pushing the user preference down the expression tree and introducing the algorithm for its implementation.

References

[1] S. Kaci and L. W. N. van der Torre, "Algorithms for a nonmonotonic logic of preferences.," in *ECSQARU* (L. Godo, ed.), vol. 3571 of *Lecture Notes in Computer Science*, pp. 281–292, Springer, 2005.

[2] G. von Wright, *The logic of preference*. Edinburgh University Press, Edinburgh, 1963.

[3] 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.

[4] 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.

[5] K. Govindarajan, B. Jayaraman, and S. Mantha, "Preference datalog," Tech. Rep. 95-50, 1, 1995.

[6] 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.

[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] W. Kießling and B. Hafenrichter, "Algebraic optimization of relational preference queries," Tech. Rep. 2003-01, Institute of Computer Science, University of Augsburg, February 2003.

[12] J. Chomicki, "Preference Formulas in Relational Queries," *ACM Trans. Database Syst.*, vol. 28, no. 4, pp. 427–466, 2003.

[13] 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.

[14] J. Chomicki and J. Song, "Monotonic and nonmonotonic preference revision," 2005.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] S. Raghavan and H. Garcia-Molina, "Complex queries over web repositories," tech. rep., Stanford University, February 2003.

[20] R. Nedbal, "Relational Databases with Ordered Relations," *Logic Journal of the IGPL*, vol. 13, no. 5, pp. 587–597, 2005.

[21] D. R. Raymond, *Partial-order databases*. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1996. Adviser-W. M. Tompa.

[22] 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.

[23] 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.

[24] 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.

[25] A. Lerner, *Querying Ordered Databases with AQuery*. PhD thesis, ENST-Paris, France, 2003.

[26] 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.

[27] L. Libkin, *Aspects of partial information in databases*. PhD thesis, University of Pensylvania, Philadelphia, PA, USA, 1995.

[28] 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.

[29] A. Eckhardt, "Methods for finding best answer with different user preferences," Master's thesis, 2006. In Czech.

[30] 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.

[31] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," in *Symposium on Principles of Database Systems*, 2001.

[32] 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.

[33] 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.

[34] 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.

[35] S. Kaci and L. van der Torre, "Non-monotonic reasoning with various kinds of preferences," in *IJCAI-05 Multidisciplinary Workshop on Advances in Preference Handling* (R. Brafman and U. Junker, eds.), (Edinburgh, Scotland), pp. 112–117, 2005.

Redakční a publikační systém založený na principech EBM a Web 2.0

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.

Abstrakt

Od počátku 90. let 20. století, kdy se systematicky začaly vyvíjet nástroje a metodika pro zavádění medicíny založené na důkazech (EBM) do klinické praxe, došlo ke značnému rozvoji informačních zdrojů a služeb zaměřených na podporu EBM. Současně docházelo k posunu ve vztahu uživatelů k internetu. Webové technologie, které dříve byly v rukou profesionálních programátorů, se přiblížily uživatelům natolik, že zanikla hranice mezi autory obsahu a čtenáři. Tento jev, v posledních letech popisovaný jako Web 2.0, je zdrojem cenného poznání (”wisdom of crowds”, ”collective knowledge”). Tato práce vychází z principů medicíny založené na důkazech a využívá nástroje Webu 2.0 pro vytvoření nového informačního zdroje, který naplňuje pevná kritéria EBM a současně umožňuje využití prvků Webu 2.0 podporujících sdílení znalostí a komunikaci jeho uživatelů. Výsledkem je systém pro budování databáze poznání vzniklého na podkladě systematického výzkumu doplňovaná názory a praktickými zkušenostmi členů dané virtuální komunity.

Klíčová slova: vědecké lékařské informace, medicína založená na důkazech, EBM, podpora klinického rozhodování, informační zdroje, Web 2.0, nová média

1. Úvod

Informační zdroje odpovídající pravidlům medicíny založené na důkazech (EBM) musí splňovat několik základních kritérií:

- Aktuálnost (nejnovější informace, pravidelná aktualizace)
- Validita (metodologická správnost)
- Klinická relevance (odpovědi na klinické otázky)
- Rychlá dosažitelnost a praktičnost (elektronická forma, snadné vyhledávání)

S ohledem na výše uvedené nároky vznikly pro potřeby EBM některé specifické dokumenty, mezi něž patří především tzv. **sekundární zdroje** odvozené analýzou a syntézou primárních časopiseckých článků, tj. originálních studií. Za nejspolehlivější primární zdroje jsou považovány randomizované kontrolované studie, které stojí na vrcholu tzv. **pyramidy důkazů** (obr. 1). Ze stejných důvodů se informační zdroje zaměřené na podporu EBM v klinické praxi postupně vyvíjely a v současnosti je lze rozdělit do pěti základních skupin (viz pyramida ”5S”, obr. 2, [6]). Při vyhledávání odpovědí na klinické otázky se doporučuje začínat u sekundárních zdrojů a postupovat od vrcholu pyramidy směrem k jejímu základu.

Obrázek 1: Pyramida důkazů.

1.1. Sekundární informační zdroje pro podporu EBM

Mezi sekundární zdroje v kontextu terminologie medicíny založené na důkazech patří: systematické přehledy, CATs (Critical Appraised Topics), BETs (Best Evidence Topics), POEMs (Patient Oriented Evidence that Matters), klinická doporučení (CPGs, Clinical Practice Guidelines) a Ekonomické analýzy.

a. Systematické přehledy (systematic reviews)
Podle současných kritérií medicíny založené na důkazech jsou systematické přehledy v daném čase nejkvalitnější zdroje informací o určitém tématu nebo klinické otázce a stojí tedy na vrcholu dříve již zmíněné pyramidy důkazů (obr. 1). Vznikají v metodicky přesně definovaném a reprodukovatelném procesu, jehož součástí je pečlivé a důkladné vyhledávání primárních vědeckých dokumentů (publikovaných i nepublikovaných), kritické posouzení jejich validity (k dalšímu zpracování jsou vybrány pouze studie odpovídající stanoveným kritériím) a často i následné statistické zpracování (metaanalýza). Cílem tohoto procesu je minimalizovat riziko systematické chyby (bias) a získat tak co možná nejspolehlivější závěry.

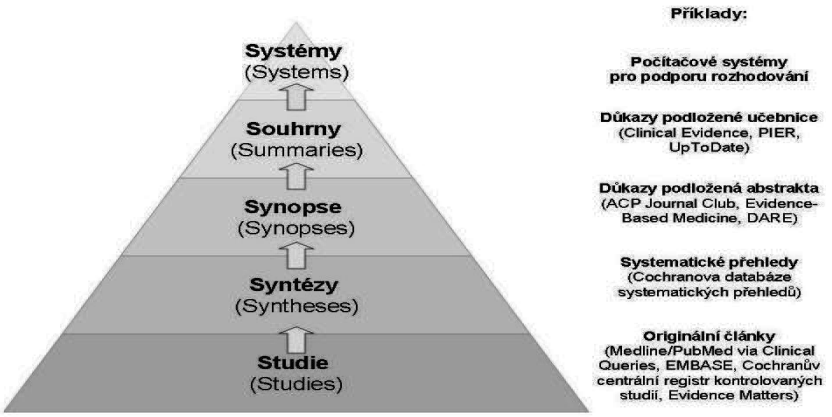
Tvorbou systematických přehledů se zabývá například Cochranova spolupráce (Cochrane Collaboration), která vytváří a čtvrtletně aktualizuje tzv. Cochranovy systematické přehledy. Jsou obsahem Cochranovy databáze systematických přehledů (The Cochrane Database of Systematic Reviews, CDSR) v Cochranově knihovně. V úvodu každého takového dokumentu najdeme datum posledního prohledávání informačních zdrojů a datum poslední podstatné provedené změny.

b. CATs, BETs, POEMs
CATs (Critical Appraised Topics) a BETs (Best Evidence Topics) jsou kratší dokumenty shrnující důkazy na úzce specializovanou klinickou otázku (terapeutický postup, diagnostický test ap.). Dokumenty tohoto typu najdeme v databázích mateřských univerzit, organizací a institucí, například v CATbank (Centre for Evidence-Based Medicine, Oxford) nebo na Evidence-Based Pediatrics Web Site (University of Michigan).

POEMs (Patient-Oriented Evidence that Matters) jsou takové důkazy, jejichž výsledky jsou významné z hlediska pacienta (morbidita, mortalita, kvalita života) na rozdíl od tzv. DOEs (Disease Oriented Evidence), které se zabývají charakteristikami nemoci (patofyziologie, etiologie). Články typu POEM vycházejí v každém čísle Journal of Family Practice a jsou základem pro Family Medicine Journal Clubs.

c. Klinická doporučení (clinical practice guidelines)
Klinická doporučení jsou systematicky vyvíjené dokumenty pro podporu rozhodování o patřičné léčebné péči v konkrétní klinické situaci. Bývají vytvářeny a aktualizovány odbornými asociacemi nebo klinickými skupinami a publikovány v odborných časopisech, na internetových stránkách odborných společností či patřičných vládních rezortů nebo pomocí účelového tisku. Najdeme je rovněž ve specializovaných databázích (např. Evidence-Based Medicine Guidelines).

d. Ekonomické analýzy (economic analyses)
Ekonomické analýzy jsou dokumenty, které pomocí formálních kvantitativních metod srovnávají alternativní postupy z hlediska nákladů a výsledků. Rovněž tento druh informací najdeme v příslušných databázích, například v NHS Economic Evaluation Databaze (NHS EED) vytvářené v Centre for Reviews and Dissemination při Univerzitě v Yorku.



Obrázek 2: Evoluce informačních zdrojů pro podporu EBM.

1.2. Evoluce informačních zdrojů pro podporu EBM

Vývoj specializovaných informačních zdrojů pro podporu EBM lze znázornit pomocí pětistupňové pyramidy (”5S”, obr. 2, [6]):

Studie (studies): jednotlivé originální články vyhledatelné v tradičních biomedicínckých databázích, jako jsou Medline, EMBASE nebo CINAHL. Klinické studie je možné vyhledávat také přímo v registrech, jako jsou Cochrane Central Register of Controlled Trials nebo Current Controlled Trials.

Syntézy (syntheses): systematické přehledy a metaanalýzy všech dostupných a srovnatelných originálních studií zabývajících se danou problematikou. Mezi prameny typu syntézy patří Cochranovy přehledy a přehledy non-Cochranova typu, jako jsou např. CATs (Critically Appraised Topics) nebo BETs (Best Evidence Topics).

Synopse (synopses): stručné (často jednostránkové), výstižné a přehledné popisy (strukturovaná abstrakta) systematických přehledů nebo originálních studií. Spolu s níže uvedenými souhrny jsou považovány za nejpraktičtější soubory informací pro lékaře v klinické praxi. Synopse najdeme ve specializovaných časopisech (např. ACP Journal Club, Evidence-Based Medicine nebo Evidence-Based Cardiovascular Medicine) a databázích (např. v databázi DARE, Database of Abstracts of Reviews of Effects, která zahrnuje studie hodnotící efektivitu léčebných postupů).

Souhrny (summaries): vycházejí ze synopsí, syntéz a studií a integrují všechny dostupné důkazy na dané klinické téma. Na rozdíl od synopsí, syntéz a studií tak poskytují informace relevantní pro danou klinickou situaci z více aspektů a jsou tedy v určitém smyslu ”EBM učebnicemi”. Patří sem například Clinical Evidence, PIER (Physicians’ Information and Education Resource) nebo UpToDate.

Systémy (systems) jsou softwarové aplikace pro podporu rozhodování, které automaticky propojují nejnovější a v dané době nejspolehlivější klinické důkazy s informacemi o konkrétním pacientovi (elektronický zdravotní záznam).
Dokumenty dosahující potřebné metodologické kvality však ještě musí být navíc **klinicky relevantní**. Je zřejmé, že vyhledání patřičně kvalitních a současně relevantních dokumentů podle výše uvedeného modelu není v tradičních biomedicínských databázích snadné a vyžaduje dobrou znalost dotazovacího jazyka dané databáze. Jistou pomůckou jsou předdefinované filtry. V databázi PubMed se jedná o tzv. PubMed Clinical Queries.

Pomocí **Clinical Queries** je možné vyhledávat jednak studie podle klinických kategorií (Clinical Study Category), jako jsou etiologie, diagnóza, terapie, prognóza a návody pro klinické předpovědi, jednak lze hledat systematické přehledy (Systematic Reviews). Kromě pravých systematických přehledů Cochranova typu tento filtr selektuje navíc také metaanalýzy, přehledy klinických studií, články zaměřené na evidence-based medicine, konference formulující shodná stanoviska a praktická doporučení (guidelines).

1.3. Web 2.0

Termín Web 2.0 byl poprvé použit Timem O’Reillym a zástupci MediaLive International při plánování konceptu pro první konferenci na téma aktuální situace a nových trendů na poli internetu, která se uskutečnila v roce 2004 [9], [11]. Konference s názvem Web 2.0 pak dala podnět pro nespočet diskuzí o tomto kontroverzním pojmu, především však ale poukázala na skutečnost, že od roku 2000 poněkud stagnující internetové podnikání nabírá nový směr.

Od prvního vyslovení termínu Web 2.0 bylo vykonáno mnoho pokusů o vyjádření jasné definice tohoto pojmu, které se - stejně jako termín samotný - vyznačují jistou vágností a provokují odbornou i laickou internetovou veřejnost k dlouhým diskuzím o jeho pravé podstatě a smysluplnosti. Podle O’Reillyho definice z října 2006 je Web 2.0 revoluce v podnikání v počítačovém průmyslu způsobená posunem k internetu jako platformě a pokus porozumět pravidlům vedoucím k úspěchu na této nové platformě. (”Web 2.0 is the business revolution in the computer industry caused by the move to the internet as platform, and an attempt to understand the rules for success on that new platform.”) [8].

Ačkoliv termín Web 2.0 navozuje dojem, že se jedná o novou verzi Webu, nejedná se o ”upgrade” celosvětové sítě z hlediska technických specifikací. Jde spíše o nové přístupy a způsoby využití stávajících webových technologií, jejichž výsledkem je tzv. **druhá generace webových služeb** a na webu založených komunit (Community 2.0), které díky aplikacím založeným na sociálním software (social software) posilují spolupráci a sdílení informací mezi uživateli (př. social networking sites, wikis nebo folksonomie) [11]. Pro Web 2.0 jsou charakteristické projekty, které používají technologie a principy **zaměřené na uživatele** služeb, a to často až do té míry, že jim umožňují podílet se na obsahu či tvorbě projektu [11]. Typická je proto **změna komunikačního modelu** z dříve běžného ”one to one” na dnes stále častější ”many to many”. Obsah webových stránek už tak není tvořen pouze webmastery

a jednotlivými autory, ale samotnými uživateli a jejich skupinami (”user-powered content”). Ruku v ruce s tím jdou aplikace, které by bylo možné souhrnně nazvat ”**reputační systémy**”. Ty umožňují uživatelům hodnotit a potažmo doporučovat (nebo naopak nedoporučovat) ostatním členům dané komunity jednotlivé produkty či příspěvky (ať už jde o výrobky nebo nejrůznější texty). Reputační systémy mají různou podobu od diskuze pod příspěvkem v blogu či jiném publikačním systému přes hlasovací systém s ikonou ”Vote it” nebo ”Digg it” apod. až po sofistikované miniaplikace automaticky analyzující počet hlasů přidělených jednotlivým příspěvkům a nabízející nejlépe hodnocené příspěvky jako další informaci navíc. Vedle vyjádření kladného hlasu některé z těchto systémů umožňují také příspěvek zavrhnout, označit jako nepřijatelný či nepatřičný (”Bury it”, ”Flag it as inappropriate”), takže ”čištění” komunitou nesystematicky přidávaného obsahu může být opravdu velmi účinné a výsledná kolekce textů, obrázků nebo jiných formátů pak může v případě dostatečné návštěvnosti webových stránek dosahovat nečekané kvality.

2. Cíl práce

Cílem této práce bylo **vytvořit platformu pro průběžně doplňovanou databázi dokumentů** naplňujících požadavky **EBM** na metodologickou kvalitu a klinickou relevanci, propojit tento obsah s prvky **Webu 2.0** a umožnit uživatelům kromě snadného sledování přírůstku do databáze a jejího prohledávání navíc také komunikovat o jednotlivých článcích, hodnotit je slovně nebo pomocí pětistupňové škály a využívat další prvky charakteristické pro Web 2.0.

3. Metodika a popis sytému

Pro ukládání a správu záznamů byl vybrán redakční systém určený pro publikování blogu od společnosti Google známý pod názvem **Blogger** (www.blogger.com). Jako základ pro obsah systému byly s ohledem na své postavení v pyramidě důkazů (viz výše) vybrány **systematické přehledy doplněné metaanalýzou**, které jsou pilířem postupně vznikající, plnotextově prohledatelné databáze. Doplňujícími informacemi systému jsou pak přehledy publikací nejnovějších **kontrolovaných klinických studií** a přehledy dalších klinicky významných článků z oblasti diagnostiky, etiologie a prognózy nemocí, **varování** publikovaná vybranými státními úřady pro kontrolu léčiv a **cílené vyhledávače** lékařských doporučení a klinických studií dostupných na internetu.

3.1. Bibliografie s abstrakty systematických přehledů a metaanalýz

Cochranovy i non-Cochranovy systematické přehledy jsou průběžně vyhledávány v databázi **MEDLINE/PubMed**. Dokumenty jsou filtrovány s ohledem na jednotlivé klinické specializace s pomocí terminologie **MeSH**. Vyhledané dokumenty jsou před samotným vložením do systému ještě zvlášť **posouzeny z hlediska relevance** a **popsány značkami (tagy)**, které charakterizují jejich obsah. Měsíčně je zakládáno několik desítek dokumentů, přičemž tento počet kolísá především v době aktualizace Cochranovy databáze systematických přehledů (4x ročně). Vedle služby nabízející **filtrovaný přehled nejnovějších článků** (resp. jejich bibliografických záznamů, ve většině případů včetně abstraktů) publikujících klinicky validní a relevantní důkazy tak vzniká navíc **kumulativní databáze**, kterou je možné prohledávat plnotextově nebo tematicky (pomocí značek/tagů přidělovaných při zakládání dokumentů do sytému).

3.2. Upozornění na články hodnocené postpublikačně s ohledem na potřeby klinické praxe

Existují dvě významné služby zaměřené na třídění a hodnocení publikovaných článků s ohledem na potřeby klinické praxe a kritéria EBM: McMaster Premier Literature Service (PLUS) a Faculty 1000 Medicine. **Na vybrané** (veřejně dostupné) **dokumenty** z těchto informačních zdrojů **je poukazováno formou citací s webovými odkazy do databáze MEDLINE/PubMed**.

a. McMaster **P**remier **L**iterat**u**re **S**ervice (**PLUS**) je knihovnicko informační servis, jehož podstatou je systematické prohledávání 110 vybraných biomedicínských časopisů, identifikace potenciálně významných článků, které z metodologického hlediska splňují kritéria EBM, a následné hodnocení těchto článků lékaři v praxi z hlediska jejich klinické relevance a praktického dopadu. Takto vybrané a ohodnocené články je poté možné získávat pomocí e-mailové alertní služby nebo vyhledávat přímo v McMasteské databázi, kde je rovněž možné procházet seznam nejvíce čtených článků. Služba je provozována ve spolupráci se známým nakladatelem odborné lékařské literatury BMJ Publishing Group a distribuována jako **BMJ Updates+** (www.bmjupdates.com). Na základě McMaster PLUS je založena i služba **Medscape Best Evidence** poskytovaná serverem Medscape, který je součástí sítě profesionálních portálů WebMD Health Professional Network. Pro účely této práce jsou ze služby McMaster PLUS vybírány nejvíce sledované články.

b. **Faculty 1000 Medicine** je služba dalšího známého nakladatele v oblasti odborné biomedicínské literatury, kterým je BioMed

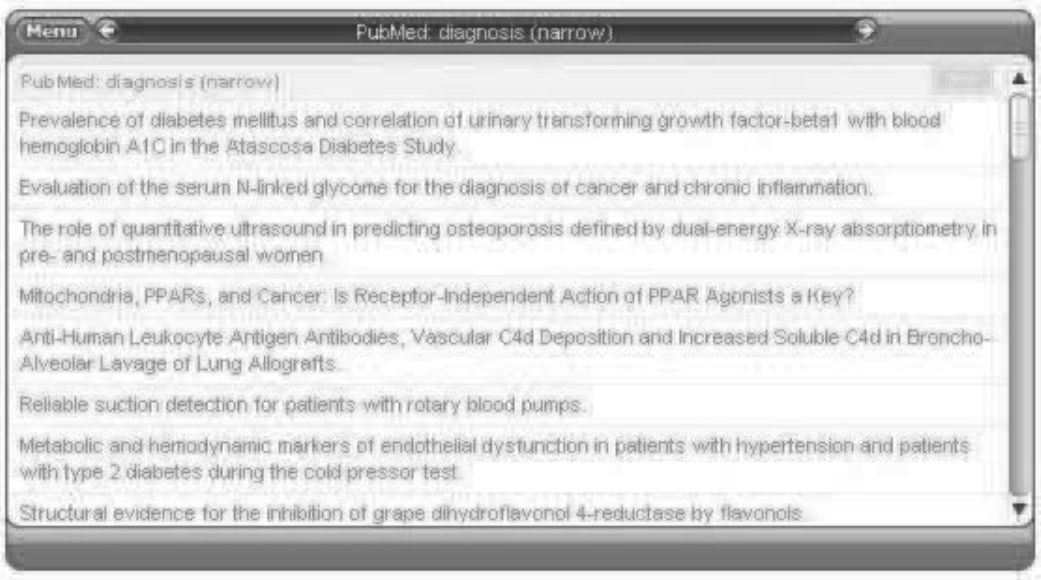
Central (www.f1000medicine.com). I ona je založena na postpublikačním vyhodnocování článků, byť pro jejich výběr i hodnocení samotné existují jiná pravidla než v případě výše uvedných BMJ Updates+. Tato služba upozorňuje na nejzajímavější a nejvlivnější články z oblasti medicíny na základě doporučení téměř 2500 předních vědců a kliniků z 18 oborů, kteří je vybírají, hodnotí a přidělují jim tzv. F1000 faktor. Pro potřeby této práce jsou využívány **články, které** jsou podle členů Faculty 1000 Medicine natolik významné, že **mění pohled na dosavadní klinickou praxi** (”articles that change clinical practice”).

3.3. Upozornění na nejnovější výsledky randomizovaných kontrolovaných klinických studií a na další klinická témata prostřednictvím technologie RSS

RSS (Really Simple Syndication) je technologie používaná k publikování (resp. také sledování) často aktualizovaného obsahu. RSS dokument (tzv. feed nebo kanál) obsahuje buď část obsahu z patřičné webové stánky nebo plný text. Aktualizovaný obsah je pak možné automaticky odebírat pomocí agregátoru neboli RSS čtečky.

Zdrojem obsahu pro tuto část systému je opět databáze **MEDLINE/PubMed** (www.pubmed.gov), přičemž výběr článků je prováděn na základě **filtrů** odpovídajících tzv. **Clinical Queries** (viz výše) [2], [3], [4], [12], [13], [14].

Pro účely popisovaného systému byl zvolen odběr názvů (titulků) nejnovějších článků vyhledávaných podle níže uvedených filtrů. Přehledy titulků jsou agregovány do webové miniaplikace (obr. 3), v níž je možné titulky prohlížet a v případě zájmu prokliknout na celý abstrakt přímo do databáze PubMed.



Obrázek 3: RSS čtečka nejnovějších článků vybraných z databáze PubMed.

Filtr pro články týkající se léčby:
(randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract]))

Filtr pro články týkající se diagnostiky:
(specificity[Title/Abstract])

Filtr pro články týkající se etiologie nemocí:
((relative[Title/Abstract] AND risk*[Title/Abstract]) OR (relative risk[Text Word]) OR risks[Text Word] OR cohort studies[MeSH:noexp] OR (cohort[Title/Abstract] AND stud*[Title/Abstract]))

Filtr pro články týkající se prognózy:
(prognos*[Title/Abstract] OR (first[Title/Abstract] AND episode[Title/Abstract]) OR cohort[Title/Abstract])

Filtr pro návody na klinické předpovědi:
(validation[tiab] OR validate[tiab])

3.4. Upozornění a varování vybraných státních úřadů pro kontrolu léčiv

Nepublikovaná data přicházející z klinické praxe formou hlášení o nežádoucích účincích léků státním úřadům pro kontrolu léčiv jednotlivých zemí jsou podchycena v popisovaném systému pomocí RSS kanálů přímo ze stránek patřičných úřadů. V této fázi byly do systému zahrnuty farmakovigilanční zprávy ze tří institucí:

a. **Státní úřad pro kontrolu léčiv** (ČR), www.sukl.cz

b. **Medicines and Healthcare products Regulatory Agency** (UK), www.mhra.gov.uk

c. **Food and Drud Administration** (USA), www.fda.gov

Obrázek 4: Oblak štítků (tag cloud) a možnost prohlížení článků podle témat.

3.5. Prvky Web 2.0

a. Štítky (tagy)
Vybrané články jsou při zakládání do systému označovány štítky (tagy), pomocí kterých mohou být prohlíženy tematicky související články. Relativní četnost štítků je vizualizována ve formě tzv. oblaku štítků (tag cloud, obr. 4), který usnadňuje orientaci v obsahu databáze.

b. Komentáře
Pod každý příspěvek mohou uživatelé vkládat své komentáře a doplňovat tak odborný obsah systému vybíraný z databáze MEDLINE/PubMed a vytvářet tak složky zvané v terminologii Web 2.0 jako ”user-generated content” a ”soft peer-review” [10] (viz také níže). Tento **uživateli vytvářený obsah** mohou zájemci sledovat jednak přímo pod články, jednak mohou názory a komentáře k článkům odebírat do svých RSS čteček prostřednictvím RSS kanálů. S ohledem na zaměření systému se očekává, že komentáře budou mít odborný charakter a budou poskytovat praktické pohledy na komentovaná témata a hodnocení článků založená na osobních zkušenostech. Diskuze je zcela otevřená pro všechny uživatele systému, z důvodu prevence zneužití či vandalizmu je však požadována registrace komentátorů prostřednictvím Gmail účtu nebo Open ID.

c. Hodnocení článků
Do systému byl implementován nástroj pro hodnocení článků z pohledu uživatelů. Články je možné hodnotit v **pětistupňové škále** (1-2 hvězdičky: špatné hodnocení, 3-4 hvězdičky: dobré hodnocení, 5 hvězdiček: vynikající hodnocení). Výsledky hodnocení při dostatečném počtu uživatelů slouží jako jistá alternativa oficiálního recenzního procesu (tzv. ”soft peer-review” [10]) a umožňují rychle **určit v množství článků, které** z nich mají nejvyšší hodnocení a tedy **nejvíce stojí za pozornost**. (Na tomto místě je však nutné připomenout, že jde o hodnocení článků, které samy již prošly oficiálním recenzním procesem a jsou z hlediska kvality na vysoké úrovni. Hodnocení komunity však přidává další aspekty, jejichž podrobný rozbor by ale byl již mimo původní zaměření tohoto článku.) Uvedený nástroj současně automaticky **vyhodnocuje nejvýše oceněné články** a nabízí jejich přehled na postranním panelu (”The most popular posts/articles”), stejně jako **doporučuje další dobře hodnocené články** přímo pod jednotlivými záznamy (”Recommended posts/articles”, obr. 5) a umožňuje tak využití dalšího prvku charakteristického pro Web 2.0, kterým je **”vytěžování společného poznání”** dané komunity (”collective knowledge”, ”wisdom of crowds”).

Obrázek 5: Hodnocení článků uživateli a nabídka dalších dobře hodnocených článků.

d. RSS kanály
Systém nabízí RSS kanály pro **nově přidané články i komentáře** k nim, které uživatelé mohou **průběžně sledovat prostřednictvím** svých **RSS čteček**.

e. Komunitní záložky
Systém je vybaven propojením každého článku s více než dvaceti službami pro **zakládání a sdílení on-line záložek** (social bookmarking websites) a umožňuje tak uživatelům jednak praktický přístup k takto založeným EBM textům z jakéhokoliv místa vybaveného připojením k internetu (a neomezuje tedy uživatele na jejich lokální programy pro ukládání a správu odborné literatury typu Reference Manager ap.) a dále umožňuje tzv. virové (česky v tomto kontextu častěji tzv. **virální**) **šíření nejvíce ceněných článků** prostřednictvím internetu. Lze předpokládat, že vzhledem k charakteru publikací vybíraných do sytému může mít tato posledně jmenovaná funkce velký význam pro šíření a zavádění nejnovějších vědeckých poznatků do klinické praxe.

4. Závěr

Přínos popisovaného systému v kontextu současné nabídky informačních zdrojů, služeb a systémů zaměřených na potřeby EBM je očekáván jednak v rovině **rozšíření nabídky specializovaných informačních zdrojů** pro podporu klinického rozhodování, jednak v rovině **propojení** tohoto zdroje s **prvky a nástroji Webu 2.0**.

Jak bylo uvedeno výše, začínají v posledních letech vznikat systémy zaměřené na **postpublikační evaluaci biomedicínské literatury**. Mezi nejvýznamnější patří Faculty 1000 Medicine, BMJ Updates+, Medscape Best Evidence nebo Ophthalmology+, jež nabízejí filtrované informace zaměřené na bezprostřední využitelnost v klinické praxi. V době exponenciálního růstu informací, kdy lékaři čelí tzv. informačnímu paradoxu (tzn. přetížení informacemi, přičemž právě potřebné informace jsou nedostupné [1]), je jejich praktický význam vysoký.

Paralelně vznikají iniciativy využívající **nástroje a služby Webu 2.0** pro potřeby vědecké komunity, biomedicínské obory nevyjímaje (Věda 2.0, Medicína 2.0). Na vzestupu jsou nová publikační a komunikační média, spolu s nimiž stoupá objem uživateli vytvářeného obsahu. Sociální software a sociální sítě umožňují snadné a rychlé sdílení informací a pružnou komunikaci, díky čemuž je rychlost šíření nových poznatků nesrovnatelně vyšší a doba od formulace vědeckých závěrů k jejich uvedení do všeobecného povědomí se zkracuje.

Tato práce kombinuje oba výše uvedené principy. Výsledkem je nástroj pro poskytování **informačního servisu** a budování kumulativní **databáze publikací** splňujících nejpřísnější **kritéria EBM**, který navíc umožňuje využití **vlastností charakteristických pro Web 2.0**. Vedle předem daných a explicitně prověřených pravidel pro výběr článků zařazovaných do systému nabízí tedy i možnost pro vyjádření názoru komunity uživatelů a v jistém smyslu tedy i další rovinu postpublikačního hodnocení článků (viz výše zmíněné ”soft peer-review”). Systém zahrnuje v současné době dva nástroje umožňující interakci s komunitou uživatelů: hodnocení pomocí pěti hvězdiček a vyjádření slovní v rámci komentářů pod články. Dále systém

zahrnuje možnost vyhledávání lékařských doporučení a randomizovaných klinických studií pomocí cílených vyhledávačů, informace o nejnovějším obsahu z vybraných informačních zdrojů, k dispozici je rovněž možnost odebírání nejnovějšího obsahu pomocí RSS kanálů a ukládání vybraných článků do osobních i sociálních webových záložek.

Literatura

[1] J.A.M. Gray, "Where's the chief knowledge officer?", *British Medical Journal*, vol. 317, pp. 832–840, 1998.

[2] R.B. Haynes et al., "Developing optimal search strategies for detecting clinically sound studies in MEDLINE", *Journal of the American Medical Informatics Association*, vol. 1, pp. 447–458, 1994.

[3] R.B. Haynes et al., "Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey", *British Medical Journal*, vol. 330, p. 1179, 2005.

[4] R.B. Haynes, N.L. Wilczynski, and Hedges Team, "Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey", *British Medical Journal*, vol. 328, p. 1040, 2004.

[5] R.B. Haynes, "Evidence-based information resources", Presentation, Oxford, 2007; on-line [cit. 2008-05-03], dostupný z: http://www.cebm.net/index.aspx?o=1480 .

[6] R.B. Haynes, "Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions", *Evidence-Based Medicine*, vol. 11, pp. 162–164, 2006.

[7] D.L. Hunt, R. Jaeschke, K.A. McKibbon, Evidence-Based Medicine Working Group, "Users' guides to the medical literature, XXI: using electronic health information resources in evidence-based practice", *JAMA*, vol. 283, pp. 1875–1879, 2000.

[8] T. O'Reilly, "Web 2.0 Compact Definition: Trying Again"; on-line [cit. 08-02-02], dostupný z: http://radar.oreilly.com/archives/2006/12/web_20_compact.html
.

[9] T. O'Reilly, "What Is Web 2.0"; on-line [cit. 08-02-02], dostupný z: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
.

[10] D. Taraborelli, "Soft peer review: Social software and distributed scientific evaluation", Proceedings of the 8th International Conference on the Design of Cooperative Systems (COOP '08), Carry-Le-Rouet: 2008; on-line [cit. 08-07-20], dostupný z: http://nitens.org/docs/spr_coop08.pdf .

[11] „Web 2.0", in Slovník internetových výrazů; on-line [cit. 08-02-02], dostupný z: http://www.symbio.cz/slovnik/web-2-0.html .

[12] N.L. Wilczynski, R.B. Haynes, and Hedges Team, "Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey", *BMC Medicine*, vol. 2 (23), 2004.

[13] N.L. Wilczynski, R.B. Haynes, and Hedges Team, "Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE", *AMIA Annual Symposium Proceedings*, pp. 719-723, 2003.

[14] S.S. Wong et al., "Developing Optimal Search Strategies for Detecting Sound Clinical Prediction Studies in MEDLINE", *AMIA Annual Symposium Proceedings*, p. 728, 2003.

Flying Amorphous Computer and Its Computational Power (Extended Abstract)

Post-Graduate Student:
RNDr. Lukáš Petrů

Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25

118 00 Prague, Czech Republic

lukas.petru@st.cuni.cz

Supervisor:
Prof. RNDr. J. Wiedermann, DrSc.

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

182 07 Prague, Czech Republic

wieder@cs.cas.cz

Field of Study:
Theoretical Informatics

Motivation. In 1999, a group of people led by Kristofer Pister from University of California, Berkeley, presented an idea of a so-called *smart dust* (cf. [5], [11], [6], [12]). The smart dust is a network of computers of extremely small scale—each computer should have size of 1 mm^3 and even smaller if technology permits. The ultimate goal was to have each computer the size of a dust mote.

It is anticipated that these motes could be easily distributed in a target area by e.g. dropping them from an airplane. The deployed motes would then serve to monitor the target area as to the temperature, humidity, amount of precipitation. This information may help in agriculture to ensure better yield. Or the motes are dropped in a perimeter of a secured area and using auditory and electromagnetic sensors they discover unwanted intruders. Other applications could be in a health industry for monitoring the physical conditions of patients and the movements of patients in a hospital building. The project concentrated mainly on technical issues connected with the need to use or develop new, very efficient components from which to build the motes.

Concurrently with the Smartdust project a similar computing paradigm appeared at MIT Computer Science and Artificial Intelligence Laboratory. The respective idea—called *amorphous computing*—was also introduced in 1999. Amorphous computing assumes that large number of simple, identical devices (thousands or millions of them) will be randomly scattered over a target area. The question connected with amorphous computing was how to organize and program these devices so that they cooperate and, as a whole, perform some useful action (see [1], [2], [3]).

The two paradigms, Amorphous computing and Smart dust, are similar in that both assume very simple devices used in large numbers. The difference between Amorphous computing and Smart dust is in the assumed method of communication. Smart dust assumes optical transmission, which is a long-range communication over uni-directional links and the light beam is sent into some direction. On the other hand, Amorphous computing assumes that the devices communicate by radio, which is a short-range communication over bi-directional links and the signal is broadcast omnidirectionally.

A recent review [10] discusses various techniques to take smartdust in sensor networks beyond millimeter dimensions to the micrometre level. For communication purposes, so-called nanoradios are considered (cf. [4]).

Our approach. Neither in the Smartdust nor in the amorphous computing project attention was paid to the recursive-theoretical, or computational complexity aspects of the underlying new computational paradigm. The core of the new paradigm can be aptly summarized as sensing, computing, communication, and mobility. In order to study the respective issues three things are needed: (i) a detailed *mathematical*, or *computational*, for that matter, model capturing the main features of the respective paradigm communication protocols; (ii) communication protocols enabling message passing within the network of processors, and (iii) proof of the universal computing power of the resulting model.

In a series of papers, we have gradually developed models of amorphous computing that, in the order of their increased generality, cover the main features of various types of amorphous computers.

All models possess universal computing power which was shown by simulating other universal computing models known from the computability theory. In our first results we have shown simulation of a cellular

automaton and parallel Turing machine ([7], [8]). Our later model simulated RAM ([9], [13], [14]). Recently, we have shown that our work can also be extended to nano computers that communicate by sending special signaling molecules ([15]). Till late it remained an open problem if it was possible to have reliable computation on so-called *flying amorphous computer*, i.e., on such a computer whose computing nodes are constantly moving causing constant changes of possible communication links. The model of the flying amorphous computer will be part of the PhD thesis that is being finished now. Here we briefly introduce the model and the obtained results.

Model. The flying amorphous computer consists of N identical nodes (a node of our amorphous computer corresponds to a mote of a Smartdust model). Each node is modelled as a RAM with a fixed number of registers of size $O(\log N)$ bits. Thus, each register can hold a number in the range 0 to N. The memory of all the nodes is initially empty. All nodes are randomly placed in a target area of a square shape. Each node is randomly assigned a direction vector determining the direction in which the node at hand moves with constant speed. All nodes move with the same speed; however, each node moves in a different direction. If a node reaches the border of the target area, it bounces off the edge like a billiard ball and continues moving in the mirrored direction. This ensures that the nodes do not leave the target area. The model of the communication captures the properties of a simple radio. There is a transmission range r. All nodes that are at distance at most r from a node are called neighbours of that node. A node can either send a message or receive. If a node is receiving and exactly one of its neighbours is sending, the message is transferred. If two or more neighbours are sending, there is a collision and the message is not transferred. No node can recognize the state of collision from the state of no transmission.

Communication protocol. For our model we have developed a special communication protocol that works under the minimal requirements as far as the computational and communication functionalities of the individual nodes are concerned. These functionalities are: finite-state memory, randomness, asynchronicity, anonymity of processors, and one-way communication without a possibility of signal reception acknowledgement.

Using this protocol, any node can start broadcasting a message to its neighbours. These neighbours will then broadcast the same message to their neighbours and so on until the message finally covers the whole amorphous computer. Due to the movement of nodes this communication mechanism is unreliable—it is not guaranteed that a message will be delivered to all nodes of the computer. There is even no guaranty that a message will be delivered at all to any node. Therefore all algorithms designed for an amorphous computer must cope with this unreliability. Of course, in the case of always failing communication our amorphous computer could not compute anything. Therefore we have to make an assumption that communication will not keep failing all the time—rather, it is always the case that from time to time communication will be successful. We say that an amorphous computer is *lively flying* if it posses the following property: a sequence of a message broadcastings from one node to some other node always succeeds in delivering the message at hand to some other node after an unknown, but finite number of attempts. Note that the lively flying condition does not allow to derive any time estimates on how long will a computation take. Nevertheless, it at least allows proving termination and correctness of our algorithms.

Universality. The property of universality is shown by giving an algorithm simulating a unit-cost RAM with a high probability. The first algorithm to be run on an amorphous computer is the address assignment algorithm. One of all nodes is selected and address 1 is given to it. Then one node from the rest is selected and address 2 is given to it. The process continues untill all addresses up to N are assigned. Then the input data can be stored into the nodes by an external operator. Now we can start the simulation of a unit cost RAM using $O(N)$ registers. There is one special node called *base node*, which simulates the control unit of the RAM. All other nodes simulate the memory registers of the RAM. When a read or write to register is required, the base node broadcasts a message to all nodes and the node with particular address performs the reading or writing of its internal memory. To cope with the unreliability in communication we require that an acknowledgement from a target node is sent and received by the base node. Therefore the cycle of sending an instruction by the base node and its waiting for an acknowledgement is repeated until the acknowledgement is obtained. Only then the simulation of a RAM computation can resume by processing the next RAM instruction. In this way it is guaranteed that there are no errors in the simulation and thanks to our assumption on the lively flying amorphous computer we also can prove simulation termination in finite, but unknown time.

References

[1] H. Abelson, et al. Amorphous Computing. MIT Artificial Intelligence Laboratory Memo No. 1665,

Aug. 1999

[2] H. Abelson, D. Allen, D. Coore, Ch. Hanson, G. Homsy, T. F. Knight, Jr., R. Nagpal, E.Rauch, G. J. Sussman, R. Weiss. Amorphous Computing. Communications of the ACM, Volume 43, No. 5, pp. 74–82, May 2000

[3] H. Abelson, J. Beal, G. J. Sussman. Amorphous Computing. Computer Science and Artificial Intelligence Laboratory, Technical Report, MIT-CSAIL-TR-2007-030, June 2007

[4] K. Bullis. TR10: NanoRadio. Technology Review. Cambridge: MIT Technology Review, 2008-02-27

[5] J. M. Kahn, R. H. Katz, K. S. Pister. Next century challenges: mobile networking for ”Smart Dust”. In Proceedings of the 5th Annual ACM/IEEE international Conference on Mobile Computing and Networking, MobiCom ’99, ACM, pp. 271–278, Aug 1999

[6] J. M. Kahn, R. H. Katz, K. S. J. Pister. Emerging Challenges: Mobile Networking for Smart Dust. Journal of Communications and Networks, Volume 2, pp 188–196, 2000

[7] L. Petru. On the Computational Power of an Amorphous Computer. WDS’04 Proceedings of Contributed Papers. Prague, CZ. pp. 156–162, June 2004

[8] L. Petrů. Amorfní počítání: model univerzálního počítače sestaveného z jednoduchých kooperujících agentů. Proceedings of Kognice a umělý život VI. Opava, CZ. pp. 309–314, May 2006

[9] L. Petrů, J. Wiedermann. A Model of an Amorphous Computer and Its Communication Protocol. In: Proc SOFSEM 2007: Theory and Practice of Computer Science. LNCS Volume 4362, Springer, pp. 446–455, July 2007

[10] M. J. Sailor, J. R. Link. Smart dust: nanostructured devices in a grain of sand, Chemical Communications, vol. 11, p. 1375, 2005

[11] B. Warneke, M. Last, B. Liebowitz, K. S. J. Pister. Smart Dust: communicating with a cubic-millimeter computer. Computer, Volume: 34, Issue: 1, pp. 44–51, Jan 2001

[12] B. Warneke, B. Atwood, K. S. J. Pister. Smart dust mote forerunners. In Proceedings of the 14th IEEE International Conference on Micro Electro Mechanical Systems, 2001, MEMS 2001, pp. 357–360, 2001

[13] J. Wiedermann, L. Petru. Computability in Amorphous Structures. In: Proc. CiE 2007, Computation and Logic in the Real World. LNCS Volume 4497, Springer, pp. 781–790, July 2007

[14] J. Wiedermann, L. Petru. On the Universal Computing Power of Amorphous Computing Systems. Theory of Computing Systems, Springer, New York. To appear.

[15] J. Wiedermann, L. Petru. Communicating Mobile Nano-Machines and Their Computational Power. In: Proc. Nano-Net 2008, LNICST, Springer. To appear

SNOMED CT a jeho 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 grantu 1ET200300413 AV ČR.

Abstrakt

Článek popisuje mezinárodní klasifikační systém SNOMED CT, jeho využití, základní komponenty a hierarchie. Dále popisuje Minimální datový model pro kardiologii a využití systému SNOMED CT v tomto datovém modelu.

Klíčová slova: klasifikační systémy, SNOMED CT, Minimální datový model pro kardiologii

1. Úvod

Vymezení, pojmenování a třídění lékařských pojmů není dosud optimální. Pro jeden pojem existuje často více než deset synonym. Vhodný kódovací systém ale rychle poskytne jednoznačný kód pro libovolný biomedicínský poznatek. Tato práce je zaměřena na klasifikační systém SNOMED Clinical Terms, pomocí jehož konceptů jsme zakódovali atributy v Minimálním datovém modelu pro kardiologii.

2. SNOMED CT

SNOMED Clinical Terms® (SNOMED CT®) [1, 2, 3] je komplexní klinická terminologie, která poskytuje klinický obsah a expresívnost pro klinickou dokumentaci a výkaznictví. Může být využit pro kódování, vyhledávání a analyzování klinických dat. SNOMED CT vzniknul sloučením terminologií SNOMED Reference Terminology (SNOMED RT), kterou vytvořila College of American Pathodologists (CAP) a Clinical Terms Version 3 (CTV3), kterou vyvinul National Health Service (NHS) ve Velké Británii. Tato terminologie obsahuje koncepty, termíny a vztahy s cílem přesně vyjadřovat klinické informace napříč celým zdravotnictvím.

3. Využití terminologie SNOMED CT

Zdravotnické softwarové aplikace se zaměřují na sběr klinických dat, na propojení klinických znalostních databází, získávání informací a také na shromažďování a výměnu dat. Informace ale mohou být zaznamenány různými způsoby v různou dobu a na různých místech.

Standardizované informace zlepšují analýzu. SNOMED CT poskytuje standard pro klinické informace. Softwarové aplikace mohou využívat koncepty, hierarchie a vztahy jako společný referenční bod pro analýzu dat. SNOMED CT slouží jako základ, na kterém mohou zdravotnické organizace vyvinout efektivní aplikace, aby mohly provádět výzkum ze závěrů, hodnotit kvalitu péče a náklady na ní a aby mohly navrhnout efektivní lékařská doporučení pro léčbu.

Standardizovaná terminologie může přinést výhody lékařům, pacientům, administrátorům, softwarovým vývojářům a plátcům. Klinická terminologie může pomoci poskytovatelům lékařské péče tak, že jim poskytne jednodušeji dostupné a kompletní informace, které náleží k procesu zdravotnické péče (chorobopis pacienta, nemoci, léčby, laboratorní výsledky, atd.) a proto vyúsťují v lepší výsledky v péči o pacienta. Klinická terminologie může umožnit poskytovateli lékařské péče identifikovat pacienty podle zakódované informace v jejich záznamech a tím usnadnit další vyšetřování a léčbu [4].

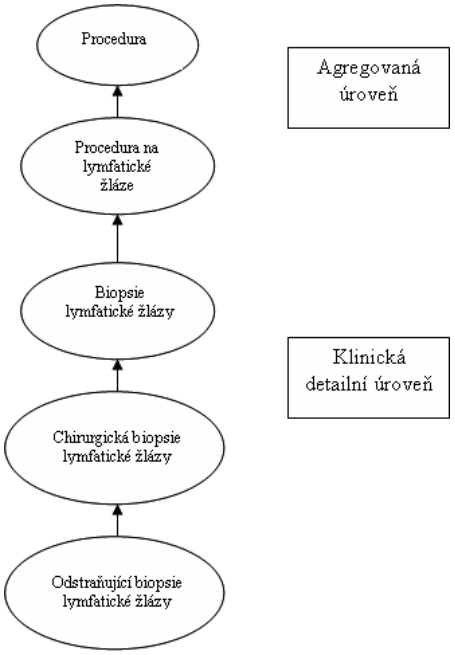
4. Základní komponenty terminologie SNOMED CT

4.1. Koncepty

V rámci SNOMED CT znamená „koncept“ klinický význam, který je identifikován jedinečným numerickým

identifikátorem (ConceptID), který se nikdy nemění. Koncepty jsou reprezentovány jedinečným, pro člověka čitelným „Zcela specifickým názvem“ (Fully Specified Name) (FSN). Koncepty jsou formálně definovány ve vztazích k dalším konceptům. Tyto „logické definice“ poskytují explicitní význam, který může počítač zpracovat a dotazovat se na něj. Každý koncept má také skupinu termínů, které pojmenovávají koncept způsobem čitelným pro člověka.

Koncepty představují různé stupně klinického detailu. Koncepty mohou být velice obecné nebo mohou představovat zvyšující se specifické úrovně detailu, kterým se také říká zvyšující se granularita. Zvyšující úrovně granularity zlepšují schopnost kódovat klinická data v náležité úrovni detailu.



Obrázek 1: Úrovně granularity

Koncepty ve SNOMED CT mají jedinečné numerické identifikátory, které se nazývají ConceptID. ConceptID neobsahuje hierarchické nebo implicitní významy. Numerický identifikátor neukazuje žádnou informaci o povaze konceptu.

Příklad: 367416001 je ConceptID pro koncept *angina pectoris (disorder)*.

4.2. Popisy (druhy, označení)

Popisy konceptu (concept descriptions) jsou termíny nebo názvy, které jsou přiděleny konceptu ve SNOMED CT. „Termín“ v tomto kontextu znamená frázi, která je použita k pojmenování konceptu. Jedinečné DescriptionID identifikuje popis. Násobné popisy mohou být spojeny s konceptem, který je identifikován svým ConceptID.

Příklad: Několik popisů spojených s ConceptID 22298006:

- Zcela specifický název: *Myocardial infarction (disorder)*
  DescriptionID 751689013
- Preferovaný termín: Myocardial infarction
  DescriptionID 37436014
- Synonymum: Cardiac infarction
  DescriptionID 37442013
- Synonymum: Heart attack
  DescriptionID 37443015
- Synonymum: Infarction of heart
  DescriptionID 37441018

Každý z výše zmíněných popisů má jedinečné DescriptionID a všechny tyto popisy jsou spojeny s jedním konceptem (a jedním ConceptID 22298006).

4.2.1 Druhy popisů:

Fully Specified Name (FSN) (Zcela specifický název)

Každý koncept má jeden jedinečný FSN, který má poskytnout jednoznačný způsob, jak pojmenovat koncept. Účelem FSN je jednoznačně identifikovat koncept a objasnit jeho význam. Neznamená to nutně, že představuje nejčastěji používanou nebo přirozenou frázi konceptu. Každý FSN je ukončen „sémantickým přívlastkem“, který je v závorce na konci konceptu. „Sémantický přívlastek“ označuje sémantickou kategorii, do které koncept patří (např. Disorder (choroba), Organism (organismus), Person (osoba), atd.). Například *Hematom (morfologická abnormalita)* je FSN, které představuje popis toho, co patologové vidí na úrovni tkáně, zatímco *Hematom (choroba)* je FSN, který označuje koncept, který by použili praktičtí lékaři pro kódování klinické diagnózy hematomu.

Preferred Term (Preferovaný termín)

Každý koncept má jeden preferovaný název, který zachycuje obvyklé slovo nebo frázi, kterou pojmenovávají koncept kliničtí lékaři. Například koncept *54987000 Repair of common bile duct (procedure) (obnova žlučovodu (procedura))* má preferovaný termín *Choledochoplasty (plastika*

žlučovodu), který představuje obvyklý název, který používají kliničtí lékaři k popisu této procedury.

Na rozdíl od FSN nemusí být preferované termíny jedinečné. Občas se může stát, že preferovaný termín pro jeden koncept může být synonymem nebo preferovaným termínem pro jiný koncept.

Příklad: *Cold sensation quality (qualifier value) (druh pocitu nachlazení (hodnota kvalifikátoru))* má preferovaný termín *Cold (nachlazení)*. *Common cold (disorder) (běžné nachlazení (choroba))* má synonymum *Cold (nachlazení)*.

V obou případech představuje *cold (nachlazení)* obvyklou klinickou frázi, která se používá k zachycení významu FSN.

Synonyma

Synonyma představují další doplňkové termíny, které představují stejný koncept jako FSN. Synonyma, stejně jako preferované termíny, nemusí být jedinečné.

Příklad: Některá synonyma, která jsou spojena s ConceptID *22298006*, který má FNS *Myocardial infarction (disorder)* jsou:

- Synonymum: Cardiac infarction
  DescriptionID 37442013
- Synonymum: Heart attack
  DescriptionID 37443015
- Synonyum: Infarction of heart
  DescriptionID 37441018

4.3. Vztahy

Koncepty ve SNOMED CT jsou propojovány pomocí vztahů. Existují čtyři druhy vztahů, které mohou být ve SNOMED CT přiřazeny konceptům:

- defining (definující),
- qualifying (vymezující),
- historical (historické),
- additional (doplňkové).

Každý koncept ve SNOMED CT je logicky definovaný svými vztahy k jiným konceptům.

5. Hierarchie

Koncepty SNOMED CT jsou organizovány do hierarchií. Koncept klasifikace SNOMED CT je „Root concept“ (kořenový koncept). Koncept zahrnuje koncept nejvyšší úrovně (supertyp) a všechny koncepty pod ním (jeho subtypy). Protože jsou hierarchie klesající, tak koncepty uvnitř nich se stávají více a více specifickými (nebo-li granulovanými). „Subtypy“ (nebo-li „potomci“) jsou potomci „supertypu“ (nebo-li „rodiče“).

Příklad: *Streptococcal arthritis (disorder) (streptokoková artritida (choroba))* je subtypem konceptu *Bacterial arthritis (disorder) (bakteriální artritida (choroba))*.

„Supertypy“ jsou předky „subtypu“.

Příklad: *Bacterial arthritis (disorder)* je supertyp *Streptococcal arthritis (disorder)*.

Mezi nejvyšší hierarchie patří:

- clinical finding (klinický nález),
- procedure (procedura),
- observable entity (pozorovatelná entita),
- body structure (struktura těla),
- organism (organismus),
- substance (substance),
- pharmaceutical/biologic product (farmaceutický/biologický produkt),
- specimen (vzorek),
- special concept (speciální koncept),
- physical object (fyzický předmět),
- physical force (fyzikální síla),
- event (událost),
- enviroments/geographical locations (prostředí/geografická místa),
- social context (sociální kontext),
- situation with explicit context (situace s explicitním kontextem),
- staging and scales (fáze a měřítka),
- linkage concept (spojovací koncept),
- qualifier value (hodnota kvalifikátoru) a
- record artifact (artefakt záznamu).

Hierarchie ***Klinický nález*** obsahuje sub-hierarchii *Disease* (nemoc). Koncepty, které jsou potomci *Disease* (nebo disorders (choroby, zdravotní potíže)), jsou vždy abnormální klinické stavy.

Koncepty ***Procedura*** představují aktivity, které jsou prováděny při péči o zdraví. Tato hierarchie představuje širokou škálu aktivit, včetně, ale ne pouze, invazních procedur *(odstranění nitrolebeční tepny (procedura))*, podávání léků *(očkování proti černému kašli (procedura))*, zobrazovací procedury *(ultrasonografie prsu (procedura))*, vzdělávací procedury *(osvěta o dietě s nízkým obsahem soli (procedura))* a administrativní procedury *(přenos lékařských záznamů (procedura))*.

Situace s explicitním kontextem byla až do července 2006 nazývaná kategorií závislou na kontextu. Tato hierarchie byla potom přejmenována, aby lépe popsala význam konceptů v této hierarchii. Koncepty v hierarchii *Procedura* a *Klinické nálezy* mohou v klinickém záznamu představovat podmínky a procedury, které ještě neproběhly (např. *plánovaná endoskopie (situace)*); podmínky a procedury, které se vztahují k někomu jinému než k pacientovi (např. *rodinná anamnéza: diabetes mellitus (situace)*) nebo podmínky a procedury, které se objevily v jiné době než v přítomnosti (např. *záznamy o dřívější splenektomii (situace)*). Ve všech těchto případech je klinický kontext upřesněný. Druhý příklad, ve kterém je důraz konceptu kladen na jinou osobu než na pacienta, může být vyjádřen ve zdravotním záznamu kombinací záznamu v „rodinné anamnéze“ s hodnotou „diabetes“. Specifický kontext (v tomto případě rodinná anamnéza) by byl vyjádřen strukturou záznamu. V tomto případě kontextově závislý koncept *Rodinná anamnéza: diabetes mellitus (situace)* by se nepoužil, protože informační model už aspekt diabetu mellitu v rodinné anamnéze zachytil.

Na koncepty v hierarchii ***Pozorovatelná entita*** můžeme pomýšlet jako na ty, které zastupují otázku nebo proceduru, které mohou podat odpověď nebo výsledek. Například *levý ventrikulární koncový diastolický tlak (pozorovatelná entita)* by mohl být interpretován jako otázka: „Co je to levý ventrikulární koncový diastolický tlak?“ nebo „Co je to měřený levý ventrikulární koncový diastolický tlak?“. Pozorovatelné veličiny jsou elementy, které mohou být použity k zakódování elementů na kontrolním seznamu nebo jakýkoli element, kterému může být přidělena hodnota. *Barva nehtu (pozorovatelná entita)* je pozorovatelná veličina. *Šedé nehty (nález)* je nález. Jedno z využití *pozorovatelných entit* v klinickém záznamu jsou záhlaví v šabloně. *Pohlaví (pozorovatelná entita)* může být využito k zakódování sekce šablony „pohlaví“, kde by si uživatel vybral „muž“ nebo „žena“. „Ženský rod“ by potom znamenal nález.

Koncepty ***Struktura těla*** zahrnují normální i abnormální anatomické struktury. Normální anatomické struktury mohou být použity ke specifikaci místa na těle, které se týká nemoci nebo procedury, např. *struktura mitrální chlopně (struktura těla)*. Morfologické změny normálních struktur těla jsou vyjádřeny sub-hierarchií *Struktura těla, změněná od své původní anatomické struktury (morfologická abnormalita)*. Příklad může být *polyp (morfologická abnormalita)*.

Hierarchie ***Organismus*** zahrnuje důležité organismy v lidské a zvířecí medicíně. Organismy se ve SNOMED CT používají také při modelování příčin nemocí. Jsou důležité ve veřejném zdravotnictví pro podmínky podléhající ohlašovací povinnosti a pro protokoly o nakažlivých nemocech v klinických systémech podpory rozhodování. Sub-hierarchie organismu zahrnují například *zvíře (organismus)*, *mikroorganismus (organismus)*, *rostlina (organismus)*. Příklad konceptu Organismus je *lišejník (rostlina) (organismus)*.

Hierarchie ***Substance*** zahrnuje koncepty, které se používají pro zaznamenávání aktivních chemických složek léků, potravin a chemických alergenů, nepříznivých účinků, toxicity nebo informací o otravě a pokynů lékařů a sester. Koncepty z této hierarchie představují obecné „substance“ a chemické složky *Farmaceutického/biologického produktu (produkt)*, který je v separátní hierarchii. Nicméně, sub-hierarchie *Substance* také zahrnují například *substanci těla (substance)* (koncepty, které vyjadřují substance těla); *dietní substanci (substance)* a *diagnostickou substanci (substance)*. Příkladem je *insulin (substance)*.

Hierarchie ***Farmaceutický/biologický produkt*** stojí odděleně od hierarchie *Substance*. Tato hierarchie má jasně rozlišovat léčiva (produkty) od jejich chemických složek (substance). Například *Diazepam (produkt)*.

Hierarchie ***Vzorek*** zahrnuje koncepty, které představují entity, které jsou získány (většinou od pacienta) během vyšetření nebo analýzy. *Vzorky* mohou být definovány atributy, které specifikují: normální nebo abnormální struktura těla, ze které jsou získány; procedura, která se používá ke sběru vzorků; zdroj, ze kterého byly sebrány a substance, ze které se skládají. Příkladem je *vzorek z prostaty získaný jehlovou biopsií (vzorek)*.

Koncepty v hierarchii ***Fyzický předmět*** zahrnují přírodní a umělé předměty. Jedním z použití těchto konceptů je modelování procedur, které používají různá zařízení (např. katetrizace). Příkladem konceptu v této hierarchii je *filtr duté žíly (fyzický předmět)*.

Koncepty v hierarchii ***Fyzická síla*** jsou zaměřeny zejména na vyjádření fyzických sil, které mohou hrát roli jako mechanismus zranění, například *střídavý proud (fyzická síla)*.

Hierarchie ***Událost*** zahrnuje koncepty, které zastupují výskyty (vyjma procedur a zásahů). Příkladem těchto konceptů je *bioteroristický útok (událost)* nebo *zemětřesení (událost)*.

Hierarchie ***Prostředí a geografická místa*** obsahuje různé druhy prostředí a také názvy míst jako jsou země, státy a regiony, například *Kanárské ostrovy (geografické místo)*, *rehabilitační oddělení (prostředí)* nebo *jednotka intenzivní péče (prostředí)*.

Hierarchie ***Sociální kontext*** obsahuje sociální podmínky a okolnosti, které jsou důležité pro zdravotnictví. Patří sem rodinný stav, ekonomický stav, etnické a náboženské dědictví, životní styl a povolání. Tyto koncepty představují sociální aspekty, které ovlivňují zdraví a léčbu pacienta. Mezi sub-hierarchie *Sociálního kontextu* patří: etnická skupina, povolání, osoba, náboženství/filosofie a ekonomický status.

Hierarchie ***Fáze a měřítka*** je rozdělena na sub-hierarchie jako jsou *hodnotící škála* a *fáze nádoru*.

Hierarchie ***Spojovací koncept*** obsahuje koncepty, které se používají pro vazby. Dělí se na sub-hierarchie *uplatnění vztahu* a *atribut*. Sub-hierarchie uplatnění vztahu umožňuje použití konceptů klasifikace SNOMED CT ve výkazech HL7, které prokazují vztahy mezi výkazy. Příkladem konceptu uplatnění vztahu je *má vysvětlení (uplatnění vztahu)*. Koncepty, které se odvozují od této sub-hierarchie jsou používány ke stavbě vztahů mezi dvěmi koncepty klasifikace SNOMED CT, jelikož ukazují druh vztahu mezi těmito koncepty. Některé atributy mohou být použity k logické definici konceptu (definující atributy). Tato sub-hierarchie také zahrnuje nedefinující atributy (jako ty, které se používají ke sledování historických vztahů mezi koncepty) nebo atributy, které mohou být užitečné k modelování definic konceptů, ale které nebyly ještě použity v modelování dřívějších konceptů ve SNOMED CT.

Hierarchie ***Hodnota kvalifikátoru*** zahrnuje některé koncepty, které se používají jako hodnoty pro atributy SNOMED CT, které nejsou zahrnuty nikde jinde ve SNOMED CT. Nicméně tyto hodnoty pro atributy nejsou omezeny pouze na tuto hierarchii a mohou být nalezeny i v jiné hierarchii. Příkladem konceptu této hierarchie je *levý (hodnota kvalifikátoru)* nebo *jednostranný (hodnota kvalifikátoru)*.

Jednou ze sub-hierarchií ***Speciálního konceptu*** je *Nečinný koncept*, který je supertypem pro všechny koncepty, které byly ukončeny a ukazují na aktivní koncept v terminologii.

Artefakt záznamu je entita, která je vytvořena osobou nebo osobami, aby poskytla dalším lidem informace o událostech a stavech různých záležitostí. Většinou je záznam nezávislý na svých jednotlivých fyzických doložených příkladech a skládá se z jednotlivých částí informací (většinou slov, slovních spojení a vět, ale také z čísel, grafů a další elementů informací). *Artefakty záznamu* nemusí být kompletní zprávy nebo kompletní záznamy. Mohou být částí větších *artefaktů záznamu*. Například celkový zdravotní záznam je *artefakt záznamu*, který také může obsahovat další *artefakty záznamu* ve formě jednotlivých dokumentů nebo zpráv, které na druhou stranu mohou obsahovat jemněji granulované *artefakty záznamů* jako jsou sekce nebo dokonce záhlaví sekcí.

6. Minimální datový model pro kardiologii

Minimální datový model pro kardiologii (MDMK) [5, 6, 7, 8] byl sestaven v letech 2000–2004 v rámci výzkumného centra EuroMISE - Kardio. Kardiologie je velice rozsáhlý obor a proto byl MDMK zaměřen pouze na aterosklerotická kardiovaskulární onemocnění. Cílem tohoto datového modelu je 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 osmi skupin znaků. Na začátku je rodinná anamnéza, následuje sociální anamnéza a toxikománie, osobní anamnéza, současné obtíže možného kardiálního původu, dosavadní léčba, fyzikální vyšetření a blok parametrů EKG.
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 Městkou nemocnicí Čáslav. V současné době jsou v databázi ADAMEK zaznamenána data o 1289 pacientech.

7. Atributy MDMK zakódované pomocí SNOMED CT

Tabulka 1 ukazuje několik příkladů atributů z Minimálního datového modelu, kterým bylo přiděleno ConceptID z klasifikačního systému SNOMED CT. Prvním předpokladem kódování, je ale přeložení názvu atributů do anglického jazyka, jelikož v současné době existuje pouze americká, britská, španělská a německá verze.

Atributy z MDMK		English equivalent	SNOMED CT (Concept ID)
rodinný stav			
	svobodný/á	Marital status: single, never married (finding)	125725006
	ženatý/vdaná	Legally married (finding)	36629006
	vdovec/vdova	Widowed (finding)	33553000
	rozvedený/á	Divorced (finding)	20295000
	jiný	Other	
žije sám		Lives alone (finding)	105529008
nejvyšší dosažené vzdělání			
	základní	Educated to secondary school level (finding)	224297003
		Continued education to sixth form (finding)	224298008
	středoškolské	Received higher education (finding)	224299000
		Received polytechnic education (finding)	224301007
		Received higher education college education (finding)	224302000
	vysokoškolské	Received university education (finding)	224300008
alergie na léky		Drug allergy (disorder)	416098002
		Allergic reaction to drug (disorder)	416093006
hypertenze		Essential hypertension (disorder)	59621000
		High blood pressure (& [essential hypertension])	194757006
		Essential hypertension NOS (disorder)	266228004
hyperlipoproteinémie		Hyperlipoproteinemia (disorder)	3744001
		Fredrickson type IV hyperlipoproteinemia (disorder)	238085009
		Fredrickson type I hyperlipoproteinemia (disorder)	238086005
		Familial type 5 hyperlipoproteinemia (disorder)	34349009
		Familial hyperlipoproteinemia (disorder)	238038003
		Familial type 3 hyperlipoproteinemia (disorder)	398796005
		Fredrickson type IIa hyperlipoproteinemia (disorder)	397915002
ischemická choroba srdeční		Ischemic heart disease (disorder)	414545008
dušnost		Asthma (disorder)	187687003
bolest na hrudi		Dull chest pain (finding)	3368006
palpitace		(Palpitations) or (awareness of heartbeat) or (fluttering of heart)	161965005
otoky		Swelling or edema (finding)	248477007
synkopa		Syncope (disorder)	271594007
klaudikace		Claudication (finding)	275520000
hmotnost		On examination - weight NOS (finding)	162770007
		Height and weight (observable entity)	162879003
výška		Body height measure (observable entity)	50373000
tělesná teplota		Body temperature finding	105723007
		Body temperature (observable entity)	276535009
obvod pasu		Abdominal girth measurement (procedure)	48094003
dechová frekvence		Respiratory rate (observable entity)	86290005

Tabulka 1: Atributy z MDMK

8. Závěr

Efektivní péče o zdraví vyžaduje dobré informace. Bezpečná a vhodná výměna klinických informací je nezbytná k zajištění kontinuity péče o pacienty a to v různých časech, na různých místech a u různých poskytovatelů zdravotní péče. Současné zdravotnické informační systémy umožňují sbírat různé klinické informace, jsou propojeny s klinickými znalostními databázemi, mohou vyhledávat data, shromažďovat data, analyzovat data, vyměňovat si data a mají i plno dalších funkcí. SNOMED CT může poskytnout základy pro tyto funkce. Informační systémy mohou využít koncepty, hierarchie a vztahy jako společný referenční bod. SNOMED CT ale může i překročit přímou péči o pacienty. Tato terminologie může, například, usnadnit podporu rozhodování, statistické zpracovávání,

sledování veřejného zdraví, zdravotnický výzkum a analýzy nákladů.

Mapování terminologie v aplikacích elektronického zdravotního záznamu na mezinárodně používané klasifikační systémy je základem pro interoperabilitu heterogenních systémů elektronického zdravotního záznamu.

Literatura

[1] http://www.ihtsdo.org/snomed-ct/.
http://www.nlm.nih.gov/research/umls/Snomed/

[2] snomed_main.html (last reviewed June 24th, 2008).

[3] The International Health Terminology Standards Development Organisation: SNOMED Clinical Terms® User Guide. January 2008 International Release.

[4] Přečková P., Zvárová J., Špidlen J., „International Nomeclatures in Shared Healthcare in the Czech Republic“, Proceedings of 6th Nordic Conference on eHealth and Telemedicine, Helsinky, Finland. pp. 45-46.

[5] Adášková J., Anger Z., Aschermann M., Bencko V., Berka P., Filipovský J., Goláň L., Grus T., Grünfeldová H., Haas T., Hanuš P., Hanzlíček P., Holcátová I., Hrach K., Jiroušek R., Kejřová E., Kocmanová D., Kolář J., Kotásek P., Králíková E., Krupařová M., Kyloušková M., Malý M., Mareš R., Matoulek M., Mazura I., Mrázek V., Novotný L., Novotný Z., Pecen L., Peleška J., Prázný M., Pudil P., Rameš J., Rauch J., Reissigová J., Rosolová H., Rousková B., Říha A., Sedlak P., Slámová A., Somol P., Svačina Š, Svátek V., Šabík D., Šimek S., Škvor J., Špidlen J., Štochl J., Tomečková M., Umnerová V., Zvára K., Zvárová J., „Návrh minimálního datového modelu pro kardiologii a softwarová aplikace ADAMEK. Interní výzkumná zpráva EuroMISE Centra - Kardio“, Praha, říjen 2002.

[6] Tomečková M., „Minimální datový model kardiologického pacienta - výběr dat“, Cor et Vasa, 2002, Vol. 44, No. 4 Suppl., s. 123.

[7] Mareš R., Tomečková M., Peleška J., Hanzlíček P., Zvárová J., „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.

[8] Přečková P., „Jazyk lékařských zpráv“, Doktorandský den 2007. Praha, MATFYZPRESS 2007, ISBN 978-80-7378-019-7, s. 75-79.

Nevyužité možnosti sémantického webu

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 podpořena projektem 1ET100300419 programu Informační společnost (Tématického programu II Národního programu výzkumu v ČR: Inteligentní modely, algoritmy, metody a nástroje pro vytváření sémantického webu), 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".

Abstrakt

Vize sémantického webu byla představena před skoro již 10 lety, avšak žádná z její aplikací prozatím nedokázala oslovit takové množství lidí, jaké dnes používá web v současné podobě. Příspěvek se věnuje možnostem sémantického webu a přínosům, které může přinést pro koncové uživatele. Nejprve podává přehled o současných technologiích i jejich použití a následně diskutuje možnosti plynoucí z použití odkazů v prostředí sémantického webu tak, jak je známe z webu současného, tedy rozšiřující, zpřesňující či udávající kontext prezentované informace.

1. Vyhledávání a vize sémantického webu

Současný web čelí mnoha problémům. Mezi ty nejstěžejnější patří problematika vyhledávání relevantních informací na webu. Ta je dnes většinou řešena pomocí tzv. *information retrieval* nástrojů [1], které pracují s inverzními indexy uchovávající (četnost) výskytu jednotlivých slov v (webových) dokumentech. Relevance dokumentu je pak stanovena pomocí kosinové míry reflektující podobnost mezi zadanými klíčovými slovy a slovy obsaženými v daném dokumentu.

Tato relevance však nic neříká o kvalitě poskytovaných dat. Proto bývá rozšířena o další nepřímou míru udávající odhadnutou kvalitu dat prezentovaných v dokumentu. Jednou z takových měr je Page-Rank [2], který je založen na předpokladu, že dokumenty prezentující kvalitní data jsou častěji odkazovány z jiných (kvalitních) dokumentů. Zavedením této míry se podařilo uspořádat (vůči klíčovým slovům relevantní) dokumenty i podle jejich kvality.

Díky značné redundanci dat na současném internetu však ani takové uspořádání nemusí vést ke zlepšení vypovídací schopnosti výsledku hledání. Na většinu dotazů dnešní vyhledávače vrátí desetitisíce odkazů; koncový uživatel mnohdy stěží analyzuje první dvacítku odkazů a ostatní, i z hlediska úspory času, zcela ignoruje. To vede k faktu, že získání *kompletní informace* pomocí současných vyhledávacích nástrojů je velmi obtížné, ne-li nemožné.

Nejen tento problém se snaží vyřešit vize sémantického webu [3, 4], která umožňuje definovat vedle samotných dat i metadata k jejich popisu. Jinými slovy nedefinuje pouze objekty jako takové, ale vymezuje popis objektu pomocí ostatních (stejným způsobem popsaných) objektů. Například popis třídy *dítě* je možné vztáhnout k popisu třídy *osoba*.

Dokumenty sémantického webu se skládají z RDF[1] trojic

$$(\text{object}, \text{predicate}, \text{subject}) \in (\mathbb{R} \cup \mathbb{B}) \times \mathbb{R} \times (\mathbb{R} \cup \mathbb{B} \cup \mathbb{L})$$

kde [5]

- $\mathbb{R}$ značí množinu tzv. *resources* identifikující popisované objekty;
- $\mathbb{B}$ značí množinu tzv. *blank nodes*, které sami o sobě nemají žádný význam, sloužících k identifikaci složitějších (vícearitních) struktur;
- $\mathbb{L}$ značí množinu literálů. Ta může být dále rozšířena o informaci o použitém přirozeném jazyku či terminologii.

[1] Resource Description Framework

Každý resource $\mathbb{R}$ je, dle definice, identifikován pomocí URI, např. ve tvaru

```
http://example.com/ontologie#dite
```

Vyhledávání v prostředí sémantického webu se primárně soustředí na vytváření indexu ukazující, který resource je popsán ve kterém dokumentu. Prohledávání takových indexů ale může být spojeno s odvozování, např. při hledání instancí třídy *osoba* zahrnout i instance třídy *dítě*.

Současný sémantický web se spíše orientuje na vytyčení pojmů pomocí ontologií; je známé nasazení vize sémantického webu v prostředí webových služeb, kdy jejich ontologický popis umožňuje kooperaci mezi dílčími webovými službami. Sémantický web je ale i odpovědí na otázku, jak najít na webu kompletní informaci samotnou, ne pouze odkazy na ní, tak, jak se dělají dnešní vyhledávače.

2. Formáty používané na webu

Za první formát webových dokumentů lze považovat HTML[2], který rozšířil formátovaná data o hypertextové odkazy. Tento formát je postaven na SGML, dnes se většinou používá jako základ striktnější XML[3]. Fragment takového HTML dokumentu může být ilustrován například pomocí:

```
<div class='item'>
   <img src='disk.samsung.spinpoint-F1-500GB.jpg'
        alt='Disk Samsung Spin Point F1 500GB'/>
   <div>Disk Samsung Spin Point F1 500GB</div>
   <ul>
     <li><b>Product No</b>:
         HD202IJ</li>
     <li><b>Interface:</b>
         SATA-II</li>
     <li><b>Space:</b>
         500GB</li>
     <li><b>RPM:</b>
         7200</li>
     <li><b>Warranty</b>:
         36 months</li>
     <li><b>Price</b>:
         1273 CZK</li>
     <li><b>Price incl. VAT</b>:
         1557 CZK</li>
     <li><b>Produced by:</b>
         <a href='http://www.samsung.com/global/business/hdd/productmodel.do?
            group=72&type=61&subtype=63&model_cd=240&ppmi=1155'>Samsung</a>
   </ul>
</div>
```

Takovýto fragment dokumentu může být zaindexován fulltextovými vyhledávači, jako relevantní je možné vybrat klíčová slova *SATA-II, HD202IJ, Samsung, Spin Point F1, 500GB*. Pakliže koncový uživatel zvolí některé z těchto klíčových slov, dříve či později by měl ve výsledku vyhledávání narazit na odkaz na dokument obsahující tento fragment. Pokud si uživatel bude chtít vybrat tento disk z nabídky všech prodejců, nezbude mu nic jiného, než projít ručně všechny tyto prodejce.

Naopak dokumenty sémantického webu jsou předurčeny pro další strojové zpracování. Vzhledem k tomu, že se prozatím nepodařilo v dostatečné míře prosadit publikování dat ve formátech sémantického webu, uchýlilo se konsorcium W3C, definující standarty v oblasti webu, v roce 2004 k návrhu rozšíření formátu HTML o další atributy RDFa[4]. Účelem rozšíření je zavést možnost sémantické anotace přímo do HTML dokumentů. Stejný fragment by pak vypadal následovně:

```
<div aboat='HD202IJ-in-my-shop' class='item'
     xmlns:disk-ont='http://example.com/disk-ont'
     xmlns:myshop='http://myshop.com'>
   <img src='disk.samsung.spinpoint-F1-500GB.jpg'
      alt='Disk Samsung Spin Point F1 500GB' rel='picture'/>
   <div property='disk-ont:Name'>Disk Samsung Spin Point F1 500GB</div>
   <ul>
      <li><b>Product No</b>:
         <span property='disk-ont:Product-ID'>HD202IJ</span></li>
      <li><b>Interface:</b>
         <span property='disk-ont:Interface'>SATA-II</span></li>
      <li><b>Capacity:</b>
         <span property='disk-ont:Capacity'500GB</span></li>
      <li><b>RPM:</b>
         <span property='disk-ont:Disk-rpm'>7200</span></li>
      <li><b>Warranty</b>:
         <span property='disk-ont:Warranty'> 36 months</span></li>
      <li><b>Price</b>:
         <span property='myshop:Price'>1273 CZK</span></li>
      <li><b>Price incl. VAT</b>:
         <span property='myshop:Price-inc-VAT'> 1557 CZK</li>
      <li><b>Produced by:</b>
      <a href='http://www.samsung.com/global/business/hdd/productmodel.do?
           group=72&type=61&subtype=63&model_cd=240&ppmi=1155'
         rel='disk-ont:Producer'>Samsung</a>
   </ul>
</div>
```

Z takto anotovaného dokumentu lze pomocí XSLT[5] transformace (obecně transformující jeden XML dokument na jiný dokument) získat přímo popis vlastností disku v RDF. Získaný fragment RDF dokumentu pak bude

```
<rdf:Description rdf:about='HD202IJ-in-my-shop'
                 xmlns:disk-ont='http://example.com/disk-ont'
                 xmlns:myshop='http://myshop.com'>
   <disk-ont:Picture rdf:resource='disk.samsung.spinpoint-F1-500GB.jpg'/>
   <disk-ont:Name>Disk Samsung Spin Point F1 500GB</disk-ont:Name>
   <disk-ont:Product-ID>HD202IJ</disk-ont:product-ID>
   <disk-ont:Interface>SATA-II</disk-ont:interface>
   <disk-ont:Capacity>500GB<disk-ont:capacity>
   <disk-ont:Disk-rpm>7200</disk-ont:disk-rpm>
   <disk-ont:Warranty>36 months</disk-ont:warranty>
   <myshop:Price>1273 CZK</myshop:price>
   <myshop:Price-inc-VAT> 1557 CZK</myshop:Price-inc-VAT>
   <disk-ont:Producer
     rdf:resource='http://samsung.com/global/business/hdd/productmodel.do?
        group=72&type=61&subtype=63&model_cd=240&ppmi=1155' />
</rdf:Description>
```

Ani toto rozšíření se prozatím nedočkalo velkého ohlasu mezi producenty dat, a tak koncoví uživatelé zůstávají bez možnosti efektivně (automaticky) zpracovávat data v současné době schovaná uprostřed formátování.

[2] HyperText Markup Language
[3] Extensible Markup Language
[4] Resource Description Framework Attributes
[5] Extensible Stylesheet Language Transformations

3. Distribuované prostředí

Web jako takový je distribuované prostředí, ve kterém kdokoliv může publikovat cokoliv. Web si koncoví uživatelé navykli používat; pakliže najdou zajímavý dokument, jisto jistě prozkoumají i odkazy vedoucí z tohoto dokumentu. I z tohoto důvodu se navigaci uživatele po webových stránkách věnuje značná pozornost a je jedním z hlavních kritérií hodnocení kvality (přístupnosti) webu.

Všimněme si, že každý resource v sémantickém webu je identifikován pomocí URI. Co by se však stalo, kdyby namísto (virtuálního) URI dokument odkazoval stejně jako je to u současného webu na jiný webový dokument obsahující detailnější informace o popisovaném objektu? Ve zvoleném případě by výrobce disků publikoval na adrese *http://example.com/sata-II-disks.rdf* dokument popisující například sérii disků. Příklad fragmentu takového dokumentu nechť je následující

```
<disk-ont:disk rdf:ID='HD202IJ'
               xmlns:disk-ont='http://example.com/disk-ont.rdf'>
   <disk-ont:picture rdf:resource='disk.samsung.spinpoint-F1-500GB.jpg'/>
   <disk-ont:Name>Disk Samsung Spin Point F1 500GB</disk-ont:Name>
   <disk-ont:product-ID>HD202IJ</disk-ont:product-ID>
   <disk-ont:interface>SATA-II</disk-ont:interface>
   <disk-ont:capacity>500GB<disk-ont:capacity>
   <disk-ont:disk-rpm>7200</disk-ont:disk-rpm>
   <disk-ont:warranty>36 months</disk-ont:warranty>
</disk-ont:disk>
```

přičemž jednotlivé vlastnosti mohou být definovány v externí ontologii *http://example.com/disk-ont.rdf*:

```
<rdfs:Property rdf:ID='disk-ont:Name'>
    <rdfs:label xml:lang='en'>Product Name</rdfs:label>
    <rdfs:label xml:lang='cs'>Označení produktu</rdfs:label>
    ...
</rdfs:Property>
```

Jak je patrné, tato ontologie může obsahovat popisy vlastností v různých jazykových mutací. Ty mohou být následně využity pro generování HTML verze dokumentu, viz předchozí příklady.

Samotný obchod pak pouze deklaruje, že prodává daný disk a tuto informaci pouze rozšíří o specifika obchodu jako jsou cena, zkušenosti nakupujících a podobně:

```
<myshop:disk rdf:ID='HD202IJ-in-my-shop'
   <myshop:ProductDetail
      rdf:resource='http://example.com/sata-II-disks.rdf#HD202IJ'/>
   <myshop:Price>1273 CZK</myshop:price>
   <myshop:Price-inc-VAT> 1557 CZK</myshop:Price-inc-VAT>
</myshop:disk>
```

Tento model distribuce dat má několik výhod. První výhodou je nižší redundance dat, v původní architektuře každý prodejce musel uvádět veškerá data. Pro poskytovatele obsahu (ať výrobce či obchodníka) pak odpadá nutnost znovu zpracovávat data - pokud obchodník bude používat značení výrobce (ontologii poskytnutou výrobcem), má výrobce jistotu, že nedochází ke klamání koncového zákazníka se strany prodejce, naopak prodejce může deklarovat (např. elektronickým podpisem výrobce), že jím zprostředkovávaná data jsou ověřena. Obecně tímto postupem může být budována důvěra mezi subjekty publikující data na webu.

Další výhoda se uplatní u vyhledávání. Pokud se zákazník rozhodne pro daný disk, hledá již pouze prodejce, kteří tento disk nabízejí. Vzhledem k tomu, že disk je vždy identifikován pomocí URL na straně výrobce, je takové vyhledávání téměř triviální.

Toto zjednodušení vyhledávání je způsobeno tím, že není potřeba (heterogenní) data od různých prodejců integrovat. Integrace dat [6], neboli hledání korespondencí mezi daty více zdrojů a jejich následné spojování, sama o sobě představuje velmi těžkou a obecně automaticky [7] neřešitelnou úlohu. Čím složitější (a expresivnější) je popis objektů, tím je složitější i integrační proces. Díky tomu, že je objekt jednoznačně identifikován cílovou URL odkazu, není potřeba data integrovat v takovém rozsahu (integrují se pouze atributy specifické pro daného prodejce).

V neposlední řadě současné prohlížeče webových dokumentů umožňují zpracovat libovolný XML dokument a zobrazit jej buďto pomocí kaskádových stylů CSS a nebo pomocí XSLT transformace. Tato funkcionalita umožňuje stáhnout XML dokument obsahující pouze prostá RDF data, v jehož hlavičce je uvedeno, jakým způsobem mají být data zformátována. V případě XSLT transformace XML dokumentu do XHTML formátu je použita následující hlavička:

```
<?xml version='1.0' encoding='utf-8'?>
<?xml-stylesheet type='text/xsl' href='rdf2html.xslt'?>
```

kde *rdf2html.xslt* je šablona popisující transformaci z RDF trojic do HTML dokumentu. Tuto transformaci provede přímo prohlížeč a zobrazí její výstup. Koncový uživatel tak vůbec nepozná, že si neprohlíží klasickou webovou stránku, ale RDF dokument. Bohužel, tato technologie, byť je již dlouhodobě podporována všemi předními webovými prohlížeči, nebývá užívána, neboť současné vyhledávače nejsou schopni takto publikovaná data zpracovat. Tento způsob značně minimalizuje objem nutných datových přenosů, což je vhodné například u mobilních zařízení.

Další výhodou distribuované architektury a potažmo celého sémantického webu je fakt, že k takovýmto dokumentům mohou velmi jednoduše přistupovat

[6] Asynchronous JavaScript and XML

aplikace označované jako *Web X.0*. Tyto aplikace postupně načítají/modifikují zobrazovanou stránku pomocí AJAX[6] technologie, na straně prohlížeče spouštěných *javascriptových* programů umožňujících interakci mezi uživatelem a poskytovanými daty. Na jednotlivé RDF dokumenty lze pohlížet jako na tzv. *REST*[7] *webové služby* [8] volané AJAX programy. Zásadní nevýhodou této technologie je nemožnost indexace obsahu (neb aktuálně zobrazená data neodpovídají žádné URL, na kterou by se mohl uživatel později odkázat).

Tuto potencionální nevýhodu lze obejít publikováním jak RDF dokumentu formátovaného pomocí XML, tak statické HTML stránky, která vznikla identickou transformací na straně serveru. Tedy uživatel má možnost získat odkaz na (přibližně) stejný obsah reprezentovaný statickou HTML verzí, u které je uvedena korespondence s původním RDF dokumentem (například i pomocí RDFa rozšíření) a další navigace (hledání podobných produktů, více detailů, konkurenční prodejci) je zprostředkována již v rámci aktivní složky obsahu stránky.

Použití distribuované architektury tak, jak je popsána výše, v praxi naráží na pomalé odezvy webových serverů (čas potřebný k navázání spojení je podstatně větší nežli čas potřebný k samotnému přenosu dat). Tento problém lze vyřešit buďto efektivním cacheováním načtených dokumentů, které navíc může být podpořeno postupným načítáním obsahu pomocí AJAX aplikace.

4. Odhad struktury dat

Sémantický web umožňuje popisovat vlastnosti objektů pomocí vztahů. Tyto vztahy jsou definovány obecně pomocí resource - každý návrhář ontologie může použít své vlastní zavedení vlastností. Tento fakt obecně velmi ztěžuje jakékoliv složitější operace, včetně integrace ontologií. Z tohoto důvodu se mnohé nástroje poohlíží po podstatně jednodušších, byť méně popisných formalismech.

Vzhledem k nedostatku dat ve formátu sémantického webu je žádoucí najít způsob, jak využít data z webových stránek a extrahovat je do formátu sémantického webu (například anotací pomocí RDFa atributů). Pro anotaci je však potřeba znát strukturu dat; ta na webových stránkách nebývá uvedena a pak nezbývá nic jiného, než se ji pokusit odhadnout.

Strukturu dat lze popsat mnohými formalismy, ilustrujme ji na příkladu formalismu inspirovaném relačními databázemi [9]. Struktura dat je odhadnuta analýzou *extensionálních funkčních závislostí* platných na dané množině dat.

Funkční závislost mezi dvěma atributy je integritní omezení zajišťující jednoznačnou odvoditelnost hodnoty atributu na pravé straně při znalosti hodnoty atributu na levé straně. Příkladem funkční závislosti je například

$$\text{Stát} \rightarrow \text{Měna}$$

Samotné záznamy jsou popsány v odpovídající relaci. Všimněme si, že unární funkční závislost[8] je možné popsat pomocí odpovídající trojice

$$(\text{Stát}, \text{implies}, \text{Měna})$$

Abychom mohli stejným způsobem zavést i vztahy mezi hodnotami atributů, je vhodné pro každou funkční závislost definovat její *instance* [10]

$$A_1 \rightarrow A_2 \in \mathscr{F} \rightsquigarrow (A_1, A_1(t)) \rightarrow (A_2, A_2(t)) \in \mathscr{I}$$

kde

- $A_1, A_2 \in \mathscr{R}$ jsou atributy relace $\mathscr{R}$
- $A_\star(t)$ je zobrazení přiřazující záznamu t hodnotu atributu $A_\star$

Nazveme-li dvojici atribut-hodnota *elementem* (A, v), pak je možné tyto instance rovněž vyjádřit jako vztahy mezi elementy, které jsou popsány pomocí trojic

$$((A_1, v_1), \text{implies}, (A_2, v_2))$$

Takováto reprezentace dat ve formátech sémantického webu je vhodná v případě, že není zajištěna korektnost odhadnuté struktury dat. Pokud je odhadnutý model označen jako korektní, je možné data transformovat do formy [11]

$$(v_1, \text{name}(A_1 \rightarrow A_2), v_2)$$

kde $name$ je funkce pojmenovávající funkční závislosti. Pokud se přidržíme zvolené funkční závislosti, příkladem výsledku trasformace instance může být trojice

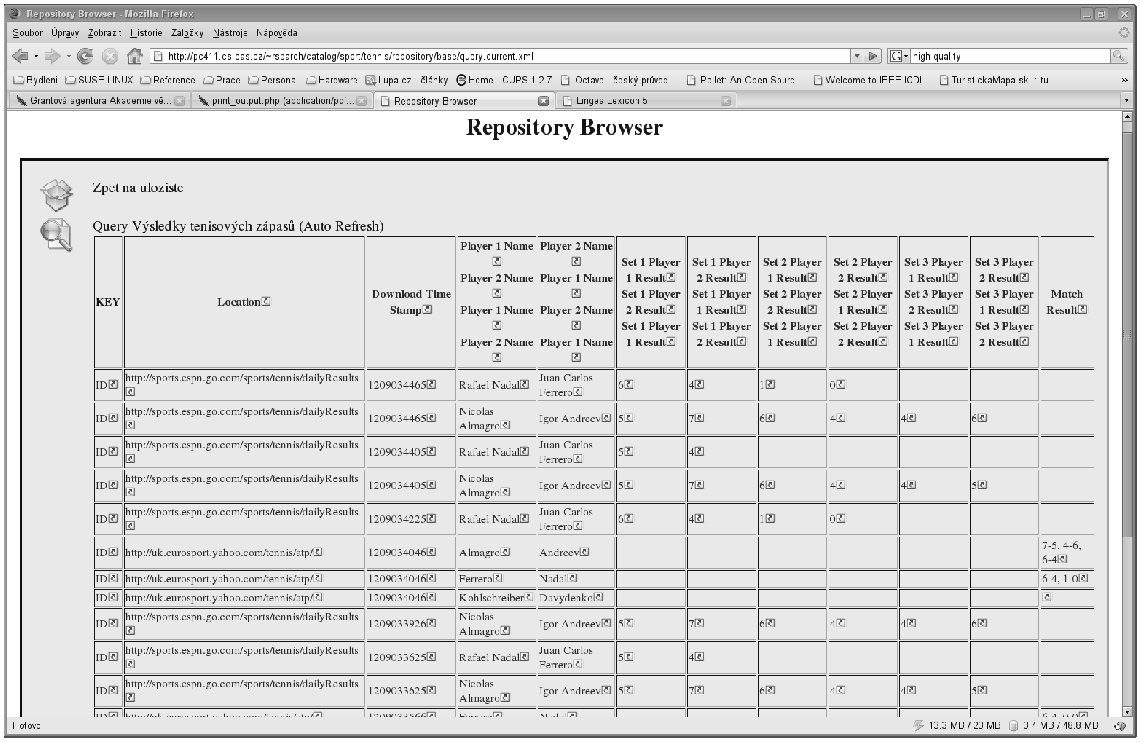
$$(\text{Česká Republika}, \text{has-a-Měna}, \text{Česká koruna})$$

Tyto trojice mnohou být uloženy do XML formátu. Například

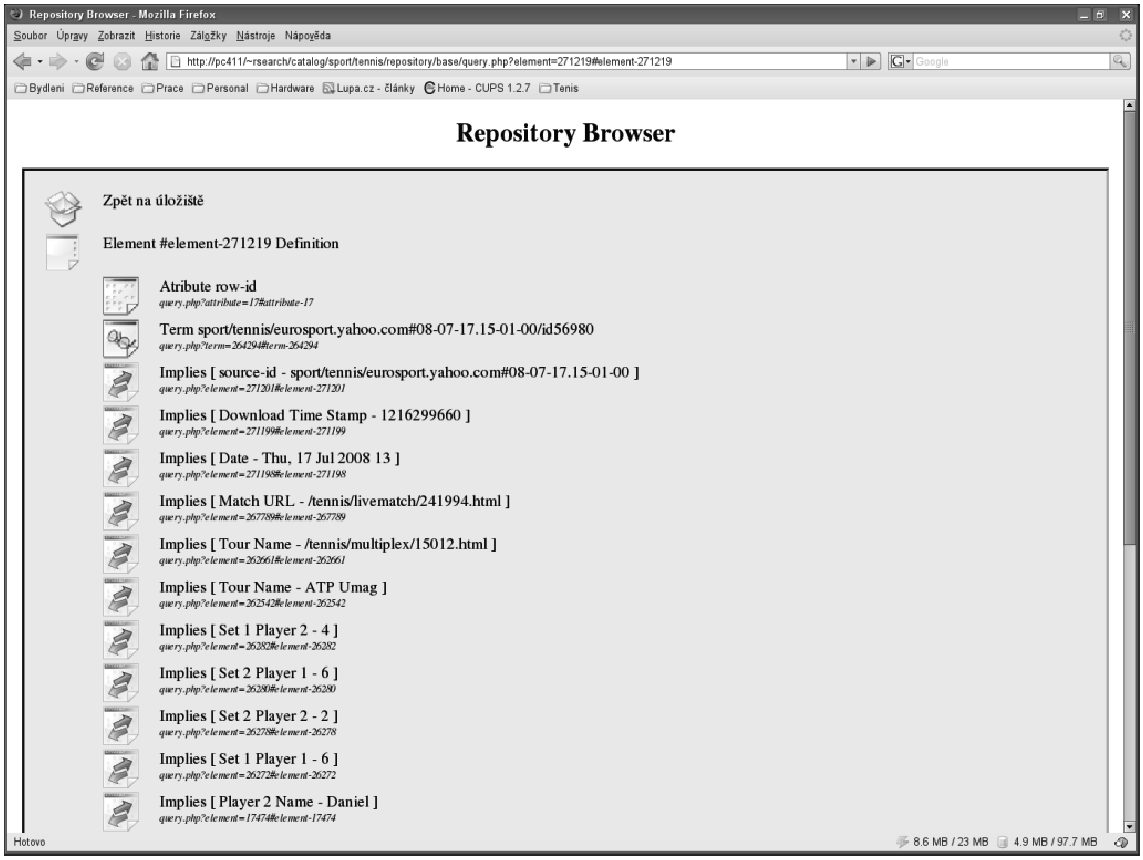
```
<state rdf:ID='CeskaRepublika'>
   <has-a-Mena rdf:resource=
     'CeskaKoruna.rdf#CeskaKoruna'/>
</state>
```

[7] Representational State Transfer

[8] Unární funkční závislost je funkční závislost mezi jednoduchými atributy (t.j. s aritou 1)



Obrázek 1: Ukázka stránky experimentálního portálu



Obrázek 2: Rekonstrukce záznamu

Jistě popis dat získaný odhadem jejich struktury z množiny vstupních dat nebude dosahovat expresivity známé z lidmi tvořených ontologií, avšak poskytuje za lehce splnitelných podmínek RDF dokumenty jistým, pro technická data postačujícím, způsobem. I takto jednoduchý popis dat může být použit pro učení extrakčních metod, které získávají anotovaná data z webových stránek [12, 13, 14]

V současné době je experimentálně provozován portál shromažďující informace o sportovních utkáních, kdy struktura dat byla odhadnuta z dat několika heterogenních zdrojů a data uložena na základě této struktury. Ilustrace portálu je na obrázcích 1 a 2.

5. Závěr

Příspěvek se snaží shrnout aktuální trendy, problémy a technologie jak na současném webu, tak v prostředí webu sémantického. Zvláště se pak věnuje problematice vyhledávání dat, diskutuje související problémy a navrhuje jejich řešení.

V sekci 2 ukazuje na příkladu fragmentu HTML dokumentu, jak může být zaindexován pro fulltextové vyhledávání. Ukazuje použití rozšíření RDFa, které umožňuje anotovat části HTML dokumentu. Pokud jsou hodnoty anotovány, je možné automaticky převést takový HTML dokument do RDF dokumentu a ten dále zpracovat další nástroji.

Sekce 3 pak inovativně diskutuje výhody distribuce dat dokumentů sémantického webu, kdy resource není reprezentován pouze URI, ale URL obsahující detailnější informace o odkazovaném objektu. Zásadní výhodou tohoto přístupu je, že odpadá nutnost jinak velmi obtížné, automaticky téměř neřešitelné, integrace dat jednotlivých zdrojů. Celý problém je ilustrován na příkladě.

Jelikož v současné době nejsou k dispozici taková data požadovaného rozsahu a zaměření, sekce 4 navrhuje problém řešit pomocí metod odhadu struktury dat a tyto metody využít pro základní definici popisu dat prostřednictvím formátů sémantického webu.

Pokud by se podařilo myšlenky prezentované v článku naplnit, celá vize by našla uplatnění pro širokou veřejnost dnes používající internet.

Literatura

[1] P. Raghavan, “Information retrieval algorithms: a survey,” in *SODA ’97: Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, (Philadelphia, PA, USA), pp. 11–18, Society for Industrial and Applied Mathematics, 1997.

[2] A. N. Langville and C. D. Meyer, *Google’s Page Rank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, July 3 2006.

[3] G. Antoniou and F. van Harmelen, *A Semantic Web Primer (Cooperative Information Systems)*. The MIT Press, April 2004.

[4] T. Lee, “Relational databases on the semantic web,” *http://www.w3.org/DesignIssues/RDB-RDF.html [on-line]*, 1998.

[5] L. Baolin and H. Bo, “Network and parallel computing, ifip international conference, npc 2007, dalian, china, september 18-21, 2007, proceedings,” in *NPC* (K. Li, C. R. Jesshope, H. Jin, and J.-L. Gaudiot, eds.), vol. 4672 of *LNCS*, pp. 364–374, Springer, 2007.

[6] M. Lenzerini, “Data integration: a theoretical perspective,” in *PODS ’02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, (New York, NY, USA), pp. 233–246, ACM Press, 2002.

[7] E. Rahm and P. A. Bernstein, “A survey of approaches to automatic schema matching,” *VLDB Journal: Very Large Data Bases*, vol. 10, no. 4, pp. 334–350, 2001.

[8] R. Battle and E. Benson, “Bridging the semantic web and web 2.0 with representational state transfer (rest),” *Web Semant.*, vol. 6, no. 1, pp. 61–69, 2008.

[9] C. J. Date, *An Introduction to Database Systems*. Addison Wesley Longman, October 1999.

[10] M. Řimnáč, “Data structure estimation for rdf oriented repository building,” in *Proceedings of the CISIS 2007*, (Los Alamitos, CA, USA), pp. 147–154, IEEE Computer Society, 2007.

[11] M. Řimnáč, “Transforming current web sources for semantic web usage,” *Proc. of SOFSEM 2006*, vol. 2, pp. 155–165, 2006.

[12] Z. Li and W. K. Ng, “Wdee: Web data extraction by example,” in *DASFAA* (L. Zhou, B. C. Ooi, and X. Meng, eds.), vol. 3453 of *LNCS*, pp. 347–358, Springer, 2005.

[13] W. Holzinger, B. Krüpl, and M. Herzog, “Using ontologies for extracting product features from web pages,” in *International Semantic Web Conference* (I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, eds.), vol. 4273 of *LNCS*, pp. 286–299, Springer, 2006.

[14] M. Nekvasil, “Využití ontologií při indukci wrapperů,” *Proc. of Znalosti 2007*, pp. 336–339, 2007.

Maximálně věrohodné odhady a lineární regrese ve výběrových šetřeních

doktorand:
MGR. MICHAELA ŠEDOVÁ, MSC.

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

sedova@euromise.cz

školitel:
MGR. MICHAL KULICH, PH.D.

Katedra pravděpodobnosti a matematické statistiky
Univerzita Karlova v Praze
Sokolovská 83
186 75 Praha 8

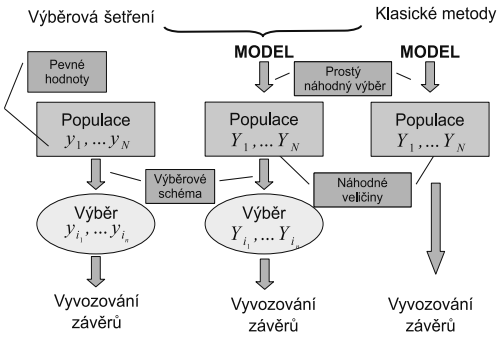
kulich@karlin.mff.cuni.cz

obor studia:
Pravděpodobnost a matematická statistika

Práce byla částečně podpořena výzkumným záměrem AV0Z 10300504 a MSM 0021620839.

V klasické teorii výběrových šetření jsou předmětem studia převážně parametry charakterizující konečnou populaci, jako např. úhrn nebo průměr N pevných hodnot. Někdy však může nastat situace, kdy bychom rádi výsledky zobecnili na jiné populace, nebo i tutéž populaci v jiném čase. Navíc, připustíme-li, že sesbíraná data nemusí být zcela spolehlivá, vidíme, že je vhodné chápat naše pozorování jako realizace náhodných veličin. Takto přistupují k datům klasické statistické metody. Ty však předpokládají, že je k dispozici prostý náhodný výběr, což v kontextu výběrových šetření zpravidla není možné.

Proto je někdy potřebné zvolit postup analýzy dat, který kombinuje oba tyto přístupy, tedy modifikovat metody tak, aby zohledňovaly dané výběrové schéma. Rozdíl v přístupu teorie výběrových šetření, klasických metodách a našem postupu (kombinace obojího) je schématicky popsán na obrázku 1.



Obrázek 1: Přístup teorie výběrových šetření, klasických metod a kombinace obojího.

V příspěvku definujeme následující výběrové schéma. Máme náhodný vektor (Y, W), kde Y představuje sledovanou veličinu a W stratum v populaci. Jedinci patřící do stejného strata mají stejnou pravděpodobnost zahrnutí do výběru, avšak mezi straty se tato pravděpodobnost může lišit.

Pořídíme-li výběr podle popsaného schématu, zastoupení jednotlivých strat neodpovídá skutečnému poměru v populaci. To je potřeba zohlednit ve stanovování odhadů parametrů a jejich vlastností. Např. odhad střední hodnoty Y má podobu váženého průměru pozorování, kde váhy jsou převrácenou hodnotou empirických pravděpodobností výběru. Rozptyl takového odhadu se skládá ze dvou členů. První z nich odpovídá rozptylu, který bychom obdrželi, kdybychom pozorovali celou populaci, druhý představuje penaltu za to, že máme k dispozici pouze výběr.

Tento přístup rozvíjíme dál a popisujeme modifikaci maximálně věrohodných odhadů. Zde je nutné vážit skórové statistiky v rovnicích pro odhad parametrů. Uvádíme konkrétní výpočet pro lineární model a výsledek ilustrujeme na malé simulační studii.

Literatura

[1] Šedová M., Kulich M. (2007): Statistical Methods for Analysis of Survey Data, in *WDS'07 Proceedings of Contributed Papers: Part I– Mathematics and Computer Sciences* (eds. J. Safrankova and J. Pavlu), Prague, Matfyzpress, pp. 181–186.

Ruled Based Analysis of Behaviour Learned by Evolutionary Algorithms and Reinforcement Learning

Post-Graduate Student:
MGR. STANISLAV SLUŠNÝ
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
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, Czech Republic
roman@cs.cas.cz

Field of Study:
Software Systems

This work deals with the problem of designing adaptive embodied agent. We have considered several adaptive mechanisms. In our previous work, we have been examining mainly Evolutionary robotics (ER). We utilized local unit network architecture called radial basis function (RBF). This network has more learning options, and (due to its local nature) better interpretation possibilities [10, 11] than multilayer perceptron networks.

The lack of theoretical insight into Evolutionary Algoritm is the most serious problem of the previous approach. We summarize our experiences and do the comparison with to Reinforcement Learning (RL) - another widely studied approach in Artificial Intelligence.

The RL is based on dynamic programming [5]. It has solid theoretical backgrounds built around Markov chains and several proven fundamental results. On the other side, theoretical assumptions cannot be often fullfiled in the experiments.

RL is focusing on agent, that is interacting with the environment by its sensors and effectors. This interaction process helps agent to learn effective behavior. These kinds of tasks are commonly studied on miniature mobile robots of type Khepera [2] and E-puck [1].

Probably the most commonly used algorithm of RL is Q-learning. However, in real life applications, state space is too big and convergence toward optimal strategy is slow with Q-learning algorithm. RL suffers from the curse of dimensionality. Therefore, several improvements have been suggested to speed up the learning process.

A lot of efforts have been devoted recently to rethinking the idea of states by using function approximators [6], defining notion of options and hierarchical abstractions [4]. Dzeroski in his work [9] suggests to combine RL with Inductive Logical Programming. In this method, called Relational Reinforcement Learning, agent can ”reason” about states. This way, complexity of state space can be reduced significantly.

The distinction between classical RL and Relational Reinforcement Learning is the way how gained experiences (knowledge) are represented. In classical Q-learning algorithm, tuples <situation, action, reward> are stored in a pure sequential manner. In relational version of the algorithm, they are stored in the structure called *Logical decision tree* [7]. We have used logical decision trees as implemented in the programs TILDE [7] from package ACE-ilProlog [8].

In the past, performance of Relational Reinforcement Learning have been experimentally evaluated on deterministic tasks and games only [9]. We are focusing on noisy environments with high degree of uncertainty. As we will show, even in these conditions, Relational Reinforcement Learning can find satisfactory solution.

We present a case study of these two approaches on maze exploring and multi-robot light searching task. Experiments with both real and simulated miniature Khepera and E-puck robots will be described and discussed. Knowledge in the form of if-then rules is extracted from the trained RBF neural networks and compared to the relational RL transition table representation. Several performance measures are studied and compared for both approaches.

Our architecture enables agent to make reactive decisions with background planning and reasoning about the states. Thus, it is combining old-fashioned planning based on logical programming with behavior based robotics.

References

[1] E-puck, online documentation. http://www.e-puck.org.

[2] Khepera II documentation. http://k-team.com.

[3] Webots simulator. http://www.cyberbotics.com/.

[4] A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. 13:341–379.

[5] R. E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[6] D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Ahtena Scientific, 1996.

[7] H. Blockeel and L. De Raedt. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101:285–297, 1998.

[8] H. Blockeel, L. Dehaspe, B. Demoen, G. Janssens, J. Ramon, and H. Vandecasteele. Improving the efficiency of inductive logic programming through the use of query packs. *Journal of Artificial Intelligence Research*, 16:135–166.

[9] S. Dzeroski, L. De Raedt, and K. Driessens. Relational reinforcement learning. *Machine Learning 43*, pages 7–52, 2001.

[10] S. Slušný and R. Neruda. Evolving homing behaviour for team of robots. *Computational Intelligence, Robotics and Autonomous Systems. Palmerston North : Massey University*, 2007.

[11] S. Slušný, R. Neruda, and P. Vidnerová. Evolution of simple behavior patterns for autonomous robotic agent. *System Science and Simulation in Engineering. - : WSEAS Press*, pages 411–417, 2007.

Dynamic Classifier Systems for Classifier Aggregation

Post-Graduate Student:
ING. DAVID ŠTEFKA
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
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, Czech Republic
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 aggregation is a method for improving quality of classification – instead of using just one classifier, a team of classifiers is created, and the outputs of the individual classifiers are aggregated into the final prediction. Common methods for classifier aggregation, such as mean value aggregation or weighted mean aggregation are *static*, i.e., they do not adapt to the currently classified pattern. In this paper, we introduce a formalism of *dynamic* classifier systems, which use the concept of dynamic classification confidence in the aggregation process, and therefore they dynamically adapt to the currently classified pattern. The results of the experiments with quadratic discriminant classifiers on four artificial and four real-world benchmark datasets show that dynamic classifier systems can significantly outperform both confidence-free and static classifier systems.

1. Introduction

Classification is a process of dividing objects (called *patterns*) into disjoint sets called *classes* [1]. Many machine learning algorithms for classification have been developed – for example naive Bayes classifiers, linear and quadratic discriminant classifiers, k-nearest neighbor classifiers, support vector machines, neural networks, or decision trees. If the quality of classification (i.e., the classifier's predictive power) is low, there are several methods we can use to improve it.

One comonly used technique for improving classification quality is called *classifier combining* [2] – instead of using just one classifier, we create and train a team of classifiers, let each of them predict independently, and then combine (aggregate) their results. It can be shown that a team of classifiers can perform better in the classification task than any of the individual classifiers.

There are two main approaches to classifier combining: *classifier selection* [3, 4, 5] and *classifier aggregation* [6, 7]. If a pattern is submitted for classification, the former technique uses some rule to select one particular classifier, and only this classifier is used to obtain the final prediction. The latter technique uses some aggregation rule to aggregate the results of all the classifiers in a team to get the final prediction.

A common drawback of classifier aggregation methods is that they are static, i.e., they are not adapted to the particular patterns that are currently classified. In other words, the aggregation is specified during a training phase, prior to classifying a test pattern. However, if we use the concept of dynamic classification confidence (i.e., the extent to which we can "trust" the output of the particular classifier for the currently classified pattern), the aggregation algorithms can take into account the fact that "this classifier is not good *for this particular pattern*".

Surprisingly, such dynamic classifier systems are not used very often in classifier combining. However, there has already been some work done in the field of dynamic classifier systems – Robnik-Šikonja and Tsymbal et al. [8, 9] study dynamic aggregation of random forests [10], i.e., dynamic classifier systems of decision trees. The authors report significant improvements in classification quality when using dynamic voting compared to simple voting. However, they study dynamic classifier systems only in the context of random forests, and they use only confidence measures based on the so-called margin.

In this paper, we provide a general formalism of dynamic classification confidence measures and

dynamic classifier systems, and we experimentally study the performance of confidence-free classifier systems (i.e., systems that do not utilize classification confidence at all), static classifier systems (i.e., systems that use only "global" confidence of a classifier), and dynamic classifier systems (i.e., systems that adapt to the particular pattern submitted for classification).

The paper is structured as follows. In Section 2, we introduce the formalism of classifier combining, namely in Section 2.1, we define basic concepts of classification, in Section 2.2 we introduce the concept of classification confidence, and we introduce three dynamic confidence measures, in Section 2.3 we deal with classifier teams and ensembles, and in Section 2.4, we finally define classifier systems and show several examples of dynamic classifier systems. In Section 3, we experimentally investigate the suitability of the proposed dynamic confidence measures, and the performance of the proposed dynamic classifier systems. Section 4 then concludes the paper.

2. Formalism of Classifier Combining with Classification Confidence

2.1. Classification

Throughout the rest of the paper, we use the following notation. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a n-dimensional *feature space*, an element $\vec{x} \in \mathcal{X}$ of this space is called a *pattern*, and let $C_1, \ldots, C_N \subseteq \mathcal{X}$, $N \geq 2$, be disjoint sets called *classes*. The index of the class a pattern $\vec{x}$ belongs to will be denoted as $c(\vec{x})$ (i.e., $c(\vec{x}) = i$ iff $\vec{x} \in C_i$). The goal of classification is to determine to which class a given pattern belongs, i.e., to predict $c(\vec{x})$ for unknown patterns.

Definition 1 *We call a* classifier *every mapping* $\phi : \mathcal{X} \to [0,1]^N$*, where* $[0,1]$ *is the unit interval, and* $\phi(\vec{x}) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$ *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 $\mu_i(\vec{x}) > \mu_j(\vec{x})$, it means that the pattern $(\vec{x})$ 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].*

Definition 2 *Classifier* ϕ *is called* crisp*, iff* $\forall \vec{x} \in \mathcal{X}\ \exists i$*, such that:*

$$\mu_i(\vec{x}) = 1, \text{ and } \forall j \neq i\ \mu_j(\vec{x}) = 0.$$

Classifier ϕ *is called* normalized*, iff*

$$\forall \vec{x} \in \mathcal{X} : \sum_{i=1}^{N} \mu_i(\vec{x}) = 1,$$

where $\phi(\vec{x}) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$.

Remark 2 *Normalized classifiers are sometimes called* probabilistic *[6]. However, they do not need to be based on probability theory, so we will call them just normalized.*

Definition 3 *Let* ϕ *be a classifier,* $\vec{x} \in \mathcal{X}$*,* $\phi(\vec{x}) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$. Crisp output *of* ϕ *on* $\vec{x}$ *is defined as* $\phi_{cr}(\vec{x}) = \arg\max_{i=1,\ldots,N} \mu_i(\vec{x})$.

2.2. Classification Confidence

Classification confidence expresses the degree of trust we can give to a classifier ϕ when classifying a pattern $\vec{x}$. It is modelled by a mapping κ_ϕ.

Definition 4 *Let* ϕ *be a classifier. We call a* confidence measure *of classifier* ϕ *every mapping* $\kappa_\phi : \mathcal{X} \to [0,1]$.

The higher the confidence, the higher the probability of correct classification. $\kappa_\phi(\vec{x}) = 0$ means that the classification may not be correct, while $\kappa_\phi(\vec{x}) = 1$ means the classification is probably correct. However, κ_ϕ does not need to be modelled by a probability measure.

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 5 *Let* ϕ *be a classifier and* κ_ϕ *its confidence measure. We call* κ_ϕ static*, iff it is constant in* $\vec{x}$*, we call* κ_ϕ dynamic *otherwise.*

Remark 3 *Since static confidence measures are constant, independent on the currently classified pattern, we will omit the pattern* $(\vec{x})$ *in the notation, i.e., we will denote them just* κ_ϕ.

Remark 4 *In the rest of the paper, we will use the indicator operator* I*, defined as* $I(true) = 1$*,* $I(false) = 0$.

2.2.1 Static confidence measures: After the classifier has been trained, we can use a testing set (i.e., a set of patterns on which the classifier has not been trained) to assess its predictive power as a whole (from global view). These methods include accuracy, precision, sensitivity, resemblance, etc. [1, 11], 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 testing set:

$$\kappa_{\phi}^{(GA)} = \frac{\sum_{\vec{y}\in\mathcal{M}} I(\phi(\vec{y}) \stackrel{?}{=} c(\vec{y}))}{|\mathcal{M}|}, \quad (1)$$

where $\mathcal{M}$ is the testing set of ϕ.

2.2.2 Dynamic confidence measures: An easy way how a dynamic confidence measure can be defined is to compute some property on patterns neighboring with $\vec{x}$. Let $N(\vec{x})$ denote a set of neighboring training or validating patterns (we can use both training and validating set for computing $N(\vec{x})$, but it is usually better to use validating set, because if we use training patterns, the results will be biased). In this paper, we define $N(\vec{x})$ as the set of k patterns nearest to $\vec{x}$ under Euclidean metric. Now we will define three dynamic confidence measures which use $N(\vec{x})$:

Euclidean Local Accuracy (ELA) measures the local accuracy of ϕ in $N(\vec{x})$:

$$\kappa_{\phi}^{(ELA)}(\vec{x}) = \frac{\sum_{\vec{y}\in N(\vec{x})} I(\phi_{cr}(\vec{y}) \stackrel{?}{=} c(\vec{y}))}{|N(\vec{x})|}, \quad (2)$$

where $\phi_{cr}(\vec{y})$ is the crisp output of ϕ on $\vec{y}$.

Euclidean Local Match (ELM) is based on the ideas from [12], and measures the proportion of patterns in $N(\vec{x})$ from the same class as ϕ is predicting for $\vec{x}$:

$$\kappa_{\phi}^{(ELM)}(\vec{x}) = \frac{\sum_{\vec{y}\in N(\vec{x})} I(\phi_{cr}(\vec{x}) \stackrel{?}{=} c(\vec{y}))}{|N(\vec{x})|}, \quad (3)$$

where $\phi_{cr}(\vec{x})$ is the crisp output of ϕ on $\vec{x}$.

Euclidean Average Margin (EAM) is defined as mean value of the margin [10, 8, 9] in $N(\vec{x})$:

$$\kappa_{\phi}^{(EAM)}(\vec{x}) = \frac{\sum_{\vec{y}\in N(\vec{x})} mg(\phi(\vec{y}))}{|N(\vec{x})|}, \quad (4)$$

where the margin is defined as $mg(\phi(\vec{y})) =$

$$\begin{cases} \mu_{c(\vec{y})}(\vec{y}) - \max\limits_{\substack{i=1,\ldots,N \\ i\neq c(\vec{y})}} \mu_i(\vec{y}) & \text{if } \phi_{cr}(\vec{y}) = c(\vec{y}), \\ 0 & \text{otherwise.} \end{cases}, \quad (5)$$

where $\phi(\vec{y}) = (\mu_1(\vec{y}), \ldots, \mu_N(\vec{y}))$, and $\phi_{cr}(\vec{y})$ is the crisp output of ϕ on $\vec{y}$.

The dynamic confidence measures defined in this section have one drawback – they need to compute $N(\vec{x})$, 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 $\phi(\vec{x})$, 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.

Remark 5 *All the previous confidence measures are* model-indifferent*, i.e., they could be used for any classifier. However, measures which take into account specific aspects of the classification method could be designed – for example, Robnik-Šikonja and Tsymbal et al. [8, 9] use dynamic confidence of a decision tree in a random forest [10] as average margin computed on instances similar to the currently classified pattern, where the similarity is based on specific aspects of random forests. Such* model-specific *measures could use the information from the classification process better than model-indifferent measures. However, due to space constraints we do not deal with model-specific measures in this paper.*

2.3. 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 confidence measure defined.

Definition 6 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.*

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*. For this reason the methods which create a team of classifiers are sometimes called *ensemble methods*. 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* [13], *boosting* [14], *error correction codes* [2], or *multiple feature subset* methods [15]. These methods try to create an ensemble of classifiers which are both *accurate* and *diverse* [16].

Since the main focus of this paper lies in studying classification confidence, we will not study these methods here, and we will just assume in the rest of the paper that we have constructed a classifier team $(\mathcal{T}, \mathcal{K})$ of r classifiers using some of these methods.

If a pattern is submitted for classification, the team of classifiers gives us two different informations – outputs of the individual classifiers (a *decision profile*), and values of classification confidences of the classifiers (a *confidence vector*).

Definition 7 *Let $(\mathcal{T}, \mathcal{K})$ be a classifier team, $\mathcal{T} = (\phi_1, \ldots, \phi_r)$, $\mathcal{K} = (\kappa_{\phi_1}, \ldots, \kappa_{\phi_r})$, and let $\vec{x} \in \mathcal{X}$. Then we define* decision profile $\mathcal{T}(\vec{x}) \in [0,1]^{r,N}$ *as*

$$\mathcal{T}(\vec{x}) = \begin{pmatrix} \phi_1(\vec{x}) \\ \phi_2(\vec{x}) \\ \vdots \\ \phi_r(\vec{x}) \end{pmatrix} = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \ldots & \mu_{1,N} \\ \mu_{2,1} & \mu_{2,2} & \ldots & \mu_{2,N} \\ & & \ddots & \\ \mu_{r,1} & \mu_{r,2} & \ldots & \mu_{r,N} \end{pmatrix}, \tag{6}$$

and confidence vector $\mathcal{K}(\vec{x}) \in [0,1]^r$ *as*

$$\mathcal{K}(\vec{x}) = \begin{pmatrix} \kappa_{\phi_1}(\vec{x}) \\ \kappa_{\phi_2}(\vec{x}) \\ \vdots \\ \kappa_{\phi_r}(\vec{x}) \end{pmatrix} \tag{7}$$

Remark 6 *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 be always followed by $(\vec{x})$.*

2.4. Classifier Systems

After the pattern $\vec{x}$ has been classified by all the classifiers in the team, and the confidences were 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 8 *Let $r, N \in \mathbf{N}$, $r, N \geq 2$. A* team aggregator *of dimension (r, N) is any mapping $\mathcal{A} : [0,1]^{r,N} \times [0,1]^r \to [0,1]^N$.*

A classifier team with an aggregator will be called a *classifier system*. Such system can be also viewed as a single classifier.

Definition 9 *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. We define an* induced classifier *of $(\mathcal{T}, \mathcal{K}, \mathcal{A})$ as a classifier Φ, defined as*

$$\Phi(\vec{x}) = \mathcal{A}(\mathcal{T}(\vec{x}), \mathcal{K}(\vec{x})).$$

The 4-tuple $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A}, \Phi)$ is called a classifier system.

Depending on the way how a classifier system utilizes the classification confidence, we can distinguish several kinds of classifier systems.

Definition 10 *Let $(\mathcal{T}, \mathcal{K})$ be a classifier team. $(\mathcal{T}, \mathcal{K})$ is called* static*, iff*

$$\forall \kappa \in \mathcal{K} : \kappa \text{ is a static confidence measure}.$$

$(\mathcal{T}, \mathcal{K})$ is called dynamic*, iff*

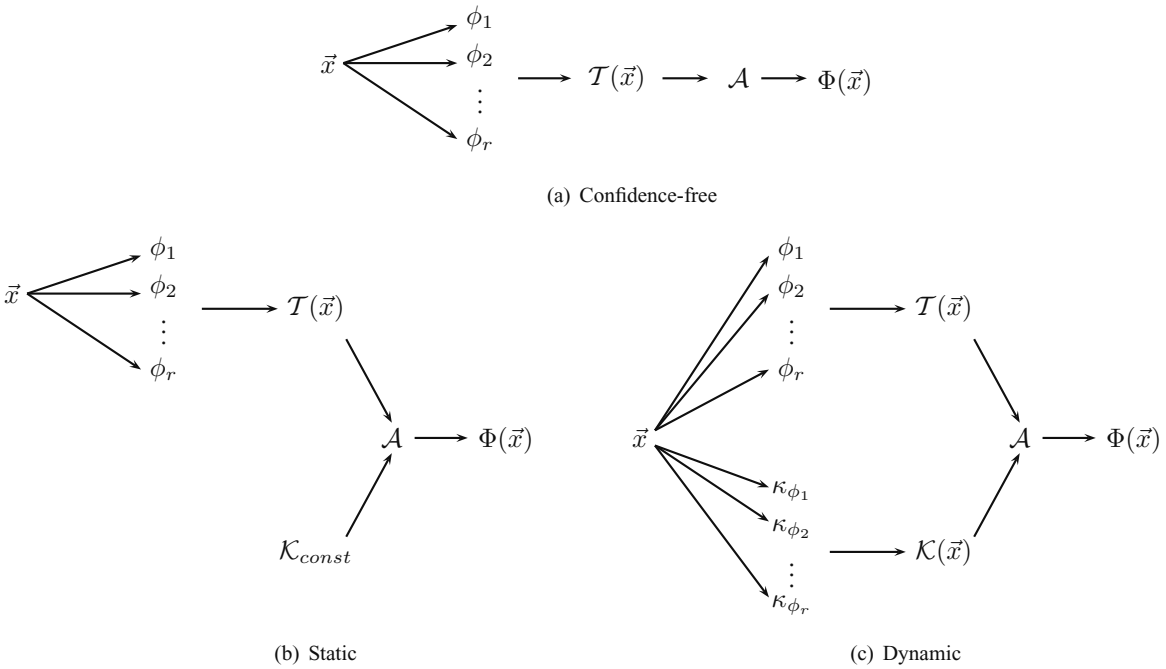
$$\forall \kappa \in \mathcal{K} : \kappa \text{ is a dynamic confidence measure}.$$

Definition 11 *Let $\mathcal{A}$ be a team aggregator of dimension (r, N). We call $\mathcal{A}$* confidence-free*, iff $\forall\, \mathbf{T} \in [0,1]^{r,N}$:*

$$(\forall \vec{k_1}, \vec{k_2} \in [0,1]^r : \mathcal{A}(\mathbf{T}, k_1) = \mathcal{A}(\mathbf{T}, k_2)).$$

Definition 12 *Let $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A}, \Phi)$ 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 systems do not utilize the classification confidence at all (for example a team of classifiers aggregated by simple voting). Static systems utilize classification confidence, but only as a global property (for example a team of classifiers aggregated by weighted voting with constant classifier weights). Dynamic systems utilize classification confidence in a dynamic way, i.e. the aggregation is adapted to the particular pattern submitted for classification (for example a team of classifiers aggregated by weighted voting with classifier weights computed for every pattern). The different approaches are schematically shown in Fig. 1.



(a) Confidence-free

(b) Static

(c) Dynamic

Figure 1: Schematic comparison of confidence-free, static, and dynamic classifier systems.

Remark 7 *Since confidence-free classifier systems do not utilize the classification confidence, we will denote them* $\mathcal{S} = (\mathcal{T}, \mathcal{A}, \Phi)$*, and their team aggregators will be defined as a mapping* $\mathcal{A} : [0,1]^{r,N} \to [0,1]^N$.

Many methods for aggregating the team of classifiers into one final classifier have been proposed in the literature. A good overview of commonly used aggregation methods can be found in [6]. These methods comprise simple arithmetic rules (voting, sum, product, maximum, minimum, average, weighted average, etc.), fuzzy integral, Dempster-Shafer fusion, second-level classifiers, decision templates, and many others.

In the following text, we define several team aggregators. We will use the notation from Def. 7 and Def. 9. Let $\Phi(\vec{x}) = \mathcal{A}(\mathcal{T}(\vec{x}), \mathcal{K}(\vec{x})) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$.

Mean value aggregation (MV) is the most common (confidence-free) aggregation technique. Its aggregator is defined as

$$\mu_j(\vec{x}) = \frac{\sum_{i=1,\ldots,r} \mu_{i,j}(\vec{x})}{r}. \tag{8}$$

If the classifiers in the team are crisp, MV coincides with voting.

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:

$$\mu_j(\vec{x}) = \frac{\sum_{i=1,\ldots,r} \kappa_{\phi_i} \mu_{i,j}(\vec{x})}{\sum_{i=1,\ldots,r} \kappa_{\phi_i}}. \tag{9}$$

Dynamic weighted mean aggregation (DWM) has the same aggregator as SWM, but the weights are dynamic classification confidences:

$$\mu_j(\vec{x}) = \frac{\sum_{i=1,\ldots,r} \kappa_{\phi_i}(\vec{x}) \mu_{i,j}(\vec{x})}{\sum_{i=1,\ldots,r} \kappa_{\phi_i}(\vec{x})}. \tag{10}$$

Filtered mean aggregation (FM) has the same aggregator as MV, but prior to computing the aggregated values, the classifiers which have (dynamic) classification confidence lower than $T \in [0,1]$ are discarded:

$$\mu_j(\vec{x}) = \frac{\sum\limits_{\substack{i=1,\ldots,r \\ \kappa_{\phi_i}(\vec{x}) > T}} \mu_{i,j}(\vec{x})}{|\{\phi \in \mathcal{T} | \kappa_{\phi_i}(\vec{x}) > T\}|}. \tag{11}$$

3. Experiments

3.1. Experiment 1 – Choosing the Right Confidence Measure

To gain a general idea to which extent the proposed dynamic confidence measures (ELA, ELM, and EAM) really express the probability that the classification of the currently classified pattern is right, we examined

the distributions of the confidence values for correctly classified and for misclassified patterns.

The confidence measures were tested on quadratic discriminant classifiers [1]. The classifiers were implemented in Java programming language and 10-fold crossvalidation was performed to obtain the results. We measured histograms of the local classification confidence values for correctly classified and for misclassified patterns from four artificial (Clouds, Concentric, Gauss_3D, Waveform) and four real-world (Breast, Phoneme, Pima, Satimage) datasets from the Elena database [17] and from the UCI repository [18]. As $N(\vec{x})$, we used the set of 20 nearest neighbors of $\vec{x}$ under Euclidean metric.

The histograms of the dynamic confidence values for the particular datasets are shown in Fig. 2. Before discussing the results, we should say a few words about how the results *should* ideally look like. We will denote the distribution of local classification confidence values for correctly classified patterns as "OK distribution", and for misclassified patterns as "NOK distribution". The OK distribution should be concentrated near one, while the NOK distribution should be concentrated near zero, and ideally, the distributions should be clearly separated. If the distributions overlap, or if the NOK distribution has high values near one, it means that the measure does not really express the probability that the classification of the currently classified pattern is right.

The results show that for some datasets, all the dynamic confidence measures provide good separation of the OK and NOK patterns, which suggests the measures are suitable for using in dynamic classifier systems. The most representative example of such behavior is the Phoneme dataset, where the OK and NOK distributions for all three dynamic confidence measures are clearly separated.

For some datasets, there are notable differences in the dynamic confidence measures – e.g., in the case of the Satimage dataset, the EAM confidence measure provides much better separation of the OK and NOK patterns than the other two measures. In the case of the Concentric dataset, the ELM confidence measure is an obvious winner. This means that the performance of a confidence measure is dependent on the particular dataset, and that the choice of a confidence measure should be always done with respect to the particular data.

For several datasets, all three dynamic confidence measures provided very poor separation of the OK and NOK patterns, which raises doubts about the suitability of the measures in dynamic classifier systems. This is the case of the Gauss_3D or the Pima dataset.

However, we cannot make direct conclusions about suitability of the measures just from the separation properties of the OK and NOK patterns. To give one example: even if the separation is good enough, the high values of dynamic classification confidence may be obtained on the "easy" patterns, and the low values on the "hard" patterns. Moreover, if the classifiers in the classifier system are "similar", all of them will have similar confidence on a particular pattern. Therefore, dynamic aggregation of the system will bring no improvement in the classification quality, since all the classifiers appear the same for the system's aggregator. This may be the explanation of the result of Exp. 2 for the Phoneme dataset, where the FM aggregation has gives very different performance for ELM and EAM confidence measures, even if the OK and NOK separation of the measures is nearly the same (see Fig. 2).

3.2. Experiment 2 – Confidence-free vs. Static vs. Dynamic Classifier Systems

In the second experiment, we compared the performance of the classifier aggregation algorithms described in Section 2.4. The main emphasis was given to comparing confidence-free vs. static vs. dynamic classifier systems. We used the same datasets as in Exp. 1.

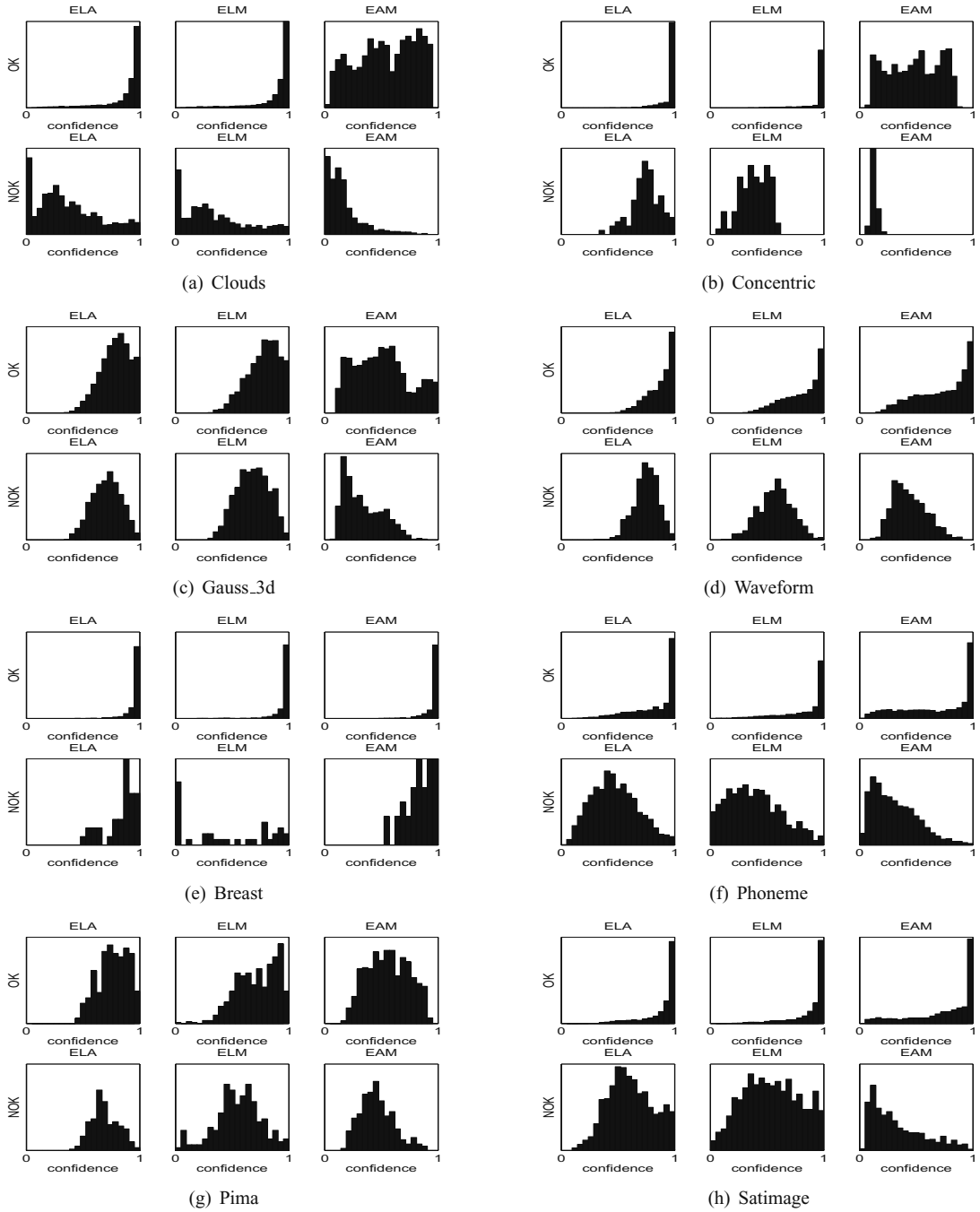
For all the classifier systems we used, the classifier team $\mathcal{T}$ was an ensemble of quadratic discriminant classifiers, created either by the bagging algorithm [13] (which creates classifiers trained on random samples drawn from the original training set with replacement), or by the multiple feature subset method [15] (which creates classifiers using different combinations of features), depending on which method was more suitable for the particular dataset.

For the comparison, we designed the following classifier systems (refer to Section 2.2 and Section 2.4 for the description of the algorithms):

MV confidence-free system aggregated by mean value aggregation

SWM cl. system aggregated by static weighted mean aggregation; as a confidence measure, we used GA

DWM cl. system aggregated by dynamic weighted mean; as a confidence measure, we used ELA, ELM, and EAM



(a) Clouds (b) Concentric

(c) Gauss_3d (d) Waveform

(e) Breast (f) Phoneme

(g) Pima (h) Satimage

Figure 2: Histograms of dynamic confidence values of a quadratic discriminant classifier (ELA - Euclidean Local Accuracy, ELM - Euclidean Local Match, EAM - Euclidean Average Margin) for correctly classified (OK) and misclassified (NOK) patterns.

Table 1: Comparison of the aggregation methods – non-combined classifier (NC), mean value (MV), static weighted mean (SWM) using GA confidence measure, dynamic weighted mean (DWM) using three confidence measures (ELA, ELM, EAM), and filtered mean (FM) using three confidence measures (ELA, ELM, EAM). Mean error rate (in %) ± standard deviation of error rate from a 10-fold crossvalidation was measured. The best result is displayed in boldface, statistically significant improvements to NC, MV, and SWM are marked by footnote signs. The (B/M) after dataset name means whether the ensemble was created by Bagging or Multiple feature subset algorithm.

	Non-Combined	Conf.-free	Static		Dynamic		
Dataset	NC	MV	κ	SWM	κ	DWM	FM
Clouds (M)	25.0 ± 1.7	25.0 ± 2.1	GA	24.7 ± 1.6	ELA	23.4 ± 1.5	22.3 ± 1.5 *†‡
					ELM	23.2 ± 1.2	$\mathbf{22.0 \pm 2.1}$ *†‡
					EAM	23.5 ± 1.5	23.3 ± 1.4
Concentric (B)	3.5 ± 1.0	3.8 ± 0.6	GA	4.0 ± 0.8	ELA	3.2 ± 1.1	2.1 ± 1.3 †‡
					ELM	2.9 ± 1.6	$\mathbf{1.8 \pm 0.8}$ *†‡
					EAM	3.8 ± 1.3	4.3 ± 1.5
Gauss_3D (B)	21.4 ± 1.7	21.6 ± 1.1	GA	21.5 ± 2.1	ELA	21.5 ± 1.4	21.7 ± 1.3
					ELM	$\mathbf{21.3 \pm 2.0}$	22.0 ± 1.3
					EAM	21.5 ± 2.0	21.7 ± 1.3
Waveform (B)	14.9 ± 2.5	15.0 ± 1.4	GA	14.8 ± 0.9	ELA	14.7 ± 1.9	15.0 ± 1.2
					ELM	14.8 ± 2.5	$\mathbf{14.5 \pm 1.2}$
					EAM	14.6 ± 2.0	15.5 ± 1.0
Breast (M)	4.8 ± 2.9	4.7 ± 2.5	GA	4.2 ± 2.4	ELA	3.0 ± 2.1	2.9 ± 1.8
					ELM	3.0 ± 1.9	3.1 ± 2.1
					EAM	3.2 ± 2.0	$\mathbf{2.9 \pm 1.7}$
Phoneme (M)	24.7 ± 1.1	23.5 ± 1.6	GA	24.0 ± 1.4	ELA	21.5 ± 1.9 *‡	17.2 ± 1.4 *†‡
					ELM	21.2 ± 1.8 *‡	$\mathbf{16.9 \pm 2.0}$ *†‡
					EAM	21.9 ± 0.9 *	20.7 ± 1.7 *†‡
Pima (M)	27.1 ± 4.4	25.4 ± 3.6	GA	25.0 ± 5.6	ELA	25.8 ± 6.5	24.0 ± 2.7
					ELM	24.0 ± 4.1	25.0 ± 7.4
					EAM	24.8 ± 6.3	$\mathbf{23.5 \pm 5.4}$
Satimage (B)	15.6 ± 1.7	15.5 ± 1.2	GA	15.5 ± 1.7	ELA	15.3 ± 1.6	15.2 ± 2.4
					ELM	15.3 ± 1.3	$\mathbf{14.4 \pm 1.0}$
					EAM	15.5 ± 1.2	15.0 ± 1.5

*Significant improvement to NC
†Significant improvement to MV
‡Significant improvement to SWM

FM cl. system aggregated by filtered mean; as a confidence measure, we used ELA, ELM, and EAM

We also compared the systems' performance with the so-called *non-combined classifier* (NC), i.e., a common quadratic discriminant classifier (the NC classifier represents an approach which we had to use if we could use only one classifier).

All the methods were implemented in Java programming language, and a 10-fold crossvalidation was performed to obtain the results. For the dynamic confidence measures, we used the same definition of $N(\vec{x})$ as in Exp. 1, and the threshold T for FM aggregators was set to $T = 0.8$ or $T = 0.9$, depending on the particular dataset (based on some preliminary testing; no fine-tuning or optimization was done).

The results of the testing are shown in Table 1. Mean error rate and standard deviation of the error rate of the induced classifiers from a 10-fold crossvalidation was measured. We also measured statistical significance of the results – at 5% confidence level by the analysis of variance using the Tukey-Kramer method (by the 'multcomp' function from the Matlab statistics toolbox).

The results show that for most datasets, the dynamic classifier systems outperform both confidence-free and static classifier systems. For three datasets, these results were statistically significant. FM usually gives better results than DWM, and if we compare the three dynamic confidence measures, we can say that ELM gives usually the best results, ELA and ELM being slightly worse. However, as we already discussed in Exp. 1, the performance of the individual confidence measures is dependent on the particular dataset. Generally speaking,

the FM-ELM was the most successfull algorithm in this experiment.

It should be noted that the experimental results from this paper are relevant only to quadratic discriminant classifiers, because for any other classifier types (k-NN, SVM, decision trees, etc.), the dynamic confidence measures could give quite different results.

4. Summary

In this paper, we have studied dynamic classifier aggregation. We have introduced the formalism of classifier systems which can be used with (dynamic) classification confidence, and we have defined confidence-free, static, and dynamic classifier systems. We have introduced three dynamic classification confidence measures (ELA, ELM, EAM), and we have shown a way how these measures can be used in dynamic classifier systems – we have introduced two algorithms for dynamic classifier aggregation.

In our first experiment, we have studied the distributions of values of the proposed dynamic classification confidence measures for correctly classified and misclassified patterns, which can give us a hint about suitability of the measures in dynamic classifier systems. The results show that the performance of the particular confidence measure is dependent of the particular dataset.

In the second experiment, we have compared the performance of confidence-free, static, and dynamic classifier systems of quadratic discriminant classifiers. The results show that dynamic classifier systems can significantly outperform both confidence-free and static classifier systems.

The main contribution of this paper is the verification that the concept of dynamic classification confidence can significantly improve the classification quality, and that it is a general concept, which can be incorporated into the theory of classifier aggregation in a systematic way.

In our future work, we plan to study dynamic classification confidence measures for other classifiers than quadratic discriminant classifier, mainly decision trees and support vector machines, and to study model-specific confidence measures for these classifier types. We will also incorporate local classification confidence into more sophisticated classifier aggregation methods, for example fuzzy t-conorm integral [19].

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. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds.), vol. 4212 of *Lecture Notes in Computer Science*, pp. 801–808, Springer, 2006.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] D. J. Hand, *Construction and Assessment of Classification Rules*. Wiley, 1997.

[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 International Conference, on Case-Based Reasoning, ICCBR 2005, Chicago, USA, Proceedings* (H. Muñoz-Avila and F. Ricci,

eds.), vol. 3620 of *Lecture Notes in Computer Science*, pp. 177–190, Springer, 2005.

[13] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[14] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, pp. 148–156, 1996.

[15] S. D. Bay, "Nearest neighbor classification from multiple feature subsets," *Intelligent Data Analysis*, vol. 3, no. 3, pp. 191–209, 1999.

[16] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[17] UCL MLG, "Elena database," 1995. http://www.dice.ucl.ac.be/mlg/?page=Elena.

[18] 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.

[19] D. Štefka and M. Holeňa, "The use of fuzzy t-conorm integral for combining classifiers.," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 9th European Conference, ECSQARU 2007, Hammamet, Tunisia* (K. Mellouli, ed.), vol. 4724 of *Lecture Notes in Computer Science*, pp. 755–766, Springer, 2007.

Combination of Methods for Ontology Matching

Post-Graduate Student:
ING. PAVEL TYL

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic

Faculty of Mechatronics and Interdisciplinary Engineering Studies
Technical University of Liberec
Hálkova 6
461 17 Liberec, Czech Republic

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, Czech Republic

stuller@cs.cas.cz

Field of Study:
Technical Cybernetics

This work was partly supported by the Research Center 1M0554 of Ministry of Education of the Czech Republic: "Advanced Remedial Technologies", 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

While (partial) ontologies usually cover a specific topic/area, many applications require much more general approach to describe their data. Ontology matching can help to transform several such partial ontological descriptions into a single unifying one.

The paper describes a case study of using different methods, compares their advantages and discusses a possibility of using particular results for the definition of the final ontology. Two trivial ontologies were created (independently of any tool) and they were matched using various selected tools.

1. Introduction

Many ontologies were, and are, created in different areas of human activities. Ontologies often contain overlapping concepts. For example companies may want to use standard ontologies of certain domain community or authority along with ontology specific for their own company. In other words creators of ontologies can use existing ontologies as a basis for creating new ones by *integration* or *merging* of the *existing ones*.

Ontology matching is the process of finding relationships or correspondences between entities of different ontologies which are somehow semantically connected. The output of a matching process is a set of these correspondences between two or more ontologies called an *ontology alignment*. The oriented version of an ontology alignment is an *ontology mapping*[1]. Relationships originated by ontology matching can be used to realize the following operations on ontologies:

- *Ontology Merging*[2]– creating a new ontology containing concepts from source ontologies (in general overlapping – see Fig. 2). Initial ontologies (see Fig. 1) remain unaltered.

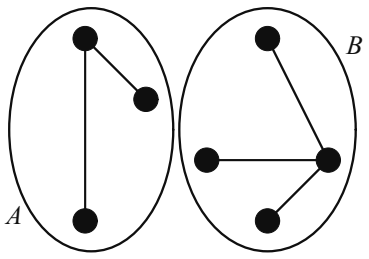


Figure 1: Initial ontologies *A* and *B*.

- *Ontology Integration* – inclusion of one ontology into another one by expressing the relationships between both of them, creating *"superontology"* connecting (partial) concepts and containing the knowledge from both source ontologies (see Fig. 3). One ontology remains unaltered while the other one is modified by knowledge of the first one.

[1] Ontology mapping can be seen as a collection of mapping rules (with some direction – from one ontology to another one, i.e. *Source* → *Target*).
[2] Ontology merging is similar to *schema integration* in databases.

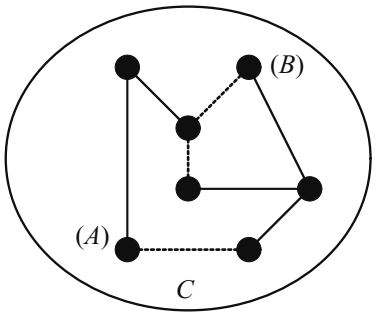


Figure 2: *Merging* – After merging the relationships between the original ontologies disappear.

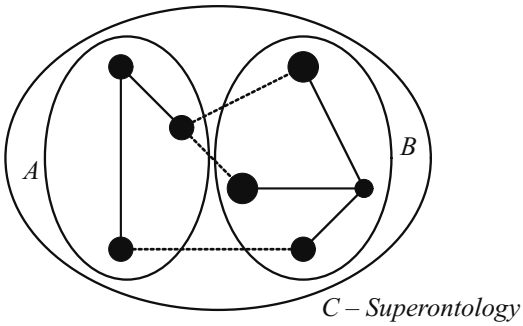


Figure 3: *Integration* – First ontology is unaltered while the second one is modified.

Whereas original ontologies are during *ontology merging* replaced by a new ontology (without initiation of direct correspondence between initial ones and the new one)[3], some documents need not reflect this replacement, but denote original ontologies. On the contrary, in the case of *ontology integration* the *superontology* is logicaly connected with the initial ontologies and in case some documents reference a concept from an original ontology, this concept is put over superontology. For this reason I prefer *ontology integration* in practise.

Ontology matching is in most cases done manually or semi-autoamtically, mostly with a support of some graphical user interface. A manual specification of ontology parts for matching is time consuming and moreover error prone process. Therefore there is a need for development of faster methods, which can process ontologies at least semi-automatically.

There are several tools that support user ontology matching. These tools use various techniques for proposal of integration rules, some advanced ones solve the question how to effectively combine results of particular techniques. These techniques unwind from the level of abstraction they work with.

Disadvantage of some of these methods is the necessity of setting numerous parameters from which sugesstions of integration rules unwind. In many of them the parameters setting plays so essential role that it can not be accomplished without deeper knowledge of concepts desribed in partial input ontologies.

Usually every matching tool innovates ontology matching on a particular aspect, nevertheless there exist several similar properties (with only minor exceptions) common to all of these tools [4]:

- Schema-based matching solutions are much more investigated than instance based solutions. This is partly caused by the fact instances may not be available during ontology matching process.
- Most of the systems focus on specific application domains (medicine, music...) as well as on dealing with particular ontology types (RDF, OWL...). Only few system are so general they can suit various application domains together with generic ones and support multiple formats. These are, for example, COMA++ [I2] or S-Match [5].
- Most approaches take as input a pair of ontologies. Only few systems take as input multiple ontologies or more general structures. These are, for example, DCM [6] or Wise-Integrator (automatical web form data integration) [7].
- Most of the approaches handle only tree-like structures. Several advanced systems handle more general graph structures. These are, for example, COMA/COMA++ [I2] or OLA [I3] (uses Alignment API [I1]).
- Most of the systems focus on discovering of *one-to-one* alignments. But it is possible to encounter more complicated relationships as *one-to-many* or *many-to-many*. These relationships can manage for example DCM (use statistical methods and is not applicable in this study) [6] or CTXMatch2 [1].
- Most of the systems identify relationships (i.e. Prompt [8]), some of them focus on computing confidence measures of these relationships (i.e. COMA++ [I2]). This is based on the assumption of equivalence relation between ontology entities. Only few systems compute logical relations between ontology entities (such as equivalence or subsumption). These are for example CTXMatch2 [1] or S-Match [5].

[3] Correspondences between ontologies, provenance and other metadata can be represented by other indirect methods [11].

2. Experiment

The folowing tools were used in the experiment as representatives of "exceptions" from the previous list – COMA++ [I2], CTXMatch [1] and Alignment API [I1]. For demonstration of automatical suggestion of alignment was used Prompt [8] (plugin for Protégé system [I6]).

Our experiments were executed with the test OWL [I4] ontologies (MyPerson.owl and MyCustomer.owl) pictured on Fig. 4. For testing the ontologies containing classes only were used.

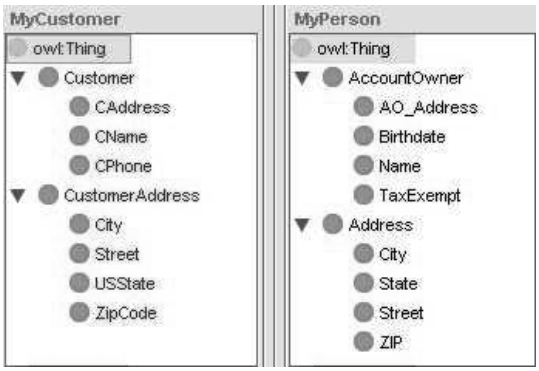


Figure 4: Test ontologies.

The test ontologies were matched directly by particular tools or by application interfaces.

Elements of the test ontologies were numbered in the following way:

Columns		Rows	
1:	AccountOwner	1:	CustomerAddress
2:	AO_Address	2:	Street
3:	Birthdate	3:	ZipCode
4:	TaxExempt	4:	City
5:	Name	5:	USState
6:	Address	6:	Customer
7:	State	7:	CPhone
8:	Street	8:	CName
9:	City	9:	CAddress
10:	ZIP		

Following table represents relationships that could be subjectively expected as "ideal" on the assumption that *Account Owners* are considered to be *Customers* ($\sim$), etc. Sign $\sqsubset$ means the relation of subsumption. Sign $\sqsupset$ denotes generalization. Values in the tables then express a confidence measure of the fact that relations mentioned above conform. If there are some missing rows or columns in the tables, they contained no data.

	1	2	3	4	5	6	7	8	9	10
1						$\sim$				
2								$\sim$		
3										$\sim$
4									$\sim$	
5							$\sim$			
6	$\sim$									
7										
8					$\sim$					
9		$\sim$								

2.1. CTXMatch

CTXMatch2.2 [1] uses a semantic matching approach. It translates the ontology matching problem into the logical validity problem and computes logical relations, such as equivalence or subsumption between concepts and properties. CTXMatch is a sequential system which, at the element level, uses only WordNet [I7] to find initial matches for classes. At structure level it uses logical reasoners (i.e. Pellet [I5]) with the help of deductive techniques and verification of performability of logical formulas to compute resulting matching.

Threshold value – Matching results can be filtered by setting the **threshold** in the $< \mathbf{0}, \mathbf{1} >$ range. Relationships rated by lower value (in case of this experiment value 0.5) are not reflected (and not displayed in tables) for inconclusiveness.

Ontology throughpass task – Deep ontology throughpass (**hierarchical task**) is denoted by the word "hierarchy", flat throughpass (**flat task**) is denoted by the word "flat".

Mapping – Mapping **one-to-one** is denoted by **1:1**. Mapping **many-to-many** is denoted by **M:M**.

The same settings for all the experiments are the following:

- threshold: **0.5**
- input format: **OWL** [I4]
- output format: **XML** [I8]
- matching method: **DL** (using of description logic for deduction of possible relationships)

2.2. Alignment API

Alignment API is Java application interface that uses methods based on processing of word strings (String-based methods). It is used by other matching tools like OLA [I3] or FOAM [3].

Levenshtein Algorithm – The Levenshtein distance [9] is defined as the minimal number of characters we have to replace, insert or delete to transform one string into another.

Smoa Algorithm – Smoa [9] is the measure dependent on the length of "common" substrings and "not common" substrings, when the second mentioned part is deleted from the first one.

WordNet Database – WordNet [I7] is a leading linguistic database of English at worldwide scale. It groups together english words into the set of synonyms called *synsets* and give their short general definitions.

2.3. COMA++

COMA/COMA++ (COmbination of Matching Algorithms) [2] is a schema matching tool based on parallel composition of matchers. In his graphical user interface it offers an extensible libraries of matching algorithms. It is possible to modify default settings and parameters for certain ontologies in order to get better results. Parameters and settings in this case are not only threshold or one default string method, but many others (for example setting of consequence of used techniques).

2.4. Prompt

Prompt [8] is an extension plugin to the Protégé editor [I6]. Among other operations with pairs of ontologies (merging, extraction, versioning...) Prompt offers also interface for transformation of one ontology to another one and therefore it uses automatic matching at first.

Table 1: Similarity Measure CTXMatch – hierarchy – M:M.

	5	6	7	8	9	10
1		$\sqsupset$ **1.0**	$\sqsupset$ 1.0	$\sqsupset$ 0.56	$\sqsupset$ 0.56	$\sqsupset$ 0.56
2		$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sqsupset$ **1.0**	$\sqsupset$ 0.56	$\sqsupset$ 0.56
3		$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sqsupset$ 0.56	$\sqsupset$ **0.56**
4		$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sqsupset$ 0.56	$\sqsupset$ **1.0**	$\sqsupset$ 0.56
5		$\sqsupset$ 1.0	$\sqsupset$ **1.0**	$\sqsupset$ 0.56	$\sqsupset$ 0.56	$\sqsupset$ 0.56
6			$\sqsupset$ 0.67			
7			$\sqsupset$ 0.67			
8	$\sqsupset$ **0.67**		$\sqsupset$ 0.67			
9		$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sqsupset$ 0.56	$\sqsupset$ 0.56	$\sqsupset$ 0.56

Table 2: Similarity Measure CTXMatch – hierarchy + Semantic Relation – M:M.

	5	6	7	8	9	10
1		$\sqsupset$ **1.0**	$\sqsupset$ 1.0	$\sim$ 0.48	$\sim$ 0.48	$\sim$ 0.48
2		$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sqsupset$ **1.0**	$\sim$ 0.4	$\sim$ 0.4
3		$\sim$ 0.62	$\sim$ 0.62	$\sim$ 0.4	$\sim$ 0.4	$\sim$ **0.4**
4		$\sqsupset$ 1.0	$\sqsupset$ 1.0	$\sim$ 0.4	$\sqsupset$ **1.0**	$\sim$ 0.4
5		$\sqsupset$ 1.0	$\sqsupset$ **1.0**	$\sim$ 0.4	$\sim$ 0.4	$\sim$ 0.4
6			$\sim$ 0.53			
7			$\sim$ 0.45			
8	$\sim$ **0.45**		$\sim$ 0.45			
9		$\sqsupset$ 1.0	$\sim$ 0.62	$\sim$ 0.4	$\sim$ 0.4	$\sim$ 0.4

Table 3: Similarity Measure CTXMatch – hierarchy – 1:1.

	1	2	3	4	5	6	7	8	9
1						$\sqsupset$ **1.0**			
2	$\sqsupset$ 0.39								
3							$\sqsupset$ 1.0		
4								$\sqsupset$ 0.56	
5									$\sqsupset$ 0.56
6				$\sqsubset$ 0.39					
7									
8					$\sqsupset$ **0.67**				
9									

Table 4: Similarity Measure CTXMatch – hierarchy + Semantic Relation – 1:1.

	1	2	3	4	5	6	7	8	9
1						$\sqsupset$ **1.0**			
2	$\sim$ 0.31								
3							$\sim$ 0.62		
4								$\sqsupset$ 0.4	
5									$\sqsupset$ 0.4
6				$\sqsubset$ 0.31					
7									
8					$\sqsupset$ **0.45**				

Table 5: Similarity Measure CTXMatch – flat – 1:1.

	5	6	7	8	9
1					
2				$\sqsubset \sqsupset \sim$ **1.0**	
3					
4					$\sqsubset \sqsupset \sim$ **1.0**
5					
6					
7			$\sqsupset$ 1.0		
8	$\sqsupset$ **1.0**				

Table 6: Similarity Measure CTXMatch – flat + Semantic Relation – 1:1.

	5	6	7	8	9
1					
2				*Equiv.* **1.0**	
3					
4					*Equiv.* **1.0**
5					
6					
7			$\sqsupset$ 1.0		
8	$\sqsupset$ **1.0**				

Table 7: Similarity Measure CTXMatch – flat – M:M.

	5		6		7		8	9
1			$\sqsupset$ **1.0**	$\sim$ **0.7**	$\sqsupset$ 1.0	$\sim$ 0.7		
2					$\sqsupset$ 1.0	$\sim$ 0.7	$\sqsubset \sqsupset \sim$ **1.0**	
3					$\sqsupset$ 1.0	$\sim$ 0.7		
4					$\sqsupset$ 1.0	$\sim$ 0.7		$\sqsubset \sqsupset \sim$ **1.0**
5					$\sqsupset$ **1.0**	$\sim$ **0.7**		
6					$\sqsupset$ 1.0	$\sim$ 0.7		
7					$\sqsupset$ 1.0	$\sim$ 0.7		
8	$\sqsupset$ **1.0**	$\sim$ **0.7**			$\sqsupset$ 1.0	$\sim$ 0.7		
9			$\sqsupset$ 1.0	$\sim$ 0.7	$\sqsupset$ 1.0	$\sim$ 0.7		

Table 8: Similarity Measure Alignment API – Levenshtein.

	1	2	3	4	5	6	7	8	9	10
1		$\sim$ **0.53**								
2								$\sim$ **1.0**		
3										$\sim$ **0.43**
4									$\sim$ **1.0**	
5							$\sim$ **0.71**			
6				$\sim$ 0.5						
7	$\sim$ 0.33									
8					$\sim$ **0.8**					
9						$\sim$ 0.87				

Table 9: Similarity Measure Alignment API – Smoa.

	5	6	7	8	9	10
1		$\sim \mathbf{0.82}$				
2				$\sim \mathbf{1.0}$		
3						$\sim \mathbf{0.6}$
4					$\sim \mathbf{1.0}$	
5			$\sim \mathbf{0.92}$			
6						
7						
8	$\sim \mathbf{0.89}$					
9		~ 0.93				

Table 10: Similarity Measure Alignment API – WordNet.

	5	6	7	8	9	10
1		$\sim \mathbf{0.64}$				
2				$\sim \mathbf{1.0}$		
3						$\sim \mathbf{0.86}$
4					$\sim \mathbf{1.0}$	
5			$\sim \mathbf{0.83}$			
6	~ 0.33					
7						~ 0.6
8	$\sim \mathbf{0.94}$					
9		~ 0.97				

Table 11: Similarity Measure COMA++ – Own settings + COMA defaults.

	5	6	7	8	9	10
1		$\sim \mathbf{0.69}$				
2		~ 0.77		$\sim \mathbf{0.78}$		
3						
4						
5			$\sim \mathbf{0.83}$			~ 0.61

Table 12: Similarity Measure Prompt – Automatic Matching.

	1	2	3	4	5	6	7	8	9	10
1										
2								$\sim$		
3										
4									$\sim$	
5							$\sim$			
6										
7										
8					$\sim$					
9		$\sim$				$\sim$				

3. Experiment evaluation

Mapping returned by tool CTXMatch with hierarchical throughpass task identified 6 from 8 possible relationships, but how it is visible from Table 1, next to these relationships are with the same coefficients of confidence detected other relationships between ontologies, which do not correspond to any facts. In other words it could be better to use this method for weighing of already detected relationships than for detection alone.

If we combine similarity measure with linguistic analysis (Semantic Relation), see Table 2, by most of wrong selected candidates comes to a downtrend of rating. This downtrend is noticed even by two nonconflicting rules, but without a detriment to correctness.

In case of choosing one-to-one mapping by the same method, we can do the selection of candidates in Table 3, respectively with linguistic analysis in Table 4 based on ratings from Table 1 and 2. By this selection initially

wrong selected relationships are reduced , but only 2 selected relationships correspond to the facts. While flat throughpass task (see Tab. 5) detected correctly 3 of 8 relationships, only one wrong relationship was selected by on-to-one mapping. The lexical analysis (see Tab. 6) does not bring any changes into the result mapping.

If we pass over the necessity of selection one-to-one mapping (as shown in Tab. 7), ambiguity of assigning ontology elements appears only by element 7 – *State*.

If we compare hierarchical throughpass task and flat throughpass task, selection of candidates is more restrictive (candidates have lower rating). If we focus on the type of analysis, then methods based on Levenshtein algorithm were able to correctly select 6 of 8 relationships with 3 wrong (see Tab. 8), methods based on Smoa algorithm selected correctly 6 relationships and 1 wrong (see Tab. 9). Methods using WordNet database selected 6 correct and 3 wrong relationships (see Tab. 10). If we compare results from these analysis with Table 1, we can see that relationships selected by these methods could help to solve ambiguities of selection (i.e. by tool CTXMatch). Not least COMA++ tool detected correctly 5 of 8 relationships (see Tab. 11) and it can be rated very positively without selection of wrong relationship. Similarly, extension Prompt, detected only 3 equivalences (see Tab. 12) without marking wrong relationship. Both last mentioned tools use combination of different methods and in term of confidence return correct mapping rules, but at the price of detecting incomplete set of these rule (some relationships are not detected at all).

4. Conclusion

In our experiment evaluation we have not found any tool that perfectly covers whole spectrum of ontology matching tasks. It tends to necessity of using more tools and combine their results.

From the performed study it follows (using given tools) that it may be appropriate to make rough outline by tool CTXMatch that can filter out candidates with less support. Ambiguity of selection can be solved by String-based methods (WordNet, Levenshtein Algorithm...), which do not need to give correct results on principle. Remarkable is really sporadic occurence of subsumption. It can be explained by the fact that subsumption can be detect in most cases by logical reasoner, when it uses information about existence of relationships between some concepts.

Our experiments reflect that tools COMA++ and Prompt offer better portfolio of methods, which are combined and return more accurate (but sometimes conservative) results with the possibility that some acceptable mappings are not proposed at all.

Therefore it is very useful to validate the results of matching process against the data used by initial ontologies. My future research direction will follow the same topics.

References

[1] BOUQUET, P. – SERAFINI, L. – ZANOBINI, S.: “Semantic Coordination: A New Approach and Application”. In *Proc. 2^{nd} International Semantic Web Conference (ISWC)*, volume 2870 of Lecture Notes in Computer Science, p. 130–145, Sanibel Island (FL US), 2003.

[2] DO, H. – RAHM, E.: “COMA – A System for Flexible Combination of Schema Matching Approaches”. In *Proc. 28^{th} International Conference on Very Large Data Bases (VLDB)*, p. 610–621, Hong Kong (CN), 2002.

[3] EHRIG, M. – SURE, Y.: “FOAM – Framework for Ontology Alignment and Mapping – Results of the Ontology Alignment Evaluation Initiative”. In *Proceedings K-CAP Workshop on Integrating Ontologies*, volume 156, p. 72–76, Banff (CA), 2005.

[4] EUZENAT, Jérôme – SHVAIKO, Pavel: “Ontology Matching”. Springer-Verlag, Berlin/Heidelberg, 2007. ISBN 978-3-540-49611-3.

[5] GIUNCHIGLIA, F. – SHVAIKO, P. – YATSKEVICH, M.: “S-Match: An Algorithm and an Implementation of Semantic Matching”. In *Proc. Dagstuhl Seminar*, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl (DE), 2005.

[6] HE, B. – CHANG, K. C.: “Automatic Complex Schema Matching Across Web Query Interfaces: A Correlation Mining Approach”. *Volume 31 of ACM Transactions on Database Systems (TODS)*, p. 346–395, ACM, New York, 2006.

[7] HE, H. – MENG, W. – YU, C. – WU, Z.: “WISE-Integrator: A System for Extracting and Integrating Complex Web Search Interfaces of the Deep Web”. In *Proc. 31^{st} International Conference on Very Large Data Bases (VLDB)*, p. 1314–1317, Trondheim (NO), 2005.

[8] NOY, F. N. – MUSEN, M.: “PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment”. In *Proc. 17^{th} National Conference of*

Artificial Intelligence (AAAI), p. 450–455, Austin (TX US), 2000.

[9] STOILOS, G. – STAMOU, G. – KOLLIAS, S.: "A String Metric for Ontology Alignment". In *Proc. 4th International Semantic Web Conference (ISWC)*, volume 3729 of Lecture Notes in Computer Science, p. 624–637, Galway (IE), 2005.

[10] STRACCIA, U. – TRONCY, R.: "oMAP: Combining Classifiers for Aligning Automatically OWL Ontologies". *Volume 3806 of Lecture Notes in Computer Science*, p. 133–147, Springer-Verlag, Berlin/Heidelberg, 2005.

[11] VRANDECIC, D. – VÖLKER, J. – HAASE, P. – TRAN, D. T. – CIMIANO, P.: "A Metamodel for Annotations of Ontology Elements in OWL DL". In *Proceedings of the 2nd Workshop on Ontologies and Meta-Modeling*, Karlsruhe (GE), 2006.

[I1] Alignment API and Alignment Server [online]: <http://alignapi.gforge.inria.fr>.

[I2] COMA++ – A System for Flexible Combination of Matching Algorithms [online]: <http://dbs.uni-leipzig.de/en/Research/coma.html>.

[I3] OLA – OWL Lite Alignment [online]: <http://www.iro.umontreal.ca/~owlola/alignment.html>.

[I4] OWL – Web Ontology Language / W3C Semantic Web Activity [online]: <http://www.w3.org/2004/OWL>.

[I5] Pellet – OWL Reasoner [online]: <http://pellet.owldl.com>.

[I6] Protégé – Ontology Editor and Knowledge Acquisition System [online]: <http://protege.stanford.edu>.

[I7] WordNet – Lexical database [online]: <http://wordnet.princeton.edu>.

[I8] XML – Extensible Markup Language / W3C XML Activity [online]: <http://www.w3.org/XML>.

Model Selection for Detection of Directional Coupling from Time Series

Post-Graduate Student:
ING. MARTIN VEJMELKA
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
vejmelka@cs.cas.cz

Supervisor:
RNDR. MILAN PALUŠ, DRSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
mp@cs.cas.cz

Field of Study:
Biocybernetics and Artificial Intelligence

This work was supported by the EC FP6 NEST initiative project BRACCIA.

Abstract

This paper deals with the problem of selecting the conditioning model in the estimation of conditional mutual information in the context of detecting directional influence from raw time series. An approach similar to model selection in model fitting to time series is presented. A numerical study illuminating the problem and showing the effectivity of the proposed procedure is summarized at the end of the paper.

1. Introduction

The discipline of nonlinear dynamics has proven fruitful as many problems from meteorology [1, 2], geology [2], life sciences [3] and physics have been more satisfactorily understood in this framework. Time series analysis is a frequent tool used to process the activity records of dynamical oscillatory processes. Methods have been developed to detect various forms of synchronization and directional coupling from time series. The detection of directional influence is an important method of examining drive-response relationships in complex dynamical systems. Paluš [4, 5] has advocated the use of the conditional mutual information functional $I(X;Y_T|Y)$ between the two time series as a measure of "net information flow" between the process $\mathcal{X}$ and the process $\mathcal{Y}$ at some point of time in the future. Conditional mutual information has been applied in the context of phase dynamics to phase time series which simplify the analysis of signals [5, 6]. In this work the problem of discovering the directionality of coupling in amplitude time series is investigated and a method to solve one of the problems is presented.

The conditional mutual information can be decomposed into several terms which are interpretable in the context of time series analysis of nonlinear dynamical systems

$$I(X;Y_T|Y) = I(X;Y_T;Y) - I(X;Y) - I(Y;Y_T), \quad (1)$$

where X and Y are the time series of the processes $\mathcal{X}$ and $\mathcal{Y}$ respectively and Y_T is the time series of the process $\mathcal{Y}_T$, which is the process $\mathcal{Y}$ shifted by T samples into the future.

The term $I(X;Y_T;Y)$ of (1) represents the total common information in all the processes $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Y}_T$.

The term $I(X,Y)$ represents the effect of common dynamics and common history. Common history can be brought about by the same noise or external influences on the two processes. If the two processes have narrowband spectra with close peaks, then their time series may have some common parameters (e.g. the period of oscillation), this increases the amount of mutual information in the first term and must be subtracted. If additionally the dynamics themselves, which are represented by the equations in case of models, share some common traits or the entire form then this may cause similar amplitude distributions. None of the above effects is brought about by the influence of directional coupling. It is therefore important to subtract these components from the term $I(X;Y_T;Y)$ to ensure that they are excluded from the estimation of "net information flow". We note here that the mutual information $I(X,Y)$ can be used to detect synchronization of the investigated processes.

The term $I(Y;Y_T)$ represents the action of the process upon itself and is connected to the predictability of the process. It is imperative that this term is estimated well and removed from the total common information. Underestimation of this term will result in false positive detections as strong action of the process $\mathcal{Y}$ upon itself

will be misinterpreted as directional influence from the process $\mathcal{X}$. Effective estimation of this term is crucial to the correct application of the framework for detecting directional influence and will be the goal of this work.

The variables X, Y represent the time series of the given processes and may in general be multidimensional. Multidimensional time series can be either directly measured by observing several aspects of the activity of a dynamical process or can be constructed from a single time series by means of an embedding technique. A frequently used embedding technique is that of time-delay embedding [7, 8], where equidistantly spaced samples of a given time series are used to construct a vector

$$\begin{aligned} \bar{x}(t) &= (x_1(t), x_2(t), ..., x_K(t)) \\ &= (x(t), x(t-\tau), ..., x(t-(K-1)\tau)), \end{aligned} \tag{2}$$

where $x(t)$ is the scalar time series of the activity of process $\mathcal{X}$ and τ is the *delay* between successive samples and K is the embedding dimension.

An important parameter is the number of samples the process $\mathcal{Y}$ is shifted into the future. In our previous work [9, 10] conditional mutual information is averaged for shifts T from 1 up to two periods of the faster process in the investigated system pair. For model systems or systems with simple structure improvements to this scheme are possible as there are clear patterns in the conditional mutual information with respect to the time shift. The estimation method used is equiquantal binning as it has shown the best properties in model tests and has been successfully applied to some experimental datasets [10, 6, 2].

1.1. The intersample delay

There are multiple established techniques for selecting the time delay τ to construct a vector representation of the state of a dynamical system from a univariate time series [8, 7, 11]. Kantz and Schreiber [12] have however argued that there is no optimal way of selecting the time delay in general. Rather the specific purpose with which the embedding is constructed allows one to discuss and gauge the optimality of an embedding method. The use of an intersample delay is a way to circumvent the problem of selecting samples that are highly correlated and thus as a set contain a lower amount of information about the structure of the system in state space [12]. The classical procedure requires the delay to be fixed first and then using another method the dimension is fixed by testing if adding more dimensions to the vector is reasonable [13]. This is simple because samples are considered sequentially. However we know of no apriori reason to restrict the selection procedure in this way.

It is important to produce a model which fits the dynamics of the time series as well as possible in a sense that will be described later. Selecting the delay greater than 1 in effect pre-filters the samples that can be included in the model. If the intersample delay is say $\tau = 2$ then only the time series samples $x(t-d), d \in \{2, 4, 6, 8, ...\}$ may be considered for inclusion in the conditioning model. Since the model search procedure is time intensive, it is advantageous to apriori restrict the set of possible delays for performance purposes. Additionally, a model utilizing samples close to each other will end up modeling the temporal structure of the time series instead of the geometrical structure of the attractor in the state space. However these are not rigorous arguments and counterexamples may be found where the optimal selected model contains samples close together.

The most frequently used method of selecting the intersample delay is to select the first minimum of the lagged mutual information $I(Y; Y_T)$ where T is the lag in samples between the original and shifted time series of the process [14]. This is the procedure that will be henceforth used to select an intersample delay. There have been many other suggestions in the literature (an overview can be found in [12]) but all of the suggested methods are based on heuristic arguments. Time lagged mutual information has been applied and found to work well in many practical settings although caution is advised as the first minimum may be spurious.

2. The model selection procedure

The purpose of this work is to select a proper vector representation of the process $\mathcal{Y}$ which enables a good estimation of the term $I(Y; Y_T)$ in (1) as explained in the Introduction. A good model is a model that maximizes the *expected* lagged mutual information $I(Y; Y_\tau)$, where τ is the intersample delay selected according to the method in the last paragraph. There are two reasons for this choice: a single lag is necessary because of the computational costs of computing the full, say 50, estimates and averaging them. Secondly, selecting too small a lag will result in temporal correlations guiding the selection and a lag too large will attenuate the deterministic structure between the lagged process and the original process. Because real dynamical processes are affected by external influences and usually are encumbered by noise, this means that the effects of the auto-structure of the process are attenuated for larger distances.

2.1. Model specification and criterion

Formally, each model M is completely specified by the indices of the samples used in constructing the state space vector $\bar{y}(t)$ as

$$\begin{aligned} M &= \{i_1, i_2, ..., i_K\} \text{ implies that} \\ \bar{y}(t) &= (y(t - i_1\tau), y(t - i_2\tau), ..., y(t - i_K\tau)), \end{aligned} \tag{3}$$

where K is the number of samples in the vector and depends on M. We will denote by Y_M the state space representation of the process Y using the vector specified by M. Then the best model M^* has the property

$$M^* = \text{argmax}_M \mathcal{E}[I(Y_M; Y_\tau)] \tag{4}$$

It is important to maximize the expectation of the mutual information over entire reconstructed space because the in-sample estimate would always increase if more samples were added to an existing state-space vector. This phenomenon is known as overfitting in the pattern recognition community. The problem can be converted to a problem of minimizing the conditional entropy

$$M^* = \text{argmin}_M \mathcal{E}[H(Y_\tau|Y_M)], \tag{5}$$

as $H(Y_\tau|Y_M) = I(Y_M; Y_\tau) + H(Y_\tau)$ and $H(Y_\tau)$ is a constant with respect to the optimization problem. In fact, due to the use of the equiquantal estimator the marginal entropy $H(Y_\tau) = H(Y) = \log B$. As usual, we assume the underlying processes to be ergodic for the duration of the analysis time window and this allows us to substitute expected values over time for expected values over the state space.

Any admissible model can be expressed as

$$M = (i_1, i_2, ..., i_K), \tag{6}$$

for $0 = i_1 < i_j < i_{j+1} \leq L, j \in \{2, .., K\}$ where L is some pre-selected maximum distance to the farthest considered sample and $K < K_{\max}$ is the number of elements in the model. It is important that $i_1 = 0$ is always included in the model because otherwise the random variables X and Y in term $I(X; Y)$ in (1) would not be taken at the same instant of time and would thus not represent the common history of the two processes. This would give the computed conditional mutual information different semantics and it would not reflect the "net information flow". This is not a significant restriction for dynamical systems because the action of noise, external influences and other factors causes the process to produce new information continuously and "forget" its initial conditions thus rendering samples further back in time less useful for constructing models. The threshold also limited by computational constraints and the number of models. The maximum size of the model $K_{\max}$ is also limited by computational constraints as the size of the model set grows combinatorially. A more important limit is the length of the time series itself which affects the maximum size K of the model M which can be reliably estimated. This however happens automatically during the estimation process as models with too many free parameters with respect to the length of the time series will be poorly estimated and the expected value of the conditional entropy will be high.

2.2. Conditional entropy and classification

It remains to show how the expectation of the criterion $[H(Y_\tau|Y_M)]$ can be computed for a given model M. First, given the number of bins B, the samples of the investigated time series are discretized using the equiquantal scheme into the B levels. The model specification M is then used to construct pairs

$$(\bar{y}_M(t), y(t + \tau)), \tag{7}$$

where the indices building the vector $\bar{y}^i_M(t)$ will be selected according to the model specification M. As the time when the training pair occurs in the time series will not be important, we will abbreviate the notation of the state vector $\bar{y}^i_M$ and the (predicted) future value to y^i_τ. When denoting the variable rather than a particular value, the index i will be omitted. The training pairs will be used to construct a classifier which will attempt to model the probability distribution function (PDF) of the state space of the underlying process. The classifier will be a simple multidimensional histogram which will aggregate all the training samples in its estimate of the PDF. The goal of the classifier is to predict the future state y^i_τ from the given vector $\bar{y}^i_M$. This process might seem crude but the key point is that in the estimation of the conditional mutual information functional (1), all the terms are estimated in exactly the same way. It follows that any problems that the classification process will have in estimating the PDF correctly are also expected in the estimation of CMI. It would thus not be useful to use a different classification scheme here because the model fitting procedure would yield a model which would not respect the advantages and disadvantages of this particular estimator and could potentially have a completely different number of free parameters.

It will now be shown that choosing a suitable loss function results in the error rate to be an estimate of the required criterion (conditional entropy)

$$L(y^i_\tau, \bar{y}^i_M) = -\log p(y^i_\tau|\bar{y}^i_M), \tag{8}$$

where the conditional entropy $p(y_\tau|\bar{y}_M)$ is unknown. We must substitute an estimate of the conditional

probability computed as

$$\hat{p}(y_\tau^i|\bar{y}_M^i) = \frac{N(y_\tau^i, \bar{y}_M^i)}{\sum_{y'_\tau} N(y'_\tau, \bar{y}_M^i)}, \tag{9}$$

where $N(\cdot,\cdot)$ is the number of occurrences of the pair in the training set. As the pair $(\bar{y}_M^i, y_\tau^i)$ is expected to be seen in a long sequence with probability $p(\bar{y}_M^i, y_\tau^i)$, the expected mean error over the state space will be

$$\begin{aligned} &- \sum_{(y_\tau, \bar{y}_M)} p(y_\tau, \bar{y}_M) \log \hat{p}(y_\tau|\bar{y}_M) \\ &- \sum_{(y_\tau, \bar{y}_M)} p(\bar{y}_M) p(y_\tau|\bar{y}_M) \log \hat{p}(y_\tau|\bar{y}_M) \\ &= \hat{H}(Y_\tau|Y_M) \end{aligned} \tag{10}$$

To further understand this result, let us relate it to the expected error assuming we would know the true distribution $p(y_\tau, \bar{y}_M)$:

$$\begin{aligned} &\hat{H}(Y_\tau|Y_M) - H(Y_\tau|Y_M) = \\ = \; &- \sum_{(y_\tau, \bar{y}_M)} p(\bar{y}_M) p(y_\tau|\bar{y}_M) \log \hat{p}(y_\tau|\bar{y}_M) + \\ + \; &\sum_{(y_\tau, \bar{y}_M)} p(\bar{y}_M) p(y_\tau|\bar{y}_M) \log p(y_\tau|\bar{y}_M) = \\ = \; &E_{Y_M} D(\hat{p}(y_\tau|\bar{y}_M) || p(y_\tau|\bar{y}_M)). \end{aligned} \tag{11}$$

The result shows that the expected error is equal to the value of the optimal expected error (conditional entropy) if the probability density function was known and the mean Kullback-Leibler divergence between the estimated and actual conditional probability density over all the states. It is clear that the conditional entropy is always overestimated. It is also clear that if the model contains a higher amount of free parameters (total histogram bins), the K-L divergence will increase as the estimate of the conditional probability density will be poorer and the bias will increase. This is behavior is favorable as it penalizes overfitting of the model.

Practically this procedure still has some unresolved problems. If a previously unseen pair $(y_\tau|\bar{y}_M)$ is encountered during the estimation of the criterion, the estimated conditional probability would be $\hat{p}(y_\tau|\bar{y}_M) = 0$ or undefined. The same would occur when a leave-one-out procedure is applied and the training pair exists only once in the training set. A regularization procedure is needed to deal with these pairs. Since the conditional probability estimate is computed from the accumulated histogram using (9). To resolve this a fixed term Δ is substituted for the unknown conditional probabilty in the loss function

$$L^*(y_\tau^i, \bar{y}_M^i) = \begin{cases} -\log \hat{p}(y_\tau^i, \bar{y}_M^i) & \text{if } N(y_\tau^i, \bar{y}_M^i) > 0 \\ -\log \Delta & \text{otherwise} \end{cases} \tag{12}$$

Obviously if no previously unseen states are encountered, the modified loss function gives identical results to the original loss function. When optimizing the model, we have elected to set $\Delta = \frac{1}{B}$, where B is the number of bins. This has the simple rationale that when the particular vector $\bar{y}_M^i(t)$ has not been seen in at all, then equal probability is assigned to all the possible future states y_τ.

Due to the form of the loss function, the same penalty is also assigned if the vector $\bar{y}_M^i(t)$ has been previously seen but not together with the given future state y_τ^i. In this case it is unclear whether $\frac{1}{B}$ is the best choice but no plausible argument has been found that would advocate selecting a different value for this situation.

A complete method for selecting a model for conditioning the CMI (1) from a given time series has now been constructed. The method connects a classification problem to the required criterion by using a suitably constructed loss function which is regularized for practical purposes.

We note here that there are many established methods for model selection in time series analysis (and elsewhere) such as the MDL principle [15], the Bayesian information criterion [16] or the Akaike information criterion [17]. These selection mechanisms however do not optimize the required criterion. These methods additionally assume a particular distribution family of the probability density function of the samples or the estimation of a likelihood function.

2.3. Including surrogates

It has been previously explained that the goal of the selection of the conditioning model was to be able to correctly determine directionality of coupling in as many cases as possible. To understand the influence of the surrogate time series on the usefulness of a particular conditioning model, it is necessary to recount the method of statistical testing of the estimates of conditional mutual information.

At the core of the directionality detection method is the estimation of conditional mutual information (1) for different lags T. These values are averaged over the selected lags to construct an index of directionality. This index reacts to an increase in coupling by increasing its value. However any directionality index also reacts to a change in other factors involving the underlying systems and the time series: noise levels, main frequencies, external influences on the systems and others. The *inverse problem* of determining directional influence is much more difficult: given a value of the index, can we infer that directional coupling exists ?

Surrogate testing is a method of verifying if there is sufficient evidence available to infer that directional coupling is present in a particular direction. The method is a simple one-sided hypothesis test with the null hypothesis of no directional coupling. The distribution of the index under the null hypothesis can be estimated by evaluating the index on as many surrogate time series as is deemed necessary and is computationally feasible. Usually 100 or 200 surrogates are used if the analysis is being performed offline. Surrogate time series are time series which preserve all of the properties of the original time series except the property being tested. Here, directional coupling is the tested property and surrogate time series are such time series that preserve the dynamical structure of the individual underlying processes but do not preserve the effect that coupling has on the time series. This is done by somewhat altering the temporal structure of the time series so that cause and effect of the coupling are separated and mixed in the time series. Common procedures which more or less accomplish this goal include Fourier transform surrogates [18], permutation surrogates [19], amplitude adjusted Fourier transform surrogates [20] or twin surrogates [21]. Each procedure is applicable in different situations and has it's advantages and disadvantages [10]. If a model of the underlying system is available, surrogate time series can be simply generated using the model by creating two pairs of time series of the coupled models and than taking the first time series from the first pair and the second time series from the second pair, these surrogates are called *equation-based surrogates*. These surrogates have the ideal properties and can be used as a standard against which other surrogate generation schemes are compared.

It is important to note that the hypothesis test is performed as if the surrogates had the ideal properties listed above. This is however only an approximation as the surrogate generation algorithm always destroys some of the dynamical structure it its random phase. This is a critical point for the model selection procedure.

Let this state of affairs now be related to the model selection procedure. Ideally when selecting a model, there would be enough data points in the source time series so that the set of data can be split into a *training* and *testing* set. The training set would be used to construct the models and the testing set would be used to obtain an unbiased estimate of the expectation of the criterion (4) for a given model. Assuming that models of the dynamical systems are available as much testing data as needed could be generated (this testing data would in fact be equivalent to the *equation-based surrogates*. This would seem to be fortuitous but in practice it is rarely the case that models of the underlying systems are available as the most interesting applications of the nonlinear dynamical framework are in areas where the physics of the analyzed systems is still poorly understood. If equation-based surrogates were available there would be no bias in the distribution under the null hypothesis stemming from the difference in the dynamics in the original and surrogate time series. In this case the conditioning model that would be optimal with respect to criterion (4) would also be optimal for use in the surrogate time series as they are for all practical purposes identical to the original time series. A leave-one-out procedure on the training set from the original time series would the suffice to select the best useable model.

In practice one of the above algorithms which does not need the underlying model is used to generate surrogates which are not identical in dynamical structure to the original time series. A possible exception to this rule are the twin surrogates which are difficult to apply in practice but do well in the preservation of the dynamical structure. Training and testing the model using a leave-one-out scheme would thus yield a model which is not the best possible for the evaluation of (1) as this model would not take into account the deformation of the dynamical structure due to the use of the surrogate generation algorithm. This is one of the most important practial caveats in the application of the above method for selecting conditioning models. It follows that creating the model on the original time series and testing the model (computing the criterion value) on the surrogates is what is required to obtain the best conditioning model. It has been found that the models selected using this procedure have less elements than those selected using a leave-one-out scheme. This is due to the fact that more complex models are more sensitive to the partial modification of the dynamical structure due to the surrogate generation algorithms. If the surrogates would have a dynamical structure identical to the original time series, then this procedure would be exactly the same as would be applied in a standard pattern recognition problem with a training and testing set.

2.4. The final procedure

The entire procedure for model selection can thus be summarized as:

- Input: time series with N points, no. of bins B, maximum model size $K_{\max}$, most distant sample L
- Compute intersample delay τ
- Generate r surrogate time series for testing

- For each possible model M:
- Build the histogram estimate $\hat{p}(y_\tau|\bar{y}_M)$ on the original time series
- Estimate the expected conditional entropy on the r surrogate time series and average the result: this is the criterion value
- Select the model M^* with the smallest criterion value

The more surrogates are used, the better will be the estimated conditional entropy. The generation of surrogates is usually fast for most surrogate generation algorithms but the estimation of the expected conditional entropy is expensive for long time series and must be repeated for each model of which there are $\binom{L-1}{K-1}$ as $i_1 = 0$ is always part of the model.

3. Numerical studies

In this section the effectivity of the presented procedure for selecting conditioning models will be shown on model systems the parameters and structure of which are known.

3.1. Rössler systems

In the first example, we will work with the famous Rössler system pair:

$$\begin{aligned} \dot{x}_{1,2} &= -\omega_{1,2}y_{1,2} - z_{1,2} + \epsilon_{1,2}(x_{2,1} - x_{1,2}) \\ \dot{y}_{1,2} &= \omega_{1,2}x_{1,2} + a_{1,2}y_{1,2} \\ \dot{z}_{1,2} &= b_{1,2} + z_{1,2}(x_{1,2} - c_{1,2}), \end{aligned} \tag{13}$$

where $a_{1,2} = 0.15$, $b_{1,2} = 0.2$, $c_{1,2} = 10$, $\omega_{1,2} = 1 \pm 0.015$ and $\epsilon_{1,2}$ is the coupling between the systems. The systems are integrated using a Runge-Kutta 4th order scheme with $dt = 0.05$ and the resulting time series is subsampled by a factor of 6 to yield 20 points per period of the system. Conditional mutual information (1) is computed for lags $T \in \{1, .., 50\}$ and averaged. The number of bins was set to 8 which is a value that works well for many systems [9, 10].

Fig. 1 shows the resulting curves of conditional mutual information against coupling strength for different selected models for the length of time series 32768 samples. The coupling strength $\epsilon_1 = 0$ while ϵ_2 was varied between $\{0, 0.2\}$. Such a long time series allows even CMI estimates with 3 elements in the conditioning model to be computed and thus negates any advantage a simpler model might have due to insufficient data. The intersample delay was set to $\tau = 5$. At the top, the model $M_0 = \{0\}$ was applied. It is clearly seen here, that a single condition is not sufficient as the CMI curve for the reverse direction is not constant but increases considerably towards $\epsilon_2 = 0.08$. In the middle the model $M* = \{0, 1\}$ was the result of the above optimization procedure. The bottom row is the model $M_L = \{0, 1, 7\}$ which was selected by using a leave-one-out estimation method without using the surrogate time series to test the model. The larger model M_L does not bring any improvement over model M^* recommended by the model selection procedure. The curve in the direction of coupling reacts to the coupling just as well as the more complicated model. In the reverse direction, the conditional mutual information is constant and close to 0 until the generalized synchronization threshold is reached. This is the desired behavior of the index.

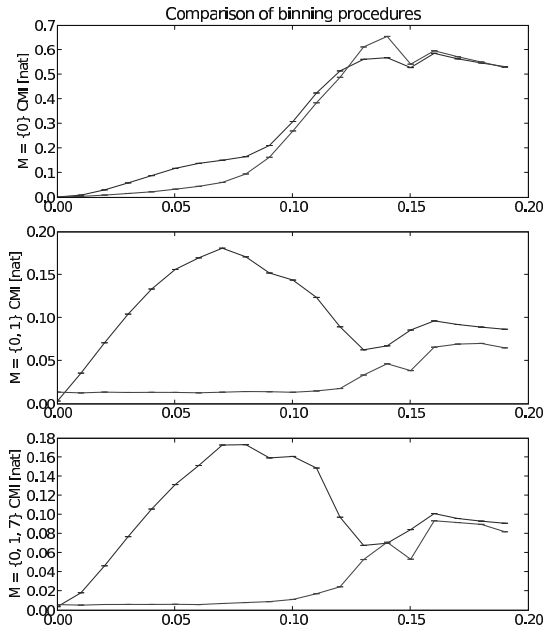


Figure 1: Conditional mutual information vs. strength of coupling for Rössler pair (13). Single condition model (top), optimal model per the selection procedure (center) and the model selected by the leave-one-out procedure on original time series data only (bottom).

Tests of detection of directionality in unidirectional coupling using all three models listed above have clearly shown that the model selected by the proposed procedure involving surrogate testing was the most effective. The proposed model has no false positives in the tested parameter range of window sizes from 256

points to 32768 points in powers of two and coupling strengths $\epsilon_1 = 0$ and $\epsilon_2 \in \{0, 0.1\}$. The model M_L had very low sensitivity and did not detect almost any directional coupling at all. The examination of the relevant histograms of the CMI indices has revealed that there is strong positive bias in the surrogates in the direction of coupling which renders all the detections negative. The model M_0 on the other hand has many false positive detections of coupling rendering the estimates unusable.

3.2. Van der Pol systems

The coupled Van der Pol equations are frequently used as example systems in nonlinear dynamics as they exhibit nonlinearity (and a stable limit cycle) but not deterministic chaos and complement other frequently used chaotic systems, such as the Rössler system or the Lorenz system. The nonlinearity of the Van der Pol system can be controlled by means of a parameter. The equations of the Van der Pol are given by

$$\ddot{x}_{1,2} - \mu(x_{1,2}^2 - 1)\dot{x}_{1,2} + \omega_{1,2}^2 x_{1,2} + \epsilon_{1,2}(x_{2,1} - x_{1,2}) + \eta_{1,2} = 0, \tag{14}$$

where $\mu = 0.2$ is the parameter affecting the nonlinearity of the model, $\omega_{1,2} = 1 \pm 0.1$ sets the main frequency of the model, $\eta_{1,2}$ are independent white zero-mean gaussian noise terms with standard deviation 0.1 and $\epsilon_{1,2}$ are the coupling strengths. The Van der Pol system pair was integrated with a Heun (reverse Euler) scheme with $dt = 0.01$ and subsampled by a factor of 20.

The intersample delay was comupted as $\tau = 5$ samples. With the parameters above the model selection procedure recommended the model $M^* = \{0, 6\}$, i.e. a two-dimensional model. The procedure was rerun without constraining the selected model to multiples of $\tau = 5$ and instead allowed to select any indices that are multiples of 2. Note that the prediction horizon $I(Y_\tau, Y_M)$ was the same in both runs. Using this less restrictive setting, the model selection procedure selected the model $M' = \{0, 12\}$ which is quite different to the previously chosen model. This shows that pre-selection can have adverse effects on the quality of the selected model.

Interestingly enough, the leave-one-out procedure selected a model $M = \{0, 12, 13\}$ with $\tau = 1$ (prediction horizon $I(Y_5, Y_M)$. The selected model is 3 dimensional, although the underlying dynamical model is only 2 dimensional. We note here that the model is not deterministic but stochastic and contains a noise input which is filtered by the dynamics of the system. Additionally, the model element 0 is forced to be a part of all models although it might not necessarily be useful in the prediction. Either of these considerations may explain why a 3 dimensional model was selected by the procedure.

The curves of conditional mutual information averaged for the lags $T \in \{17, 22\}$ (in case of the Van der Pols it is clear that coupling has most effect at these lags) is shown in Fig. 2.

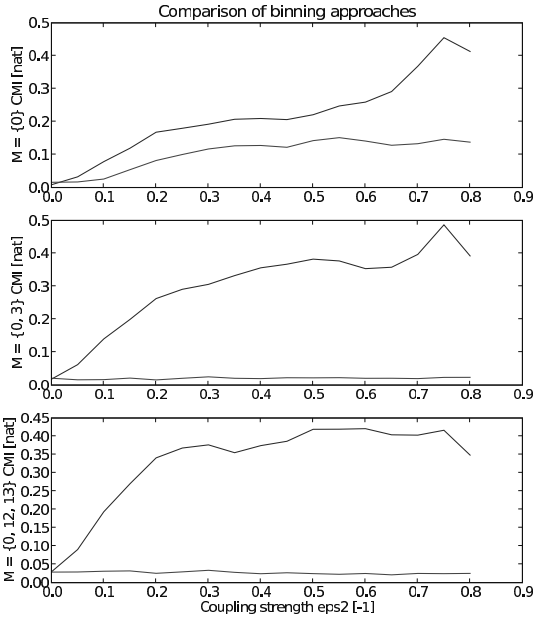


Figure 2: Conditional mutual information vs. strength of coupling for the Van der Pol pair (14). Single condition model (top), optimal model per the selection procedure (center) and the model selected by the leave-one-out procedure on original time series data only (bottom).

4. Conclusion

The selection of a conditioning model for processing amplitude series is a difficult problem and requires careful consideration. A method for selecting a conditioning model has been presented which attempts to select the optimal model with respect to the problem of detecting directional coupling.

The error of the prediction of a considered model was connected to the criterion (time lagged mutual information or conditional entropy) by selecting a suitable loss function. It has been shown that the error is positively biased with respect to the true expected value of the conditional entropy. The bias and variance that surrogates introduce into the directionality detection

method have been replicated in the model selection method by using generated surrogate time series to estimate the criterion instead of leave-one-out cross-validation or splitting the original time series into a training and testing set.

Some positive results have been shown on well-known and frequently used model systems. The recommended models have worked better than other reasonable choices. This has been verified by testing the conditioning models on the actual directionality detection problem for the considered systems. There are still however unresolved issues such as pre-filtering of the allowable samples to be included in the model and the methods is still very much a work in progress.

References

[1] D. Maraun and J. Kurths, "Epochs of phase coherence between El Niño/Southern Oscillation and indian monsoon," *Geophysical Research Letters*, vol. 32, no. 15, 2005.

[2] M. Paluš and D. Novotná, "Quasi-biennial oscillations extracted from the monthly NAO index and temperature records are phase-synchronized," *Nonlinear Processes in Geophysics*, vol. 13, pp. 287–296, 2006.

[3] C. Schäfer, M. Rosenblum, J. Kurths, and H.-H. Abel, "Heartbeat synchronization with ventilation," *Nature*, vol. 392, 1998.

[4] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, "Synchronization as adjustment of information rates: Detection from bivariate time series," *Physical Review E*, vol. 63, 2001.

[5] M. Paluš and A. Stefanovska, "Direction of coupling from phases of interacting oscillators : An information-theoretic approach," *Physical Review E*, vol. 67, 2003.

[6] B. Musizza, A. Stefanovska, P. V. E. McClintock, M. Paluš, J. Petrovcic, S. Ribaric, and F. F. Bajrovič, "Interactions between cardiac, respiratory, and eeg-delta oscillations in rats during anaesthesia," *Journal of Physiology*, vol. 580, pp. 315–326, 2007.

[7] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence* (D. Rand and L. Young, eds.), vol. 898, (Berlin), pp. 366–381, Springer, 1981.

[8] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, no. 3–4, pp. 579–616, 1991.

[9] M. Paluš and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," *Physical Review E*, vol. 75, p. 056211, 2007.

[10] V. M. and M. Paluš, "Inferring the directionality of coupling with conditional mutual information," vol. 77, p. 026214, 2008.

[11] T. Sauer, "Reconstruction of dynamical systems from interspike intervals," *Physical Review Letters*, vol. 72, no. 24, pp. 3811–3814, 1994.

[12] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*. Cambridge: Cambridge University Press, 1997.

[13] M. B. Kennel, B. R., and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45, p. 3403, 1992.

[14] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, pp. 1134–1140, 1986.

[15] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[16] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[17] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[18] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer, "Testing for nonlinearity in time series: The method of surrogate data," *Physica D*, vol. 58, pp. 77–94, 1992.

[19] A. Stefanovska, H. Haken, P. V. E. McClintock, M. Hožič, F. Bajrovič, and S. Ribarič, "Reversible transitions between synchronization states of the cardiorespiratory system," *Physical Review Letters*, vol. 85, pp. 4831–4834, 2000.

[20] T. Schreiber and A. Schmitz, "Improved surrogate data for nonlinearity tests," *Physical Review Letters*, vol. 77, pp. 635–638, 1996.

[21] M. Thiel, M. Romano, J. Kurths, M. Rolfs, and R. Kliegl, "Twin surrogates to test for complex synchronisation," *Europhysics Letters*, vol. 75, pp. 535–541, 2006.

Katalog lékařských doporučených postupů v ČR

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

Tento článek pojednává o projektu webového Katalogu lékařských doporučených postupů v ČR, jehož účelem je shromažďovat informace o dokumentech lékařských doporučení publikovaných českými odbornými autoritami prostřednitvím Internetu a poskytovat je odborné veřejnosti pro užití a podporu rozhodování v klinické praxi a pro další výzkum v oblasti tvorby a formalizace lékařských doporučení. Součástí projektu je struktura databáze a návrh webových rozhraní, které ve zkušebním provozu fungují na internetové adrese http://neo.euromise.cz/ddp .

1. Úvod

Lékařskými doporučenými postupy (lékařskými doporučeními, guidelines, dále jen LD) nazýváme popis uspořádání jednotlivých částí daného pracovního procesu lékaře, resp. ve zdravotnictví obecně. Jedná se o právně nezávazné, vysoce odborné dokumenty zaměřené na diagnostiku a terapii daného onemocnění, mnohdy jsou ovšem pojaty komplexně a snaží se postihnout celou šíři problematiky daného onemocnění, skupiny onemocnění nebo diagnostického či terapeutického zákroku. Jejich cílem je sjednotit, zjednodušit a zefektivnit péči o pacienta s použitím nejnovějších a nejkvalitnějších vědeckých poznatků, kterou je jejich prostřednictvím možno poskytovat jednotně v daném územním či společenském celku. Věcně a obsahově podobnými dokumenty jsou pak standardy léčebné péče či protokoly léčebné péče, které bývají formulovány konkrétněji, stručněji, se zaměřením na praktické použití a jeho efektivitu. Terminologicky jsou ovšem pojmy guideline, standard a doporučení v literatuře i chápání zdravotnických institucí mnohdy zaměňovány. [1, 2, 3, 4]

LD jsou vytvářena lékařskými autoritami na různých úrovních - od celosvětově působících autorit typu Světové zdravotnické organizace, přes odborně vyhraněné mezinárodní společnosti (např. Evropská kardiologická společnost) a národní organizace ať již odborné (např. Česká kardiologická společnost) nebo zaměřené na vývoj LD napříč odbornostmi (NICE - National Institute for Health and Clinical Excellence, UK), až po dokumenty vytvářené v rámci menších územních celků nebo jednotlivých pracovišť. Pro potřeby Katalogu lékařských doporučení a tohoto textu budou dále zmiňovány pouze dokumenty s minimálně celonárodní působností. [5, 6, 7, 8]

Zavádění LD do použití v praxi má smysl pouze v případě masového rozšíření těchto textů v cílové skupině lékařů. Jejich publikace se proto soustřeďuje především na odborná periodika specializovaná na konkrétní tematiku, na tématické nebo přehledové sborníky, či monografie. Publikace v elektronické podobě přináší mnohé výhody, především v rychlosti, ekonomice a celkové efektivitě zavedení do klinické praxe. V případě publikace prostřednictvím služby World Wide Web se dosahuje rychlé dostupnosti dokumentů pro velmi širokou skupinu uživatelů neomezenou geograficky ani počtem, nevýhodou je nutnost zajištění kvalitativních měřítek na Internetu dostupných dokumentů.[9]

1.1. Stav tvorby a publikace lékařských doporučených postupů v ČR

Pro proces vytváření LD byly v zahraničí vypracovány některé metodiky (SIGN, COGS, NICE) a nástroje pro ověřování kvality (AGREE). Odborné společnosti systematicky vyvíjející větší množství LD si také vytvořily vlastní metodiku. V České republice byl v roce 1998 zahájen projekt centrálně řízené tvorby LD zaštítěný ČLS JEP, v rámci kterého vzniklo kolem tří

set dokumentů LD, byl však velmi brzy ukončen. V současnosti některé skupiny publikující LD a obdobné dokumenty vytvářejí vlastní základy metodologie jejich tvorby (SVL, NRMSČR). [10, 12, 13]

LD jsou v současnosti v České republice vytvářeny odbornými lékařskými společnostmi, Českou lékařskou společností Jana Evangelisty Purkyně a Národní radou pro medicínské standardy ČR. Tvorbu doporučených postupů a jejich kvalitu nekoordinuje žádný společný orgán. Pro publikaci doporučených postupů slouží buď tištěná odborná periodika (např. Cor et Vasa, Vnitřní lékařství, Moderní gynekologie a porodnictví) nebo samostatné publikace (viz Společnost všeobecného lékařství). Většina odborných lékařských společností publikuje doporučené postupy také prostřednictvím Internetu v rámci vlastní webové prezentace.

Nejvíce doporučených postupů je (červen 2008) publikováno na serveru ČLS-JEP (305 doporučených postupů), dále Česká onkologická společnost (281 včetně zahraničních), Společnost všeobecného lékařství (34), Česká kardiologická společnost (34), Česká dermatovenerologická společnost (36), Česká neurologická společnost (17), Česká diabetologická společnost (12), Česká pneumologická a ftizeologická společnost (14), Česká revmatologická společnost (10) a další. Některé společnosti LD nepublikují, nebo je publikují pouze v tištěné podobě.

1.2. Zahraniční organizace a portály věnující se problematice lékařských doporučených postupů a jejich katalogizaci

Z důvodu rostoucí obliby Internetu jako publikačního média pro odborné texty, jmenovitě pro LD, vznikly v zahraničí jednak specializované národní či mezinárodní instituce pro tvorbu a katalogizaci LD, jednak těmito organizacemi nebo přímo státními knihovnickými, či zdravotními institucemi spravované webové katalogy zaměřené na shromažďování a zprostředkování informací o dokumentech LD a přímo na tyto dokumenty odkazující.

Metodice tvorby a obsahu LD se věnují například projekty:

- Conference on Guideline Standardization [10, 11]
- Scottish Intercollegiate Guidelines Network [12]
- National Institute for Health and Clinical Excellence [8]

Mezi nejvýznamější portály a katalogy shromažďující informace o zahraničních dokumentech LD patří:

- National Guideline Clearinghouse [15]
- National Library of Guidelines Specialist Library [16]
- Ärtzliches Zentrum für Qualität in der Medizin - Leitlinien.de [17]

2. Koncepce Katalogu lékařských doporučení v ČR

Tvorba LD v České republice je roztříštěna mezi jednotlivé zájmové skupiny, odborné lékařské společnosti, jejich specializované sekce a další odborné autority nebo organizace (například organizace záchranné služby, jednotliví odborníci). Dokumenty LD jsou publikovány nejen na Internetu v elektronické podobě, ale obecně na různých místech, v různé formě a mají různé kvalitativní parametry.

V současné době (červen 2008) neexistuje žádná webová ani bibliografická služba, která by monitorovala výskyt a kvalitu textů českých LD. Zahraniční služby jsou zaměřeny na cizojazyčné dokumenty, které ne vždy mají plnohodnotné použití pro specifické prostředí zdravotního systému v České republice. Navíc pro část odborné a hlavně laické veřejnosti, které by nemělo být v přístupu k informacím souvisejícím s kvalitou zdravotnické péče bráněno, představují cizojazyčné publikace jazykovou překážku.

Cílem vytvoření Katalogu lékařských doporučení v ČR bylo na jednom místě koncentrovat informace o českých LD, které mohou být využity například praktickými lékaři pro všeobecný přehled, lékaři specialisty, autory vědeckých publikací, doporučených postupů a zdravotní politiky, provozovateli zdravotních zařízení, další odbornou veřejností, případně i pacienty pro zpětnou kontrolu kvality zdravotní péče či kvalitní sebepéči. V neposlední řadě pak mohou být informace shromažďované v Katalogu lékařských doporučení v ČR využity pro vývoj systémů pro podporu rozhodování a dalších aplikací lékařské informatiky a výzkum v této oblasti.

2.1. Výběr kritérií pro katalogizaci, včetně formalizace

Při tvorbě konceptu databázového záznamu pro každý dokument LD byla brána v úvahu následující kritéria:

- výběr některých nejdůležitějších a v místních podmínkách aplikovatelných kritérií obsažených v návrhu COGS [10]
- výběr kritérií podle National Guideline Clearinghouse [15]

- nejčastěji se vyskytující identifikační a katalogizační údaje uváděné u českých LD
- zařazení informace o existující formalizaci (formálním modelu postupu LD) nebo webové aplikaci, která tuto formalizaci využívá
- informace o všech výskytech a variantách textu každého LD na Internetu

Seznam parametrů sledovaných pro každý dokument LD je podrobně uveden v části 2.3.

2.2. Technické řešení katalogu

Softwarové řešení katalogu vycházelo z několika předpokladů:

- databáze by měla zajistit co nejkomplexnější pohled na každý jednotlivý dokument LD a provázanost poskytovaných informací
- obsah databáze bude třeba průběžně doplňovat a aktualizovat
- strukturu databáze a vzhled a funkčnost webového rozhraní bude třeba průběžně upravovat a doplňovat, protože v České republice neexistuje norma ani obecný konsenzus nad kvalitativními kritérii LD
- minimálně ve fázi vývoje nebudou k dispozici prostředky na zajištění provozní SW technologie, proto je třeba volit zdarma dostupné technologie na bázi volně šiřitelného software
- cílová skupina uživatelů obsahuje české lékaře, vědce a specialisty (případně pacienty), proto jediným jazykem webového rozhraní i obsahu databáze by měla být čeština a aplikace tudíž jednojazyčná

Softwarové řešení Katalogu lékařských doporučení tedy tvoří:

- Webový server Apache 2.0
- PHP verze 5.2.3-1+b1
- Databázový systém MySQL 5.0.41-Debian$_2$-log
- Webová aplikace v aktuální verzi 1.3 (červen 2008) v jazyce PHP a databáze MySQL

2.3. Databáze

Databáze obsahuje v aktuální verzi 1.3 celkem 32 tabulek, z nichž nejdůležitější jsou uvedeny v Tabulce 1.

Název tabulky	*Popis obsahu*
guid	informace o jednotlivých dokumentech LD
auth	informace o jednotlivých autorech publikujících LD
socs	informace o jednotlivých odborných společnostech publikujících LD
link	informace o jednotlivých odkazech na konkrétní umístění textu LD

Tabulka 1: Hlavní databázové tabulky

Důležité pomocné číselníky databáze Katalogu doporučených postupů v ČR:

- číselník klasifikačního systému MKN 10
- číselník nomenklaturního systému MeSH
- číselník kódovacího systému DRG
- číselník akademických titulů autorů LD
- číselník lékařských specializací
- číselník geografického určení pro použití LD
- číselník jazykových mutací LD
- číselník druhu poskytované lékařské péče
- číselník úrovně klasifikace důkazů
- číselník typu doporučení (diagnostika, terapie, prevence apod.)

Tabulka GUID obsahuje pro každý dokument LD následující informace:

- Název LD
- Plný název tištěné podoby LD
- Seznam autorů, kteří se na tvorbě LD podíleli
- Doplňující poznámka k autorům LD (například seznam oponentů, vymezení kompetence autorů a pod.)
- Kontakt na autory dokumentu (např. adresa pro korespondenci)

- Seznam angažovaných odborných lékařských společností, či jiných národních či nadnárodních institucí
- Datum vzniku dokumentu LD
- Datum poslední úpravy dokumentu LD
- Status dokumentu LD ve smyslu jeho aktuálnosti
- Typ LD
- Související kódy klasifikačního systému MKN 10
- Související kódy nomenklaturního systému MeSH včetně indexů udávajících relativní relevanci
- Související kódy systému DRG
- Seznam tématem LD dotčených lékařských specializací
- Seznam specializací, kterým je LD speciálně určeno
- Popis či definice cílové populace
- Cílová geografická oblast pro užití LD
- Seznam odkazů na konkrétní umístění textu LD na Internetu
- Úroveň klasifikace použitých důkazů
- Textová poznámka k dokumentu LD
- Abstrakt dokumentu LD
- Seznam jiných dokumentů LD, které tematicky nebo obsahově souvisejí s dokumentem LD a index udávající hierarchický vztah k původnímu dokumentu LD
- Klíčová slova související s dokumentem LD
- Informace o existenci formalizované verze (dále jen formalizace) dokumentu LD, či volně přístupné aplikaci, která LD zobrazuje, nebo informace a znalosti v LD obsažené používá
- Odkaz na umístění formalizace v Internetu
- Seznam autorů formalizace LD
- Datum formalizace LD, resp. vzniku formalizované verze
- Poznámku k formalizaci dokumentu LD

2.4. Klasifikační systémy

Ke každému dokumentu LD je v Katalogu lékařských doporučení možno přiřadit libovolný počet kódů nejběžněji v ČR používaných klasifikačních a nomenklaturních systémů. Těmito systémy jsou (ve verzi 1.3 Katalogu lékařských doporučených postupů v ČR):

- MKN 10, česká verze desáté revize mezinárodní klasifikace International Classification of Diseases
- MeSH - Medical Subject Headings v českém překladu z roku 2000
- DRG - klasifikační systém Diagnosis-Related Groups

2.5. Základní rozhraní webové aplikace

Rozhraní aplikace pro nepřihlášeného uživatele umožňuje následující funkce:

- nahlížet seznam a detaily informací o zadaných a aktivních dokumentech LD
- vyhledávání v tomto seznamu za použití jednoduchého filtru, ve kterém lze zadat:
  - * hledaný řetězec (prohledává se název LD, klíčová slova, související pojmy z číselníku MKN 10, MeSH a DRG)
  - * související kód MKN 10
  - * související kód MeSH
  - * související kód DRG
  - * rozsah data poslední aktualizace dokumentu LD
  - * status dokumentu LD ve smyslu jeho aktuálnosti
  - * striktní cílová specializace pro dokument LD
  - * odborná společnost, která se podílela na tvorbě dokumentu LD
- procházení seznamu autorů s možností zobrazení detailních informací včetně výpisu dokumentů LD, na jejichž tvorbě se podíleli
- procházení seznamu odborných společností s možností zobrazení detailních informací včetně výpisu dokumentů LD, na jejichž tvorbě se podílely

- procházení stromovou strukturou klasifikačního systému MKN 10 s možností vyhledávání zadáním části názvu a zobrazení seznamu souvisejících záznamů o dokumentech LD
- procházení stromovou strukturou nomenklaturního systému MeSH s možností vyhledávání zadáním části názvu a zobrazení seznamu souvisejících záznamů o dokumentech LD
- procházení stromovou strukturou nomenklaturního systému DRG s možností vyhledávání zadáním části názvu a zobrazení seznamu souvisejících záznamů o dokumentech LD
- seznam odkazů na větší zahraniční i české zdroje LD
- kontaktní informace k projektu
- odeslání krátké textové zprávy editorům/administrátorům projektu
- odeslání návrhu na zařazení nového doporučeného postupu, který dosud není v databázi
- u každého detailního výpisu informací o dokumentu LD odeslání upozornění o chybných nebo chybějících údajích
- prohlížení projektů/tematických celků doporučených postupů

2.6. Administrační rozhraní

Rozhraní aplikace pro přihlášeného uživatele s právy editora umožňuje následující funkce:

- krom zaslání zprávy editorům/administrátorům projektu a zaslání návrhu na zařazení nového dokumentu LD všechny funkce jako základní rozhraní (2.5)
- editaci údajů o dokumentech LD, o autorech a odborných společnostech
- přidávání interních poznámek k dokumentům LD
- procházení seznamu chyb v dokumentech LD hlášených uživateli
- procházení seznamu zpráv od uživatelů
- procházení seznamu návrhů na zařazení nového LD a vytvoření záznamu o dokumentu LD z každého návrhu
- vložení informací o zcela novém dokumentu LD
- vytvoření požadavku na ověření informací o dokumentu LD, procházení seznamem požadavků k ověření a jejich správa
- procházení seznamem odkazů na texty LD umístěných v Internetu
- prohlížení číselníku specializací
- nastavení systémových proměnných
- zobrazení výpisu přístupů na stránky
- procházení seznamem uživatelů a úprava vlastních údajů včetně hesla
- kontrola správnosti odkazů mimo Katalog lékařských doporučení
- vytváření a editace projektů/tematických celků LD
- vytváření a správa seznamu ”oblíbených” dokumentů jednotlivě pro každého registrovaného uživatele

Rozhraní aplikace pro přihlášeného uživatele s právy administrátora umožňuje navíc oproti editorovi následující funkce:

- přidávání uživatelů
- editaci údajů všech registrovaných uživatelů
- prohlížení výpisu událostí v administračním rozhraní

2.7. Zadávání obsahu uživateli a komunikace s autory - systém ověřování informací

Založení nového záznamu o dokumentu do katalogu je možné několika způsoby:

- registrovaný uživatel s právy administrátora nebo editora vytvoří v Administračním rozhraní nový záznam
- návštěvník stránek použije formulář pro zadání návrhu na zařazení dokumentu, následně editor nebo administrátor návrh přijme, či doplní
- návštěvník stránek použije formulář pro zaslání zprávy, například pokud nemá sám dostatek informací o umístění LD v Internetu, následně editor či administrátor na základě této zprávy se pokusí informace dohledat a záznam vytvořit

Pokud editor či administrátor nemá dostatek informací o dokumentu LD nebo tyto informace nepovažuje za věrohodné, vytvoří záznam typu "neověřen", který se nezobrazuje nepřihlášeným uživatelům, a pokusí se informace dohledat. Použít může Odeslání podnětu k ověření, kdy je autorovi dokumentu LD na jeho e-mailovou adresu odeslána zpráva s jedinečným odkazem na speciální ověřovací webové rozhraní Katalogu lékařských doporučení, kde může všechny informace o dokumentu LD upravit a potvrdit jejich správnost. Editorovi (administrátorovi) se pak v administračním rozhraní ověření zobrazí jako potvrzené a on ho může přijmout a informace o dokumentu LD potom publikovat (zobrazit i pro nepřihlášené uživatele).

Také ve chvíli, kdy je uživatelem, editorem či administrátorem hlášena chyba v již zobrazovaných informacích, je možné žádost o ověření údajů autorovi znovu odeslat.

2.8. Projekty a tematické celky

Pro možnost vytváření skupin záznamů o dokumentech LD na základě tematické či jiné souvislosti obsahuje Katalog lékařských doporučení v ČR sekci Projekty/tematické celky, kde může editor či administrátor vytvářet jednotlivé projekty (skupiny) a přidávat, či odebírat provázání s informacemi o jednotlivých dokumentech LD. Nad těmito projekty sdružujícími například mezioborové dokumenty týkající se jedné orgánové soustavy, věkové skupiny obyvatelstva, nebo životní situace lze vést diskusi vkládáním textových poznámek.

3. Diskuse

V současné době je Katalog lékařských doporučených postupů v ČR ve zkušebním provozu přístupný na webové adrese http://neo.euromise.cz/ddp a obsahuje celkem 166 záznamů o dokumentech LD. Pro dlouhodobý provoz je nutné hmotné, personální a odborné zaštítění projektu, které je v jednání, v ideálním případě by se v projektu angažovala některá státní zdravotní autorita, která by mohla garantovat a prosazovat kvalitu poskytovaných informací.

Obsahové doplnění a pravidelná aktualizace katalogu je otázkou výše uvedeného dlouhodobého provozu. Výčet parametrů sledovaných u každého dokumentu LD je v jistém smyslu kompromisem mezi co nejpodrobnějšími informacemi a skutečně autory poskytovanými, či dohledatelnými informacemi. Neexistuje totiž žádná národní norma, kterou by bylo možné použít, výčet parametrů lze ovšem v dalších verzích systému měnit nebo rozšiřovat. Stejně tak lze přidávat další funkce a z Katalogu lékařských doporučení v ČR vytvořit plnohodnotný zdravotnický informační portál, případně systém do již existujícího portálu včlenit.

Katalog lékařských doporučení v ČR může být chápán také jako nástroj pro týmovou spolupráci v oblasti sledování publikační aktivity v oblasti LD, analýzu a vyhledávání dokumentů vhodných pro další zpracování, například formalizaci LD a vyhledávání informací pro vytváření komplexních systémů pro podporu rozhodování ve zdravotnictví.

4. Závěr

Za účelem shromažďování informací o dokumentech LD byl vytvořen systém Katalog lékařských doporučení v ČR ve verzi 1.3, který se skládá z databáze a webové aplikace ve dvou variantách rozhraní - pro nepřihlášené uživatele a pro editory/administrátory. Katalog eviduje údaje o dokumentech lékařských doporučení publikovaných v České republice, o jejich autorech a lékařských společnostech, které je vytvářejí. Katalog v současné verzi 1.3 pracuje ve zkušebním provozu v umístění http://neo.euromise.cz/ddp a obsahuje údaje o 166 dokumentech LD.

Literatura

[1] J.M. Grimshaw, I.T. Russell, "Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations", *Lancet*, 1993 Nov 27;342(8883):1317-22.

[2] D.A. Scalzitti, "Evidence-Based Guidelines: Application to Clinical Practice", *Physical Therapy*, 81 (10), 1622-1628, 2001

[3] K. Filip, T. Sechser, "Doporučené postupy - guidelines - standardy - 3. část", *Remedia*, vol. 15, 4-5, 2005.

[4] Ministerstvo zdravotnictví ČR, "Stand. léčebné péče", http://portalkvality.mzcr.cz/Pages/13-Standardy-lecebne-pece.html

[5] "Guidelines for the Treatment of Malaria", World Health Organization, 2006, ISBN 9241546948

[6] European Society of Cardiology, "Full list of ESC Clinical Practice Guidelines", http://www.escardio.org/guidelines-surveys/esc-guidelines/Pages/GuidelinesList.aspx

[7] "Guidelines České kardiologické společnosti", Česká kardiologická společnost

http://www.kardio-cz.cz/index.php?&desktop=clanky&action=view&id=81

[8] National Institute for Helath and Clinical Excellence, "Published clinical guidelines", http://www.nice.org.uk/Guidance/CG/Published

[9] J.R. Rosalki, S.J. Karp, "Guidance on the Creation of Evidence-Linked Guidelines for COIN", *Clinical Oncology*,11 ,1 ,28 - 32, 1999

[10] "COGS: The Conference On Guideline Standardization", http://gem.med.yale.edu/cogs/

[11] R.N. Shiffman, P. Shekelle, J.M. Overhage, J. Slutsky, J. Grimshaw, A.M. Deshpande, "Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization", *Ann Intern Med*, 2003 Sep 16;139(6):493-8

[12] Scottish Intercollegiate Guidelines Network, "Guideline Development Process", http://www.sign.ac.uk/methodology/index.html

[13] S. Býma, "Metodika CDP-PL pro tvorbu doporučeného postupu (DP)", http://www.svl.cz/default.aspx/cz/spol/svl/default/menu/doporucenepostu/centrumprosprav/metodikacdpplpr

[14] European Society of Cardiology, "Full list of ESC Clinical Practice Guidelines", http://www.escardio.org/guidelines-surveys/ esc-guidelines/Pages/GuidelinesList.aspx

[15] National Guideline Clearinghouse "NGC - Template of Guideline Attributes", http://www.guidelines.gov/submit/template.aspx

[16] "National Library of Guidelines Specialist Library ", http://www.library.nhs.uk/GuidelinesFinder/

[17] "Leitlinien.de", http://www.leitlinien.de/leitlinie

[18] P.R. Wraight, S.M. Lawrence, D.A. Campbell, P.G. Colman "Creation of a multidisciplinary, evidence based, clinical guideline for the assessment, investigation and management of acute diabetes related foot complications", *Diabetic Medicine*, 22 (2), 127-136, 2005

Ústav informatiky AV ČR, v. v. i.
DOKTORANDSKÉ DNY '08

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
Univerzity Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou 248. publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2008

ISBN 978-80-7378-054-8