LOGIC OF DISCOVERY, DATA MINING AND SEMANTIC WEB
Position Paper

Jan Rauch
Faculty of Informatics and Statistics, University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic

Keywords: Logic of discovery, Association rules, Logic of association rules, Data mining, Semantic web.

Abstract: Logic of discovery was developed in 1970's as an answer to questions "Can computers formulate and justify scientific hypotheses?" and "Can they comprehend empirical data and process it rationally, using the apparatus of modern mathematical logic and statistics to try to produce a rational image of the observed empirical world?". Logic of discovery is based on two semantic systems. Observational semantic system corresponds to observational data and statements on observational data. Theoretical semantic system concerns suitable state dependent structures. Both systems are related via inductive inference rules corresponding to statistical approaches. An attempt to modify logic of discovery to data mining was made and a framework making possible to deal with domain knowledge in data mining was developed. Possibility of enhancement of this framework for presenting results of data mining through Semantic web is suggested and discussed.

1 INTRODUCTION

Logic of discovery is developed in book (Hájek and Havránek, 1978) as an answer to questions Q_1, Q_2: (Q_1) – *Can computers formulate and justify* scientific *hypotheses?* (Q_2) – *Can they comprehend empirical data and process it rationally, using the apparatus of modern mathematical logic and statistics to try to produce a rational image of the observed empirical world?* Answers to these questions are based on a scheme of inductive inference:

$$\frac{\text{theoretical assumptions, observational statement}}{\text{theoretical statement}} .$$

Logic of discovery deals with two semantic systems - observational semantic system and theoretical semantic system. Observational semantic system has a language for speaking about observational data. Theoretical semantic system concerns state dependent structures, both systems are connected by inductive inference rules based on statistical approaches.

An attempt to modify logic of discovery for needs of data mining resulted into a suggestion of system *4ft-Discoverer* (Rauch, 2010) which is intended to be an experimental framework making possible to deal with domain knowledge when mining in particular data set. The goal of this paper is to discuss a possibility of enhancement of this framework to serve as a basis for disseminating results of data mining through Semantic web.

System *4ft-Discoverer* is based on logic of association rules (Rauch, 2005). The association rule is understood here as a general relation of two Boolean attributes. Main features of logic of discovery are summarized in section 2. The logic of association rules is introduced in section 3. Important features of the *4ft-Discoverer* are described in section 4. Possibilities to enhance *4ft-Discoverer* to a framework for disseminating results of data mining through Semantic web are discussed in section 5.

2 LOGIC OF DISCOVERY

The schema of inductive inference introduced in section 1 inspired additional five questions (Hájek and Havránek, 1978):
L0: In what languages does one formulate observational and theoretical statements?
L1: What are rational inductive inference rules bridging the gap between observational and theoretical sentences? (What does it mean that a theoretical statement is justified?)
L2: Are there rational methods for deciding whether a theoretical statement is justified (on the basis of given theoretical assumptions and observational statements)?

Rauch J..
LOGIC OF DISCOVERY, DATA MINING AND SEMANTIC WEB - Position Paper.
DOI: 10.5220/0003117203420351
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 342-351
ISBN: 978-989-8425-28-7

L3: What are the conditions for a theoretical statement or a set of theoretical statements to be of interest with respect to the task of scientific cognition?
L4: Are there methods for suggesting such a set of statements which is as interesting (important) as possible?

Answering questions (L0) – (L2) leads to logic of induction, answers to questions (L3) and (L4) lead to logic of suggestion. Answers to questions (L0) – (L4) constitute a logic of discovery developed in (Hájek and Havránek, 1978). The rational inductive inference rules bridging the gap between observational and theoretical sentences are based on statistical approaches, i.e. estimates of various parameters or statistical hypothesis tests are used.

Semantic system is defined to formalize languages for observational and theoretical statements: Semantic system $\mathcal{S} = \langle Sent, \mathtt{M}, V, Val \rangle$ is determined by a non-empty set *Sent* of *sentences*, a non-empty set $\mathtt{M}$ of *models*, a non-empty set V of *abstract values* and an *evaluating function* $Val : (Sent \times \mathtt{M}) \rightarrow V$. If it is $\varphi \in Sent$ and $\mathcal{M} \in \mathtt{M}$ then $Val(\varphi, \mathcal{M})$ is the value of φ in $\mathcal{M}$. Semantic system $\mathcal{S} = \langle Sent, \mathcal{M}, V, Val \rangle$ is *observational* if *Sent*, $\mathcal{M}$, V are recursive sets and *Val* is a partial recursive function.

Two semantic systems – obbservational semantic system $\mathcal{S}^O = \langle Sent^O, \mathtt{M}^O, V^O, Val^O \rangle$ corresponding to analyzed data and theoretical semantic system $\mathcal{S}^T = \langle Sent^T, \mathcal{U}^T, V^T, Val^T \rangle$ corresponding to the whole set of objects we are interested in are developed. The analyzed data can concern only a part of this whole set. Rationality of inductive inference rules is based on statistical approaches. It leads to observational semantic systems with formulas corresponding to statistical hypothesis tests. An example of observational system is related to logical calculus of association rules, see section 3.

3 LOGIC OF ASSOCIATION RULES

The most in (Hájek and Havránek, 1978) studied observational semantic systems are based on *observational predicate calculi* which are introduced in section 3.1. Logical calculi of association rules can be understood as modifications of observational predicate calculi, they are informally defined in section 3.2. Very important are deduction rules in logical calculi of association rules, some practically important deduction rules are mentioned in section 3.3.

3.1 Observational Predicate Calculi

Observational predicate calculus is a result of modifications of classical predicate calculus – only finite models are allowed and generalized quantifiers are added. Finite models correspond to data resulting from observation and generalized quantifiers make it possible to express various assertions on analyzed data including assertions corresponding to statistical hypothesis tests.

Set $Sent^{\mathcal{P}}$ of all closed formulas of observational predicate calculus $\mathcal{P}$ can be used to build observational semantic system $\mathcal{S}^{\mathcal{P}} = \langle Sent^{\mathcal{P}}, \mathtt{M}^{\mathcal{P}}, V^{\mathcal{P}}, Val^{\mathcal{P}} \rangle$ where $\mathtt{M}^{\mathcal{P}}$ is the set of all models (i.e. finite data structures) of $\mathcal{P}$, $V^{\mathcal{P}} = \{0,1\}$ and $Val^{\mathcal{P}}$ is a function assigning a value from $\{0,1\}$ to each couple $\langle \mathcal{M}, \Phi \rangle$ where $\mathcal{M} \in \mathtt{M}^{\mathcal{P}}$ and $\Phi \in Sent^{\mathcal{P}}$. If $Val^{\mathcal{P}}(\mathcal{M}, \Phi) = 1$ then Φ is true in $\mathcal{M}$, otherwise Φ is false in $\mathcal{M}$.

If we use predicate calculus $\mathcal{P}$ with only unary predicates $P_1, \ldots P_n$, then each model $\mathcal{M} \in \mathtt{M}^{\mathcal{P}}$ of $\mathcal{S}^{\mathcal{P}}$ is a $\{0,1\}$ – data matrix with n columns. Expression $\forall(x)P_1(x)$ and $\exists(x)(P_1(x) \vee P_2(y))$ are examples of formulas with classical quantifiers $\forall$ and $\exists$.

Expressions $\Rightarrow^{!}_{p,\alpha,B}(x)(P_1(x), P_2(x))$ and $\Leftrightarrow_{p,B}(x)(P_1(x) \wedge P_3(x), P_2(y) \vee P_4(x))$ are examples of formulas with generalized quantifiers $\Rightarrow^{!}_{p,\alpha,B}$ and $\Leftrightarrow_{p,B}$ which are introduced in table 1. These expressions concern couples of derived predicates $\langle P_1(x); P_2(x) \rangle$ and $\langle P_1(x) \wedge P_3(x); P_2(y) \vee P_4(x) \rangle$, they can be understood as generalization of association rules.

3.2 Logical Calculi of Association Rules

The boom of association rules in the 1990's (Agrawal et al., 1993) was the start of a new effort in the study of association rules as formulas of observational calculi. The syntax of used formulas of predicate observational calculi has been significantly simplified, only calculi with monadic predicates are further studied. Free and bound variables are omitted and basic Boolean attributes are used instead of predicates. Resulting calculi can be understood as logical calculi of association rules (Rauch, 2005; Rauch, 2008; Rauch, 2009).

We are going to informally outline definition of semantic system $\mathcal{AR}^{\mathcal{T}} = \langle Sent^{\mathcal{T}}_{\mathcal{AR}}, \mathtt{M}^{\mathcal{T}}, \{0,1\}, Val^{\mathcal{T}}_{\mathcal{AR}} \rangle$ of type $\mathcal{T}$ concerning association rules. Elements of $Sent^{\mathcal{T}}_{\mathcal{AR}}$ are association rules $\varphi \approx \psi$ where φ and ψ are Boolean attributes derived from columns of analyzed data matrix $\mathcal{M}$ of type $\mathcal{T}$ and $\approx$ is a 4ft-quantifier.

Such association rules are closed formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$ of association rules which is outlined in

section 3.2.1. $\mathtt{M}^{\mathcal{T}}$ is a set of all data matrices of type $\mathcal{T}$, see section 3.2.2. $Val^{\mathcal{T}}_{\mathcal{AR}}$ is an evaluating function assigning value $Val^{\mathcal{T}}_{\mathcal{AR}}(\varphi \approx \psi, \mathcal{M}) \in \{0,1\}$ to each couple $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ and $\varphi \approx \psi \in Sent^{\mathcal{T}}_{\mathcal{AR}}$. It is introduced in section 3.2.3.

3.2.1 Language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$

Association rule is expression $\varphi \approx \psi$ where φ and ψ are Boolean attributes derived from columns of an analyzed data matrix and $\approx$ is a 4ft-quantifier. Boolean attribute φ is called *antecedent* and Boolean attribute ψ is called *succedent*.

Basic Boolean attributes are created first. The basic Boolean attribute is an expression $A(\alpha)$ where $\alpha \subset \{a_1, \ldots a_t\}$ and $\{a_1, \ldots a_t\}$ is the set of all categories of the attribute A. The basic Boolean attribute $A(\alpha)$ is true in row o of $\mathcal{M}$ if it is $A(o) \in \alpha$ where $A(o)$ is the value of the attribute A in row o. Examples of basic Boolean attributes are in figure 1. These Boolean attributes are derived from columns of data matrix $\mathcal{M}$ with columns corresponding to attributes $A_1, \ldots, A_K$.

$\mathcal{M}$	A_1	$\ldots$	A_K	$A_1(1)$	$A_K(2,6)$
o_1	1	$\ldots$	6	1	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
o_n	3	$\ldots$	1	0	0

Figure 1: Data matrix $\mathcal{M}$ and basic Boolean attributes.

Boolean attributes φ and ψ are derived from basic Boolean attributes using connectives $\vee$, $\wedge$ and $\neg$ in the usual way. Expression

$$A_1(1) \wedge A_2(4,5) \approx A_K(2,6)$$

is an example of an association rule.

We consider data matrices with values – natural numbers only. The natural numbers represent categories i.e. possible values of observed attributes $A_1, \ldots, A_K$. Columns of data matrix correspond to attributes and rows correspond to observed objects, e.g. to patients. An example of such a data matrix is in figure 1.

There is only finite number of categories i.e. possible values for each attribute. Let us assume that the number of possible values of a column is t and that the possible values in this column are natural numbers $1, \ldots, t$. All possible values in the data matrix are then described by the numbers of possible values for each column. The whole information on number of columns and possible values in the data matrix is then given by type of data matrix: A *type of data matrix* is a K-tuple $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ where $t_i \geq 2$ are natural numbers for $i = 1, \ldots, K$.

Symbols of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$ of association rules of type $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ are attributes $A_1, \ldots, A_K$, 4ft-quantifiers $\approx_1, \ldots, \approx_Q$, propositional connectives $\wedge, \vee, \neg$ and parentheses. The basic Boolean attributes $A(\alpha)$ are defined in the above given way. Each basic Boolean attribute is a Boolean attribute, if φ and ψ are Boolean attributes, then $\neg\varphi$, $\varphi \wedge \psi$ and $\varphi \vee \psi$ are Boolean attributes.

Set $Sent^{\mathcal{T}}_{\mathcal{AR}}$ of semantic system $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ of association rules of type $\mathcal{T}$ is the set of all association rules i.e. closed formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$. Formal definition of language of association rules is e.g. in (Rauch, 2005).

3.2.2 Data Matrices $\mathtt{M}^{\mathcal{T}}$

A more formal definition of a data matrix with the number of columns and the numbers of possible values in particular columns given by the type $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ is used: Let $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ be the type of data matrix. Then a *data matrix of type* $\mathcal{T}$ is a $K+1$-tuple $\mathcal{M} = \langle M, f_1, \ldots, f_K \rangle$, where M is a non-empty finite set and f_i is the unary function from M to $\{1, \ldots, t_i\}$ for $i = 1, \ldots, K$. Set M is a *set of rows* of data matrix $\mathcal{M}$. Set M is called a *domain* of data matrix $\mathcal{M}$, we write $M = Dom(\mathcal{M})$. An example of data matrix $\mathcal{M} = \langle M, f_1, \ldots, f_K \rangle$ is figure 2. We assume that $M = \{o_1, \ldots, o_n\}$.

object	f_1	$\ldots$	f_K
o_1	$f_1(o_1)$	$\ldots$	$f_K(o_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	$f_1(o_n)$	$\ldots$	$f_K(o_n)$

Figure 2: Data matrix $\mathcal{M} = \langle M, f_1, \ldots, f_K \rangle$.

3.2.3 Evaluation Function $Val^{\mathcal{T}}_{\mathcal{AR}}$

Association rule $\varphi \approx \psi$ can be true or false in given data matrix $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$. Rule $\varphi \approx \psi$ is verified on the basis of *four-fold table* $4ft(\varphi,\psi, \mathcal{M})$ of φ and ψ in $\mathcal{M}$, see figure 3.

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	a	b

Figure 3: 4ft-table 4ft($\varphi, \psi, \mathcal{M}$).

Here a is the number of objects (i.e. rows of $\mathcal{M}$) satisfying both φ and ψ, b is the number of objects

satisfying φ and not satisfying ψ, etc. $4ft(\varphi,\psi,\mathcal{M})$ is also written as $\langle a,b,c,d\rangle$ and called *4ft-table*.

Evaluation function $Val^{\mathcal{T}}_{\mathcal{AR}}$ assigns a value 0 or 1 to each couple $\langle \varphi \approx \psi, \mathcal{M}\rangle$ where $\varphi \approx \psi$ is the association rule and $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$. If $Val^{\mathcal{T}}_{\mathcal{AR}}(\varphi \approx \psi, \mathcal{M}) = 1$ then we say that *rule* $\varphi \approx \psi$ *is true in* $\mathcal{M}$ and if $Val^{\mathcal{T}}_{\mathcal{AR}}(\varphi \approx \psi, \mathcal{M}) = 0$ then we say that *rule* $\varphi \approx \psi$ *is false in* $\mathcal{M}$. $Val^{\mathcal{T}}_{\mathcal{AR}}(\varphi \approx \psi, \mathcal{M})$ is defined using 4ft-table $4ft(\varphi,\psi,\mathcal{M})$ of φ and ψ in $\mathcal{M}$ and associated function $F_{\approx}$ of ≈.

Associated function $F_{\approx}$ *of the 4ft quantifier* ≈ is a $\{0,1\}$ - valued function defined for all quadruples $\langle a,b,c,d\rangle$ of natural numbers. Value of association rule $\varphi \approx \psi$ in data matrix $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ is defined such that $Val^{\mathcal{T}}_{\mathcal{AR}}(\varphi \approx \psi, \mathcal{M}) = F_{\approx}(a,b,c,d)$ where $\langle a,b,c,d\rangle = 4ft(\varphi,\psi,\mathcal{M})$. Examples of 4ft-quantifiers ≈ and associated functions $F_{\approx}(a,b,c,d)$ are in table 1.

Table 1: Examples of 4ft-quantifiers.

≈	$F_{\approx}(a,b,c,d) = 1$ iff
$\Rightarrow_{p,B}$	$\frac{a}{a+b} \geq p \wedge a \geq B$
$\Rightarrow^{!}_{p,\alpha,B}$	$\sum_{i=a}^{r} \binom{r}{i} l^{i}(1-p)^{r-i} \leq \alpha \wedge a \geq B$
$\equiv_{p,B}$	$\frac{a+d}{a+b+c+d} \geq p \wedge a \geq B$
$\approx_{\alpha,B}$	$\sum_{i=a}^{\min(r,k)} \frac{\binom{k}{i}\binom{n-k}{r-i}}{\binom{r}{n}} \leq \alpha \wedge a \geq B$
$\sim^{2}_{\alpha,B}$	$\frac{(ad-bc)^2}{rkls} n \geq \chi^2_{\alpha} \wedge a \geq B$
$\sim^{+}_{q,B}$	$\frac{a}{a+b} \geq (1+q)\frac{a+c}{a+b+c+d} \wedge a \geq B$

The 4ft-quantifiers $\Rightarrow_{p,B}$ of *founded implication*, $\Rightarrow^{!}_{p,\alpha,B}$ of *lower critical implication*, *Fisher's quantifier* $\approx_{\alpha,B}$ and χ^2*–quantifier* $\sim^{2}_{\alpha,B}$ are defined in (Hájek and Havránek, 1978), the quantifier $\equiv_{p,B}$ of *founded equivalence* is defined in (Hájek et al., 1983) and the 4ft-quantifier of *above average dependence* $\sim^{+}_{q,B}$ is defined in (Rauch, 2005).

3.3 Deduction Rules in Logical Calculus of Association Rules

Language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$, set of data matrices $\mathtt{M}^{\mathcal{T}}$ and evaluation function $Val^{\mathcal{T}}_{\mathcal{AR}}$ constitute logical calculus of association rules (Rauch, 2005). There are various theoretically interesting and practically useful results related to logical calculus of association rules. Most of them are related to classes of 4ft-quantifiers (Rauch, 2008).

An example of a class of 4ft-quantifiers is the class of implicational 4ft-quantifiers. 4ft-quantifier ≈ is implicational if $F_{\approx}(a,b,c,d) = 1 \wedge a' \geq a \wedge b' \leq b$ implies $F_{\approx}(a',b',c,d) = 1$. Both 4ft-quantifiers $\Rightarrow_{p,B}$ and $\Rightarrow^{!}_{p,\alpha,B}$ (see table 1) are implicational.

Important results concerning soundness of deduction rules of the form $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ were achieved (Rauch, 2008). Here both $\varphi \approx \psi$ and $\varphi' \approx \psi'$ are association rules. We outline these results for the class of *interesting implicational quantifiers*: If $\Rightarrow^{*}$ is an interesting implicational quantifier then there are formulas ω_{1A}, ω_{1B}, ω_2 of propositional calculus created from φ, ψ, φ′, ψ′ so that the deduction rule $\frac{\varphi \Rightarrow^{*} \psi}{\varphi' \Rightarrow^{*} \psi'}$ is sound if and only if at least one of the following conditions (1), (2) are satisfied: (1) – both ω_{1A} and ω_{1B} are tautologies, (2) – ω_2 is a tautology.

All practically important implicational 4ft-quantifiers are interesting implicational quantifiers. Similar theorems are proved for additional classes of 4ft-quantifiers (Rauch, 2005; Rauch, 2008).

4 4FT-DISCOVERER

4ft-Discoverer $4ft\mathcal{D}^{\mathcal{T}}$ is system $4ft\mathcal{D}^{\mathcal{T}} = \langle \mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}, \mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}, \textit{4ft-Miner}, \textit{4ft-Filter}, \textit{4ft-Synt} \rangle$ where $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ and $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$ are two semantic system intended to be able to express results of observation, properties of particular data matrices and various items of domain knowledge. Here $\mathcal{T}$ is type of data matrix, it is $\mathcal{T} = \langle t_1,\ldots,t_K\rangle$, see section 3.2.1. We say that *4ft-Discoverer* $4ft\mathcal{D}^{\mathcal{T}}$ is of type $\mathcal{T}$.

$\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ and $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$ are briefly described in section 4.1. They are related each other by function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ assigning to each item of domain knowledge a set of its atomic consequences, see section 4.2.

4ft-Miner is a GUHA procedure i.e. data mining procedure which mines for association rules - couples of Boolean attributes created from columns of data matrices $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ (Rauch and Šimůnek, 2005). It has very fine tools to define a set of association rules to be generated and verified. It is introduced in section 4.3. Procedures *4ft-Filter* and *4ft-Synt* are intended to interpret results of *4ft-Miner* using domain knowledge expressed by semantic system the $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$. Both procedures are introduced in section 4.4.

4.1 Semantic Systems $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ and $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$

Semantic system $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ of type $\mathcal{T} = \langle t_1,\ldots,t_K\rangle$ is defined as $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}} = \langle \mathtt{M}^{\mathcal{T}}, Sent^{\mathcal{T}}_{\mathcal{AR}}, Val^{\mathcal{T}}_{\mathcal{AR}}, Sent^{\mathcal{T}}_{\mathtt{M}}, Val^{\mathcal{T}}_{\mathtt{M}}\rangle$, where:

- $\mathtt{M}^{\mathcal{T}}$ is the set of all data matrices $\mathcal{M}$ of type $\mathcal{T}$ see section 3.2.2.

- $Sent^{\mathcal{T}}_{\mathcal{AR}}$ is the set of all association rules $\varphi \approx \psi$ of type $\mathcal{T}$ i.e. the set of all closed formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$, see section 3.2.1.
- $Val^{\mathcal{T}}_{\mathcal{AR}}$ is the evaluation function defined in section 3.2.3.
- $Sent^{\mathcal{T}}_{\mathtt{M}}$ is a set of (closed) formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathtt{M}}$ which is a language intended to express features of particular data matrices. Informal examples of such formulas are: *data matrix* $\mathcal{M}_1 \in \mathtt{M}^{\mathcal{T}}$ *concerns only pathological patients* and *data matrix* $\mathcal{M}_2 \in \mathtt{M}^{\mathcal{T}}$ *concerns patients from a given town* (we assume that data matrices from $\mathtt{M}^{\mathcal{T}}$ concern patients). This language is not defined in details in (Rauch, 2010), more details are in section 5.
- $Val^{\mathcal{T}}_{\mathtt{M}}$ is an evaluation function for features of data matrices, $Val_{\mathtt{M}^{\mathcal{T}}} : (Sent^{\mathcal{T}}_{\mathtt{M}} \times \mathtt{M}^{\mathcal{T}}) \to \{0,1\}$. If $\theta \in Sent^{\mathcal{T}}_{\mathtt{M}}$ and $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ then $Val_{\mathtt{M}^{\mathcal{T}}}(\theta, \mathcal{M})$ is the value of feature θ for $\mathcal{M}$. If $Val_{\mathtt{M}^{\mathcal{T}}}(\theta, \mathcal{M}) = 1$ then $\mathcal{M}$ has feature θ, otherwise $\mathcal{M}$ has not feature θ.

Please note that we use here the notion *semantic system* in a broader sense than defined in (Hájek and Havránek, 1978), the same is true for system $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$ introduced below. We call system $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ *observational* to express that $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ concerns results of observation.

Semantic system $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$ of type $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ is defined as $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}} = \langle \mathtt{U}, Sent^{\mathcal{T}}_{\mathtt{U}}, Cons^{\mathcal{T}}_{\mathcal{AR}} \rangle$ where

- $\mathtt{U} = \bigcup \{Dom(\mathcal{M}) \mid \mathcal{M} \in \mathtt{M}^{\mathcal{T}}\}$ is a union of domains of all data matrices $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$, see also section 3.2.2.
- $Sent^{\mathcal{T}}_{\mathtt{U}}$ is a set of (closed) formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathtt{U}}$ which is a language intended to express various items of knowledge related to set $\mathtt{U}$ or items of general knowledge. Thus each $I \in Sent^{\mathcal{T}}_{\mathtt{U}}$ is an item of knowledge. An example of item of knowledge related to set $\mathtt{U}$ is information on specific vaccination applied to all patients in a given region. We assume that each data matrix $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ concerns only patients from this region. An example of an item of general knowledge is a commonly accepted fact that if weight increases then blood pressure increases too. Examples of formulas from $Sent^{\mathcal{T}}_{\mathtt{U}}$ are given below.
- $Cons^{\mathcal{T}}_{\mathcal{AR}}$ is a function assigning to each $I \in Sent^{\mathcal{T}}_{\mathtt{U}}$ a set of association rules which can be understood as consequences of item I. This function is intended to connect observational semantic system $\mathcal{S}^{\mathcal{T}}_{\mathcal{AR}}$ and theoretical semantic system $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$ by adding semantics to items of domain knowledge. More information is in section 4.2.

System $\mathcal{U}^{\mathcal{T}}_{\mathcal{AR}}$ is called *theoretical* because of it talks about the whole set of objects we are interested in.

Language $\mathcal{L}^{\mathcal{T}}_{\mathtt{U}}$ is intended to express items of knowledge related to set $\mathtt{U}$ or items of general knowledge. Some examples of general knowledge follow. Here A is one of attributes $A_1, \ldots, A_K$ of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$, the same is true for B. In addition, ω, ω_1, ω_2 are Boolean attributes of $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$ and ω does not contain attribute A.

- $A \uparrow\uparrow B$ means that if A increases then B increases
- $A \uparrow\downarrow B$ means that if A increases then B decreases
- $A \rightarrow^{+} \omega$ means that if A increases then relative frequency of ω increases
- $A \rightarrow^{-} \omega$ means that if A increases then relative frequency of ω decreases
- $\omega_1 \rightarrow^{+} \omega_2$ means that if ω_1 is satisfied then relative frequency of ω_2 increases
- $\omega_1 \rightarrow^{-} \omega_2$ means that if ω_1 is satisfied then relative frequency of ω_2 decreases.

We can imagine that there is an additional parameter making possible to express that a formula is opinion of expert X or an assertion from a paper Y.

4.2 $Cons^{\mathcal{T}}_{\mathcal{AR}}$ – Atomic Consequences

Function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ is used instead of the statistical approaches used in (Hájek and Havránek, 1978) to connect observational semantic system $\mathcal{S}^{O}$and theoretical semantic system $\mathcal{S}^{T}$. It is assumed that function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ is defined with help of domain expert. It adds semantics to items of domain knowledge expressed by formulas from $Sent^{\mathcal{T}}_{\mathtt{U}}$.

We show how function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ creates a set $Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\uparrow B, \mathcal{M})$ of association rules – formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$ which can be considered as a set of all atomic consequences of item $A \uparrow\uparrow B$ of knowledge in data matrix $\mathcal{M}$. Function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ can be seen as a family of functions $Cons^{\mathcal{T}}_{\approx}$ where $\approx$ is a 4ft-quantifier of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$. Function $Cons^{\mathcal{T}}_{\approx}$ creates a set $Cons^{\mathcal{T}}_{\approx}(A \uparrow\uparrow B, \mathcal{M})$ of association rules – formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$ such that this set can be considered as a set of all atomic consequences of $A \uparrow\uparrow B$ of the form $\varphi \approx \psi$ in data matrix $\mathcal{M}$. Then $Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\uparrow B, \mathcal{M})$ is defined as a union

$$\bigcup \{Cons^{\mathcal{T}}_{\approx}(A \uparrow\uparrow B, \mathcal{M}) \mid \approx \text{ belongs to } \mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}\} .$$

We outline how function $Cons^{\mathcal{T}}_{\Rightarrow_{p,B}}$ works for 4ft-quantifier $\Rightarrow_{p,B}$ of founded implication (see table 1)

and item $A \uparrow\uparrow B$ of domain knowledge. The functions $Cons^{\mathcal{T}}_{\approx}$ for additional 4ft-quantifiers and formulas of $\mathcal{L}^{\mathcal{T}}_{M}$ are defined using similar principles, see also (Rauch, 2009).

We assume that attribute A has categories $1,\ldots,u$ and attribute B has categories $1,\ldots,v$. Our task is to define a set of rules $\rho \Rightarrow_{p,B} \sigma$ which can be naturally considered as a set of all consequences of item $A \uparrow\uparrow B$ and which are as simple as possible. We assume the simplest rules in form $A(\alpha) \Rightarrow_{p,B} B(\beta)$ where $\alpha \subset \{1,\ldots,u\}$ and $\beta \subset \{1,\ldots,v\}$.

The rule $A(low) \Rightarrow_{p,B} B(low)$ saying *if A is low then B is low* can be understood as a natural consequence of $A \uparrow\uparrow B$. The only problem is to define coefficients α and β which can be understood as *low*. We choose natural A_{low}, $1 < A_{low} < u$ and natural B_{low}, $1 < B_{low} < v$ and then we consider α as *low* if and only if $\alpha \subset \{1,\ldots,A_{low}\}$ and β as *low* if and only if $\beta \subset \{1,\ldots,B_{low}\}$, see also section 4.3.

Also the rule $A(high) \Rightarrow_{p,B} B(high)$ saying that *if A is high then B is high* can be understood as a natural consequence of $A \uparrow\uparrow B$. We choose natural A_{high}, $1 < A_{low} < A_{high} < u$ and natural B_{high}, $1 < B_{low} < B_{high} < v$ and then we consider α as *high* if and only if $\alpha \subset \{A_{high},\ldots,v\}$ and β as *high* if and only if $\beta \subset \{B_{high},\ldots,v\}$.

It remains to define values of parameters p and B of $\Rightarrow_{p,B}$. We can choose each $p \geq 0.9$ and $B \geq \frac{n}{20}$ where n is the number of rows of data matrix $\mathcal{M}$. However, boundaries of p and B as well as values A_{low}, A_{high}, B_{low}, B_{high} should be determined by a domain expert.

Set of all rules $A(low) \Rightarrow_{p,B} B(low)$ and $A(high) \Rightarrow_{p,B} B(high)$ satisfying the above given conditions can be considered as $Cons^{\mathcal{T}}_{\Rightarrow_{p,B}}(A \uparrow\uparrow B, \mathcal{M})$ – a set of atomic consequences of $A \uparrow\uparrow B$ of the form $\rho \Rightarrow_{p,B} \sigma$ in $\mathcal{M}$.

Set $Cons^{\mathcal{T}}_{\Rightarrow_{p,B}}(A \uparrow\uparrow B, \mathcal{M})$ can be defined in a finer way by rules $A(medium) \Rightarrow_{p,B} B(medium)$ with a suitable definition of "medium". Rules $A(low, medium) \Rightarrow_{p,B} B(medium)$, etc. can also be added.

There is a natural requirement on consistency of set $Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\uparrow B, \mathcal{M})$ of atomic consequences, detailed discussion is however without the scope of this paper.

4.3 GUHA Procedure *4ft-Miner*

4ft-Miner procedure mines for association rules of the form $\varphi \approx \psi$ where $\varphi \in \Phi$, $\psi \in \Psi$, and φ and ψ have no common attributes. Input parameters define analyzed data matrix $\mathcal{M}$, 4ft-quantifier $\approx$, set of relevant antecedents Φ and set of relevant succedents Ψ.

Each antecedent is a conjunction $\tau_1 \wedge \cdots \wedge \tau_m$ of *partial antecedents* $\tau_1,\ldots,\tau_m$. Each partial antecedent is either conjunction $\lambda_1 \wedge \cdots \wedge \lambda_q$ or disjunction $\lambda_1 \vee \cdots \vee \lambda_q$ of *literals* $\lambda_1,\ldots,\lambda_q$. Each literal is a basic Boolean attribute $A(\alpha)$ or its negation $\neg A(\alpha)$. Definition of set of relevant antecedents Φ consists of definitions of relevant partial antecedents $\Phi_1,\ldots,\Phi_m$, $\tau_1 \wedge \cdots \wedge \tau_m$ is a relevant antecedent if $\tau_1 \in \Phi_1,\ldots,\tau_m \in \Phi_m$.

Definition of a relevant partial antecedent is given by list $A'_1,\ldots,A'_u$ of attributes, by a minimal and maximal number of literals in particular partial antecedents and by a type of partial antecedent i.e. *conjunctions* or *disjunctions*. In addition, for each attribute A' a set of relevant basic Boolean attributes which are automatically generated is defined. There are various detailed possibilities how to define all relevant basic Boolean attributes $A'(\alpha)$ (Rauch and Šimůnek, 2005). We outline only one of them. We use attribute A with categories 1, 2, 3, 4, 5. Option *intervals of length 2-3* gives basic Boolean attributes $A(1,2)$, $A(2,3)$, $A(3,4)$, $A(4,5)$, $A(1,2,3)$, $A(2,3,4)$, $A(3,4,5)$. This way we can get basic Boolean attributes $A(low)$, $A(high)$, $B(low)$, $B(high)$, see section 4.2.

Set Ψ of relevant succedents is defined analogously. The output of *4ft-Miner* is set Ω of association rules $\varphi \approx \psi$ which are true in $\mathcal{M}$ and both $\varphi \in \Phi$ and $\psi \in \Psi$. The *4ft-Miner* procedure does not use apriori, its implementation is based on representation of analyzed data by suitable strings of bits (Rauch and Šimůnek, 2005).

Let us note that the *4ft-Miner* procedure mines also for conditional association rules $\varphi \approx \psi/\chi$ where φ, ψ and χ are Boolean attributes. The association rule $\varphi \approx \psi/\chi$ is true in data matrix $\mathcal{M}$ if and only if the rule $\varphi \approx \psi$ is true in data matrix $\mathcal{M}/\chi$ where $\mathcal{M}/\chi$ is a data matrix consisting from all rows of $\mathcal{M}$ satisfying χ.

The input of *4ft-Miner* can contain also a definition of set Ξ of relevant conditions in addition to definitions of set of relevant antecedents Φ and set of relevant succedents Ψ. The set Ξ is defined analogously to sets Φ and Ψ.

The output of *4ft-Miner* is then set Ω of conditional association rules $\varphi \approx \psi/\chi$ true in $\mathcal{M}$ which are true in $\mathcal{M}$ and both $\varphi \in \Phi$, $\psi \in \Psi$ and $\chi \in \Xi$.

4.4 Procedures *4ft-Filter* and *4ft-Synt*

The *4ft-Filter* procedure filters out consequences of given item of domain knowledge from the output of *4ft-Miner*. Item of domain knowledge is expressed by

a formula from $Sent^{\mathcal{T}}_{\mathfrak{U}}$. The *4ft-Synt* recognizes groups of patterns which can be considered as a consequence of a (yet unknown) item of knowledge.

Function $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M})$ defined for all formulas $I \in Sent^{\mathcal{T}}_{\mathfrak{U}}$, association rules $\varphi \approx \psi \in Sent^{\mathcal{T}}_{\mathcal{AR}}$ and data matrices $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ can be used to realize 4ft-Filter procedure. It is defined such that $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M}) = 1$ if rule $\varphi \approx \psi$ can be considered as a consequence of I, otherwise $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M}) = 0$.

Value $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M})$ is computed using function $Cons^{\mathcal{T}}_{\mathcal{AR}}$, see section 4.2 and using deduction rules $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$, see section 3.3. There are criteria of correctness of rules $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ for each 4ft-quantifier $\approx$ of 4ft-Miner procedure (Rauch, 2005; Rauch, 2008; Rauch and Šimůnek, 2005). Function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ is defined for all $I \in Sent^{\mathcal{T}}_{\mathfrak{U}}$, and $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ such that $Cons^{\mathcal{T}}_{\mathcal{AR}}(I, \mathcal{M}) = \Lambda$ and Λ is a set of all association rules $\rho \approx \sigma$ which can be considered as atomic consequences of I in $\mathcal{M}$.

Value $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M})$ is computed in two steps. In the first step we compute set $\Lambda = Cons^{\mathcal{T}}_{\mathcal{AR}}(I, \mathcal{M})$. In the second step we test correctness of $\frac{\rho \approx \sigma}{\varphi \approx \psi}$ for each $\rho \approx \sigma \in \Lambda$. If there is such a correct rule, then $\varphi \approx \psi$ is considered as consequence of I in $\mathcal{M}$ and $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M}) = 1$. Otherwise $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M}) = 0$.

Function $Is4ftConsequence(I, \varphi \approx \psi, \mathcal{M})$ can also be used to realize the procedure *4ft-Synt* which recognizes groups of rules $\varphi \approx \psi$ which can be considered as a consequence of a (yet unknown) items of knowledge. We assume that each, even yet unknown, item of knowledge is represented by a formula of $Sent^{\mathcal{T}}_{\mathfrak{U}}$. The procedure *4ft-Synt* can be then realized such that we choose formula $\omega \in Sent^{\mathcal{T}}_{\mathfrak{U}}$ and using function $Is4ftConsequence(\omega, \varphi \approx \psi, \mathcal{M})$ we pick up all consequences of ω from output of 4ft-Miner procedure. However, we have somehow to limit set of tested formulas $\omega \in Sent^{\mathcal{T}}_{\mathfrak{U}}$. A more detailed study of this problem is out of the scope of this paper.

5 4FT-DISCOVERER AND SEMANTIC WEB

One of *10 challenging problems in data mining research* (see http://www.cs.uvm.edu/∼icdm/) is characterized as *mining complex knowledge from complex data*. It is emphasized that all the current data mining systems can do is hand the results back to the user. However, it is necessary to relate results to real world decisions they affect. A way how to do it is to arrange results of data mining into an analytical report structured both according to the analyzed problem and to the user's needs. Core of such a report is a set of assertions on analyzed data together with some explanation comments. Such analytical report can be considered as a formal structure. An idea of indexing such reports by logical formulas corresponding to patterns resulting from data mining is outlined in (Rauch, 1997). It means that such analytical reports are natural candidates for Semantic Web.

Project SEWEBAR concerning these ideas is described in (Rauch and Šimůnek, 2007). It is assumed there are various institutions (e.g. hospitals) storing data in their databases. There are automatically or semi-automatically produced *local analytical reports* giving answers to various *local analytical questions*. It is further assumed that these reports are presented on Internet. It is natural to try to get answers to various *global analytical questions* using these local analytical reports. It is again assumed that answers to *global analytical questions* will be presented on Internet in a form of analytical reports. We call such reports *global analytical reports*. Various aspects of the SEWEBAR project are discussed in (Rauch, 2007; Rauch and Šimůnek, 2009; Kliegr et al., 2009) including formulation of analytical questions using various items of domain knowledge. Some experiments are presented at http://sewebar.vse.cz/.

The SEWEBAR project is based on dealing with analytical reports which are considered as formal structures. No unified formal framework is given to the project till now. The goal of this section is to discuss possibilities of enhancement of the 4ft-Discoverer to serve as a formal framework for the SEWEBAR project. We are going to identify main related problems and to sketch possible ways of their solution.

Overview of currently known main problems related to enhancement of 4ft-Discoverer is given in section 5.1. Possibilities of solution of particular problems are discussed in sections 5.2 – 5.4.

5.1 Enhancing 4ft-Discoverer

The core of 4ft-Discoverer is formal framework for dealing with domain knowledge and association rules (i.e. interesting couples of Boolean attributes related in a given way in a given data matrix). We have formulas expressing items of domain knowledge, procedures *4ft-Miner*, *4ft-Filter*, and *4ft-Synt*, and function $Is4ftConsequence$, see section 4.4.

By these tools we are able to achieve interesting results in solving various local analytical questions related to a given data matrix. Our task is to arrange

these results into a local analytical report such that it will be possible to deal with the report as with a formal object. Some remarks to this problem are in section 5.2.

The current version of 4ft-Discoverer is tailored to analysis of one particular data matrix using domain knowledge expressed by formulas from $Sent^{\mathcal{T}}_{\mathcal{AR}}$, see section 4.1. Only very few attention is given to knowledge related to particular data matrices. This knowledge is assumed to be formalized by $Sent^{\mathcal{T}}_{\mathcal{M}}$ i.e. the set of (closed) formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{M}}$ which is a language intended to express characteristics of particular data matrices, see section 4.1. This requires more attention even when mining in one particular data matrix. Some remarks to knowledge related to particular data matrices are in section 5.3.

Our goal is to get answers to various *global analytical questions* using *local analytical reports* presented on Internet. The answers to *global analytical questions* will be presented as *global analytical reports*. It is further assumed that such global analytical reports will be used as input for answering additional global analytical reports. This approach brings lot of various problems. Initial comments to them are in section 5.4.

Very important is application of classical Semantic web technologies in the SEWEBAR project. In this paper we are not interested in this topic. Let us however emphasize that there are various activities in this directions, see e.g. (Kliegr et al., 2009). The current state is presented at http://sewebar.vse.cz/.

Let us also note that 4ft-Discoverer is tailored to association rules mined by the *4ft-Miner* GUHA procedure. There are six additional GUHA procedures mining for various types of patterns (Hájek et al., 2010). Similar formal framework can be developed for these procedures.

5.2 Local Analytical Reports

An example of local analytical question is the question: *Are there any association rules which can be considered as exceptions from the generally accepted fact $A \uparrow\uparrow B$ in given data matrix $\mathcal{M}$? We assume that the exception concerns a subset of rows defined by attributes $C_1,\ldots,C_L$ – columns of $\mathcal{M}$.* Informally speaking, this task can be solved in following steps:

1. We identify exceptions with conditional association rules $\tau \approx \sigma/\chi$ satisfying
   - $\tau \approx \sigma \in Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\downarrow B)$ i.e. $\tau \approx \sigma$ is an atomic consequence of $A \uparrow\downarrow B$ which is a contradiction to $A \uparrow\uparrow B$.
   - χ is a Boolean attribute derived from attributes $C_1,\ldots,C_L$.
2. We take into account that it is possible that $Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\downarrow B)$ and $Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\uparrow B)$ have common rules. For example it can happen $A(medium) \Rightarrow_{p,B} B(medium) \in Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\downarrow B)$, $A(medium) \Rightarrow_{p,B} B(medium) \in Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\uparrow B)$, see section 4.2.
3. We use *4ft-Miner* with input parameters such that
   - set Φ of relevant antecedents is the set of all τ where $\tau \approx \sigma \in Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\downarrow B)$. It can be done due to the possibility to use option *intervals* for set of all relevant basic Boolean attributes derived from attribute A, see section 4.3.
   - set Ψ of relevant succedents is the set of all σ where $\tau \approx \sigma \in Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\downarrow B)$.
   - set Ξ of relevant conditions is defined as a set of Boolean attributes derived from attributes $C_1,\ldots,C_L$ in a suitable way
   - we use quantifier $\Rightarrow_{p,B}$ with $p = 0.9$ and $B \geq \frac{n}{20}$ where n is the number of rows of data matrix $\mathcal{M}$, see section 4.2.
4. Function $Is4ftConsequence(A \uparrow\uparrow B, \varphi \approx \psi, \mathcal{M})$ (see section 4.4) is used to filter out from Ω all rules $\tau \approx \sigma/\chi$ satisfying $\tau \approx \sigma \in Cons^{\mathcal{T}}_{\mathcal{AR}}(A \uparrow\uparrow B)$.
5. The remaining conditional association rules correspond to searched exceptions.

The above informally described steps can be formalized and automatized. In addition they can be described such that it will be possible to understood this description as a local analytical report answering the given analytical question. This approach differs from that introduced in (Suzuki, 2004).

Such local analytical reports are formal structures and they can be indexed for automatized search. Formulas like $\tau \approx \sigma/\chi$ and $A \uparrow\uparrow B$ can be also used for indexing and searching to deal with semantics. Lot of similar local analytical questions can be formalized and answered by local analytical reports in the above outlined way. Some of them are sketched in (Rauch and Šimůnek, 2009). Detailed elaboration of this topic is a subject of current research.

5.3 Knowledge on Data Matrices

Properties of analyzed data are crucial for analysis and interpretation of results. It is ideal when the data satisfies all requirements for correct application of statistical approaches. However in the case of data mining it is only rare situation. Our goal is to use properties of analyzed data both to formulation and

solution of suitable analytical questions in a similar way the knowledge expressed by formulas of $Sent^{\mathcal{T}}_{\mathtt{U}}$ is used.

Language $\mathcal{L}^{\mathcal{T}}_{\mathtt{M}}$ is intended to express characteristics of particular data matrices and it is assumed to use set $Sent^{\mathcal{T}}_{\mathtt{M}}$ of closed formulas of this language to deal with knowledge on particular data matrices in the same way as formulas of $Sent^{\mathcal{T}}_{\mathtt{U}}$ are used, see section 4.1. It means we have to:

- get formulas of $Sent^{\mathcal{T}}_{\mathtt{M}}$ expressing important properties of data matrices in a similar way the formulas $A \uparrow\uparrow B, A \uparrow\downarrow B, \ldots,$ of $Sent^{\mathcal{T}}_{\mathtt{U}}$ express important items of domain knowledge, see section 4.1.
- define function $ConsM^{\mathcal{T}}_{\mathcal{AR}}$ adding semantics to formulas from $Sent^{\mathcal{T}}_{\mathtt{M}}$, similarly to the way function $Cons^{\mathcal{T}}_{\mathcal{AR}}$ gives semantics to formulas from $Sent^{\mathcal{T}}_{\mathtt{U}}$; $ConsM^{\mathcal{T}}_{\mathcal{AR}}(I,\mathcal{M})$ is a set of association rules – formulas of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$ which can be considered as a set of all atomic consequences of item I of knowledge on data matrix $\mathcal{M}$.
- define function

  $$Is4ftConsequenceM(I,\varphi \approx \psi,\mathcal{M})$$

  for all $I \in Sent^{\mathcal{T}}_{\mathtt{M}}$, association rules $\varphi \approx \psi \in Sent^{\mathcal{T}}_{\mathcal{AR}}$ and data matrices $\mathcal{M} \in \mathtt{M}^{\mathcal{T}}$ such that $Is4ftConsequenceM(I,\varphi \approx \psi,\mathcal{M}) = 1$ if rule $\varphi \approx \psi$ can be considered as a consequence of I in data matrix $\mathcal{M}$ and $Is4ftConsequenceM(I,\varphi \approx \psi,\mathcal{M}) = 0$ otherwise; $Is4ftConsequenceM(I,\varphi \approx \psi,\mathcal{M})$ is analogous to $Is4ftConsequence(I,\varphi \approx \psi,\mathcal{M})$, see section 4.4.

We give a very simple example of a formula from $Sent^{\mathcal{T}}_{\mathtt{M}}$. It is formula $Fr_{\geq 0.9}(A_1(1))$ saying that at least 90 per cent of rows of data matrix satisfy basic Boolean attribute $A_1(1)$. It is $Val_{\mathtt{M}^{\mathcal{T}}}(Fr_{\geq 0.9}(A_1(1)),\mathcal{M}) = 1$ if at least 90 per cent of rows of $\mathcal{M}$ satisfy basic Boolean attribute $A_1(1)$, otherwise it is $Val_{\mathtt{M}^{\mathcal{T}}}(Fr_{\geq 0.9}(A_1(1)),\mathcal{M}) = 0$.

Function $ConsM^{\mathcal{T}}_{\mathcal{AR}}$ can be seen as a family of functions $ConsM^{\mathcal{T}}_{\approx}$ where $\approx$ is a 4ft-quantifier of language $\mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}$, it is analogous to $Cons^{\mathcal{T}}_{\approx}$. Then $ConsM^{\mathcal{T}}_{\mathcal{AR}}(Fr_{\geq 0.9}(A_1(1)),\mathcal{M})$ is defined as a union

$$\bigcup\{ConsM^{\mathcal{T}}_{\approx}(Fr_{\geq 0.9}(A_1(1)),\mathcal{M}) \mid \approx \text{ belongs to } \mathcal{L}^{\mathcal{T}}_{\mathcal{AR}}\}\,.$$

We outline function $ConsM^{\mathcal{T}}_{\Rightarrow_{p,B}}$ for 4ft-quantifier $\Rightarrow_{p,B}$ of founded implication (see table 1). We can define $ConsM^{\mathcal{T}}_{\Rightarrow_{0.9,B}}(Fr_{\geq 0.9}(A_1(1)),\mathcal{M})$ as a set of all rules $\varphi \Rightarrow_{p,B} A_1(1)$ where $0.85 \leq p \leq 0.95$ and $B \geq \frac{n}{20}$ where n is the number of rows of data matrix $\mathcal{M}$. However, boundaries of p and B should be determined by a domain expert.

$Is4ftConsequenceM(Fr_{\geq 0.9}(A_1(1)),\varphi \approx \psi,\mathcal{M})$ is computed in two steps, see also section 4.4. In the first step we compute set $\Lambda = ConsM^{\mathcal{T}}_{\mathcal{AR}}(Fr_{\geq 0.9}(A_1(1)),\mathcal{M})$ of rules $\rho \approx \sigma$ which can be considered as atomic consequences of $Fr_{\geq 0.9}(A_1(1))$ in $\mathcal{M}$. In the second step we test correctness of deduction rule $\frac{\rho \approx \sigma}{\varphi \approx \psi}$ for each $\rho \approx \sigma \in \Lambda$. If there is such a correct rule, then $\varphi \approx \psi$ is considered as consequence of $Fr_{\geq 0.9}(A_1(1))$ in $\mathcal{M}$ and $Is4ftConsequenceM(I,\varphi \approx \psi,\mathcal{M}) = 1$. Otherwise $Is4ftConsequenceM(I,\varphi \approx \psi,\mathcal{M}) = 0$.

Detailed elaboration of the outlined approach is a subject of current research.

5.4 Global Analytical Reports

The goal is to get answers to various *global analytical questions* using *local analytical reports* presented on Internet. The answers to *global analytical questions* will be presented as *global analytical reports*. It is further assumed that such global analytical reports will be used as input for answering additional global analytical reports. It means that the global analytical reports must be again treated as formal objects.

The global analytical questions are formulated on the basis of available local analytical reports. Thus the research of global analytical questions must start with preparing variety of local analytical questions and corresponding analytical reports. An example of local analytical question together with a sketch of its solution by means of *4ft-Discoverer* are in section 5.2. Additional examples of local analytical questions are in (Rauch and Šimůnek, 2009).

Each local analytical question leads to several global analytical question. We denote as LAQ_1 the local analytical question introduced in section 5.2: *Are there any association rules which can be considered as exceptions from the generally accepted fact $A \uparrow\uparrow B$ in given data matrix $\mathcal{M}$? We assume that the exception concerns a subset of objects defined by attributes $C_1,\ldots,C_L$ concerning data matrix $\mathcal{M}$.* Then we can formulate e.g. the following global analytical questions GAQ_1 and GAQ_2:

GAQ_1**:** Which data matrices are similar to the given data matrix $\mathcal{M}$ what concerns solutions of LAQ_1?

GAQ_2**:** Which data matrices differ from the given data matrix $\mathcal{M}$ what concerns solutions of LAQ_1?

Lot of additional global analytical questions can be formulated. The core problem related to solution of such global analytical questions is comparison of results concerning two data matrices $\mathcal{M}_A$ and $\mathcal{M}_B$. There are two possibilities:

1. Both $\mathcal{M}_A$ and $\mathcal{M}_B$ belong to one *4ft-Discoverer* $4ft\mathcal{D}^{\mathcal{T}}$.
2. $\mathcal{M}_A$ belongs to *4ft-Discoverer* $4ft\mathcal{D}^{\mathcal{T}_A}$ and $\mathcal{M}_B$ belongs to *4ft-Discoverer* $4ft\mathcal{D}^{\mathcal{T}_B}$ where $\mathcal{T}_A \neq \mathcal{T}_B$.

There is a research effort to solve problem of comparison of $\mathcal{M}_A$ and $\mathcal{M}_B$ for both possibilities. However its description is out of the scope of this paper.

6 CONCLUSIONS

Logic of discovery was introduced in (Hájek and Havránek, 1978) and modified in (Rauch, 2010). The modification resulted into a system *4ft-Discoverer* $4ft\mathcal{D}^{\mathcal{T}}$ which is a framework for mining association rules and application of domain knowledge in the mining process. We have briefly introduced the *4ft-Discoverer* $4ft\mathcal{D}^{\mathcal{T}}$ and then we have shown that it can be enhanced for needs of the SEWEBAR project which aims to disseminating results of data mining in the form of analytical reports answering reasonable analytical questions.

We have identified several research problems related to this enhancement and outlined possibilities of their solution. Further work concerns solution of these problems.

ACKNOWLEDGEMENTS

This paper was prepared with the support of Institutional funds for support of a long-term development of science and research at the Faculty of Informatics and Statistics of The University of Economics, Prague.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *19 ACM SIGMOD Conf. on the Management of Data, Washington,DC*.

Hájek, P. and Havránek, T. (1978). *Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory*. Springer–Verlag, Berlin Heidellberg New York, 1st edition.

Hájek, P., Havránek, T., and Chytil, M. (1983). *The GUHA Method (in Czech)*. Academia, Praha, 1st edition.

Hájek, P., Holeňa, M., and Rauch, J. (2010). The guha method and its meaning for data mining. *Journal of Computer and System Science*, 76(1):34–48.

Kliegr, T., Ralbovský, M., Svátek, V., Simunek, M., Jirkovský, V., Nemrava, J., and Zemánek, J. (2009). Semantic analytical reports: A framework for post-processing data mining results. In Rauch, J., Ras, Z. W., Berka, P., and Elomaa, T., editors, *ISMIS*, volume 5722 of *Lecture Notes in Computer Science*, pages 88–98. Springer.

Rauch, J. (1997). Logical calculi for knowledge discovery in databases. In Komorowski, J. and Zytkow, J., editors, *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, volume 1263 of *LNAI*, pages 47–57, Berlin. Springer.

Rauch, J. (2005). Logic of association rules. *Applied Intelligence*, 22(1):9–28.

Rauch, J. (2007). Project SEWEBAR considerations on semantic web and data mining. In *IICAI*, pages 1763–1782.

Rauch, J. (2008). Classes of association rules: An overview. In Lin, T. Y., Xie, Y., Wasilewska, A., and Liau, C.-J., editors, *Data Mining: Foundations and Practice*, volume 118 of *Studies in Computational Intelligence*, pages 315–337. Springer.

Rauch, J. (2009). Considerations on logical calculi for dealing with knowledge in data mining. In W., R. Z. and Dardzinska, A., editors, *Advances in Data Management*, volume 118 of *Studies in Computational Intelligence*, pages 177–199. Springer.

Rauch, J. (2010). Considerations on logic of discovery and data mining. In *Suggested for publication at CLA 2010*.

Rauch, J. and Šimůnek, M. (2005). An alternative approach to mining association rules. In Lin, T. Y., Ohsuga, S., Liau, C.-J., Hu, X., and Tsumoto, S., editors, *Foundations of Data Mining and knowledge Discovery*, volume 6 of *Studies in Computational Intelligence*, pages 211–231. Springer.

Rauch, J. and Šimůnek, M. (2007). Semantic web presentation of analytical reports from data mining - preliminary considerations. In *Web Intelligence*, pages 3–7. IEEE Computer Society.

Rauch, J. and Šimůnek, M. (2009). Dealing with background knowledge in the sewebar project. In Berendt, B., Mladenič, D., de Gemmis, M., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., and Železný, F., editors, *Knowledge Discovery Enhanced with Semantic and Social Information*, volume 220 of *Studies in Computational Intelligence*, pages 89–106. Springer.

Suzuki, E. (2004). Discovering interesting exception rules with rule pair. In *In J. Fuernkranz (Ed.), Proceedings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning*, pages 163–178.