Logical Calculi for Knowledge Discovery in Databases

Jan Rauch

Laboratory of Intelligent Systems, Faculty of Informatics and Statistics, University of Economics, W. Churchill Sq. 4, 13067 Prague, Czech Republic

Abstract. Observational calculi were defined in relation to GUHA method of mechanising hypotheses formation. Formulae of observational calculi correspond to statistical hypothesis tests and various further assertions verificated in the process of data analysis. An example of application of the GUHA procedure PC-ASSOC is described in the paper. Logical relation among formulae of observational calculi are discussed and some important results concerning deduction rules are shown. Possibilities of applications of logical properties of formulae corresponding to hypotheses tests in the field of KDD are suggested.

1 Introduction

The goal of this paper is to introduce special logical calculi as a useful tool for Knowledge Discovery in Databases (KDD). We start with the following two facts:

- Each database can be understood as a formally described data structure. We refer to a fact that particular relations and fields have their own names. Results of methods of data mining are assertions dealing with these names. Assertions are in various form, e.g. association rules [1], results of statistical hypotheses tests or presentation graphs. Anyway, each such assertion can be understood as a *formal expression concerning a formal data structure*.
- Mathematical logic studies formal languages and formal data structures as their models. It is defined what does it mean that a sentence of formal language is true/false in a model. A very known example is first-order predicate calculus. There is lot of interesting results concerning universally valid formulas, deduction rules, an axiomatization, a decidability, etc. see e.g. [6].

We are going to argue that some of these logical concepts are or could be useful from the point of view of KDD.

a) Observational calculi were defined and studied in relation to GUHA methods of mechanising hypotheses formation [2]. GUHA is a method of exploratory data analysis and it is also successfully used as a method of KDD [10]. The goal of GUHA method is to offer all interesting facts following from the analysed data to the given problem. GUHA is realised by GUHA-procedures. GUHA-procedure is a computer program, the input of which consists of the

analysed data and a few parameters defining a very large set of potentially interesting hypotheses (usually 10^4 - 10^6). GUHA procedure automatically generates each particular hypothesis from the given set and tests if it is supported by analysed data. The output of the procedure consists of all hypotheses supported by the given data. GUHA deals with hypotheses based on statistical tests (e.g. Fisher's test or Chi-square test) as well as with hypotheses of a different nature (e. g. in the form of an association rule [1]). Logical aspects of GUHA procedures are discussed in Sect. 2.

b) Special deduction rules belong to important features of observational calculi. Deduction rules concern hypotheses generated and verified by GUHA procedures. An example of such a hypothesis is

$$A \wedge B \Longrightarrow^* C \wedge D,$$

where A, B, C and D are basic attributes, $A \wedge B$ and $C \wedge D$ are derived attributes and $\Longrightarrow^*$ corresponds to an implicational relation of derived attributes. Informally speaking, such a deduction rule says that if a hypothesis Φ is supported by the analysed data than also a hypothesis Ψ is supported by these data, a relatively simple condition concerning Φ and Ψ must however be satisfied. It is possible to show, that this condition is the same both for simple association rule and for complicated statistical tests. More information is in Sect. 3.

c) The above mentioned deduction rules are interesting not only from the point of view of GUHA procedures. They could be useful also in the process of interpretation of results of data mining. One of trends in this area is to arrange results into an analytic-synthetical report structured both according to the analysed problem and to the reader's needs. Such a report is possible to understand as a chain of formal expressions concerning formal data structure. In other words the report can be considered as a chain of formulas of an appropriate logical calculus. It opens further possibilities to applications of observational calculi and their logical properties. Some remarks are in the Sect. 4.

2 GUHA Method

There are several implementations of GUHA method see e.g. [4], [7]. The most frequently used GUHA procedure is the procedure ASSOC. Its last implementation is the system PC-GUHA [3] for personal computers. The core of PC-GUHA is the procedure PC-ASSOC. First we show an example of its application and then we will discuss its logical aspects. We show an example concerning truck reliability. It is a modified version of the example given in [10]. It concerns a data matrix shown in Tab. 1.

The data matrix describes warranty failures. Each row corresponds to a failure. Each column corresponds to an attribute describing the failure. The first row describes a failure of the PART 15 (starter), the TYPE of the truck was 1 (lorry), the failure happened in January 1992 (MONTH = 9201). It was the

Table 1. Analysed data matrix

numb.	PART	TYPE	MONTH	COUNTRY	GARAGE
1	15	1	9201	191	19116
2	35	5	9209	427	42701
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
950	24	8	9203	663	66303

failure of the lorry owned by a garage in Prague (GARAGE = 19116) in the Czech Republic (COUNTRY = 191), see also Tab. 2.

Table 2. Number of distinct values for particular attributes

attribute	distinct values	examples of values
PART	40	`15 = starter, 35 = pump, ...`
TYPE	8	`1 = lorry, 2 = TIR, ...`
MONTH	9	9201, 9202, ..., 9209
COUNTRY	17	`191 = CZ, 427 = GE, ...`
GARAGE	53	`19116 = Prague, 42739 = Berlin, ...`

There are 950 failures. Data matrix was analysed in the frame of a pilot study. The analysed data matrix was only a small part of the complete warranty failures set. The goal of the pilot study was to search for all calamities hidden in the data matrix. Occurrence of at least 90% failures of one part under specific circumstances was considered as a calamity. Occurrence of 93% of failures of all pumps at only TIR trucks from the garage in Prague is an example of such a calamity which must be further investigated. However not each such situation is a real problem. If a part is used only in the TIR truck then 100% of failures of this part concern TIR truck. This is not a crisis but only a well known fact.

The only way to find all calamities in the above given sense is to find all occurrences of at least 90% failures of one part under specific circumstances. Each found case must be then judged using expert knowledge. In other words we are searching for all assertions like

$$\text{PART(pump)} \Longrightarrow_{90\%} \text{TYPE(TIR)} \wedge \text{GARAGE(Prague)}$$

true in analysed data matrix. This assertion we can read *"At least 90% of failures of pumps concern TIR trucks from the garage Prague"* . There is large number of

possible interesting assertions. Let us suppose we are searching only for assertions of the form

$$\text{PART}(?) \Rightarrow_{90\%} \text{TYPE}(?) \wedge \text{MONTH}(?) \wedge \text{GARAGE}(?)$$

where "?" could be substitute by a particular value. There is 152 640 of such possible assertions (40 x 8 x 9 x 53, see number of distinct values in Tab. 2). However we are interested in some other assertions too, thus the number of all interesting assertions is greater than 152 640. This is a typical situation when the GUHA procedure PC-ASSOC is useful. We shall formulate our problem as a task for it.

Procedure PC-ASSOC deals with **relevant questions** (potentially interesting hypotheses) of the form

antecedent $\sim$ **succedent**

where **antecedent** and **succedent** are conjunction of properties derived from attributes described by analysed data matrix (e.g. TYPE(8) $\wedge$ PART(15). The symbol $\sim$ is called a **generalised quantifier**. It defines a **kind of dependency of antecedent and succedent** (e.g. $\Rightarrow_{90\%}$ used above or the dependency given by Fisher's test). The relevant question **antecedent** $\sim$ **succedent** corresponds to a question if **antecedent** and **succedent** are in relation given by the generalised quantifier $\sim$.

The PC-ASSOC procedure generates each of the relevant questions from the given set and verifies if the corresponding answer is "yes". In such a case the relevant question is a **relevant assertion** (hypothesis supported by the given data).

The set of the relevant questions is in the case of the procedure PC-ASSOC given by:

- a set of antecedent attributes (attributes to be used in an antecedent) and by minimal and maximal number of attributes in an antecedent,
- a set of succedent attributes (attributes to be used in a succedent) and by minimal and maximal number of attributes in a succedent,
- a generalised quantifier,
- some further facultative possibilities (e.g. syntactical restrictions on antecedent or succedent or a mode of dealing with missing information).

The generalised quantifier could be based on statistical tests (e.g. Fisher's test or Chi-square test) or on some simple numerical condition. The verification of a relevant question **antecedent** $\sim$ **succedent** is based on frequencies from the contingency table Tab. 3.

The frequency a means number of objects satisfying both **antecedent** and **succedent**, b is number of objects satisfying **antecedent** and non satisfying succedent, r is number of objects satisfying **antecedent**, etc.

We shall use a generalised quantifier $\Rightarrow_{5;90\%}$ of founded implication. The relevant question

Table 3. Contingency table of antecedent and succedent

	succedent	¬ succedent	
antecedent	a	b	r
¬ antecedent	c	d	s
	k	l	n

$$\textbf{antecedent} \Rightarrow_{5;90\%} \textbf{succedent}$$

is true in analysed data if the condition

$$\frac{a}{a+b} \leq 0.9 \wedge a \geq 5$$

is satisfied, a and r are frequencies from the contingency table.

Our task is to find all circumstances under which at least 90% of failures of a part are cumulated. We can use the procedure PC-ASSOC with the following parameters:
antecedent: attributes: PART,
minimal number of attributes: 1, maximal number of attributes: 1,
succedent: attributes: TYPE, MONTH, COUNTRY, GARAGE,
minimal number of attributes: 1, maximal number of attributes: 3,
generalised quantifier: $\Rightarrow_{5;90\%}$
further restriction: COUNTRY and GARAGE not in one succedent.

More than 200 000 of relevant questions are given in this way, e.g.:

$$\text{PART}(25) \Rightarrow_{5;90\%} \text{MONTH}(9201) \wedge \text{COUNTRY}(427) ,$$

$$\text{PART}(35) \Rightarrow_{5;90\%} \text{TYPE}(1) \wedge \text{MONTH}(9201) \wedge \text{GARAGE}(42739) .$$

Solution of this task took less than 10 seconds (PC 486, 100 MHz). It was found 89 relevant assertions. Two examples of relevant assertions follow:

$$\text{PART}(\text{mirror}) \Rightarrow_{25;93\%} \text{TYPE}(\text{lorry}) ,$$

it means that 93% of failures of mirror concern lorries and that there are 25 of failures of mirrors at lorries.

$$\text{PART}(\text{starter}) \Rightarrow_{33;100\%} \text{TYPE}(\text{lorry}) \wedge \text{COUNTRY}(\text{Germany}),$$

it means that 100% of failures of starter concern lorries in Germany and that there are 33 of such failures.

Some of relevant assertions were found very interesting from the point of view of quality management. Very important fact is that GUHA gives **all** relevant assertions of the given type. If no relevant assertion is found we can conclude that nothing interesting in a given sense is hidden in the analysed data.

Let us remark that the solution time is approximately linearly dependent on number of rows in analysed data matrix. It means that the solution time of the same task in the data matrix of 1900 rows is about 18 seconds, in the data matrix of 6650 rows is solution time about 61 seconds etc.

We shall now discuss some logical aspects of GUHA procedures. We use the following facts:

1. The relevant questions manipulated by GUHA procedure are formal expressions concerning formal data structures.
2. The output of GUHA procedure can be in some cases unsuitable large. In the above used example this situation can occur when we use the level 80% instead of 90%. In such a case it is reasonable to search a way how to express the output set of relevant assertion in a more comprehensive way.
3. GUHA procedure has to generate and to verify a large number of relevant questions. It is reasonable to search methods how to decrease number of actually generated and verificated relevant questions.

According to point 1 it is not difficult to define a logical calculus such that relevant questions correspond to formulae of these calculus, analysed data matrices correspond to models of calculi, a relevant assertion is a formula true in model etc. Such calculi were defined and further studied in [2]. They are called *observational calculi* . The above mentioned generalised quantifiers are used both as a formalisation of simple relations like $\Rightarrow_{5;90\%}$ and as a formalisation of complex statistical hypotheses test. An informal definition of a simple observational calculus is given in the next paragraph.

One of ways how to decrease number of output relevant assertions (see point 2) is to study logical dependencies among formulae of corresponding logical calculus. It is possible that the output assertion Ψ logically follows from the output assertion Φ. It means that if the formula Φ is true than we know that Ψ is true too without testing Ψ in the analysed data matrix. In some cases it is easy to recognise that the formula Ψ logically follows from the formula Φ. Thus, if Φ is a part of the output and Ψ logically follows from Φ it is possible to omit Ψ from the output.

More sophisticated logical dependencies of this kind are described in [2]. Some special logical dependencies can be used to decrease the number of actually generated and verificated relevant questions, see above given point 3. Useful logical dependencies often are in the form of relatively simple deduction rules. Such deduction rules were studied also in [8]. Several related results are briefly described in the next paragraph. We focus on deduction rules because of they can be used not only in GUHA method but also in interpretation of results of data mining, see paragraph 4.

3 Observational Calculi and Deduction Rules

At first, we indicate a definition of an observational calculus. We will proceed in a very informal way, the formal one is in [2]. Relevant questions have to correspond

to formulae of this calculus. Models of a defined calculus have to correspond to analysed data matrices. Thus it is reasonable to define calculi of different types, each type corresponding to a data matrix type. We will focus on data matrices like that in Tab. 1. Their type we will denote M-FAILURES and we will define a particular observational calculus **FAILURES** of the type M-FAILURES.

Data matrix type M-FAILURES is given by five columns: PART, TYPE, MONTH, COUNTRY, GARAGE and its possible values. Naturally, there is a finite number of possible values for PART, TYPE, COUNTRY and GARAGE. If we consider only the period of years 1990 - 2000 than MONTH has also a finite number of values. Let us suppose that possible values for the column PART are 1, 2, ..., p, values for TYPE are 1, 2, ..., t, values for MONTH are 1, 2, ..., m, etc. We say that the type M-FAILURES is $< p, t, m, c, g >$.

Language of the calculus **FAILURES** is defined in this way:

Basic symbols:

Basic attributes: PART[1], ..., PART[p], ..., GARAGE[1], ..., GARAGE[g]

Propositional connectives: $\wedge$, $\vee$, $\neg$

Generalised quantifier: $\Longrightarrow_{5;90\%}$

Derived attributes:

- Each basic attribute is an attribute.
- If ϕ and ψ are attributes than also $\phi \wedge \psi$, $\phi \vee \psi$ and $\neg\phi$ are attributes.
- Usual conventions concerning parenthesis are valid.

Formulae:

If ϕ and ψ are attributes than $\phi \Longrightarrow_{5;90\%} \psi$ is a formula. The attribute ϕ is here called *the antecedent* and the attribute ψ is called *the succedent.*

Models of the calculus **FAILURES** are all data matrices of the type M - FAILURES. We consider each such data matrix M with n rows to be a result of observation of n objects (failures). We say that i-th object has the basic attribute PART[1] if in i-th row of M is the value 1 in the column PART. We say that i-th object has the derived attribute PART[1] $\wedge$ TYPE[2] if in i-th row of M is both the value 1 in the column PART and the value 2 in the column TYPE. Similarly for other basic and derived attributes. We define a function $Fr(\phi, M)$ where ϕ is an attribute and M is a data matrix of the type M-FAILURES as a number of objects having attribute ϕ.

Values of formulae are defined using associated function $F_{\Longrightarrow_{5;90\%}}$ of a generalised quantifier $\Longrightarrow_{5;90\%}$. It is a $\{0,1\}$ - valued function and it is defined for all quadruples $< a, b, c, d >$ of non-negative integer numbers such that $a + b + c + d > 0$. Usually we write only $\Longrightarrow_{5;90\%} (a, b, c, d)$ instead of $F_{\Longrightarrow_{5;90\%}}(a, b, c, d)$. We define:

$\Longrightarrow_{5;90\%} (a, b, c, d) = 1$ if $a/(a + b)(0.9$ and $a \geq 5$,

$\Longrightarrow_{5;90\%} (a, b, c, d) = 0$ otherwise.

A value $Val(\phi \Longrightarrow_{5;90\%} \psi, M)$ of a formula $\phi \Longrightarrow_{5;90\%} \psi$ in the data matrix M is defined as the value

$$\Longrightarrow_{5;90\%} (Fr(\phi \wedge \psi, M), Fr(\phi \wedge \neg\psi, M), Fr(\neg\phi \wedge \psi, M), Fr(\neg\phi \wedge \neg\psi, M)) \ .$$

Let us emphasise that the frequencies $Fr(\phi \wedge \psi, M)$, $Fr(\phi \wedge \neg\psi, M)$,

$Fr(\neg\phi \wedge \psi, M)$, $Fr(\neg\phi \wedge \neg\psi, M)$ correspond to frequencies a, b, c, d, see contingency table Tab. 3.

The above indicated definition is very informal. The formal definition of observational predicates calculi using predicates, free and bound variables etc. is given in [2]. The main goal of this paragraph is to briefly describe results concerning logical dependencies among formulae of observational calculi like **FAILURES** calculus. We will focus not only on the generalised quantifier $\Longrightarrow_{5;90\%}$.

Let $\mathbb{C}$ be an observational calculus of the type $< p_1, \ldots, p_m >$ with basic attributes

$$P_1[1], \ldots, P_1[p_1], \ldots, P_m[1], \ldots, P_m[p_m]$$

and a generalised quantifier $\sim$. Thus,

$$P_1[1] \wedge P_2[3] \sim P_3[7] \vee P_6[4] \quad \text{and} \quad (P_4[1] \vee P_7[3]) \wedge \neg P_3[7] \sim P_9[4] \vee P_6[4]$$

are examples of formulae. Models of the calculus $\mathbb{C}$ are data matrices with columns $P_1, \ldots, P_m$. Possible values for the column P_1 are 1, ..., p_1, possible values for the column P_m are 1, ..., p_m etc.

We are interesting in logical dependencies among formulae of the calculus $\mathbb{C}$. We focus on the problem when the formula $\gamma \sim \delta$ logically follows from the formula $\phi \sim \psi$. By the general definition $\gamma \sim \delta$ logically follows from $\phi \sim \psi$ if for each model M of the calculus $\mathbb{C}$ holds: If $\phi \sim \psi$ is true in M than also $\gamma \sim \delta$ is true in M (symbolically: if $Val(\phi \sim \psi, M) = 1$, than also $Val(\gamma \sim \delta, M)) = 1$). Let us remember that $Val(\phi \sim \psi, M) = F_\sim(a, b, c, d)$, where $a = Fr(\phi \wedge \psi, M)$, $b = Fr(\phi \wedge \neg\psi, M)$, $c = Fr(\neg\phi \wedge \psi, M)$, $d = Fr(\neg\phi \wedge \neg\psi, M)$. A function $F_\sim(a, b, c, d))$ is the associated function of a generalised quantifier $\sim$. Usually we write only $\sim (a, b, c, d)$ instead of $F_\sim(a, b, c, d))$.

It is obvious that the behaviour of formulae $\phi \sim \psi$ and $\gamma \sim \delta$ depends on the properties of the function $\sim (a, b, c, d)$. There are several interesting classes of generalised quantifiers. We shall deal with **implicational quantifiers** [2]. A generalised quantifier $\Longrightarrow^*$ is implicational if it satisfies the following condition:

Let a, b, c, d, a', b', c', d' be non-negative integers such that $a+b+c+d > 0$ and $a'+b'+c'+d' > 0$. Than
$\Longrightarrow^* (a, b, c, d) = 1$ and $a' \geq a$ and $b' \leq b$ implies $\Longrightarrow^* (a', b', c', d') = 1$.

The above used generalised quantifier $\Longrightarrow_{5;90\%}$ is implicational. In [2] is defined a generalised quantifier $\Longrightarrow^!_{p,\alpha,s}$ of lower critical implication for $0 < p < 1$, $0 < \alpha < 1$ and $s > 0$:
$\Longrightarrow^!_{p,\alpha,s} (a, b, c, d) = 1$ if $\sum_{i=a}^{a+b} \frac{r!}{i!(r-i)!} p^i (1-p)^{r-i} \leq \alpha$ and $a > s$,
$\Longrightarrow^!_{p,\alpha,s} (a, b, c, d) = 0$ otherwise.

Let us remark that the formula $\phi \Longrightarrow^!_{p,\alpha,s} \psi$ corresponds to a test (on the level α) of the null hypothesis $H_0 : P(\psi|\phi) \leq p$ against the alternative one $H_1 : P(\psi|\phi) > p$. Here $P(\psi|\phi)$ is the conditional probability of the validity of ψ under the condition ϕ.

It is proved in [2] that $\Longrightarrow^!_{p,\alpha,s}$ is also implicational. An associational rule is also possible to understand as an implicational generalised quantifier. It is

easy to prove for an implicational quantifier $\Rightarrow^*$ that the value $\Rightarrow^* (a, b, c, d)$ does not depend neither on c nor on d. Thus in such a case we shall write only $\Rightarrow^* (a, b)$ instead of $\Rightarrow^* (a, b, c, d)$.

A theorem concerning the problem when $\gamma \sim \delta$ logically follows from $\phi \sim \psi$ is proved in [8]. It deals with the class of all **interesting implicational quantifiers**. We say that the implicational quantifier $\Rightarrow^*$ is interesting if $\Rightarrow^*$ is both *a-dependent* and *b-dependent* and if $\Rightarrow^* (0, 0) = 0$. A generalised quantifier $\sim$ is *a-dependent* if there are non-negative integers a, a', b, c, d such that $\sim (a, b, c, d) \neq \sim (a', b, c, d)$. Similarly for the *b-dependent* generalised quantifier.

The theorem deals with the notion of **associated propositional formula** to a given attribute. If ϕ is an attribute than associated propositional formula $\pi(\phi)$ is the same string of symbols but the particular basic attributes are understood as the propositional variables. For example: $P_4[1] \wedge P_7[3]$ is a derived attribute and $\pi(P_4[1] \wedge P_7[3])$ is a propositional formula $P_4[1] \wedge P_7[3]$ with propositional variables $P_4[1]$ and $P_7[3]$. Further we shall use the symbol $\rightarrow$ for the propositional connective of implication.

The mentioned theorem says:
If $\Rightarrow^*$ is an interesting implicational quantifier than $\gamma \Rightarrow^* \delta$ logically follows from $\phi \Rightarrow^* \psi$ if and only if at least one of the following conditions a), b) is satisfied:

a) $\pi(\phi) \wedge \pi(\psi) \rightarrow \pi(\gamma) \wedge \pi(\delta)$ and $\pi(\gamma) \wedge \neg\pi(\delta) \rightarrow \pi(\phi) \wedge \neg\pi(\psi)$ are tautologies,
b) $\pi(\phi) \rightarrow \neg\pi(\psi)$ is a tautology.

Let us remark that this theorem gives an easy way how to decide if $\gamma \sim \delta$ logically follows from $\phi \sim \psi$.

There are two useful deduction rules: **Rd** - dereducing deduction rule and **Sp** - despecifying deduction rule [2]. They concern any attributes ϕ, ψ and χ and an interesting implicational quantifier $\Rightarrow^*$. We shall write them in the form:

$$\mathbf{Rd} = \frac{\phi \Rightarrow^* \psi}{\phi \Rightarrow^* \psi \vee \chi} \quad \text{and} \quad \mathbf{Sp} = \frac{\phi \wedge \neg\chi \Rightarrow^* \psi}{\phi \Rightarrow^* \psi \vee \chi} .$$

These deduction rules are direct consequences of the above given theorem. It is proved in [8] that the generalised quantifier $\Rightarrow^!_{p,\alpha,s}$ of lower critical implication is an interesting implicational quantifier for $0 < p < 1$, $0 < \alpha < 1$ and $s > 0$. For example, according to despecifying deduction rule we know that if $P_4[1] \wedge \neg P_3[7] \Rightarrow^!_{p,\alpha,s} P_2[5]$ is true in a data matrix M, than also $P_4[1] \Rightarrow^!_{p,\alpha,s} P_2[5] \vee P_3[7]$ is true in the data matrix M.

The above indicated logical dependencies among formulae of observational calculi are used in the GUHA method. These dependencies could also be used in the way indicated in the next paragraph.

4 Analytic-synthetical Reports

One of trends in interpretation of results of data mining is to arrange results into an analytic-synthetical report structured both according to the analysed

problem and to the readers needs, see e.g. [5] or [9]. An example of a structure of such a report concerning calamities hidden in the above mentioned reliability data (see Sect. 2) follows:

1. Introduction
   ...
2. The 10 worst parts
   ...
3. Critical situations in particular countries
   ...

$$\text{PART(starter)} \Longrightarrow_{33,100\%} \text{TYPE(TIR)} \wedge \text{COUNTRY(Germany)}$$

$$\text{PART(wheel)} \Longrightarrow_{23,90\%} \text{TYPE(lorry)} \wedge \text{COUNTRY(Poland)}$$

   ...
4. Critical situations in particular garages
   ...

$$\text{PART(packing)} \Longrightarrow_{28,90\%} \text{GARAGE(Prague)}$$

   ...
5. Critical situations in particular months
   ...
6. Conclusions
   ...

It is very important that the core of such a report is a set of various relevant assertions concerning analysed data (not necessary in the form of $\phi \Longrightarrow_{s,p\%} \psi$). Thus, we can understand the whole report as a finite ordered set of formulas of an appropriate observational calculus. It means that we can deal the whole report as a formal object.

For example, we can ask about what is a logically minimal skeleton of such a report. Knowledge of logical dependencies among formulae of corresponding observational calculus is necessary to solve this task.

We can use logically minimal skeleton of a report to index content of a report in a similar way as index terms are used in information retrieval to represent the content of a textual document. Unlike index terms in information retrieval, the logically minimal skeleton will describe the content of the report in the very precise way. Let us suppose we have a large set of analytical reports, each of them characterised by logically minimal skeleton. Thus we can formalise and automatically solve e. g. the task to find all reports dealing with a given problem in the same way as a given report. Such a task is not possible to solve using usual index terms.

It is important in relation to this idea that some formulae of observational calculi could be in relatively simple way translated to a sentence of a natural language. Let us have a formula

$$\text{PART(starter)} \Rightarrow_{90\%} \text{TYPE(TIR)} \wedge \text{COUNTRY(Germany)} \ .$$

It can be expressed for example as the following natural language sentences: *"At least 90% of the starter failures happen to the TIR trucks in Germany."* or *"If a starter failure occurs, then at 90% it concerns a TIR truck in Germany."*. We can also ask if there is a way how to convert the natural language sentences to formulae of some observational calculus.

There are some research activities in the field of analytic-synthetical reports and their logical properties. More detailed description is out of the scope of this paper.

This paper is supported by grant 47160008 of Ministry of Education, Youth and Sports and by grant 201/96/1445 of the Grant Agency of the Czech Republic.

References

1. Aggraval, R. et al: Fast Discovery of Association Rules. in Fayyad, U. M. et al.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, 1996. 307–328
2. Hájek, P., Havránek T.: Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Berlin - Heidelberg - New York, Springer-Verlag, 1978, 396 p.
3. Hájek, P., Sochorová, A., Zvárová, J.: GUHA for personal computers. Computational Statistics & Data Analysis **19**, (1995) 149–153
4. Havránek, T.: The present state of the GUHA software. International Journal of Man-Machine Studies, **15**, (1981), 253–264
5. Matheus, J. et al: Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data in Fayyad, U. M. et al.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, 1996, 495–515
6. Mendelson, E.: Introduction to Mathematical Logic. Princeton, D. Van Nostrand Company, Inc., 1964
7. Rauch J.: Some Remarks on Computer Realisations of GUHA Procedures. International Journal of Man-Machine Studies, **10**, (1978) 23–28
8. Rauch, J.: Logical foundations of mechanising hypotheses formation from databases (in Czech). Thesis, Mathematical institute of Czechoslovak Academy of Sciences Prague, 1986, 133 p.
9. Rauch J.: Logical problems of Statistical Data Analysis in Data Bases. in Proceedings of the Eleventh International Seminar on Data Base Management Systems. Budapest, Central Statistical Office, 1988, 53–63
10. Rauch, J.: GUHA as a Data Mining Tool. In: Practical Aspects of Knowledge Management. Schweizer Informatiker Gesellshaft Basel, 1996