Effective Data Mining for a Transportation Information System

P. Haluzová

This paper describes the application of data mining methods in the database of the DORIS transportation information system, currently used by the Prague Public Transit Company. The goal is to create knowledge about the behavior of objects within this information system. Data is analyzed partly with the help of descriptive statistical methods, and partly with the help of association rules, which may discover common combinations of attributes that occur most frequently within a given data set. Two types of quantifiers were used when creating the association rules; namely "founded implication" and "above average". The results of the analysis are presented in the form of graphs and hypotheses.

Keywords: data mining, knowledge discovery in databases, association rule

1 Introduction

Data mining can be defined as the non-trivial extraction of implicit, previously unknown, yet potentially useful information from data, and may be defined as the science of extracting useful information from large data sets or databases. With the help of data mining, derived knowledge, relationships and conclusions are often represented as models or patterns. For example, any data cluster, tree structure, or set of rules, etc., can form a model or a pattern. The whole process is sometimes referred to as knowledge discovery in databases (KDD). Data mining only implies modeling or an analytical method in this application sense, and is considered to be a part of the KDD process [1].

One standard, named CRISP-DM (Cross-Industry Standard Process for Data Mining), describes this process step by step. It develops each phase of the KDD process and, in addition, helps to avoid common mistakes. The particular phases of the CRISP-DM methodology are [1]:

Business understanding – the first objective is to thoroughly understand what is really to be accomplished. We have to begin by uncovering important factors that can influence the outcome of the project. This task involves also more detailed fact-finding about all of the resources, assumptions and other factors that should be considered in determining the data analysis goal.

Data understanding – orientation of data. This step usually investigates a variety of descriptive data characteristics (count of entities in tables, frequency of attribute values, average values, etc.).

Data preparation – this is the most difficult and most time-consuming element of all KDD processes. The goal of data preparation is to choose (or create) relevant data from the available data, and to represent it in a form which is suitable for the analytical methods that are applied (the date often needs to be in the same data table with all other values of the object attributes). Data preparation includes activities like data selection, filtering, transformation, creation, integration and formatting.

Modeling – modeling is the use of analytical methods (algorithms) sometimes referred to as self-data mining. There are many different methods, and the most suitable one must be chosen to solve a given task; efficient settings of the parameters must also be found. This phase includes verifying the quality of the model (e.g. testing in the independent data matrix, cross validation, and so on).

Evaluation – interpretation and evaluation of the discovered knowledge. The main aspects are novelty, interest, utility and comprehensibility of the descriptive tasks. The derived knowledge is divided into the following categories:

- evident knowledge, which is comparable to "horse sense", or to the common knowledge of an expert. Even if such knowledge does not offer anything new, it can show us that the method works well and that it is able to discover knowledge.
- interesting knowledge that yields a new point of view. This is the main aim of KDD.
- knowledge that seems to be unclear or is at variance with expert knowledge. This knowledge may have been created by coincidence, and should be ruled out. Nevertheless, it can expose a new point of view that applies to all the problems, and this must be taken into consideration.

Deployment – the acquired knowledge should be modified into applicable forms, which can involve simply writing a final report or specific actions.

2 Analytic methods

The core of all KDD processes is the use of analytic methods. The input to the analytic procedures is the prepared data, and the output is discovered knowledge. Analytic methods include regression analysis, discriminatory analysis, cluster analysis, decision trees and association rules, among others. It is possible to use standard methods and also modern methods, like neural networks. Data analysis is based on association rules, so they will be described in more detail here.

2.1 Association rules

This method is based on determining of connections (associations) among attributes; in particular, it is used to discover common combinations of attributes that occur most frequently within a given data set. Association rules, which originally came from the Czech GUHA method (General Unary Hypotheses Automaton), can be divided into two categories. The rules fall into the *Ant*~*Suc* form in the first class, and the so-called contingent rules fall into the

Ant~*Suc*/*Cond* form in the second class, where *Ant* (antecedent or precondition), *Suc* (succedent or conclusion, consequent) and *Cond* (condition) are a logical conjunction of literals; and symbol ~ means a generalized quantifier indicating a type of relationship between *Ant* and *Suc*. The literal is defined as an attribute (positive literal); or its negation (a negative literal). A generalized quantifier can be in the form of an implication, an equivalence or a statistical test, etc. Regarding contingent association rules, only objects that satisfy a given condition may be included in the hypothesis [1].

A contingent table for the *Ant*~*Suc* rule and for n instances can be made [1]:

Table 1: Contingent table

	Suc	$\neg$*Suc*	Σ
Ant	a	b	r
$\neg$*Ant*	c	d	s
Σ	k	l	n

where a is the number of rows in the analyzed data matrix that satisfy the concurrent precondition and conclusion, b is the number of instances that satisfy the precondition and not the conclusion, c is the number of instances that do not satisfy the precondition but satisfy the conclusion; and d is the number of instances that satisfy neither the precondition nor the conclusion. We can determine various rule characteristics from these numbers, and with this we can quantitatively evaluate the knowledge that has been found.

The basis of all searching algorithms in association rules is the generation of combinations (conjunctions) of attribute values. They work only with categorical data, and therefore, it is necessary to separate the numerical attributes into intervals.

Quantifiers

The quantifier (generalized quantifier) characterizes the type of relationship between antecedent and succedent. There are many types of quantifiers, e.g. lower critical implication, upper critical implication, above average, below average, founded equivalence, Fisher quantifier, etc. Two types of quantifiers were used to create the hypotheses with the LISp-Miner program; namely "founded implication", and the relationship "above average".

Founded implication

This quantifier has two parameters p and B; p is called confidence and B is base. Always $p \in \langle 0, 1 \rangle$ and $B > 0$. We say that antecedent and succedent have a relation of founded implication with a minimum value of parameters p and B in the data matrix, if the equation that follows is true [1]:

$$\frac{a}{a+b} \geq p \wedge a \geq Base$$

(for a, b see the contingent table – Table 1).

This can be written formally as Ant $\Rightarrow_{p,\,B}$ *Suc* and can be interpreted as: "At least $p \cdot 100\%$ of objects satisfying the antecedent also satisfy the succedent, and at least B objects satisfy both antecedent and succedent". Or "antecedent implies succedent with probability $p \cdot 100\%$".

AA quantifier (above average)

This quantifier also has two parameters; they are confidence p, and base B. Always $p > 0$ and $B > 0$. We say that antecedent and succedent have a relation of above average, with minimum value of parameters p and B in the data matrix, if the equation that follows is true [1]:

$$\frac{a(a+b+c+d)}{(a+b)(a+c)} - 1 \geq p \wedge a \geq Base$$

(for a, b, c, d see the contingent table – Table 1).

This can be written formally as $Ant \approx_{p,\,B} Suc$ and can be interpreted as: "Among objects satisfying the antecedent, there is a relative frequency of objects that satisfy the succedent at least $p \cdot 100\%$ higher than the relative frequency of objects satisfying the succedent among all the objects in the whole data matrix; and there are at least B objects that satisfy both antecedent and succedent." Another way to express this rule is: "If we add one to p, we discover how many times the probability of the succedent increases when the antecedent is satisfied, as compared to a case where we don't know whether the antecedent is or is not satisfied." [2]

The values of the confidence and base parameters are very important for the resultant hypotheses, because the higher the parameter, the more gravity will be assigned to these hypotheses. If we set the values of the parameters too high, only the strongest hypotheses will be found. Their numbers are mostly low, and can even be zero. Consequently, it is necessary to repeatedly set parameters; and to try various values.

2.2 Software for knowledge discovery in databases

Systems for KDD can be divided into the research sphere and the commercial sphere. Most of them include components for modeling, data preparation, visualization and interpretation. Some examples of these commercial systems are Clementine, Enterprise Miner, Intelligent Miner, Knowledge Studio, and the Statistical Data Miner. LISp-Miner and Weka are examples of non-commercial tools.

The LISp-Miner system was used to generate the association rules, which have been under development for research and education purposes since 1996 at the Faculty of Informatics and Statistics of the University of Economics in Prague. This system patterns itself on the GUHA method, and it is possible to closely specify the character of each association rule. In this way, only really interesting hypotheses are found. This brevity of the selection process hastens the generation of hypotheses. LISp-Miner works with the MS Access database, which is appropriate, considering that data used only to be available in this database system.

3 Data understanding and preparation

3.1 The DORIS system

The analyzed data created between April and December 2003 was taken from the DORIS tram dispatch control system. This system observes tram-cars in the electric tramway

Table 2: Attributes of daytime traffic tables

array name	data type	meaning
Rec_id	text	primary key (integer)
Datum	date and time	date and time (in the general form dd.mm.yyyy; hh:mm:ss)
Kod	text	Identification of the type of record (log-in to the system, passage round the infra-pylon, departure from route, etc.)
LnNo	number	line number
PorNo	number	line order
EvCislo	text	tram-car license number
Vozovna	number	number of the depot (1–8), from which the tram departed
A	number	ID of tram stop
B	number	number of the stop post
C	number	delay (overtake recorded as minus)
D	number	reserve column, without data

net of the Prague Public Transit Company, and supports tram-car control in real time. DORIS ensures basic functions such as localization of all tram-cars logged into the system; evaluation of the time divergence of tram-car as the pass the tram stops; information about the departure of tram-cars compared with the timetable; control of data and phonic traffic; control of departure of tram-cars from the terminus; unified digital information about exact times, etc. [3].

The data is archived by months; three databases were available for each month: the database of daytime traffic; the database of night traffic; and the database containing emergency records (traffic accidents). The database of daytime traffic includes a table of daily records for each day of the month. There are approximately 50 000 records daily, and each record has 11 attributes (see Table 2).

3.2 Data preparation

It is necessary to integrate data from several tables to create a single database that includes all required data. Furthermore, duplicate records have to be deleted, because the existence of such records has a negative influence on the statistics; the greater the duplication, the greater the distortion. As well as duplicate records, extreme values can influence the resultant statistics and the computation of new attributes (average values). Records with extreme values of delay/overtake attributes were therefore deleted. Such values occur due to so-called “reduced operating effectiveness” events (when a tram is diverted from its route, towed to the depot because of technical problems or an accident, or is being handled in the depot). If a data table, which usually has 1.5 million records

Table 3: Meaning and categories of attributes

name of attribute	Meaning	name of category	values of attributes
weekday	days in a week	Monday to Sunday	1 to 7
hour	hour interval (e.g. 7 means an interval from 7:00 to 7:59)	morning	5, 6
		morning rush hour	7, 8
		before noon	9, 10
		noon	11, 12
		afternoon	13, 14
		afternoon rush hour	15, 16
		evening	17, 18, 19, 20
		late evening	21, 22, 23
average delay	average delay [s]	under 1 min	$\langle 0, 59\rangle$
		from 1 to 2 min	$\langle 60, 119\rangle$
		from 2 to 3 min	$\langle 120, 179\rangle$
		over 3 min	> 179

for each month, generates about 20 000 duplicate records, this corresponds to 1.3 % of the total.

If a dispatcher diverts a tram from its route, there is no applicable delay/overtake attribute for this tram, because there is no defined timetable on its "new" route, and therefore, it is not possible to determine a departure. So a 'NULL' value arises in these attributes; records with such values were also deleted in the data preparation phase. Missing values can, in general, be handled in various ways.

Because attributes like 'day of the week' or 'hour intervals' were used, it was necessary to create these from the 'date' attribute.

3.3 Creation of attributes for generating of association rules

Association rules are generated and hypotheses are created with groups of attributes, and must therefore be collected in a simple data table, mostly with the help of the SQL language aggregation function. Relevant categories must be created for the values of each attribute.

An authentic group of attributes is offered here as an example. The attribute categories: 'weekday', 'hour', 'line', 'stop' and 'average delay' are shown in the data table for the generation of hypotheses. The meanings of these attributes and their division into categories in Table 3.

The tram direction should be reflected in the survey of the delays at the stops; whether or not it is inbound towards, or outbound from the city center (this can be distinguished with the help of the 'stop post' attribute in the database). The attributes 'line' and 'stop' have one category for each value. The resultant data table for generating the association rules includes about 30 000 rows.

4 Results

We will now give some real examples and show how they have been modeled.

4.1 Description

We can obtain information from the database, which allows us to follow any leg of a selected line, either in the long term, or in one day, or in hourly segments. The "long term" pattern has a higher deposed value, and can be used for modifying timetables for adjusting the driving time between stops or as a basis for drafting new timetables.

The statistics of accidents divulge much interesting statistical information. About 800 accidents were recorded in the database between July and December 2003. With the help of simple a SQL query it was possible to determine that the most common type of accident was between a tram and a car, without resultant injury. Injury occurred in only 4.4 % of all such accidents.

Fig. 1 shows that the greatest number of accidents occurred on Monday and on Wednesday. The greatest number of accidents during the day may be assumed to have occurred during the rush hour. This is confirmed by the second graph (Fig. 2).

The greatest number of accidents occurred between 3 and 4 p.m. On the other hand, the total number of accidents in the evening, and at night, is below twenty per hour, appreciably fewer than in other hours of the day.

Another example that illustrates the use of data mining in the DORIS database is the influence of one accident on increasing the delays to other trams in the affected area. An accident has been chosen from the database that happened to the second tram in the line order for line #18, between Chotkovy sady and Malostranská tram stops. The tram struck a pedestrian, and the tram was then taken to the depot. This accident influenced the progress of trams not only on line 18, but also on lines 22 and 23. Fig. 3 describes the trams operating on lines 22 and 23, and the increase in delays to them at the time of the accident. Each tram is affixed with a line number and an ordinal number. At first sight, the ordinal numbers of both lines appear to be numbered similarly in the graph, but in reality they are slightly different.

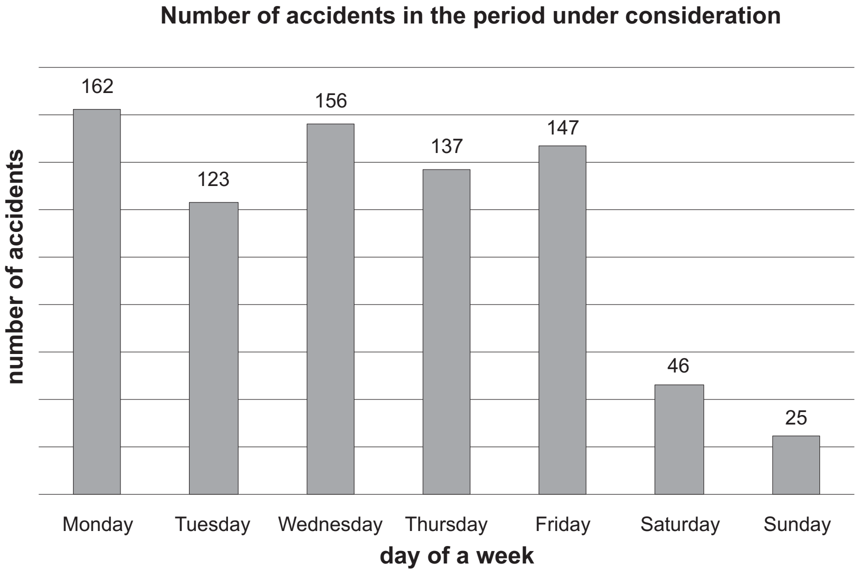

Fig. 1: Number of accidents in the study period

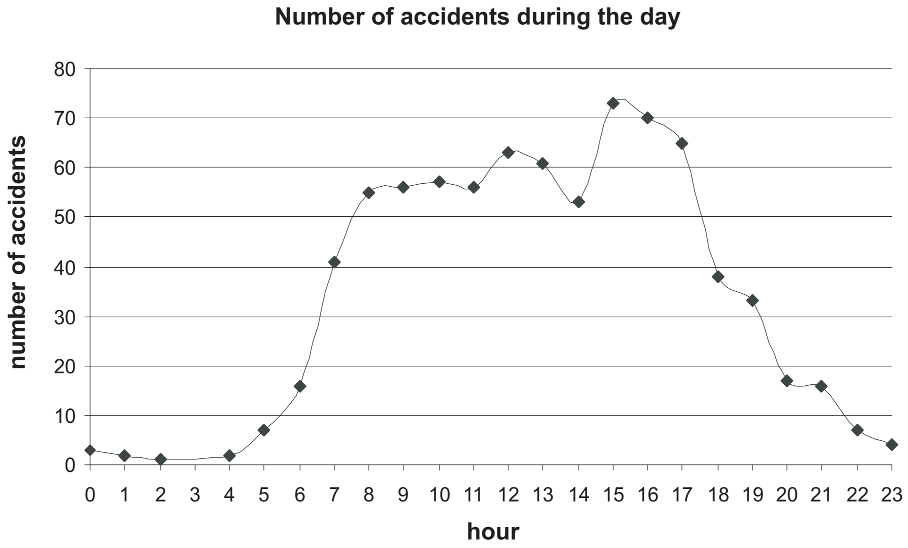

Fig. 2: The number of accidents during the day in the study period

As shown in Fig. 3, the first trams on both lines ran without any major delay. The accident occured between the passage of trams 23/1 and 22/2. The delays increased rapidly after this event. The trams most impacted were 22/2, 23/2, 22/3 and 23/3, and the delay reached as much as 700 s. Afterwards, the flow of traffic slowly returned to the normal state, but several following trams were still delayed by about 200 s.

4.2 *Searching for association rules*

The attributes 'weekday', 'hour' and 'number of accidents' were placed as a precondition to an association rule; and the attributes 'percentage of delayed trams', 'average delay', 'average overtake' were assigned to a succedent. The founded implication with the parameters $B = 15$ and $p = 0{,}700$ was used as a quantifier. These hypotheses were then searched: among those that have the relationship of the founded implication between *Ant* and *Suc*, at least 15 objects satisfy both *Ant* and *Suc* and confidence of the implication is at least 0,700. With this setting of parameters, for example, a strong hypothesis appeared as follows:

Number of accidents (high) $\Rightarrow_{0.89,\ 23}$ percentage of delayed trams (high), where the confidence is 89 % and the

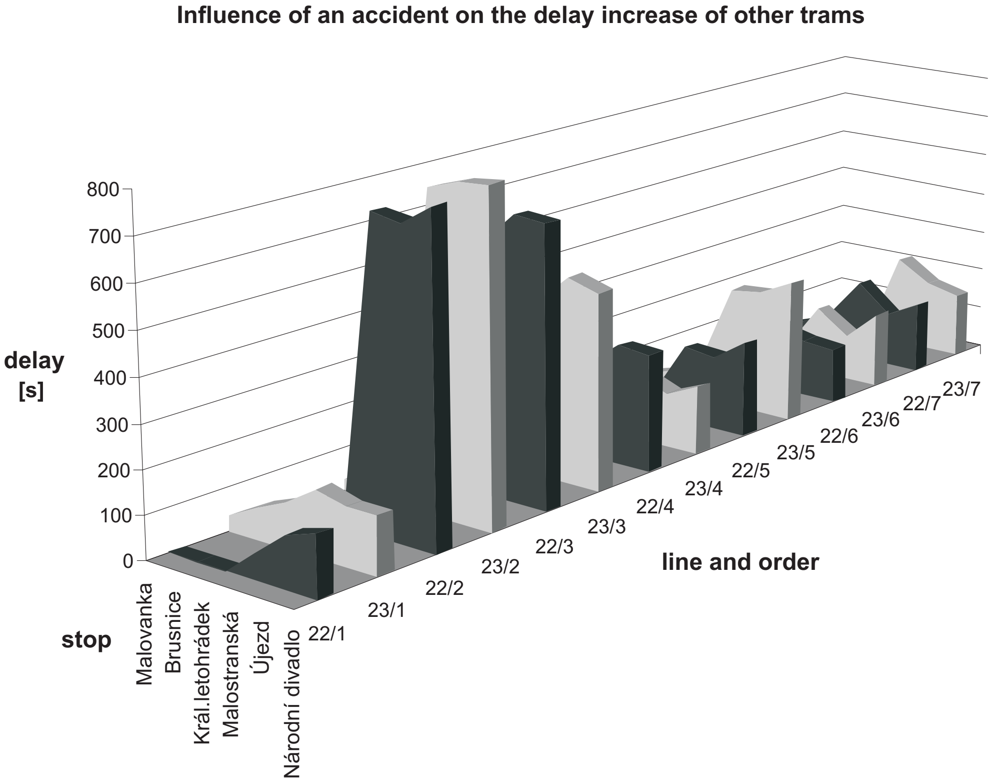

Fig. 3: The influence of an accident on increasing the delays on of lines 22 and 23

support $a = 23$ (number of objects which satisfy the concurrent precondition and conclusion). The hypothesis can be interpreted as: if a high number of accidents occurs in a given day and hour then a high percentage of trams delayed more than 180 seconds will appear with a probability rate of 89 % in the same day and hour. This is confirmed by the study of 23 entities in the data matrix. Examples of other hypotheses:

- Number of accidents (zero) $\Rightarrow_{0.79,\,15}$ percentage of delayed trams (low) & average delay (low)
- Number of accidents (low) $\Rightarrow_{0.82,\,44}$ average delay (low)

From the hypotheses mentioned above it can be unambiguously deduced that the number of accidents influences the delay rate. Such a conclusion can be intuitively assumed, and we can draw this conclusion without the use of association rules. No new elements appear, but we know that the method used for generating association rules works well.

Other searches for contingent rules followed. We used the attributes 'line', 'weekday,' 'hour' in the antecedent, and 'average delay' in the succedent. The 'stop' attribute (the direction towards city center) was used as a condition. There was an "above average" quantifier with the parameters $p = 2$ and $B = 20$. Thirty hypotheses satisfied the given conditions. The attribute 'stop' constrains the validity of founded rules at the individual stops. These founded hypotheses reveal:

Line (25) $\approx_{3,\,20}$ average delay (over 3 min)/stop (Vltavská)

This hypothesis holds true for Vltavská tram stop, in the direction towards the city center, for 20 instances in the data matrix, and p is approximately 3. This is interpreted as: there is a delay of over 3 minutes on line 25, four times more often at Vltavská tram stop than for trams of all other numbered lines running through this tram stop. The rules can be interpreted in various ways. While intelligibility and simplicity should be the most important elements in the explication, it is sometimes difficult to explain a hypothesis simply and realistically in the same instance. A given hypothesis can be simply interpreted as: trams of line 25 are those that are most often delayed at Vltavská tram stop.

As in the case of generating association rules from the previous attributes group, evident or already known conclusions appeared. As an example:

Hour (late evening) $\approx_{2,\,64}$ average delay (under 1 min)/stop (Štěpánská). Simply stated: In comparison to any other hour of the day at Stěpánská tram stop, the smallest number of delayed trams occur between 21:00 and 24:00.

5 Conclusions

The results of data analysis can serve for modifying the running times between stops, or as a basis for drafting of new timetables. They also draw attention to the existence of black spots with a high rate of accidents or a high percentage of delayed trams (or tram overtake incidence). The conclusions drawn from such analyses provide an exact tool for recognizing black spots, and suitable measures can be taken, e.g., decisions about tram preference at traffic lights or about constructing some suitable passive elements.

As a result of applying the association rules, hypotheses appeared which in most cases could have been intuitively assumed. Basically, it can be inferred that the method used here for searching for association rules in the LISp-Miner program works well. No surprising or new hypotheses were found, above all due to the interdependence among the attributes that were used (for example, there is a correlation coefficient of 0.89 for the pair of attributes 'number of accidents' and 'average delay'). It would be necessary to variegate the data base with other independent attributes in order to find more interesting hypotheses. For example, vehicle occupancy rate could be used as another new attribute to record at the stops, if the required equipment for reading the weight on an axle were available. If the vehicle occupancy were known, it would be possible to optimize the scheduled intervals between trams, to reduce or increase the frequency of connections.

The most interesting hypotheses are those predicting the probability of a delay on a line at a selected stop. Not such statistics had been created previously at the Prague Public Transit Company. When considering transport options, reliability of transfer connections and information about the probability of a delay are decisive elements when choosing a mode of transport.

References

[1] Berka, P.: *Dobývání znalostí z databází*. Praha, Academia, 2003.

[2] Kejkula, M.: *4ft-Miner pro začátečníky. Získávání znalostí z databází*. Praha, VŠE, 2004.

[3] Internal materials of the Prague Public Transit Company: *Dispečerský řídicí systém DORIS*.

[4] *The official site of the LISp-Miner project*, http://lispminer.vse.cz

[5] Hájek, P., Havránek, T.: *Metoda GUHA. Automatická tvorba hypotéz*. Praha, Academia, 1983.

[6] Ducháček, M.: *Nástroj pro správu databází s využitím pro multi-relační data mining*. Diploma thesis MFF UK, Praha, 2005.

[7] Burian, J: Datamining a AA (Above Average) kvantifikátor. *Sborník 2. ročníku konference Znalosti*. Ostrava, 2003. ISBN 80-248-0229-5.

[8] Rauch, J., Šimůnek, M.: Systém LISp-Miner. *Sborník 2. ročníku konference Znalosti*. Ostrava, 2003. ISBN 80-248-0229-5.

Ing. Petra Haluzová
email: p.haluzova@seznam.cz

Department of Informatics and Telecommunications

Czech Technical University in Prague
Faculty of Transportation Sciences
Konviktská 20
110 00 Prague 1, Czech Republic