Utilization of Selected Data Mining Methods for Communication Network Analysis

Zuzana VRANOVA, Vojtech ONDRYHAL

Communication and Information Systems Department, University of Defense, Brno, Czech Republic

zuzana.vranova@unob.cz, vojtech.ondryhal@unob.cz

Abstract. *The aim of the project was to analyze the behavior of military communication networks based on work with real data collected continuously since 2005.*

With regard to the nature and amount of the data, data mining methods were selected for the purpose of analyses and experiments. The quality of real data is often insufficient for an immediate analysis. The article presents the data cleaning operations which have been carried out with the aim to improve the input data sample to obtain reliable models.

Gradually, by means of properly chosen SW, network models were developed to verify generally valid patterns of network behavior as a bulk service.

Furthermore, unlike the commercially available communication networks simulators, the models designed allowed us to capture nonstandard models of network behavior under an increased load, verify the correct sizing of the network to the increased load, and thus test its reliability. Finally, based on previous experience, the models enabled us to predict emergency situations with a reasonable accuracy.

Keywords

Communication network (CN), data mining (DM), data preparation, decision trees.

1. Introduction

Within the research the methods of knowledge discovery in databases (data mining) for behavior analysis of the real communication network have been applied. The customized data mining methodology has been used to develop models for several individual tasks we have recognized in communication network behavior. So far, we have mastered the first two phases of the methodology, the problem analysis and the data analysis and data understanding [1]. Now, we are planning to create and exploit the data mining models of data preparation and data modeling phases. We would like to interpret and apply these models to individual tasks, such as:

- Classification and prediction of selected network parameters like short-term and long-term traffic load.
- Standardized network behavior for individual network lines in selected intervals.
- Comparison of traffic-load on different days of the week.
- Discovering trends in network behavior and its usability for prediction improvement.
- Proper network dimension verification.
- Finding groups of line with similar characteristics using segmentation algorithm.
- Extreme values detection.
- Prediction of the increase in traffic load due to irregular users' activities. This capability will distinguish our models from common commercial simulators.

The already developed models will be described in the paper and the influence of entering data quality on possible model interpretation will be discussed. For modeling the database (SQL language and Relational Database Management System) and data mining tools were used. SPSS PASW Modeler and Microsoft Access are used for experiments.

A similar problem of communication networks behavior analysis through an analysis of operational load addresses many works, such as [2], [3], [4], [5].

All of the mentioned publications work with other types of networks in terms of their principles of work; paper [2] analyzes the Internet. Similarly to our contribution, the paper focuses on the dimensioning of the network. As the evaluation parameter the percentage distribution of the traffic load was selected; the authors deal with 4 lines for the period less than 6 months. Our project verifies the dimensioning of the data on backbone network (22 lines), collected continuously since 2005. As an assessment option chosen in our project is the peak busy hour (PBH) that can express not only the intensity of losses, but also distribution of maximal values in the network over time. As in our paper the authors address the increasing trend in the load on the network. The analysis is performed only on data from four lines observed for a period of about 6 months, in our opinion, data are not statistically significant enough for such type of task.

Paper [3] deals with the type of WLAN networks, publication [4] deals with Private Mobile Radio Network.

Work [5] performs an analysis of network traffic using analytical modeling techniques. They create the appropriate models based on the same simplifying assumptions as we do; it means the stationarity of processes examined at selected intervals in mean and variance. The analyzed network is a complex queuing system, which cannot be easily described mathematically, even not for a suitably chosen initial conditions and simplifying assumptions. Therefore, we have decided to apply computer simulations in our work.

Our network differs from the others according to the theory of queuing. We analyzed the network that operates as a system with losses, works [2], [3], [4] and [5] dealt with the queuing systems.

Furthermore, unlike the work of mentioned papers and unlike the commercially available communication networks simulators, the selected methods and models allow us to detect nonstandard behavior of the network in focus and moreover the models are able to predict such situation with sufficient accuracy.

The organization of this paper is as follows. In section 2 the analyzed network is described in terms of the network topology, the technology and the principles of work. In addition, the parameters chosen for evaluation for analysis of the network behavior are defined with the justification of their choice. In section 3 we address the process of data preparation for data mining. Individual used operations and their importance for the improvement of input data are briefly characterized. As a final output of this part the analytical table with well defined inputs and outputs for modeling is created.

In section 4 we describe analysis tasks processed on the network, created models and simulation results are discussed here as well as the results of simulations and the resulting recommendations for network optimization.

In section 5 we give our conclusions.

2. Communication Network in Focus

If we want to model the behavior of a system in focus precisely, we have to be acquainted with its structure, relations of its elements, processes inside the system and their time sequence. There is a short characteristic of the communication infrastructure of the analyzed Stationary Military Communication Network (SMCN) [6].

2.1 The Topology of the SMCN

The communication infrastructure of the SMCN has been built on the basis of modern digital principles since the beginning of 1990s in accordance with world's trends and civilian and military standards. The principles of work are TDM (Time Division Multiplex) and PCM (Pulse Code Modulation), the protocols used are ISDN (Integrated Services Digital Network). The SMCN is mainly designed for voice transfer.

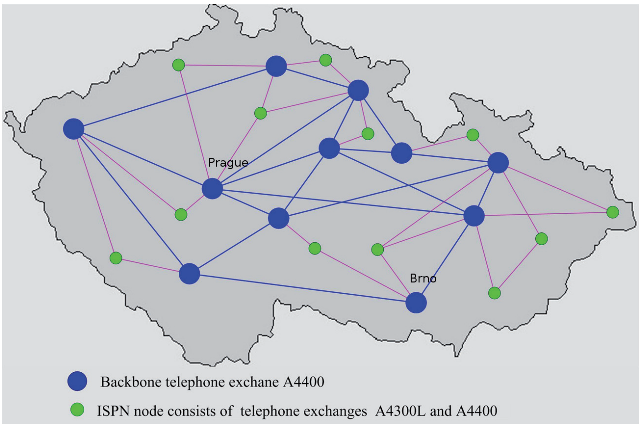

Fig. 1. A fictitious topology of the SMCN.

The SMCN consists of the modern digital exchange A4400 and A4300L designed by the French firm Alcatel. They are connected by E1 flows with the business signalization ABC, version F1 and F2. The backbone consists of the digital exchange A4400 only. ISPN node is created by twenty PBX (Private Branch Exchange) A4300L or A4400 maximally. These are connected by a special ISPN signaling protocol [7]. That is why one ISPN node appears to other network as only one big PBX. The ISPN nodes are connected by two E1 flows at least to the backbone [8].

We are processing real data acquired in NMC (Network Management Control) supervisory centers. During the first processing we limit ourselves to a Backbone Network only. In Fig. 1, the fictitious topology of the network is presented. The real network nodes and connections differ in numbers and location from the ones in the figure.

2.2 Network Variables

We have selected permeability as a basic evaluation criterion for SMCN behavior and it can be expressed as a function of the traffic load in the individual backbone trunks. The traffic load depends on the type of the system, kind and range of services provided, users' activity, period of the day and night, days, weeks and months in a year, working hours, etc. Generally, the knowledge of the traffic load enables us to examine the use of connecting lines and their real need, to assess the quality of the telecommunication system operation, to compare the system quality with other telecommunication systems, to compare and harmonize all the parts of the connecting system.

The traffic load is a function of many variables. The most frequently monitored traffic load values follow:

The intensity of the carried traffic load, Y, can be expressed as:

$$Y = \frac{1}{T}\int_{t_o}^{t_o - T} N_x dt \qquad \text{[erl]} \qquad (1)$$

where T is observation time, t_o defines observation starting time and N_x is the total of used connecting lines, or the intensity of the carried traffic load, Y, can be expressed as:

$$Y = \frac{C_r t_{os}}{T} \qquad \text{[erl]} \qquad (2)$$

where C_r is a number of calls carried out at time T and t_{os} is a mean time of line occupation by one call.

The intensity of the offered traffic load, A, can be expressed as:

$$A = \frac{1}{T}\sum_{i=1}^{m} t_{oi} \qquad \text{[erl]} \qquad (3)$$

where m is a number of calls offered at time T, t_{oi} is time of the call i.

The range of the processed traffic load, O_Y, can be expressed as:

$$O_Y = C_r t_{os} = YT\,. \qquad \text{[erlh]} \qquad (4)$$

The loss, Z, can be expressed as:

$$Z = \frac{C_Z}{C} \qquad \text{[-]} \qquad (5)$$

where C_Z is a total number of calls rejected during time T and C is a total number of processed calls.

Intensity of the carried traffic load Y, as a parameter chosen for the next analysis, expresses utilization of traffic lines. It is defined as a mean value of used traffic lines in a bundle. We use the expression (2) in our research.

The NMC enables to monitor data and voice flows in the long term, and store information essential for traffic loads evaluation. The monitoring can be carried out in half-an-hour intervals; the length of monitoring could be set (working days from-to, all days in a week, 24-hour monitoring, non-stop). The data on traffic load have been collected on our request and specification since the end of 2005 and are being continuously updated [5]. In the time being, we are concentrating on the data that are suitable for the carried traffic load calculation.

3. Data Preparation Process

The main output of the data preparation process is a so called analytical table, which contains quality data for modeling.

3.1 Original Data Description

Since the actual outcomes of the NMC system are in the form of confused text files and contain a large amount of redundant information, the data were transformed into form more suitable for further processing. For the purpose of transformation, the program in C++ was created.

```
Tuesday, April, 1, 2008
Comment: Yesterday date
Halfhour  Trunk group number  Date  Number Pbx Occupation  > R1    Name  Trunk group
count     Occupation / Trunk group count    Date (7) Allocation  Overflow
Failures Alloc.duration(Inc.)  Alloc.duration(Out.)  Total out. requests
Out.calls (node)  Out.call (other nodes)  Out.calls Inc.calls (node)       Inc.calls
(other nodes)     Inc.calls Occupation                                    0:30:00  178      31.3.08
0:30:00  1103     00:00:00          Olomouc        30       00:00:00          31.3.08
0:30:00  2        0         0       00:00:00       00:00:01          1        1       0       1       0       0       0
0:30:00  1102     00:00:00          OPAVA   30     00:02:00          31.3.08
0:30:00  0        0         0       00:00:00       01:00:00          0        0       0       0       0       0       0
```

Fig. 2. Original report in .txt format.

The example of the log file beginning fragment is displayed in Fig. 2 (log date 1.4.2008). Each record consists of 23 attributes. This file consists of 3572 records. So far, we have stored logs for 845 days; it means that 3 million records for processing or approximately 69 million of data fields are available for processing.

The program carries out data sorting based on the kind of information (date, time, etc.) filtering the separate data according to their importance (backbone, other nodes) and selecting the data suitable for the carried traffic load calculation. An example of the program outcome is shown in Fig. 3. The total count of records is reduced to 800 thousand.

```
day-of-week      day      month   year     circuit   halfhour incoming-calls-duration    incoming-call
s-duration-sec    incoming-calls-count        outgoing-calls-duration    outgoing-calls-durati
on-sec   outgoing-calls-count       missing-value
ÚT       20       12       2005    30       7:30      4:41:54  16914    58       2:27:34  8854     71       A
ÚT       20       12       2005    30       8:00      6:18:11  22691    73       2:50:06  10206    75       A
ÚT       20       12       2005    30       8:30      4:59:17  17957    66       2:26:22  8782     73       A
ÚT       20       12       2005    30       9:00      4:13:12  15192    87       2:10:50  7850     46       A
ÚT       20       12       2005    30       9:30      4:01:26  14486    65       2:49:00  10140    50       A
```

Fig. 3. The program outcome.

Such prepared data are directly imported into a target table in the database system. For each ordered couple of nodes in the network we have one text file available containing 48 records per day (one record per half an hour). Finally, the table contains 48 records per day for each ordered couple of trunks. The description of fields in the imported table follows:

- *Source-node*. Code for source node in network topology.
- *Target-node*. Code for target node in network topology.
- Day-of-week. Day of week (PO=Monday, UT=Tuesday, ST=Wednesday, CT=Thursday, PA= Friday, SV=Holiday).
- *Day*. Day part of date when the record was stored.
- *Month.* Month part of date when the record was stored.
- *Year.* Year part of date when the record was stored.
- *Circuit.* The count of circuits in a trunk (30 or 60).
- *Half-hour.* Time when the record was stored.
- *Incoming-calls-count.* The total number of incoming calls during half-hour for a line.
- *Incoming-calls-duration.* The total length of incoming calls during half-hour (timestamp format).

- *Incoming-calls-duration-sec.* The total length of incoming calls during half-hour (total of seconds).
- *Outgoing-calls-count.* The total number of outgoing calls during half-hour for a line.
- *Outgoing-calls-duration.* The total length of outgoing calls during half-hour (timestamp format).
- *Outgoing-calls-duration-sec.* The total length of outgoing calls during half-hour (total of seconds).
- *Missing-value.* The value indicates whether call related fields are empty.

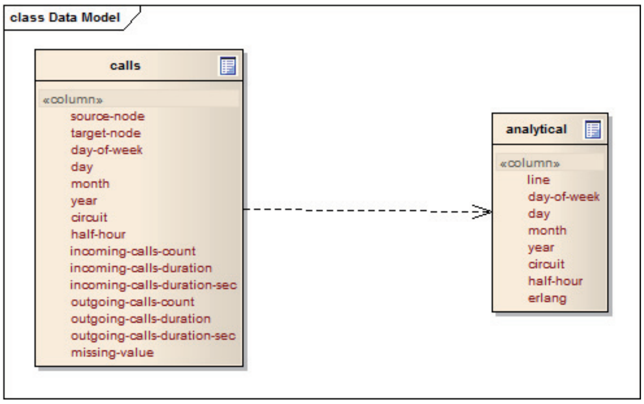

Fig. 4. Data model of imported data (calls table).

3.2 Data Preparation Methods

Data obtained from real word are usually very noisy, missing, and even inconsistent. The data should be usually pre-processed to receive true and valuable models that can be used for analysis. There are a lot of data pre-processing techniques [1] and we will apply the selected ones to improve data quality provided in our domain by communication network.

- *Data cleaning.* This technique is usually applied to remove noise and correct inconsistencies.
- *Data integration* is used for merging data from multiple heterogeneous sources into one stable and coherent data storage, such as transactional relational database (RDBMS – Relational Database Management System) or data warehouse (EDW – Enterprise Data Warehouse). Alternatively other data store types can be utilized, like stores based on XML (eXtensible Markup Language) applications.
- *Data transformation* is another data preparation technique used for substitution of current data with new improved data better fit to modeling algorithm. Normalization is one of the transformation techniques applied for methods, where distance measures are critical (e.g. clustering or segmentation).
- *Data reduction* techniques provide methods for the reduction of the data size. For example, it involves data aggregation, removing redundant or related data and clustering.

Data preparation is a complex and time consuming process. Data pre-processing takes significant time during the data mining project (from 20% to 60%). This step should not be omitted or underestimated, as the modeling results can distinguish from expected results.

3.3 Data Cleaning

The main aim of data cleaning is filling out missing values, finding outliers, smoothing out noise and correcting data inconsistencies.

For the missing value analysis many methods can be applied, from simple ones, like ignoring records with missing attributes, filling out values manually or filling out values automatically with global constant or attribute mean, to more sophisticated, like using attribute mean according to classes or filling out the most probable value generated by regression or decision tree induction.

Noise, the random error or variance in a measured variable, should be smoothed in data source before the selected data mining algorithms are applied (some methods like neural networks are able to smooth data while processed, but in general, the data sources should avoid such discrepancies). The binning, regression or clustering techniques are usually utilized for data smoothing.

3.4 Data Integration

The data for analysis usually come from more than one data source. One real-world object can be stored in different data sources; therefore this object should be identified in all the data sources and is matched up. In our project data comes from three different sources. The main data source is a collection of the log files provided by network management system, the second data source consists of manually created data records for irregular user activities analysis. The last data source contains tables with connection between loss, offer and traffic load.

Redundancy is the next important issue. An attribute derived from other attributes or attribute correlated with another existing attribute is usually redundant for background algorithms and can be removed from the table without losing data meaning. The correlation analysis, e.g. Pearson's correlation coefficient for numerical or χ^2 (chi-square) test for categorical variables, is usually applied at this step for decreasing data redundancy.

3.5 Data Transformation

During data transformation, the data are transformed into more appropriate form for further analysis. The basic data transformation tasks are aggregation, generalization, normalization and attribute construction. As an example, the traffic attribute is constructed according to expression (2) for each record.

3.6 Data Reduction

Data analysis and knowledge discovery in huge amount of data can take a long time. The reduced representation of data obtained from original data set makes analysis easier and more efficient. Data reduction task concentrates on obtaining a reduced data set without losing the original data meaning and integrity. The following methods for data reduction are available: attribute subset selection, dimensionality reduction, numerosity reduction and discretization and concept hierarchy generation.

Attribute subset selection removes irrelevant or redundant attributes. The important attributes are determined by statistical significance (e.g. information gain measure). Dimensionality reduction applies data encoding techniques for obtaining lossless (the original data can be reconstructed) or lossy (original data can only be approximated) data representation of original data.

3.7 Analytical Table Construction

Analytical table is a final result of the preparation phase. It contains clean, transformed, reduced data suitable for analysis in data mining discipline. The data model of data preparation process is displayed in Fig. 5.

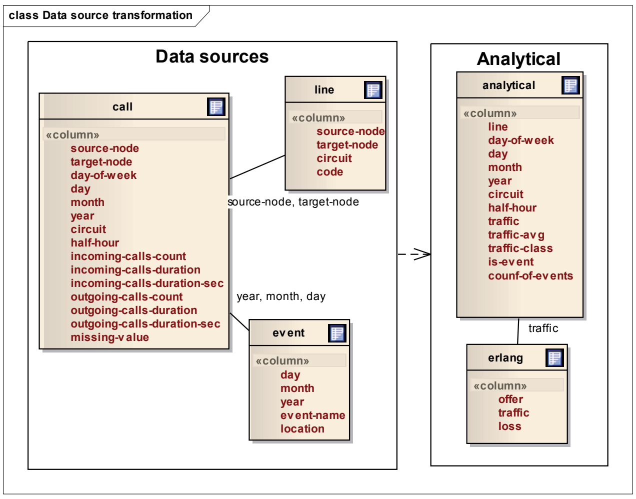

Fig. 5. Analytical table construction - data model.

The summary of newly created attributed during the data preparation process follows.

- *Line*. Code for connection between source and target nodes in network. Codes were introduced for hiding the network topology.
- *Count-of-events*. The total of significant events for a day.
- *Is-event*. Attribute indicates the existence of a significant event throughout the selected day. When count-of-events is less than 1, the value of is-event is false, otherwise the value is true.
- *Traffic*. Traffic is a carried traffic load as defined by expression (2).

$$\text{traffic} = \frac{\text{incoming-calls-duration-sec} + \text{incoming-calls-duration-sec}}{30*60} \quad (6)$$

- *Traffic-avg* and *traffic-class* are newly created values of the carried traffic load for missing values in the original data source. Two different methods were used for the calculation of a new value. Firstly, it is a mean value of a carried traffic load in a defined line (traffic-avg); secondly, it is a value based on kind of day, line, year and month (traffic-class). The second method should be more precise, but calculation is more time consuming. The usage of the new values is discussed in the description of mining models.

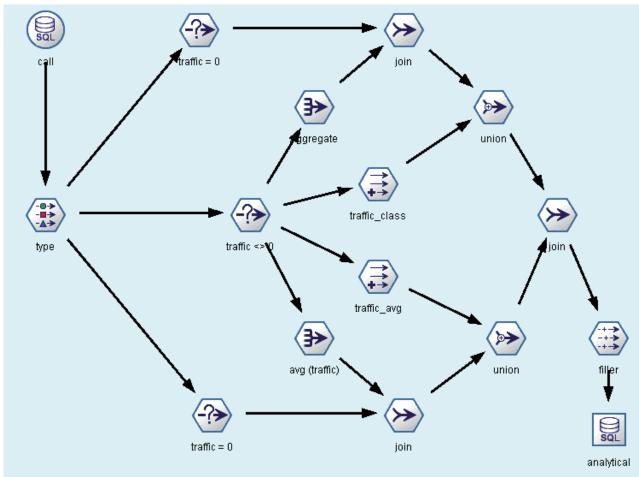

Fig. 6. Analytical table construction – process (call table part).

4. Tasks

The set of tasks for analysis have been introduced in the first section. Now, the tasks will be described in detail.

4.1 Analysis of the Traffic Load

As we already stated in section 2.1 (The topology of the SMCN) the traffic load depends on the type of the system, kind and range of services provided, users' activity, period of day and night, days, weeks and months in a year, working hours, etc. According to these common rules, the daily traffic load graphs for different situations, such as the types of days, interval of days, months and even years, were created. The outputs are significant for proper network sizing verification and for predictions.

The average traffic load for individual lines throughout the day can be seen in Figs 8, 9, 10. The high load has been identified on several lines (1, 14, 15 and 3), even at night time. After the detail analysis of the facts, it was confirmed that the night traffic contains mainly data transfers, not voice. It is necessary to note that the lines can have different capacities (30 or 60 channels), thus line 1 (60 information channels) is loaded at 50 percent at peak hours as well as line 14 or 3 (30 channels).

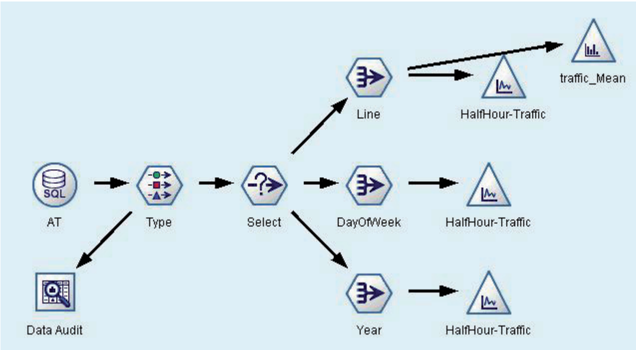

Fig. 7. The process of daily traffic load graphs creation.

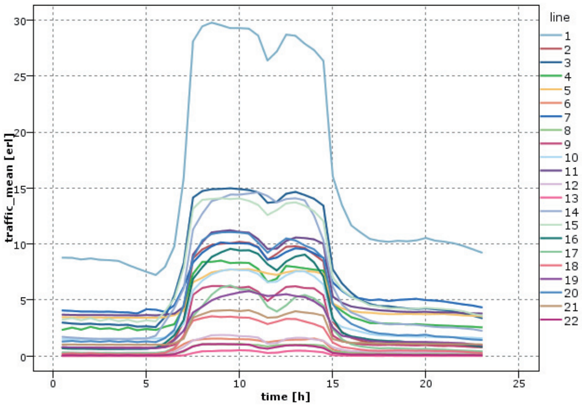

Fig. 8. The average traffic load for individual lines.

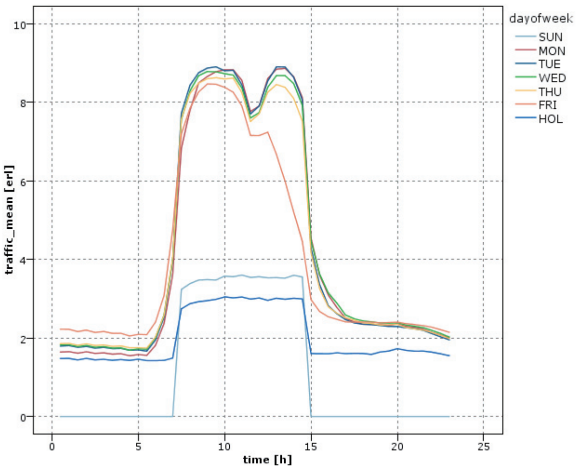

Fig. 9. The average traffic load for different type of days.

It is obvious (see Fig. 9) that the permanent data traffic is present on the lines. The level of traffic can be uncovered from the holiday traffic load. Sundays (abbr. SUN) do not show the permanent traffic due to incorrect settings in management system only for working hours (as was discovered later).

Daily traffic (traffic_Mean) differs in years. Traffic load in 2006 is higher than the load in 2007 or 2008 probably due to organization changes, reductions in our organization and due to technology updates using different network types. Permanent data transfers show a continual increase.

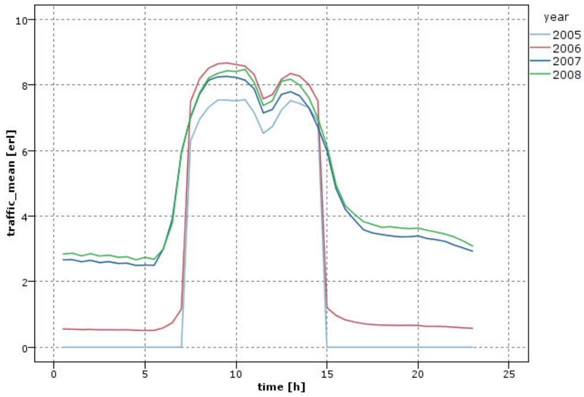

Fig. 10. The average traffic load according to years.

4.2 Daily Traffic Load Comparison

As an example of different behavior of traffic load, two days of week (Friday /FRI and Tuesday/TUE) were selected for comparison. The highly loaded line 1 from 2006 to 2008 was analyzed. Our hypothesis of the different traffic load and of daily minimums and maximums was verified (Fig. 12). The graph shown is an output of the stream in Fig. 11.

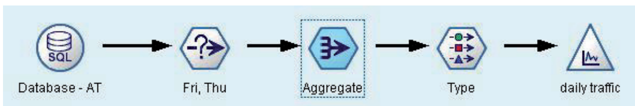

Fig. 11. Traffic load comparison – process.

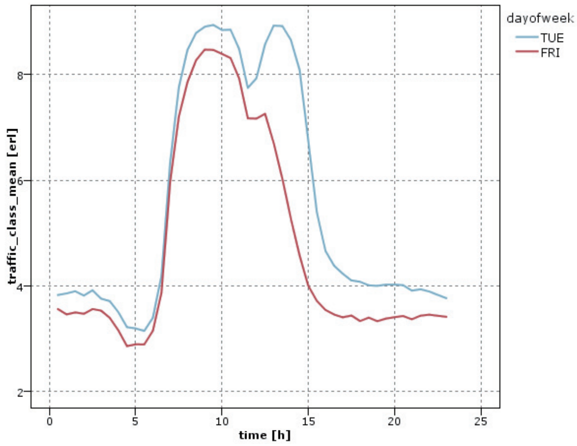

Fig. 12. Traffic load comparison - Friday vs. Tuesday.

The traffic load peak on Friday ends at 1pm, whereas the peak on Tuesday ends at 3pm. The presumption of lower load on Friday was verified as well. The rules for night data transfers can also be localized from the graphs, e.g. distribution in time, traffic load size. The decrease of load at 5am is caused by the backup procedures runs.

4.3 Standardized Network Behavior

The main goal of this task is finding the standardized network lines' behavior; it means the behavior under common conditions, e.g. regular working day, regular users' activities with the exclusion of extreme events.

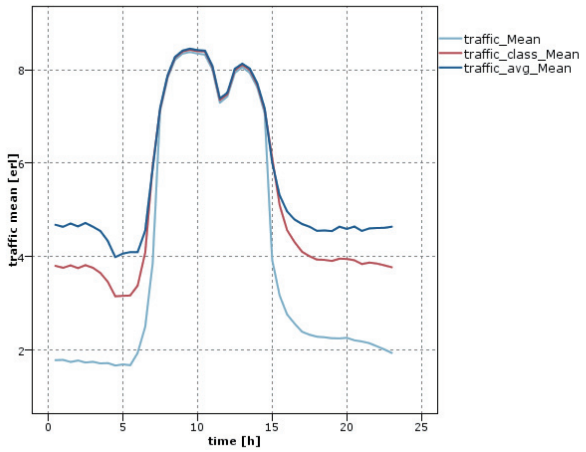

Fig. 13. Standardized network behavior model.

The graph contains three curves. The light blue one (traffic_Mean, lower one) is based on original data. It is obvious, with a good knowledge of analyzed network, that the output does not reflect the real network behavior, especially at night hours. According to this, probably mistaken graph results, the detailed analysis of original data have been undertaken. As was discovered, due to the setting in management system in 2006 the logging at night was disabled. The missing values were filled using two methods (see section 3.3). The red curve (traffic_class_Mean, middle curve) is more accurate. We use improved, better quality data with filled missing values in our research now. Original data were replaced. The dark blue (traffic_avg_Mean, top curve) was an intermediate step during the process of data quality improvement. The prediction models also give better results with a new data collection. The result is used for the detection of non-standard behavior in network.

4.4 Extreme Values Detection

Extreme values can change the meaning of data and decrease the prediction accuracy. The extreme data can be produced:

- Due to an error in the recording, e.g. carried traffic load is greater than theoretical value, or is close to this value.
- By irregular activities in network.

The first category of extreme values can be detected easily. For example, the value of carried traffic load greater than maximum indicates a defect in supervision system or logging system. Such records are resolved in the same way like records with the missing values. It means that the values are replaced by new ones.

Only 5 defect records, with value recorded by dispatching software higher than the theoretical value of the potential load on the volume, have been found in our data set. Always it was the only line (in analytical table the line was labeled as line 14). Subsequent analysis confirmed the software incompatibility with the used technology type at the observation point.

count	line
5	14

Tab. 1. Extreme values – greater than 100% of maximum.

The carried traffic load greater than 75% of maximum available value is very unusual in common communication network and the supervision system evaluates such values as critical. The higher values of the carried traffic load can indicate the troubles with network overload in the future, especially when the larger data volume transmission is requested, due to irregular network condition or user activities. Tab. 2 displays measured critical values; the line number 14 requires more investigation. Such extreme values can decrease the success of prediction and it is wise to replace them with new values. It is required for 0.01% of records.

count	line
659	14
102	3
79	5
67	15
59	11
57	2
32	10
23	16
9	7
4	9
1	19

Tab. 2. Extreme values – greater than 75% of maximum.

The carried traffic load greater than 30% of maximum value can be critical for collision type networks, like the packet switching network. The network in focus is a circuit switching network and if the routing mechanism is set properly, the values over 30% are not very critical. The 14.5 % records of the data set have carried traffic load greater than 30% of the maximum.

The critical values of the second type are generated due to irregular network conditions (usually defects) or irregular user activities. In this case, it is necessary to extend an analysis into other related domains. For example, the collection of events like planned activities, expected important visits, environmental disasters can have an impact on the analyzed data. Such connections were already mined from the data sets.

4.5 Non-standard Behavior Detection

We focused on testing of possible network overload based on a non-standard behavior of users in this task. Created models can verify network reliability during critical situations and can help not only with lines sizing but predict such activities in the future based on current data. For the analysis a new table with past events was created (already discussed in section 3.8).

An example of irregular activities detected can be found in Fig. 14 and Fig. 15. The blue curve (smooth one) is a standardized model for lines 3 and 15, red curve displays increased traffic load, especially at night hours. From the knowledge of event type, it is clear that the data transfer is increased during night hours. The event has higher impact on line 15 than on line 3. From the knowledge of event location the difference can be verified, and vice versa, we can locate an event from the knowledge of behavior on lines.

Based on subsequent analysis we can find out that the load on line 15 was increased up to 40 percent. The histogram of change in the traffic load compared to expected values is displayed in Fig. 16.

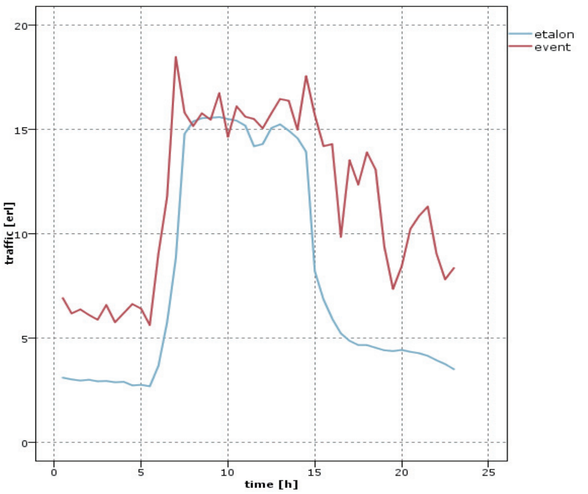

Fig. 14. Irregular activities of users (line 3).

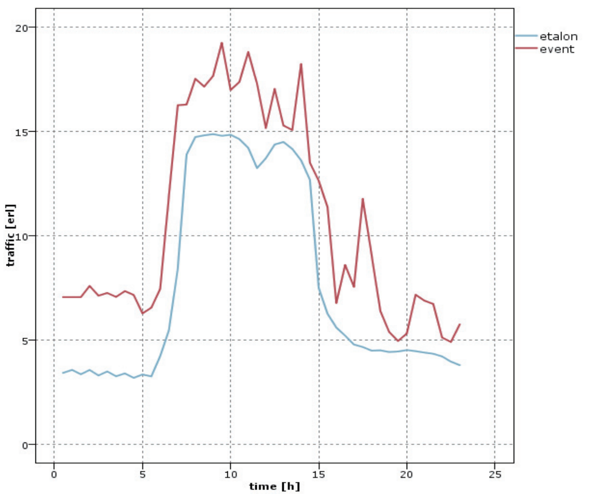

Fig. 15. Irregular activities of users (line 15).

4.6 Proper Network Dimension Verification

The goal of the task is a verification of proper network sizing and its reliability. The analyzed network differs from commercial networks; the operational cost is not a key attribute. It is necessary that the network operates in extreme situations, and therefore the system is over-designed.

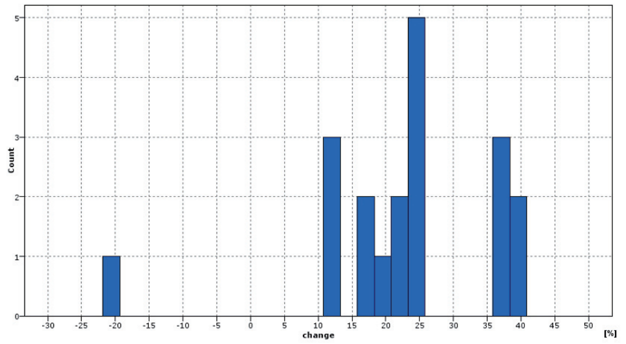

Fig. 16. Histogram of change compared to expected behavior.

The analyzed network is an example of the kind of a loss queuing system; the quality of such a system depends on a number of rejected requests. The analysis of losses was used for the analysis of proper sizing. The Erlang loss expression, describing the probability of call loss on a group of circuits together with Erlang tables, forms the theory for our research on proper sizing:

$$E_{1,n}(A) = \frac{\frac{A^n}{n!}}{1 + A + \frac{A^2}{2!} + \ldots + \frac{A^n}{n!}} \tag{7}$$

where A is the flow of traffic offered expressed in erlang, n is the number of circuits; it means the number of time slots on the PCM line.

The loss probabilities were calculated based on the knowledge of traffic offered and the number of circuits (30 or 60). It was possible to approximate the offered traffic with carried traffic load for a low network load. Network dimensioning was verified using the peak busy hour.

In the extreme values detection task we have determined the frequency of load limits for 30 percent and 75 percent. We have uncovered the lines 14, 3, 5 a 15 as the most loaded. These lines were analyzed in more detail using the loss probabilities.

The process of verification was made in the following steps.

- The most quality data collection was used (traffic_mean).
- The PBH values were calculated for all monitored lines including the PBH time location.
- The value of loss was attached to PBH using Erlang table according to the corresponding circuits.

The highest PBH values and corresponding losses were calculated for the lines 3, 14 and 15 (30 circuits) and line 1 (60 circuits) as expected.

line	PBH (Erlh)	hour	Loss int (Erl)
1	31,05908	8:30	4,28204E-05
2	10,54011	10:0	5,71874E-06
3	15,60248	9:00	0,006142887
4	8,854484	8:30	1,38742E-07
5	8,034059	2:30	1,66567E-08
6	1,64009	8:30	1,61878E-27
7	10,80951	9:00	8,35788E-06
8	1,104452	9:30	2,40874E-32
9	6,466588	8:30	6,14376E-11
10	8,198617	9:30	2,20473E-08
11	11,46917	9:00	2,90795E-05
12	1,999286	9:00	1,09567E-24
13	0,832805	21:3	1,67762E-36
14	15,37852	10:3	0,005028981
15	14,83234	8:30	0,002672779
16	10,09124	9:30	2,10828E-06
17	6,648857	9:00	1,60569E-35
18	3,738296	8:30	3,83446E-17
19	6,082595	10:0	1,87227E-11
20	11,91593	9:30	5,62591E-05
21	4,243127	10:0	1,18313E-15
22	1,144511	9:00	2,40874E-32

Tab. 3. PBH and loss values for all lines.

For modeling purposes the process stream was created (Fig. 17). The Erlang tables for 30 and 60 circuits were imported into database for simplifying the loss values calculation.

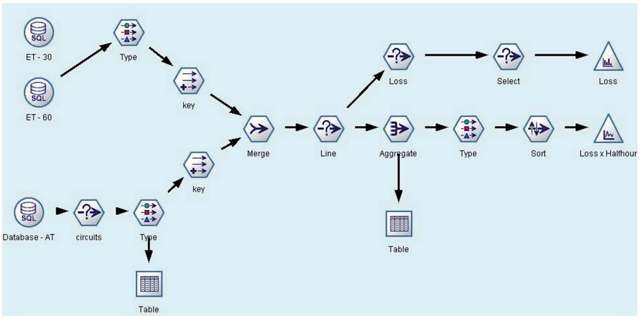

Fig. 17. Lines sizing – process.

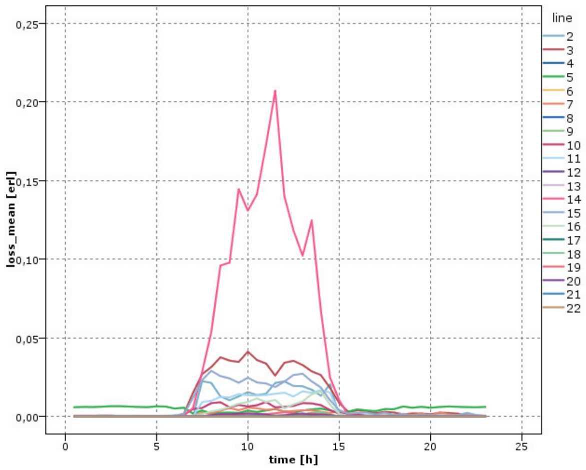

Fig. 18. Loss distribution for 30 circuits.

During the analysis the lines with higher loss values were detected in comparison with the rest of lines. The selected lines are not sized properly and in the future they can be vulnerable to extreme loads.

4.7 Prediction of the Carried Traffic Load

Prediction of selected parameters (short-time and long-time carried traffic load) for various settings of starting conditions is one of the main tasks solved in our project. Several streams were created for the tasks; data set was split into training and testing subsets. The decision tree algorithms C5.0 [1], [10] and CART [1] were utilized for the predictions. C5.0 (like its predecessor C4.5) builds decision trees from a set of training data using the concept of information entropy. CART (Classification and regression trees) builds the tree using Gini coefficient. The C5 algorithm was more successful algorithm in the project. An example of output for C5.0 and CART algorithms is in Tab. 4. The Accuracy column defines the tolerance of what is accepted as a valid prediction. (e.g. 3 means – the prediction is correct if the difference between the predicted and current values is in the interval (-3;3) [erl].

Model	Training			Testing			Accuracy			
	L	Year	Mn	L	Year	M.	0	3	5	10
C5	1	2006	all	1	2006	all	80,3	90,3	97,8	99,8
CART	1	2006	all	1	2006	all	59,5	84	92,8	98
C5	1	2006	all	1	2007	<7	2,9	18,5	28,1	61
CART	1	2006	all	1	2007	<7	3,7	21,6	31,3	61,4
C5	1	2006	all	1	2007	>6	4,6	29,6	42,8	62,6
CART	1	2006	all	1	2007	>6	5,2	31,3	41,5	59,6
C5	1	2006	all	1	2008	<7	2,8	18,3	24,8	34,2
CART	1	2006	all	1	2008	<7	3,5	22,1	28,1	34,3
C5	1	2006	all	1	2008	>6	3,2	18,9	26,4	38,7
CART	1	2006	all	1	2008	>6	3,8	21,7	28,1	37,2
C5	1	2008	<7	1	2008	<7	59	69,9	95,7	99,6
CART	1	2008	<7	1	2008	<7	13	64,2	83	97,2
C5	1	2008	<7	1	2006	all	3,3	19,8	28,4	41,5
CART	1	2008	<7	1	2006	all	3,5	23	31	41,5
C5	1	2008	<7	1	2007	<7	6,4	37,4	53,3	80,6
CART	1	2008	<7	1	2007	<7	6,2	39,3	54	79,4
C5	1	2008	<7	1	2007	>6	7,9	49,3	66,7	89,3
CART	1	2008	<7	1	2007	>6	10,7	60,2	80,2	97,1
C5	1	2008	<7	1	2008	>6	7,9	49,5	69,9	94,2
CART	1	2008	<7	1	2008	>6	9,7	57,2	76	96

Tab. 4. First phase of predictions.

The algorithms do not always provide the expected values of prediction. One would expect that a longer distance between current values and predicted values will automatically mean less accuracy of prediction. Variances can be explained by more detailed analysis of behavior for individual months and years. One of the main results of the first phase was that the prediction of new values highly depends on attribute month. Both C5.0 and CART algorithms used this attribute for splitting data collection in trees. For next analysis only C5.0 was selected, attribute month was excluded from prediction model.

The following streams (Fig. 19, Fig. 21) provided more valuable results. For training and testing, the first model uses the sets of the following pattern:

- The first setting for training set: line=3, year=2008, month in <1;3>.
- The next settings for training set: line=3, year=2008, month in <1+*n*;3+*n*> for *n* in <1;7>. Seven different settings were tested in total.
- The testing set was fixed: line=3, year=2008, month=10.

Line 3 was selected, as it is one of the most loaded lines. Results are displayed in Tab. 5; values are also displayed on the Fig. 20.

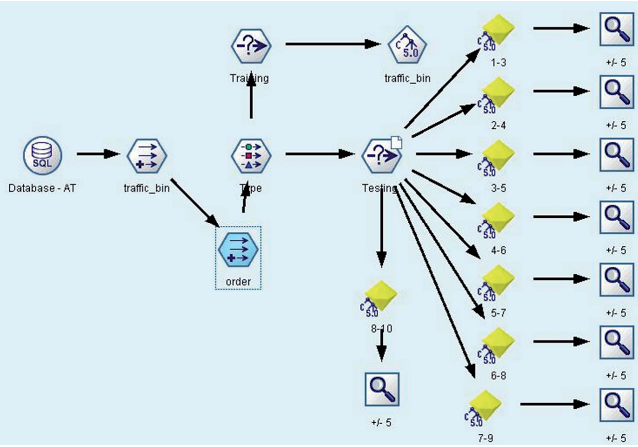

Fig. 19. Prediction for months – process.

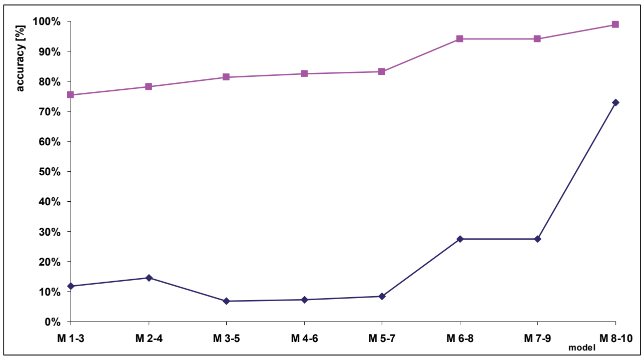

Fig. 20. Accuracy of predictions for individual models.

Node traffic_bin in the stream creates a discrete set (rounding by 1 erlang) for the continuous output value.

Model	Exactly	+/- 5
M 1-3	11,72	75,52
M 2-4	14,65	78,16
M 3-5	6,71	81,38
M 4-6	7,37	82,51
M 5-7	8,51	83,17
M 6-8	27,6	94,05
M 7-9	27,5	94,05
M 8-10	72,97	98,77

Tab. 5. Summary of predictions for stream Months.

Note. Exactly in Tab. 5 means that the rounded predicted value is equal to rounded measured value of load; +/- 5 means that predicted value is in interval +/- 5 erlangs.

The second model creates a discrete set for output value per 5 erlangs, thus for 60 circuit line 12 intervals were created and for 30 circuit line 6 intervals were created in the traffic_bin node. Each interval is substituted by a category value.

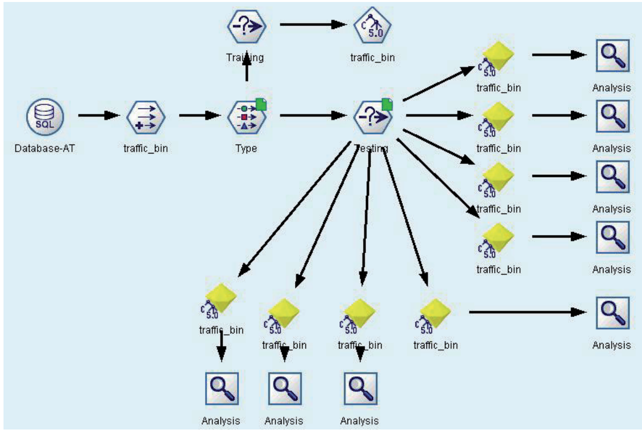

Fig. 21. Interval predictions – process.

5. Conclusion

In the paper the focus was on the analysis of a communication network using data mining techniques. The possibility to deal with a large amount of data (millions of data values) uncovers hidden dependencies in data. The entry data quality has a significant impact on further analysis. The data preparation techniques were applied to original data, and new data collections and analytic tables with pre-processed data were produced.

Gradually, by means of properly chosen software, network models were developed to verify generally valid patterns of network behavior as a queuing system and the network dimensioning was confirmed. During the analysis the lines with higher loss values in comparison with the rest of lines were detected. The selected lines are not sized properly and in the future they can be vulnerable to extreme loads.

Furthermore, unlike the commercially available communication networks simulators, the selected methods and models allow us to detect nonstandard behavior of the network in focus. Nonstandard network behavior is always associated with defects in the network, the operator intervention in its topology; it can also be related to natural disasters or current international situation. It always reflects the behavior of the network users. This might vary depending on the period of the day, the type of the day (holidays and working days), and also in connection with planned activities.

Next, the analysis of the load increase trends has been carried out. Based on the outputs, the topology changes were suggested to improve the network reliability of the current network.

The set of tasks were solved during the analysis of the communication network in focus. The results will help improve to optimize of the network.

Every modern communications system is nowadays supervised and controlled from the MNC center. During our work with the data we examined the shortcomings in our specific configuration of the monitoring system, which initially collected data only during working hours and working days. It was not possible to detect high levels of traffic during non-working time (database upgrades, PBXs) or continuous data transfers that can be found in the data collected at weekends. Feedback was our recommendation to reconfigure the system of surveillance reports. Our experience is useful and generally applicable to the supervisory staff of any commercial network.

Based on the findings the network can be better dimensioned, whether in terms of excess loads and inefficient in terms of removing of the directions in the network. Network management on the basis of the results can reinforce the overloaded lines and thereby increase network security.

SNMC is overdesigned in terms of demands for its security and reliability. Detection of non-standard behavior of network users and observations of its influence on the current network bandwidth problem is not only an issue of the military network, but also the commercial networks, which serves as a backup communications for emergency services and military during emergency situations (natural disasters, the international situation).

Prediction, which is performed in one of the sub-task, is the subject of further work and will focus on finding the trend in the data, so that we can increase accuracy of predictive models to specify the prediction, which are intended primarily for the correct sizing of the newly built on the principles of VoIP networks [9].

References

[1] HAN, J., KAMBER, M. *Data Mining – Concepts and Techniques*. Morgan Kaufmann Publishers, 2006. ISBN 1-55860-901-3.

[2] HIBA, M. O., SAMI, M. S., ELHAG, H. M. A backbone Internet traffic intensities and statistics in Sudan. In *3rd International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA 2006)*. Orlando (USA), 2006.

[3] GHAZALI, M., AZMINBIN, M. *Analysis on the Traffic Load Pattern of Unikl WLAN*. Thesis. University Utara, Malaysia , 2007.

[4] CAN, B. *Traffic Analysis and Modelling in PMR Systems*. Thesis. Electrical and Electronics Engineering Department, Bilkent University, 2003.

[5] LIU, N. X., BARAS, J. S. Statistical modeling and performance analysis of multi-scale traffic. In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2003*. San Francisco (CA, USA), 2003, p. 1837-1847.

[6] ONDRASIK, J., MAZALEK, A. New trends in ACR field communication systems In *Proceedings of New Technologies in Radio-Communications*. University of Defense, Brno (Czech Republic), 2006, 7 p.

[7] *Technical Documentation for A4300 and A4400*. Alcatel-Lucent, avenue Kléber - 92707 Colombes France, 1996, 2009.

[8] ONDRYHAL, V., VRANOVA, Z. Using data mining methods for communication net behaviour analysis. *EE Časopis pro elektrotechniku a energetiku*, 2008, vol. 14, no. 5, p. 282-286. ISSN 1335-2547.

[9] ONDRYHAL, V., VRANOVA, Z. Different ways to identify trends in network traffic. In *Proc. of Networking and Electronic Commerce Research Conference 2010*. Lake Garda (Italy), October 2010.

[10] *RuleQuest Research Data Mining Tools - Sample Applications Using See5/C5.0* [online] Cited 2011-04-12. Available at http://www.rulequest.com/see5-examples.html.

About Authors ...

Vojtech ONDRYHAL, Ph.D is a lecturer at the University of Defense, Military Technology Faculty, Communication and Information Systems Department, Brno, Czech Republic. His skills include database systems, data mining, software engineering (Java, PHP and JavaScript languages) ontology engineering and information systems modeling with UML.

Zuzana VRANOVA, Ph.D. is a lecturer at the University of Defense, Military Technology Faculty, Communication and Information Systems Department, Brno, Czech Republic. Her skills include transmission networks (ATM, SDH etc.) and data, voice and VoIP communication networks and corresponding technologies, structured cabling systems.