Document Title: **The Telephone Traffic Data Analysis**

Author(s): **Aleksander Pur and Igor Belic**

Document No.: **208041**

Date Received: **December 2004**

This paper appears in *Policing in Central and Eastern Europe: Dilemmas of Contemporary Criminal Justice,* edited by Gorazd Mesko, Milan Pagon, and Bojan Dobovsek, and published by the Faculty of Criminal Justice, University of Maribor, Slovenia.

ALEKSANDER PUR, IGOR BELIČ

THE TELEPHONE TRAFFIC DATA ANALYSIS

In everyday life enormous amount of data are collected and stored in the different databases. Such data are for example credit card transactions, activities on the web, phone calls, and many others. This large amount of data can be analysed for various purposes.

The data analysis research was directed to analyse the mobile phone activities (traffic) data. The analysis could be useful to provide additional information of various criminal activities. However there is a problem how to quickly and efficiently analyse the large amount of data. For that purpose the traditional query and the data mining methods were tested.

INTRODUCTION

More and more data about our everyday life activities is automatically collected and stored in the different databases. Such activities are for example: credit card transactions, activities on the web, phone calls and many others. The analysis of this data can provide useful information from different points of view (Pur, Bohanec, 2003). For example, by scanning each sale into a databases and analysing this data, grocery stores have determined that men in their 20s who purchase beer on Fridays after work, are also likely to buy a pack of diapers (Hayes, 1999). That kind of analysis is called Market Basket Analysis. The main purpose of this analysis is to find which articles and on what circumstances are associated with buying certain product or with other words, what is characteristically for buying certain product. In chapter 2.2 we have described theory and possibilities of using that kind of analysis in the telephone traffic data. The goal of that analysis is to find phone calls, which are characteristic or associated with certain events in which a caller was involved. The analysis could be the way to get useful information about crime events, which are reflected in this data.

METHODS AND RESULTS OF DISCOVERING ASSOCIATIONS BETWEEN TELEPHONE CALLS AND CERTAIN EVENTS

The telephone traffic data is the data about telephone calls, which is automatically stored at every telephone exchange. Such data include caller phone numbers, IMEI (International Mobile Equipment Identifier) called phone numbers, date of calls, time of calls, duration of calls and used base stations. Our usual behaviour is reflected in this data. If some event changes this behaviour then change can also be reflected in this data. That kind of the events are birthdays, holidays, trip to the mountains, sick leave or even crime acts in which caller was involved. In the presented example we analyse telephone call data of tested person. In this data we discover associations between telephone calls and certain event and demonstrate how the event is reflected in this phone call data. This event is a period of time, which tested person spent away from work because of illness (sick leave).

DISCOVERING ASSOCIATIONS WITH TABLE AND GRAPH

Aggregate values of telephone call data of tested person are presented in the table 1. Table 1 is composed of next columns: ID of called phone numbers, numbers of calls in time of the event, total number of calls and the parameter in the last column is the ratio

This item was translated into English by the source and not subject to subsequent editing. Views, opinions, and conclusions are those of the author and do not imply endorsement, recommendation, or favor by the U.S. Government.

of number of phone calls in the event time to number of total calls to phone number in first column. Last parameter is the conditional probability that the caller calls in the event time if he calls the certain number. We describe this parameter (confidence) in greater detail in the next chapter. Expected confidence is the other interesting parameter and it shows the probability that the phone call was done in the event time. We describe this parameter (expected confidence) in greater detail in the next chapter. The discrepancy between those two parameters is telling us how characteristic each group of phone calls for the event time really is.

In the last row of table 1 is the highest value of the parameter confidence. Hence the probability that caller calls phone number 22 in the event time, is 67%. Because the value of expected confidence is constant (9%) and the value of confidence is the highest for phone calls to number 22, these calls are the most characteristic for the event time. This characteristic group of calls has only one attribute, called phone number 22.

Table 1. Comparison between the phones calls in time of the event and total calls

Called numbers	Calls in time of the event	Total calls	Confidence
Number 1	13	204	0.064
Number 2	0	32	0.000
Number 3	0	28	0.000
Number 4	5	28	0.179
Number 5	2	27	0.074
Number 6	3	27	0.111
Number 7	4	26	0.154
Number 8	5	25	0.200
Number 9	0	25	0.000
Number 10	0	19	0.000
Number 11	3	15	0.200
Number 12	0	15	0.000
Number 13	0	13	0.000
Number 14	1	12	0.083
Number 15	3	11	0.273
Number 16	1	9	0.111
Number 17	1	8	0.125
Number 18	0	7	0.000
Number 19	1	6	0.167
Number 20	3	6	0.500
Number 21	1	6	0.167
Number 22	4	6	0.667

In Figure 1 new characteristic groups of phone calls, described with two attributes, can be found. The graph is based on the same phone call data. Each bar in the graph represents a group of phone calls, which is described with called number and day of the week of the phone call occurrence. The height of the bar represents the value of parameter confidence. The discrepancy between the value of parameter confidence and value of expected confidence tells us how characteristic each group of phone calls for the event time is. For example, one of the characteristic groups is composed of phone calls to number 4 on Wednesday. In graph the height of the bar that represents this group (0.67), means probability that caller is on the sick leave when number 4 is called on Wednesday.

We can expect existence of other interesting groups, which are described with more attributes. But that increases the number of that kind of groups. In this example, the number of groups will increase to 110880 if they are described with next attributes: called phone number (22), time of call (24), base station (30) and day of the week (7). In such a large number of parameters, it is very hard to find, the groups connected with the observed event time. To find those groups we can use technique, which is called association rules discovery and is described in greater detail in chapter 2.2.

Because these groups of phone calls are characteristic for the time of the event, we can also say that these groups are connected or associated with the event time. All associations between the event time and some groups of phone calls can be indications that caller is involved in that event. Attributes values, which describe this groups, can give more information on, which way caller is involved in this event.

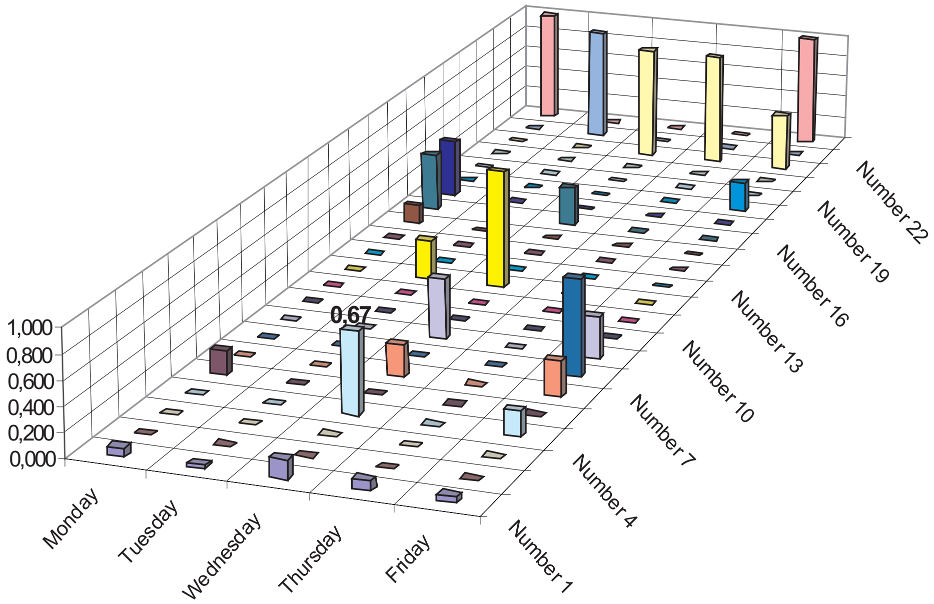

Figure 1. Confidence of the groups of phone calls

DISCOVERING ASSOCIATION RULES

The goal of the association rules discovery is to detect unexpected associations between specific values of categorical variables in a large data sets. Unexpected means, that we don't need to predict, which attributes will compose the association rules. Association rule shows attribute value conditions that occur frequently together in a given dataset. These rules provide information in the form of "if-then" statements. Association rules are computed from the data, and unlike to the if-then rules of logic, association rules are probabilistic in nature. The rule consists of left hand side (LHS) and right hand side (RHS). If a conjunction of attributes value on the left hand side occur, then with some probability, attribute value in right hand side also occurs (Han, Kamber, 2001).

A typical and widely used example of association rule discovering is Market Basket Analysis. In this example, data are collected using bar-code scanners in supermarkets. Such žmarket basket' databases consist of a large number of transaction records. Each record lists the items bought by a certain customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting stores layouts (placing items optimally with respect to each other), for cross selling, for promotions, for catalogue design and to identify customer segments based on buying patterns (MIT, 2003). Association rules look like this:

Association rule 1:

[Buy = computer] AND [Buy = monitor] ==> [Buy = printer]

Rule 1 means that if the customer buys computer and monitor then, with some probability, customer also buys printer.

Association rules discovery also enables analysts to uncover hidden patterns in telephone traffic data. Next rules are derived from the example of telephone traffic data of tested person.

Association rule 2:

[Called number = 22] ==> [Sick leave = yes]
Support = 0.597 Confidence = 66.67 Lift = 7.4

Association rule 3:

[Day of the week = Wednesday] AND [Called number = 22] ==>
===> [Sick leave = yes]
Support = 0.597 Confidence = 66.67 Lift = 7.4

The second rule means that more then 66 % of phone calls to number 22 happened when the caller was on sick leave. The third rule means that more then 66% of phone calls to number 4 on Wednesday were on sick leave. Because the association rules are probabilistic in nature the parameters such as support, confidence and lift, is needed for their validation (Srikant, Agrawal, 1998).

Support describes how prevalent the rule is throughout the dataset, or how often left and right sides of the rule occur together in the dataset as a fraction of the total number of records (formula 1). This parameter influences the validity of that rules. For example in the second rule number 22 was called from sick leave 4 times or 0.597% of the total number of phone calls. If the number off calls is smaller then the validity of the rules could be under question.

$$Support\ (A \Rightarrow B) = \frac{|A \wedge B|}{|S|} \qquad 1$$

A - left hand side
B - right hand side
S - total number of records

Confidence is the ratio of the number of records in which both sides of the rule appear together in the number of records in which the only left hand side of rule appears. For example in the second rule the confidence is 66.67%. This means from sick leave occurs 66.67% of all calls to number 22.

$$Confidence\ (A \Rightarrow B) = \frac{|A \wedge B|}{|A|} \qquad 2$$

A - left hand side
B - right hand side

Expected confidence measures the confidence of a rule as if there were no relationship between the left and right hand side of the rule. It is the ratio of the number of records in which the right hand side of the rule appears to all records in dataset. In the second rule it is the ratio of the phone calls from sick leave to all calls.

Lift is the ratio of confidence to expected confidence. It tells how much additional information the left hand side of a rule gives when trying to determine whether the right hand side is present in any given record. In the second rule lift is 7.4.

$$Lift\ (A \Rightarrow B) = \frac{\frac{|A \wedge B|}{|A|}}{\frac{|B|}{|S|}} = \frac{Confidence(A \Rightarrow B)}{\frac{|B|}{|S|}} \qquad 3$$

DISCUSSION

A lot of data about our everyday life activities is automatically collected and stored in the different databases. Because the analysis of this data can provide the useful information, we are trying to analyse the phone call data of the tested person. Our analysis is oriented to discover groups of phone calls, which are characteristic for time of some event (for example, sick leave). In the analysis we were using different methods, tables, graphs an association rules. Significant difference is between using tables or graphs and association rules. With association rules unexpected characteristic groups can be found. Unexpected means that we don't need to predict attributes for describing these groups, in graphs or tables the attributes must be known before. Another advantage of using association rules is their ability of discovering the groups, which are described with more attributes. Graphs and tables are limited in number of attributes.

ABOUT THE AUTHORs

Aleksander Pur is working in General police directorate, Informatics and telecommunications service, Ljubljana. His research interests include, DSS and data mining.

Igor Belič holds the BS degree from Computer Engineering and works as the senior lecturer at Faculty of Criminal Justice, University of Maribor.

REFERENCES

Han, M., Kamber, M. (2001). Data Mining: Concepts and Techniques. Simon Fraser University, Morgan Kaufman Publishers.

Hayes, H. (1999). The Knowledge Banks. Internet: 15.06.2004,
www.govexec.com/archdoc/0396/0396s5.htm.

MIT, (2003). Discovering Associations Rules in Transaction Databases. Massachusetts Institute of Technology (MIT). Internet: 10.05.2004,
ocw.mit.edu/NR/rdonlyres/Sloan-School-of-Management/15-062DataMiningSpring2003/F42B67E-7394-4B5F-B394-C6C542EF1DE2/0/Lecture16.pdf.

Pur, A., Bohanec, M. (2003). Knowledge Discovery from Data Bases - Possibility of Using Association Rules in Police. Varstvoslovje, FPVV, Ljubljana.

Srikant, R., Agrawal, R. (1998). Mining Quantitative Association Rules in Large Relation Tables. San Jose, IBM Almaden Research Center.