Data Mining Applications in Public Organizations

Mirjana Pejić Bach
Faculty of Economics, University of Zagreb
Trg J.F. Kennedya 6, 10000 Zagreb, Croatia
mpejic@efzg.hr

Abstract. *The purpose of the paper is to present a survey on data mining applications in public organizations. Search of the scientific data bases and Internet has revealed that most of the applications are described in the current year at the business web sites. Finance and economics, healthcare, criminal justice and defense are the most popular application areas. Classification and prediction, concept/class description and evolution analysis are the most often used methods.*

Key words. data mining, public organizations, government

1. Introduction

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand et.al, 2001). In other words, data mining is another way to find worth information in data, besides statistics, on-line processing (OLAP), spreadsheets, and basic data access.

Data mining techniques exist for a number of years and its roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning. The term itself was only introduced relatively recently, in the 1990s as a result of increased computer power and improved data collection and management techniques.

The goal of data mining methods is to find interesting patterns representing knowledge. Data mining methods can be categorized into two groups: descriptive and predictive tasks. Descriptive tasks find human-interpretable patterns that describe the data, and predictive tasks use some variables to predict unknown or future values of other variables. Following kinds of patterns can be discovered by the data mining techniques (Han *et.al.,* 2001): concept/class description, association analysis, classification and prediction, cluster analysis, outlier analysis, and evolution analysis.

In the last ten years, data mining has become one of the most popular hypes of the business world. Common uses of data mining are bioinformatics (Gusfield, 1997; Waterman, 1995), financial data analysis and financial modelling (Benninga et.al, 1997; Higgins, 1997), retail data mining and customer relationship management (Berry et.al, 1999), telecommunications (Mattison, 1997), and science (Valdes-Perez, 1999). However, public organizations only recently bring up use of data mining (Cahlink, 2000; Carbone, 1998).

Goal of the paper is to explore the possibility of using data mining in public organizations as a tool for improving their efficiency. In the paper we shall focus on areas of applications of data mining for public organizations. Summary of applications of data mining in public organizations will be presented.

2. Areas of application of data mining for public organizations

In this paper we examine possible areas of application in public organizations. Areas of application are divided into: (1) finance and economy, (2) healthcare, (3) criminal justice and defense, (4) labor and social welfare, (5) E-Government, (6) education, and (7) transport.

For each area we shall describe current applications and propose new ones. Current applications are found from the Internet search (Google) with the use of words: data mining, government, and public. Following databases are also searched: Emerald, EBSCOhost, Proquest, Science Direct, Springer Verlag, Kluwer, Engineering Village 2 & Compendex, Wiley Interscience, and ProQuest Digital Dissertations.

2.1. Finance and economy

Government tax agencies use Clementine and Intelligent Miner to build a predictive model that could improve collections management and audit selection by answering questions such as "Who is likely to become delinquent and by how much?" and "Which tax returns are likely to be non-compliant?" (SPSS, 2003; IBM, 2003). Such models reduce the opportunity for fraud. Possible good or bad tax payers could be identified by the decision tree method. Also, association analysis could detect groups of taxes that bad tax payers usually try to evade. Four similar examples are mentioned at the other pages of SPSS and IBM web sites, and one on the SAS Enterprise Miner Web site (SAS, 2003). Michael (2003) reports on one additional application of data mining on detecting tax fraud.

Neural networks have provided valuable insights for analysts forecasting tax revenues, which are critically important since agency budgets, support for education, and improvements to infrastructure all depend on their accuracy (Hansen *et.al*, 1997).

Ministry of agriculture funds essential public services like water and sewer systems, housing, health clinics and public utilities as well as investing in businesses and agricultural cooperatives. The data mining system is specifically designed to help top policy-makers to navigate and access the data in a way that is generally used by banks to analyse the effectiveness of lending programs (Makulowich, 1999).

Data mining system could be used to increase effectiveness of programs that encourage small business and innovations. Ministry of economy usually offers credit lines for small business owners and innovators. Decision tree method could be used to identify whether the applicant is a good or bad credit risk.

2.2. Healthcare

Sund (2002) describes two uses of data mining in Finish health care system. Generalized event sequences method is used to develop and implement register based performance indicators to measure the effectiveness of surgical treatment of hip fracture. Also, concept/class description is used to evaluate and compare the effectiveness of health care providers.

Data mining is often used in detecting health care fraud. IBM Fraud and Abuse Management System is used for detecting health care fraud and abuse which ranks as one of the nation's leading law enforcement frustrations (IBM, 2003). The system scores each component and processes data to generate a "suspicion index" of all providers within the group. To identify suspicious providers, users select from numerous behavior patterns appropriate to a particular peer group, then combine patterns to build an analysis model. The system can be also deployed to profile patients–facilitating "link analysis" between physicians and patients engaged in fraud. Clementine data mining system is used for: (1) detecting false claims to the state using ID numbers stolen from Medicaid patients and (2) payment error prevention like tracking indicators such as patients admitted unnecessarily, patients discharged and readmitted the same day, and, because different diagnoses are paid in different ways, incorrect diagnosis codes (SPSS, 2003). Michael (2003) reports three additional applications of data mining in healthcare fraud in US, Australia, end Europe.

Data mining is a part of scenarios for health care development in next ten years in countries like Germany, Austria (Haux *et.al.*, 2002) and Czech Republic (Zvarova *et.al.*, 2002). It is also integral part of Continuous quality improvement and research of the The National Emergency Medical Extranet Project which aims to improve emergency clinical care through real-time information support, and provide benefit through information support for public health initiatives (Barthell *et.al.*, 2003).

Some of the projects still aim to demonstrate the utility of data mining techniques. For example, DTOX (Data Mining for Toxic Hazard Analysis) evaluates the usefulness of data mining techniques for food chemical hazard assessment. Data mining techniques are used to generate toxic hazard prediction rules. Data mining system Clementine is used to apply machine learning techniques to produce prediction rules from databases (DSS Consulting, 2003).

2.3. Criminal justice and defense

Data mining could be used in finding patterns in burglaries that are done by the same and perpetrators. West Midlands Police uses descriptive information about the thieves and the description of their modus operandi. By matching unsolved cases with known perpetrators, police officials hope to clear up old cases, and determine patterns of behavior (DSS Consulting, 2003). Another data mining system – Coplink, has been successfully deployed at the Tucson Police Department, where crime analysts, officers, detectives, and sergeants from 16 departmental units use the technology voluntarily as part of their daily investigative routine (Hauck, 2002).

Usage of data mining in detecting frauds in finance and health insurance has already been described in Finance and Economics part of the paper.

Cutting maintenance costs and in the same time improving readiness to respond quickly to smaller conflicts is one the many challenges that modern day armed forces face. High costs of the equipment are one of the major problems. The defense agency used exploratory data mining to understand the relationship between part failure and tank design, manufacture and usage. The agency built predictive models to streamline the maintenance processes by fixing more parts from the same tank at a time, increasing the amount of time the vehicle can be deployed in the field (SPSS, 2003).

Government intelligence agency focused on most likely security threats, and used Clementine to develop predictive intrusion models and deployed those models into an early warning system to focus personnel on most likely security threats. The data mining system answers questions like "What specific event is most likely to be a security threat?" (SPSS, 2003).

Defense Advanced Research Projects Agency (DARPA) is developing the database callled the Total Information Awareness System with the aim to monitor consumer purchases and government transactions as part of its effort to track terrorists and their activities. However, many critics feel that the project could threaten citizens' privacy, compromise the future of electronic commerce, and is a threat for security because such a massive database would be very attractive for hackers to go after (Onley, 2002).

However, IT infrastructure intrusions are one of the problems of government intelligence agencies, and human experts are usually employed in detecting unusual activity on the network. Still, too much activity data is collected each day and data mining system is used for identifying suspicious network activity that enables personnel to focus on investigation, rather than detection. The classification tree algorithm, clustering and outlier analysis are used to continually build predictive intrusion models on the latest data and deploy them into early warning systems (Bloedorn, 2003).

2.4. Labor and social welfare

Census data is one the most comprehensive databases. It usually collects population and other statistics essential to those that have to plan and allocate resources. Major customers include departments of local and national government, and providers of services such as health and education. Data mining system is used to calculate deprivation indices that are for example use d to measure correlation between the level of deprivation and a variety of health indicators (Klosgen et.al., 2003).

KESO (Knowledge Extraction for Statistical Offices) is a from Eurostat. Its goal was to produce a prototype Data Mining System that solves the needs of the analysts of statistical datasets. One of the examples is unemployment (Siebes, 1996). Similar application is reported by IBM (2003).

2.5. E-Government

The Internet offers a tremendous opportunity for government to better deliver its contents and services and interact with citizens, businesses, and other government partners. It is proposed that new database and data mining technologies could become the catalyst for encouraging information sharing and supporting collaboration and investigation among police departments, corrections offices, social services, and courts, which previously have been difficult to conduct (Chen, 2003). SPSS proposes use of Clementine in providing better online service by determine who the visitors are, why they visit the e-

government site, and how they use it (SPSS, 2003).

2.6. Education

A state's department of education wanted to explore the relationship between curriculum structure and standardized test performance to understand how course sequence affects test scores. Data mining system is used to uncover patterns in classes to identify the effects of curriculum structure on learning, to explore the relationship between the sequences of classes' ad test scores, and to maximize curriculum structure to ensure more effective learning (SPSS, 2003).

Data mining system predicted the possibility of returning to school for every student currently enrolled at a community college in Silicon Valley. The project applies neural network, C&RT and C5.0 to choose the best prediction followed by a clustering analysis (Luan, 2001). Similar application developed by SAS Enterprise Miner is used at the Baylor University (Campanelli, 2002).

2.7. Transport

Clementine was used to predict what type of transport people would use to make particular journeys. The data used was a detailed survey of the means of transport used by the population of the Ile-de-France, with 400,000 records. As the data was not originally intended to be used for data mining, a lot of pre-processing work had to be done. The rules were generated using the C4.5 algorithm, and proved both accurate and robust, covering a large population. Testing with a validation data set confirmed their quality (DSS, 2003).

3. Summary of data mining applications in public organizations

We are aware that it is not possible to find every one data mining application in public organizations by the search of the scientific data bases or Internet. However, the presented survey can give substantial insight into the current practice of data mining in public organizations. Most of the applications are described in articles published in 2003, and we can conclude that application of data mining in public organizations grows exponentially.

Table 1. shows areas of application. Finance and economy had the largest number of applications followed by healthcare, criminal justice and defense. Other areas have rather small number of applications.

Table 1. Area of application

Area of application	#	%
Finance and Economy	10	29%
Healthcare	8	24%
Criminal justice and defense	8	24%
Labour and social welfare	2	6%
E-Government	2	6%
Education	3	9%
Transport	1	3%
Total	34	100%

Most of the applications are described at business web sites, and the leader is SPSS followed by IBM (Table 2). It should be emphasized that only particular applications, and not advertisements, described at their web sites are taken into account. Only 21% of applications are described in scientific journals.

Table 2. Source of information

Source of the information	#	%
Business web site	21	62%
News web site	3	9%
Scientific journal	7	21%
Working paper	3	9%
Total	34	100%

Method used is only described at 18 sources. Classification and prediction is most often used, and is followed by concept/class description and evolution analysis.

Table 3. Method used

Method used	#	%
Concept/class description	3	17%
Classification and prediction	8	44%
Cluster analysis	2	11%
Outlier analysis	1	6%
Evolution analysis	4	22%
Total	18	100%

4. Conclusions

This paper has reviewed data mining applications in public organizations. Readers should be cautious in interpreting the results of the survey, since the findings are based on data collected from the business web sites, journal articles, news web sites and working papers. Such approach is employed because data mining applications in public organizations are still rarely described in journal articles. However, we feel that even such a survey can describe the current state in data mining applications in public organizations.

Most of the applications are described in 2003, which shows that in the years ahead data mining applications will be more often in public organizations. Most of the applications are in the area of finance and economics followed by healthcare, criminal justice and defense. Applications are in the most cases described at the business web sites, and only 20% of applications are published in scientific journals. The most often used methods are classification and prediction, concept/class description and evolution analysis.

It can be concluded that data mining methods and other related techniques of knowledge discovery in databases and intelligent data analysis are indispensable in public organizations. However, public organizations should give more attention to the privacy and data security issues than business organizations.

5. References

[1] Barthell, E.N., Pemble, K.R. The National Emergency Medical Extranet Project, 2003; Journal of Emergency Medicine, 24(1): 95-100.

[2] Benninga, S., Czaczkes, B. Financial Modeling. Cambridge, MA: The MIT Press; 1997.

[3] Berry, M.J.A., Linoff, G.S. Mastering Data Mining. New York: John Wiley & Sons; 2000.

[4] Bloedorn, E. Data mining for improving network intrusion and detection; 2003. http://www.spss.com/s5209/ppt/MITRE.Bloedorn.ppt

[5] Cahlink, G. Data Mining Taps the Trends. Government Executive Magazine; 2000. http://207.27.3.29/tech/articles/1000managetech.htm [10/01/2000]

[6] Campanelli, M. Baylor Makes the Grade With Recruitment Analysis; 2002. DM News. [10/17/2002]

[7] Carobne, P.L. Data Mining and the Government: Is There a Unique Challenge? The On-Line Executive Journal for Data-Intensive Decision Support;1998. http://www.tgc.com/dsstar/98/0519/980519.html [3/19/1998]

[8] Chen, H. Digital Government: technologies and practices; 2003. Decision Support Systems, 34(3), 223-227.

[9] DSS Consulting. Data Mining for Toxic Hazard Analysis, 2003. http://www.datamining.hu/angol/alk_egesz.html

[10] DSS Consulting. IAURIF – Traffic flow prediction in the Paris region, 2003. http://www.datamining.hu/angol/alk_koz.html

[11] DSS Consulting. West Midlands Police Crime Detection, 2003. http://www.datamining.hu/angol/alk_bun.html

[12] Gusfield, D. Algoritms on Strings, Trees and Sequences, Computer Science and Computational Biology. New York: Cambridge University Press; 1997.

[13] Han, J., Kamber, M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers; 2001.

[14] Hand, D., Mannila, H., Smyth, P. Principles of Data Mining. Cambridge, MA: The MIT Press; 2001.

[15] Hansen, J.V., Nelson, R.D. Neural networks and traditional time series methods: a synergistic combination in state economic forecasts. IEEE Transactions on Neural Networks 1997; 8(4): 863-873.

[16] Hauck, R.V., Atabakhsb, H., Ongvasith, P., Gupta, H., Hsinchun Chen. Using Coplink to analyze criminal-justice data; 2002. Computer, 35(3), 30-37.

[17] Haux, R., Ammenwerth, E., Herzog, W., Knaupp, P. Health care in the information society. A prognosis for the year 2013, 2002. 66(1-3), 3-21.

[18] Higgins, R.C. Analysis for Financial Management. Columbus: Irwin/McGraw-Hill; 1997.
[19] IBM. Business intelligence for revenue and fiscal management; 2003. http://www-1.ibm.com/industries/government/doc/content/solution/262217109.html
[20] IBM. DecisionEdge for Fraud and Abuse Management, 2003; http://www-3.ibm.com/software/data/bi/decisionedge/defam.htm
[21] IBM. Business Intelligent Services for Government Finance, 2003; http://www-3.ibm.com/software/data/bi/press/bistore.htm
[22] IBM. University of Pennsylvania; 2003. http://www-3.ibm.com/e-business/doc/content/casestudy/24289.html
[23] Klosgen, W., May, M. Census Data Mining: An Application; 2003. http://www.di.uniba.it/~malerba/activities/mod02/pdfs/kloesgen.pdf
[24] Luan, J. Data Mining as Driven by Knowledge Management in Higher Education; 2001. http://www.cabrillo.cc.ca.us/oir/oir_reports/UCSFpaper.pdf
[25] Makulowich, J. Government Data Mining Systems Defy Definition. Washington Technology; 1999. http://216.70.54.91/news/13_22/tech_features/393-1.html [02/22/1999]
[26] Mattison, R. Data Warehousing and Data Mining for Telecommunications. Fitchburg, MA: Artech House; 1997.
[27] Michael, H. Teradata Takes Data Mining Beyond Beer and Diapers; 2002. Transforming Transactions into Relationships. http://www.ncr.com/media_information/2002/jul/pr073102b.htm [06/31/2002]
[28] Onley, D. S. DARPA's plans for data mining draw criticism. Government Computer News; 2002. http://www.gcn.com/21_34/storage/20637-1.html [12/16/2002]
[29] R. Haux, E. Ammenwerth, W. Herzog, P. Knaup, Health Care in the Information Society: a Prognosis for the Year 2013, 2002; International Journal of Medical Informatics, 66 (1-3): 3-21.
[30] SAS. Public Sector News and Events; 2003. http://www.sas.com/solutions/public_sector/news_events/
[31] Siebes, A. Data Mining for Professional Statisticians; 1996. http://www.ercim.org/publication/Ercim_News/enw25/siebes.html [06/31/2002]
[32] SPSS. Business Intelligence in Government Security; 2003. http://www.spss.com/spssbi/applications/government/brochures.htm
[33] SPSS. Clementine for Public Sector; 2003. http://www.spss.com/spssbi/applications/government/brochures.htm
[34] SPSS. Deliver effective online services to citizens; 2003. http://www.spss.com/applications/government/e_government.htm
[35] SPSS. Fight fraud, waste and abuse with powerful analytics, 2003. http://www.spss.com/applications/fraud/
[36] SPSS. Generate maximum return on data in minimum time with Clementine; 2003. http://www.spss.com/spssbi/clementine/
[37] Sund, R. Utilization of Administrative Registering using Statistical Knowledge Discovery. National Research and Development Centre for Welfare and Health, 2002; Working Paper.
[38] Valdes-Perez, P. Principles of Human-Computer Collaboration for Knowledge Discovery in Science. Artificial Intelligence 1999; 107: 335-46.
[39] Waterman, M.S. Introduction to Computational Biology: Maps, Sequences, and Genomes (Interdisciplinary Statistics). London: CRC Press; 1995.
[40] Zvarova, J., Pribik, V. Information society in Czech healthcare `starting point' to prognosis for the year 2013, 2002; International Journal of Medical Informatics, 66 (1-3): 59-68.