Analysis of the possible application of Data Mining, Text Mining and Web Mining in Business Intelligent Systems

T. Atanasova, M. Kasheva., S. Sulova and J. Vasilev

Department Informatics
University of Economics - Varna
Bulgaria, Varna, 77 Kniaz Boris I blvd
Phone: (359 52) 660-458 Fax: (359 52) 660-475 E-mail: t_atanasova@ue-varna.bg

Abstract

The design and implementation of a contemporary Business Intelligent System (BIS) requires processing of all types of information – structured, semi-structured, unstructured and real time. The basic functionality of BIS is implemented in an enterprise local area network (LAN) or by web access to web sites. Business results are bettered by the application of the following technologies: Data Mining (DM), Text Mining (TM) and Web Mining (WM). As a result companies get competitive advantage. Applications of BIS in enterprises are structured in a functional matrix. Its dimensions are managerial duties and subsystems of a company. Several problems are analyzed. They may be solved by the application of technologies DM, TM and WM in BIS. This processes lead to innovative solutions.

I.INTRODUCTION

New information technologies allow computers to extract meaning from unstructured information and directly process content. This will enable new kinds of innovative solutions to business problems. Such technologies that have been under development for the last few years and are shown great potential are Web mining and Text mining. They use methods from the well known technology DM to analyze unstructured data.

Web Mining (WM) is a process of application of DM for extraction of knowledge from the web – documents, hyperlinks, tags, http logs, app server logs and others. WM is studied in three categories:

Web content mining - information extraction from web documents. Needs of Internet users are studied. People searching patterns are constructed.

Web structure mining-Web sites structure (topology) is studied. Main research may be grouped in two subcategories: pattern extraction from hyperlinks and document structure studies. A tree-like structured is used for the description of tags (HTML, XML).

Web usage mining - provides patterns search within the flows of clicks or generated by end-user interaction with web sites. The gained information is in a set of sites, object and other web resources which are mainly accessed by groups of people with common interest.

Text mining (TM) is used for automated analysis of unstructured texts. Structured information is extracted and integrated with other structured data for further analysis and knowledge mining [6].

Figure 1 shows the common and different characteristics of DM, TM and WM.

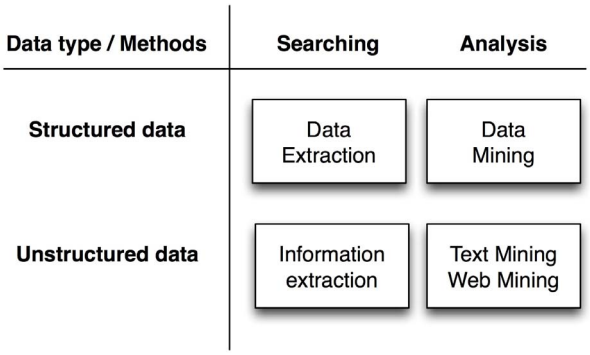

Fig. 1. Diagram of common and different characteristics of DM, TM and WM

The common characteristic of the three technologies is the searching and finding of unknown dependencies and patterns in data and the application of DM techniques in a certain stage of execution of each of them.

II. DM AND BASIC FUNCTIONALITY OF BIS

Many organizations have recognized the potential of these technologies as a way to increase their business value. DM technologies are the basic tool for analysis in business intelligent systems. If we look deeper the link between the basic functionality of BIS, DM technology is applied for knowledge management, analysis and forecasting of information (fig. 2).

Function "Knowledge management: is intended for processing of unstructured and semi-structured information in small and medium sized enterprises (SMEs). Structured information is used for pattern extraction or for a fact database. Technologies DM, WM and TM are applied. The gained results may be automatically converted in a knowledge representation model or they may be used as a fact database.

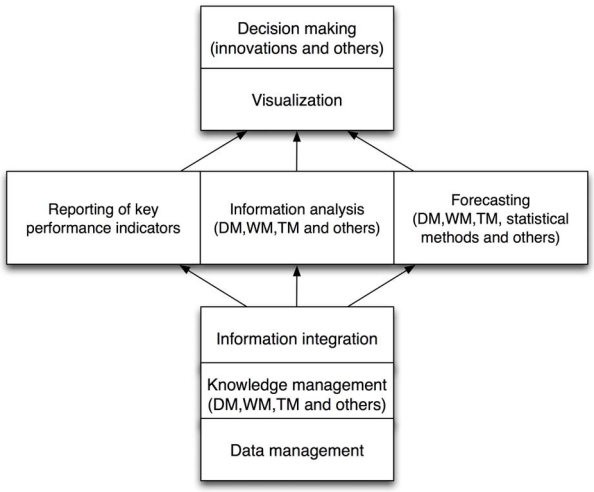

Fig.2. Hierarchy of basic functionality of BIS

The function "Information analysis" presents multi-dimensional analysis, ad-hoc queries and knowledge and data analysis. It uses data from lower levels. Some of the popular applications of DM are: analysis of manufacturing, analysis of marketing campaigns, revealing of unauthorized access to computer systems and others.

DM is used in function "Forecasting". It is used for deriving of trends on the basis of historical data. The basic types of tasks, which are solved by DM, are:

1) Classification

2) Clustering (segmentation)

3) Estimation, identification and finding out of rare cases, which stay significantly away from norms.

4) Forecasting and trend discovery

5) Association

6) Analysing

III. FUNCTIONAL MATRIX FOR THE APPLICATION OF DM, WM AND TM

A complete functional matrix is created (Table I) to examine managerial functions (by rows) and subsystems of a company (by columns) by the application of DM, TM and WM. Each cells of the matrix describes tasks which may be solved by DM, WM or TM. Table 1 shows world applications about surveyed companies – companies about which we have information.

The application of DM, WM and TM are mainly in marketing. This article presents the solution of some problems in the marketing subsystem.

The existence of a date warehouse for customers and activity of competitors is the ideal starting point for the application of DM. Data about potential customers is also a good starting point. If a company uses e-commerce, components of a successful DM are present: a lot of reliable structured data and a good possibility of estimation of return of investments (ROI).

TABLE I
FUNCTIONAL MATRIX FOR THE APPLICATION OF DM, WM AND TM

Managerial functions	Subsystems				
	Marketing	Accounting	Basic activity	Logistics	Management of human resources
Aiming	Estimation		Estimation		Classification, Estimation
Planning and forecasting	Forecasting, classification, clustering	Forecasting, estimation	Forecasting, estimation	Forecasting, estimation	
Organizing	Clustering, classification, consequent event discovery, WM, TM		Estimation	Estimation, association	
Control	Estimation, classification	Identification	Estimation, classification	Consequent event discovery, Estimation	Estimation, forecasting, association
Leading (projects, investments)	Forecasting, estimation **WM,TM**	Forecasting, estimation	Estimation	Identification	

Tasks which are solved using DM in marketing are the following:

- classification of customers. The consumer behavior is analyzed and recommendations for new products and customization of products are offered;

- clustering (segmentation). One of the basic application of DM in marketing is market segmentation. Its purpose is creation of products, services and marketing messages for each segment. Another common task is customization of goods. The product list and the list of services may be optimized. Classification and

clusterization of customers may lead to purposeful market campaigns and identification of target groups.

- estimation. The consumer behavior is modeled and an estimation is made about the adoption of a certain product, shop or marketing campaign.

We made an experiment for the evaluation of customer preferences to local chains of supermarkets. The survey covers 100 real customers in three of these chains. The survey data is used as a learning set for the preparation of a neural network. The model consists of seven factors which influence the customer choice: price, assortment, availability of goods, quality of goods, atmosphere, and quality of direct service and image of the supermarket. Factors are chosen on the basis of expert estimation method. Two output variables are defined "Supermarket preference" and "Shopping frequency".

The neural network is a multi-layer perceptron (7-6-2). It is learned by the back propagation method (fig. 3). The learning error is 0.05. After a successful testing of the network, the relative importance of input variables is reported starting with the most important ones.

1) Prices of goods
2) Atmosphere in the supermarket
3) Quality of goods
4) Availability of goods
5) Image of the supermarket
6) Assortment
7) Servicing

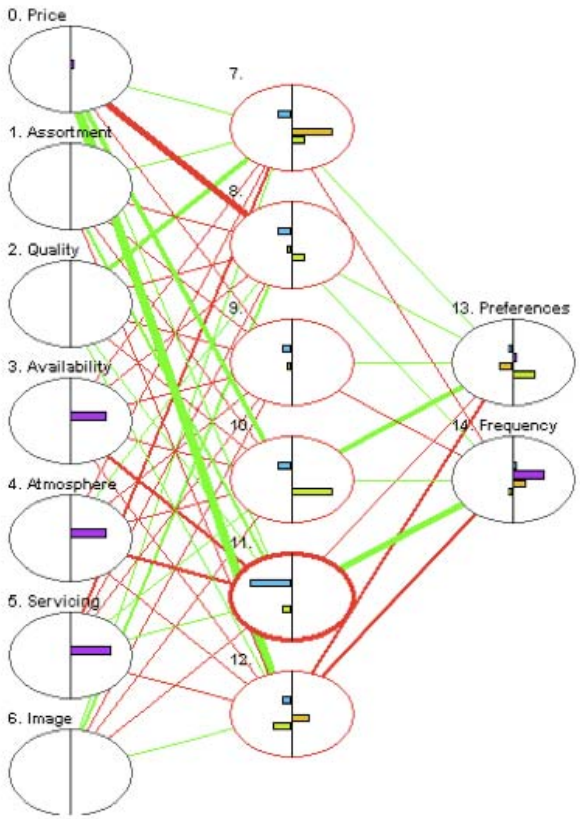

Fig.3. Neural network, used in the experiment

The results are logical and they describe correct dependencies between the input variables and the output ones – supermarket preference and shopping frequency.

The experiment shows the application of DM techniques for the estimation of the relative importance of the chosen variables in marketing.

Another solved task is the estimation of customer satisfaction about delivery conditions, trade strategy – prices, discounts and promotions. Possibilities for decreasing costs about marketing campaigns are revealed. Supplier choice, channel and time of distribution are performed.

- forecasting. DM technique is used for forecasting market supply and demand. The assortment of products and the set of services are planned (fig. 4).

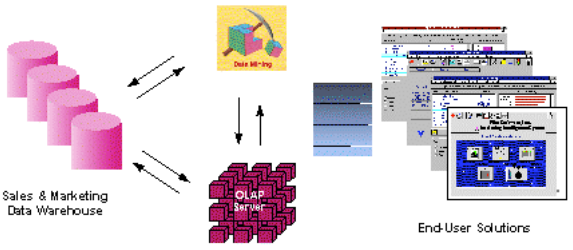

Fig.4. Integrated DM architecture [8]

- association. A common task is the identification of goods, which are apparently not connected but are sold together. This technique allows managers to provide them together and to keep them in stock.

- analyzing. The consumer behavior is analyzed. Orders for new supplies are generated. As a result, customers make regular orders and the get loyal. On-line stores such as Amazon.com and Barnesandnoble.com are good examples. When a customer orders a stock, he gets a recommendation to buy other stocks.

Companies which are working in the sphere of electronic commerce and services apply mainly web usage mining (WM). Consumer behavior and top selling products are analyzed on the basis of web sites usage.

Web content mining is used for searching and filtering of information, categorizing of documents, searching of similar web pages on different servers, identification of themes in different web documents.

The gained model (by means of WM (Web structure mining)) may be used for categorization of web pages and generation of information about similarity and links between web pages.

Text mining (TM) is used for navigation in text databases. TM helps the work with great text databases electronic flows of news, e-mail and others. TM is used for automatic classification of incoming e-mail messages for different groups and tasks. By clusterization of documents, TM tools may help the effectiveness of published news about a company.

III. CONCLUSION

Several conclusions about the application of DM, TM and WM may be made on the basis of the analysis of their application.

Technologies have an important role in the basic functions of BIS for enterprises. They may be used for distribution of investments and costs. As a result they achieve better financial results.

The availability of structured and unstructured data is the main reason for the application of DM, TM and WM for the sake of solving a great number of managerial tasks in subsystem “Marketing”.

REFERENCES

[1] A. Abraham, Business Intelligence from Web Usage Mining. <http://www.softcomputing.net/jikm.pdf>

[2] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making. *Wiley*, 2009.

[3] M. Berry and G. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, *Wiley Computer Publishing*, 2004.

[4] T. Nadeau and others Data Mining. Know It All. *Morgan Kaufmann*, 2008.

[5] Y. Ohsawa and K. Yada, Data Mining for Design and Marketing. *Chapman & Hall*, 2009.

[6] Z. Zhou, Applications of Data Mining in E-business and Finance, *IOS Press*, 2008.

[7] An Introduction to Data Mining. <http://www.thearling.com/text/dmwhite/dmwhite.htm>

[8] Data Mining – problems and new solutions (in Russian), <http://www.inftech.webservis.ru/it/database/datamining/ar1.html>.

[9] The BI project – a step to a higher analytical business level, CIO, 2008 (in Bulgarian), <http://www.imecominc.com/downloads/datasheets/dmac.pdf>.

[10] Web Content Mining. <http://www.cs.uic.edu/~liub/WebContentMining.html.

[11] Web Mining: Machine learning for Web Applications. <http://ai.arizona.edu/intranet/papers/WebMining.pdf>.

[12] http://www.thearling.com/.

[13] http://www.cs.uic.edu/~liub/WebMiningBook.html.

[14] http://hwanjoyu.org/teaching/2009-1-introdm/schedule/01.ppt.

[15] http://robotics.stanford.edu/users/ronnyk/integratingEcom.pdf.