Integration of Text- and Data-Mining Technologies for Use in Banking Applications

Jacek Maslankowski

Department of Information Systems, University of Gdańsk, Poland.
jacek@univ.gda.pl

Introduction

Unstructured data, most of it in the form of text files, typically accounts for 85% of an organization's knowledge stores, but it's not always easy to find, access, analyze or use (Robb 2004). That is why it is important to use solutions based on text and data mining. This solution is known as duo mining. This leads to improve management based on knowledge owned in organization. The results are interesting. Data mining provides to lead with structuralized data, usually powered from data warehouses. Text mining, sometimes called web mining, looks for patterns in unstructured data – memos, document and www. Integrating text-based information with structured data enriches predictive modeling capabilities and provides new stores of insightful and valuable information for driving business and research initiatives forward.

Methods for Integration Data- and Text-Mining

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases (Hand et al. 2001). Data mining has become useful over past decade in business to gain more information, to have a better understanding of running a business, and to find a new ways and ideas to extrapolate a business to other markets (Bargain et al. 2002). Data mining involves extraction, transformation and presentation of data in useful form. Creating a mining model can be compared to manufacturing process - from data by the algorithm towards the mining model (Paul et al.2002). Integrate methods of text analysis with methods for data analysis may take more profits to organizations. Data and text mining is the element of business Intelligence system, which contains software for supporting decisions. Business Intelligence means using data assets to make better business decisions. It is about access, analysis, and uncovering new opportunities (Almeida et al. 1999). The source to powered Business Intelligence systems is metadata. Several factors have triggered the need for metadata in businesses today. These include the following:

- Current systems are inflexible and nonintegrated.
- Existing data warehouses and data marts need to grow.
- Business users needs are not being fulfilled.
- Companies need to reduce the impact of employee turnover.
- Businesses need to increase user confidence in data (Marco 2000).

A major difficulty with the dataset usually used in the data mining model is its relational structure (Grant 2003). The problem has been solved by leading with dimensional structures. The data mining model describes where the source of data is that is used to train the model is stored. This source can be an OLAP cube in which the model is called MOLAP (Multidimensional OLAP). The second type is called ROLAP (Relational OLAP). A third option which attempts to take the best of each is called HOLAP (Hybrid OLAP). Data Mining model algorithms include classification, clustering, descriptive and predictive models (Bain et al.2001). The heart of data mining systems is the data warehouse. OLAP allows users to view information from many angles, drill down to atomic data, or roll up into aggregations to see trends and historical information. It is the final and arguably most important piece of the data warehouse architecture (Moeller 2001). A data warehouse is subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision (Inmon 2002). Subject areas are major grouping of physical items, concepts, events, people, and places of interest to the enterprise (Imhoff 2003).The data warehouse contains fact and dimension tables. Facts represents a business measure, while the dimension tables contain the textual descriptors of the business (Kimball and Ross 2002). The results of requirements analysis and source system audit serve as inputs to the design of the warehouse schema. The schema details all fact and dimension tables and fields, as well as data sources for each warehouse field (Humphries et al. 2001). The design of the data warehouse includes star and snowflake schemas (Scalzo 2003). Functions in text analysis are to select features for further processing. Text mining is needed to process text into a form that data mining procedures can use (Weiss 2004). This usually contains language identification, feature extraction, clustering and categorization. The language identification tool can automatically discover the language in which a document is written. Feature extraction recognizes significant vocabulary items in text. Clustering is a process which divides a collection of documents into groups. Categorization tools assign documents to preexisting categories, sometimes called "topics" or "themes" (Tkach 1998). A document warehousing is a technology which leads with text documents. It is characterized by following attributes: there is no single document structure or document type, documents are drawn form multiple sources, essential features of documents are automatically extracted and explicitly stored in the document warehouse, which are designed to integrated semantically related documents (Sullivan 2001).

Data preparation is one of the most important step in the model development process. From the simplest analysis to the most complex model, the quality of the data going in is key of success of the project (Rud 2001).

DEA (Data Envelopment Analysis) is the mathematical programming approach developed to evaluate the relative efficiency of a set of units that have multiple performance measures – inputs and outputs. DEA is particularly useful when the relationships among the multiple performance measures are unknown (Wang 2003).

Research Method

There are three main aspects of the text- and data- mining integration. First aspect concerns source of data. The data warehouse is the best source used for integration data from various OLTP systems and for using to transform into business information easily understood by tools and decision workers. This leads to create a source for data mining application. Document warehouse can be used as the source for text mining tools. Second aspect is to find a logical model, which integrates both technologies and provides useful information for decision workers. Vendors, such as SAS Institute, provides solutions to build data mining models in its enterprise miner, and text mining models in text miner. There is no simple solution to join both applications. Third aspect is how to build the presentation layer, which can visualize data taken from both sources.

The data warehouse build for using in banking organizations should cover all aspects of its organizational structure. In it's operating with data mining model particularly static activity should be included. The next step should begin process of determining sources for document warehouse. These sources are often documents, applications and any notes from clients. The most important thing is that the warehouse must properly process the documents. The least efficient warehouse contains errors in its data. Research leads by The Data Warehousing Institute has argued that bad quality of data in USA wasted about 600 milliards dollars per year. The proposal is to implement methods to prevent data errors in warehouses in this first layer of data- and text- mining integration.

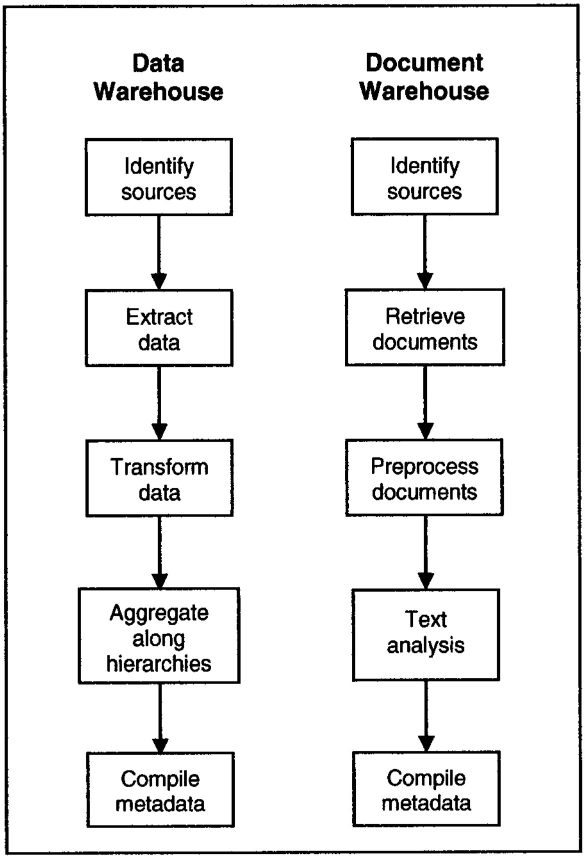

Fig. 1. Steps in document warehouse construction compared to data warehousing

As showed in figure above, the last step in building data and document warehouses is compiling metadata. In banking applications metadata used with both warehouses are similar. It is important that metadata types used in document warehouse should be similar to data warehouse. This can improve process of modeling integration tools for supporting both technologies. Using common description for metadata can provide a model which can be split by external application.

The next step is to build a solution for supporting both data- and text mining technology. The main task is to find a pattern in data or text and then use it to provide to upper layers of this proposal. This model should support reporting based on both technologies. In authors' opinion, the best way is to use external application. It can safe from losing coherence in data and document warehouses. This model has been tested based on SAS 9.1.3 software. For data mining technology was used SAS Enterprise Miner, text mining was supported by SAS Text Miner. Both technologies were split based on Sun Java architecture, supported by SAS application. In this solution there has been added indexing table, which split indexing tables from text and data mining and a search engine has been developed.

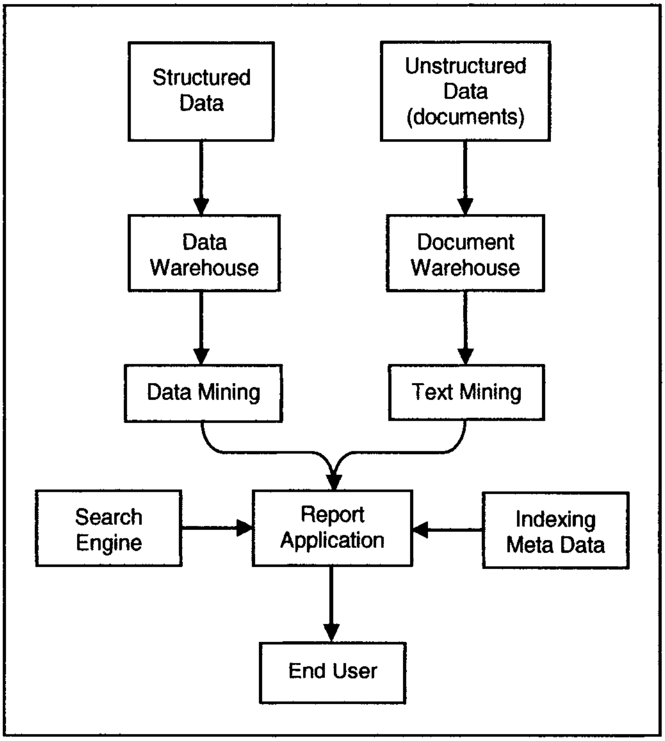

Fig. 2. Integration of data- and text-mining technologies on upper layer

The last step is to provide visualization layer for this technologies. The technology used in this model was Java Swing, based on Sun Microsystems applications. End user can request for any data from any criteria.

Results

The result is that there is a multiplier effect going on. By using data mining and text mining together, enterprises have been able to improve its functioning to around 20 percent, with the range being from 5 to 50 percent

(Creese 2004). The effect is particularly visible while loan analysis. The documents from customer can be used as a pattern for text mining tools, while his credit account can be analyzed by using data mining technology. For example a customer cannot pay rate because of a random accident. He could write a document with explaining this situation. By using only data mining tools it will not be as simple to discover the problem in payment. By using both technologies the system can compare these two patterns and discover why the client cannot pay this payment. It can take small amount of time to do this, while using only one technology can lead to necessary analyze both documents by the person.

Conclusion

The present paper has identified aspects in modeling text- and data mining technologies in tandem. Nowadays duo mining technologies are not widely used in organization. There is no literature and experience for using this technology in organizations of all kinds. In authors' opinion, this will be increasing. This requires a new approach to develop text and data mining tools. The banking organization is the best practice for using the solutions proposed by author. The application for this solution has been developed and tested in SAS 9.1.3 system.

References

Almeida M, Ishikawa M, Reinschmidt J, Roeber T (1999) Getting Started with Data Warehouse and Business Intelligence. IBM Press San Jose

Bain T, Benkovich M, Dewson R, Ferguson S, Graves C, Joubert TJ, Lee D, Scott M, Skoglund R, Turley P, Youness S (2001) Professional SQL Server 2000 Data Warehousing with Analysis Services. Wrox Press Ltd

Baragoin C, Chan R, Gottschalk H, Meyer G, Pereira P, Verhees J (2002) Enhance Your Business Applications. IBM Redbooks

Berry MW (ed) (2004) Survey of Text Mining. Springer

Creese G (2004) Volume Analytics: Duo-mining: Combining Data and Text Mining. DMReview September 2004

Grant G (2003) ERP & Data Warehousing in Organizations: Issues and Challenges. Idea Group Publishing

Hand D, Mannila H, Smyth P (2001) Principles of Data Mining. The MIT Press Cambridge

Humphries M, Hawkins MW, Dy MC (2001) Data Warehousing: Architecture and Implementation. Pearson

Imhoff C, Galemmo N, Geiger J (2003) Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley & Sons

Inmon WH (2002) Building the Data Warehouse. Third Edition. John Wiley & Sons

Kimball R, Ross M (2002) The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Second ed. John Wiley & Sons

Marco D (2000) Building and Managing the Metadata Repository: A Full Lifecycle Guide. John Wiley & Sons

Moeller RA (2001) Distributed Data Warehousing using Web Technology: How to Build a More Cost-Effective and Flexible Warehouse. Amacom

Paul S, Guatam N, Balint R (2002) Preparing and Mining Data with Microsoft® SQL Server™ 2000 and Analysis Services. MS Press

Rob D (2004) Text Mining Tools Take on Unstructured Data. Computerworld June 21 2004

Rud OP (2001) Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management. John Wiley & Sons

Scalzo B (2003) Oracle DBA Guide to Data Warehousing and Star Schemas. Prentice Hall

Sullivan D (2001) Document Warehousing and Text Mining. Wiley

Tkach D (1998) Text Mining Technology: Turning Information into Knowledge. IBM Corporation

Wang J (ed) (2003) Data Mining: Opportunities and Challenges. Idea Publishing Group

Weiss SM, Indurkhya N, Zhang T, Damerau F (2004) Predictive Methods for Analyzing Unstructured Information. Axel Springer