Karolina Kuligowska, Mirosława Lasek

Chair of Informatics and Economic Analysis,
Faculty of Economic Sciences, Warsaw University

TEXT MINING APPLICATIONS IN BUSINESS

Abstract: Text Mining methods combine Data Mining techniques and extraction of information from natural language text. Those methods enable to uncover interesting patterns from large text databases rather than from structured databases of facts. Previously unknown, discovered knowledge can enrich predictive modeling of business endeavors.

Key words: text mining, natural language processing, information extraction, knowledge discovering.

1. Introduction

Since the beginning of research in the field of Artificial Intelligence people have been trying to construct a software capable to explore data in an intelligent way. As the technology developed, the more and more powerful computers conducted the more and more advanced data analyses. Programs could explore and analyse large databases faster and better than people did. Nevertheless, computers were still unfamiliar with one of the most important human domain: natural language understanding and processing. Further scientific research in computational linguistics turned out to be successful enough in order to introduce on market software that analysed text. Natural language engineering combined with statistics and Data Mining methods enabled to create Text Mining tools. These tools help extracting pertinent information from text documents and discovering unknown patterns and linkages between those documents. Text Mining turned to be a very promising research topic. Text Mining tools can save time and earn money that would be spend on reading and exploring large document collections. Those tools have already many business applications, and some of them will be discussed in this article.

2. Text Mining methods

In Data Mining the patterns are extracted from structured databases. To the contrary, in Text Mining the information is explored from unstructured documents that are composed in natural language. In order to conduct an analysis, text documents must first be represented in a form that can be mined. This preliminary step is called text preprocessing. During preprocessing, textual data is imported from various external text repositories and transformed into single, easy to read data set.

When raw documents are transformed into the uniform coding format, text is parsed and explored. This is the next step to analyse the document collection. During parsing, low-information words such as articles, conjunctions and prepositions are excluded from the analysis. Stop lists, that contain low-information words, help to control which words to remove. Parsing step enables to break documents into terms

and to extract from these terms particular entities that are meaningful to further application. For example, parsed terms may include words, phrases, entities and numbers.

Stemming is another domain of parsing, and means finding base form of a word. It enables to better analyze more abstract terms. If stemming is applied, all declined words in document collection are grouped under the root form of a noun. Nevertheless, there exist terms that do not have the same root form but share the same meaning in relation to the context. In this case synonym lists are indispensable. They consist of words that should be treated equivalently but are not direct stems.

Finally, the process of parsing enables to generate a quantitative representation for the document repository. It can be covered by simple statistical methods based on frequency and co-occurence. In order to consider the word counts in the document collection, a term-by-document frequency matrix serves as the base for further analysis of the document collection. Frequency weights depends on the term frequency - alone and in the document collection as a whole. Those methods can opionally be followed by text filtering, question answering and fact extraction.

The purpose of Text Mining is to process documents containing unstructured text, extract meaningful terms from the text, and make the information contained in the text accessible to the various Data Mining algorithms. Information can be extracted to create summaries for the documents based on the words contained in them, to determine similarities between many documents or to show how they are related to other variables of interest in the data set. Hence, Text Mining is a powerful tool that turns text into meaningful indices, which can then be later incorporated in other analyses such as predictive Data Mining modelling or the business application of clustering and classification.

3. Business applications of Text Mining

The goal of Text Mining is data exploration, prediction and business decision making. To reach that goals, business analysts use documents clustering and documents classification.

Performing clusters within the document collection enables to summarize it without reading every document. Clustered set of thousands of documents can reveal pertinent topics and key ideas that exist in the collection. Applications of clustering include analyzing the contents of survey data, customer comments or set of e-mail to discover previously unknown knowledge. Clustering can also give an indication of what words tend to be used together or what kind of vocabulary is used in a collection.

During classification the documents are sorted into predefined categories. In fact, classification is already a form of prediction. Classification method is often applied to filter e-mails and customize them or automatic detect of a spam.

The most spectacular and promising use of Text Mining prediction covers such areas as: stock exchange, call center services pricing, e-mail filtering and anti-spam actions, customer satisfaction and behaviour predicting, identification of authorship, market intelligence.

Some of typical applications of Text Mining include:

- price changes on stock exchange predicted from news announcements about companies' financial conditions;
- cost of services predicted from description of the problem;

- identifying specific words and phrases for the process of filtering e-mails in order to detect spam;
- customer satisfaction predicted from analyses of survey data and customer on-line comments;
- examining samples of articles written by one person can prove the authorship of another fragment that has several potential authors.

Other applications of Text Mining can cover: analyzing open-ended survey responses, automatic processing of messages, analyzing warranty or insurance claims, and diagnostic interviews.

4. Further applications: Web Mining

Web mining tools enable to explore the data spread over the World Wide Web. Web mining can be divided into three areas of research: Web content mining, Web structure mining and Web usage mining. The first approach focuses on retrieval of the useful information from the Web contents and documents. The area of Web structure mining is a category which discovers how to model the link structures. The last approach refers to the techniques that try to predict the users' behaviour on the basis of previously discovered patterns specific to each of the users.

The Web has an enormous impact on today human society. It has changed ways of exchanging information and data collection. The Internet is considered as the largest information source on the earth. One can call it an incredibly huge repository of unstructured data. It is not surprising therefore, that in the consequence of applying Data Mining techniques to the textual data, researchers also tried to apply those techniques to mine data stored in the Web.

The Web users take advantage of searching tools: they find important information precisely and effectively. On the other hand, the Interenet services providers try to predict the users' behaviours and navigation patterns to reduce the traffic load and personalize provided information. The business analysts aim on learning, understanding and predicting the users' preferencies and needs. All of three mentioned groups expect to find suitable Web mining tools that will help them to solve the problems concerning enormous amount of data stored in the Web.

Web mining contributes to business activities by effectively exploring Web portals. It enables to presonalize web sites precisely by tracking users' navigating patterns and customizing Web content especially for each of them. The usage of knowledge about customers and their preferencies is relevant to make marketing decisions. On the basis of on-line polls, customer satisfaction can be measured and analyzed.

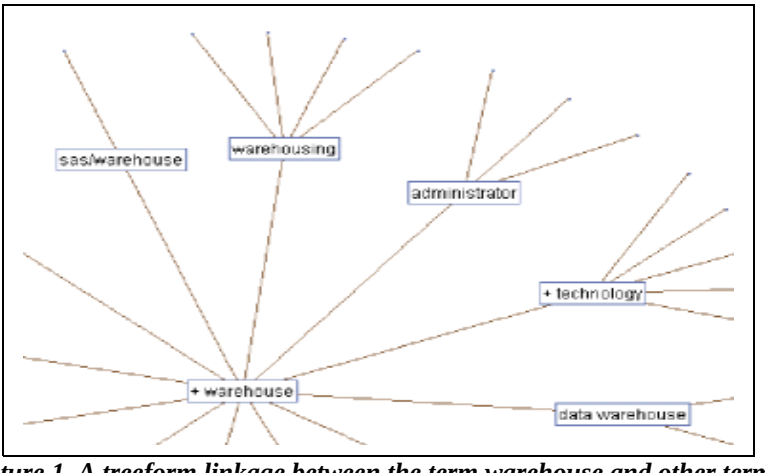
The tremendous growth of information sources available on the Web and the interest in e-commerce makes Web mining a very dynamic scientific area, converging from several research fields such as information retrieval, and artificial intelligence especially from natural language processing.

5. Text Mining tools

Among Text Mining software the most frequently used tools are: Text Miner (SAS), Text Miner (StatSoft), Text Analyst (Megaputer Intelligence), and TxtKit (Schoenerwissen). Each of the mentioned tools analyses text that exists in a document

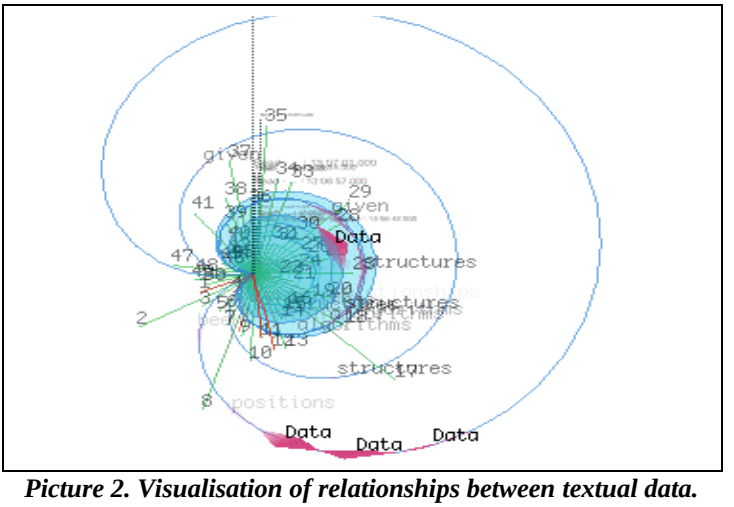
collections, and sets up the data for predictive mining and further exploration of the textual data. They enable efficiently navigate through large textbases and view accurate summaries before plunging into full documents. The capabilities of those tools include also: clustering documents into meaningful groups, automatic recognition of multiple-word terms, and finding similar terms or documents.

Using SAS Text Miner one can explore the linkage between terms. The map of words that commonly occur together is displayed in a treeform, as shown on the picture below.



Picture 1. A treeform linkage between the term warehouse and other terms.

TxtKit is an Open Source visual text mining tool for exploring large amounts of multilingual texts. It's a multiuser-application which mainly focuses on the process of reading and reasoning as a series of decisions and events. The visualization is based on the users' actions, statistical information about the data as well as collaborative filtering schemes. The picture below presents one of the graphical visualisation of relationships between textual data.



Picture 2. Visualisation of relationships between textual data.

The output visualisation is a useful outcome because it clarifies the underlying structure of what is contained in the input documents. For marketing, biomedical and many other research purposes, it can be a useful and significant result. All the above mentioned functionalities makes that nowadays Text Mining software become very flexible tool that can be used to solve a variety of problems and business applications.

ZASTOSOWANIA TEXT MINING W BIZNESIE

Streszczenie: Metody Text Mining stanowią połączenie technik eksploracji danych (Data Mining) z przetwarzaniem zawartości różnorodnych dokumentów zapisanych w języku naturalnym. Metody te stosuje się do odkrywania i wykorzystywania nie znanej wcześniej wiedzy z nieustrukturyzowanych danych o charakterze tekstowym. Wyłuskane informacje tekstowe są w coraz większym stopniu wykorzystywane w modelowaniu i prognozowaniu oraz podejmowaniu decyzji biznesowych.

Słowa kluczowe: text mining, przetwarzanie języka naturalnego, wyłuskiwanie informacji, odkrywanie wiedzy.

References

[1] Filipowska A.: Jak zaoszczędzić na czytaniu? Automatyczne tworzenie abstraktów z dokumentów, Gazeta IT, nr 3 (22), 2004.
[2] Gaizauskas R., Saggion H.: Multi-Document Summarization by Cluster/Profile Relevance and Redundancy Removal, HLT/NAACL Document Understanding Workshop, 2004, 1-8.
[3] Hearst M.A.: Untangling Text Data Mining, Proceedings of ACL-99: the 37th Annual Meeting of the Association for Computational Linguistics, 1999, 3-10.
[4] Night K.: Mining Online Text, Communications of the ACM 42, Nr 11, 1999, 58–61.
[5] Węcel K.: SAS, rejestry państwowe i text mining, Gazeta IT, nr 9 (39), 2005.
[6] Weiss S. et all.: Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer Science+Business Media, New York, 2005.