Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume II
Data Pro–I
Discovering Unknown Patterns in Free Text

Jan H Kroese  
*University of Pretoria, South Africa*

Machdel C Matthee  
*University of Pretoria, South Africa*

**INTRODUCTION**

A very large percentage of business and academic data is stored in textual format. With the exception of metadata, such as author, date, title and publisher, this data is not overtly structured like the standard, mainly numerical, data in relational databases. Parallel to data mining, which finds new patterns and trends in numerical data, text mining is the process aimed at discovering unknown patterns in free text. Owing to the importance of competitive and scientific knowledge that can be exploited from these texts, “text mining has become an increasingly popular and essential theme in data mining” (Han & Kamber, 2001, p. 428).

Text mining is an evolving field and its relatively short history goes hand in hand with the recent explosion in availability of electronic textual information. Chen (2001, p. vi) remarks that “text mining is an emerging technical area that is relatively unknown to IT professions”. This explains the fact that despite the value of text mining, most research and development efforts still focus on data mining using structured data (Fan et al., 2006).

In the next section, the background and need for text mining will be discussed after which the various uses and techniques of text mining are described. The importance of visualisation and some critical issues will then be discussed followed by some suggestions for future research topics.

**BACKGROUND**

Definitions of text mining vary a great deal, from views that it is an advanced form of information retrieval (IR) to those that regard it as a sibling of data mining:

- Text mining is the discovery of texts.
- Text mining is the exploration of available texts.
- Text mining is the extraction of information from text.
- Text mining is the discovery of new knowledge in text.
- Text mining is the discovery of new patterns, trends and relations in and among texts.

Han & Kamber (2001, pp. 428-435), for example, devote much of their rather short discussion of text mining to information retrieval. However, one should differentiate between text mining and information retrieval. Text mining does not consist of searching through metadata and full-text databases to find existing information. The point of view expressed by Nasukawa & Nagano (2001, p. 969), to wit that text mining “is a text version of generalized data mining”, is correct. Text mining should “focus on finding valuable patterns and rules in text that indicate trends and significant features about specific topics” (ibid., p. 967).

Like data mining, text mining is a proactive process that automatically searches data for new relationships and anomalies to serve as a basis for making business decisions aimed at gaining competitive advantage (cf. Rob & Coronel, 2004, p. 597). Although data mining can require some interaction between the investigator and the data-mining tool, it can be considered as an automatic process because “data-mining tools automatically search the data for anomalies and possible relationships, thereby identifying problems that have not yet been identified by the end user”, while mere data analysis “relies on the end users to define the problem, select the data, and initiate the appropriate data analyses to generate the information that helps model and solve problems those end-users uncover” (ibid.). The same distinction is valid for text mining. Therefore, text-mining tools should also “initiate analyses to create knowledge” (ibid., p. 598).

In practice, however, the borders between data analysis, information retrieval and text mining are not always quite so clear. Montes-y-Gómez et al. (2004)
proposed an integrated approach, called *contextual exploration*, which combines robust access (IR), non-sequential navigation (hypertext) and content analysis (text mining).

**THE NEED FOR TEXT MINING**

Text mining can be used as an effective business intelligence tool for gaining competitive advantage through the discovery of critical, yet hidden, business information. As a matter of fact, all industries traditionally rich in documents and contracts can benefit from text mining (McKnight, 2005). For example, in medical science, text mining is used to build and structure medical knowledge bases, to find undiscovered relations between diseases and medications or to discover gene interactions, functions and relations (De Bruijn & Martin, 2002, p. 8). A recent application of this is where Gajendran, Lin and Fyhrie (2007) use text mining to predict potentially novel target genes for osteoporosis research that has not been reported on in previous research. Also, government intelligence and security agencies find text mining useful in predicting and preventing terrorist attacks and other security threats (Fan et al., 2006).

**USES OF TEXT MINING**

The different types of text mining have the following in common: it differs from data mining in that it extracts patterns from free (natural language) text rather than from structured databases. However, it does this by using data mining techniques: “it numericizes the unstructured text document and then, using data mining tools and techniques, extracts patterns from them” (Delen and Crossland, 2007:4). In this section various uses of text mining will be discussed, as well as the techniques employed to facilitate these goals. Some examples of the implementation of these text mining approaches will be referred to. The approaches that will be discussed include categorisation, clustering, concept-linking, topic tracking, anomaly detection and web mining.

**Categorisation**

Categorisation focuses on identifying the main themes of a document after which the document is grouped according to these. Two techniques of categorisation are discussed below:

**Keyword-Based Association Analysis**

Association analysis looks for correlations between texts based on the occurrence of related keywords or phrases. Texts with similar terms are grouped together. The pre-processing of the texts is very important and includes parsing and stemming, and the removal of words with minimal semantic content. Another issue is the problem of compounds and non-compounds - should the analysis be based on singular words or should word groups be accounted for? (cf. Han & Kamber, 2001, p. 433). Kostoff et al. (2002), for example, have measured the frequencies and proximities of phrases regarding electrochemical power to discover central themes and relationships among them. This knowledge discovery, combined with the interpretation of human experts, can be regarded as an example of knowledge creation through intelligent text mining.

**Automatic Document Classification**

Electronic documents are classified according to a pre-defined scheme or training set. The user compiles and refines the classification parameters, which are then used by a computer program to categorise the texts in the given collection automatically (cf. Sullivan, 2001, p. 198). Classification can also be based on the analysis of collocation (“the juxtaposition or association of a particular word with another particular word or words” (The Oxford Dictionary, 9th Edition, 1995)). Words that often appear together probably belong to the same class (Lopes et al., 2004). According to Perrin & Petry (2003) “useful text structure and content can be systematically extracted by collocational lexical analysis” with statistical methods. Text classification can be applied by businesses, for example, to personalise B2C e-commerce applications. Zhang and Jiao (2007) did this by using a model that anticipates customers’ heterogeneous requirements as a pre-defined scheme for the classification of e-commerce sites for this purpose (Zhang & Jiao, 2007).
Clustered

Texts are grouped according to their own content into categories that were not previously known. The documents are analysed by a clustering computer program, often a neural network, but the clusters still have to be interpreted by a human expert (Hearst, 1999). Document pre-processing (tagging of parts of speech, lemmatisation, filtering and structuring) precedes the actual clustering phase (Iiritano et al., 2004). The clustering program finds similarities between documents, e.g. common author, same themes, or information from common sources. The program does not need a training set or taxonomy, but generates it dynamically (cf. Sullivan, 2001, p. 201). One example of the use of text clustering in the academic field is found in the work of Delen and Crossland (2007) whose text-mining tool processes articles from three major journals in the management information systems field to identify major themes and trends of research in this area.

Concept Linking

In text databases, concept linking is the finding of meaningful, high levels of correlations between text entities. Concept linking is implemented by the technique called link analysis, “the process of building up networks of interconnected objects through relationships in order to expose patterns and trends” (Westphal, 1998, p. 202). The user can, for example, suggest a broad hypothesis and then analyse the data in order to prove or disprove this hunch. It can also be an automatic or semi-automatic process, in which a surprisingly high number of links between two or more nodes may indicate relations that have hitherto been unknown. Link analysis can also refer to the use of algorithms to build and exploit networks of hyperlinks in order to find relevant and related documents on the Web (Davison, 2003). Concept linking is used, for example, to identify experts by finding and evaluating links between persons and areas of expertise (Ibrahim, 2004). Yoon & Park (2004) use concept linking and information visualisation to construct a visual network of patents, which facilitates the identification of a patent’s relative importance: “The coverage of the application is wide, ranging from new idea generation to ex post facto auditing” (ibid, p. 49).

Topic Tracking

Topic tracking is the discovery of a developing trend in politics or business, which may be used to predict recurring events. It thus involves the discovery of patterns that are related to time frames, for example, the origin and development of a news thread (cf. Montes-y-Gómez et al., 2001). The technique that is used to do this is called sequence analysis. A sequential pattern is the arrangement of a number of elements, in which the one leads to the other over time (Wong et al., 2000). An example of topic tracking is a system that remembers user profiles and, based on that, predict other documents of interest to the user (Delen & Crossland, 2007).

Anomaly Detection

Anomaly detection is the finding of information that violates the usual patterns, e.g. a book that refers to a unique source, or a document lacking typical information. Link analysis and keyword-based analysis, referred to above, are techniques that may also be used for this purpose. An example of anomaly detection is the detection of irregularities in news reports or different topic profiles in newspapers (Montes-y-Gómez et al., 2001).

Web Mining

“Text mining is about looking for patterns in natural language text…. Web mining is the slightly more general case of looking for patterns in hypertext and often applies graph theoretical approaches to detect and utilise the structure of web sites.” (New Zealand Digital Library, 2002)

In addition to the obvious hypertext analysis, various other techniques are used for web mining. Marked-up language, especially XML tags, facilitates text mining because the tags can often be used to simulate database attributes and to convert data-centric documents into databases, which can then be exploited (Tseng & Hwung, 2002). Mark-up tags also make it possible to create “artificial structures [that] help us understand the relationship between documents and document components” (Sullivan, 2001, p. 51). Such tags could, for example, be used to store linguistic analyses regarding the various language modules of a text, enabling
the application of data warehousing and data mining concepts in the humanities (cf. Kroeze, 2007).

Web-applications nowadays integrate a variety of types of data, and web mining will focus increasingly on the effective exploitation of such multi-faceted data. Web mining will thus often include an integration of various text mining techniques. One such an application of web mining where multiple text mining techniques are used is natural language queries or “question answering” (Q&A). Q&A deals with finding the best answer to a given question on the web (Fan et al., 2006). Another prominent application area of web mining is recommendation systems (e.g. personalisation), the design of which should be robust since security is becoming an increasing concern (Nasraoui et al., 2005).

INFORMATION VISUALISATION

Information visualisation “puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching” (Fan et al., 2006:80). Although this definition categorises information visualisation as an information retrieval technique and aid, it is often referred to as visual text mining (Fan et al., 2006; Lopes et al., 2007). A broader understanding of information visualisation therefore includes the use of visual techniques for not only information retrieval but also to interpret the findings of text mining efforts. According to Lopes et al. (2007), information visualisation reduces the complexity of text mining by helping the user to build more adequate cognitive models of trends and the general structure of a complex set of documents.

CRITICAL ISSUES

Many sources on text mining refer to text as “unstructured data”. However, it is a fallacy that text data are unstructured. Text is actually highly structured in terms of morphology, syntax, semantics and pragmatics. On the other hand, it must be admitted that these structures are not directly visible: “… text represents factual information … in a complex, rich, and opaque manner” (Nasukawa & Nagano, 2001, p. 967).

Authors also differ on the issue of natural language processing within text mining. Some prefer a more statistical approach (cf. Hearst, 1999), while others feel that linguistic parsing is an essential part of text mining. Sullivan (2001, p. 37) regards the representation of meaning by means of syntactic-semantic representations as essential for text mining: “Text processing techniques, based on morphology, syntax, and semantics, are powerful mechanisms for extracting business intelligence information from documents…. We can scan text for meaningful phrase patterns and extract key features and relationships”. According to De Bruijn & Martin (2002, p. 16), “[l]arge-scale statistical methods will continue to challenge the position of the more syntax-semantics oriented approaches, although both will hold their own place.”

In the light of the various definitions of text mining, it should come as no surprise that authors also differ on what qualifies as text mining and what does not. Building on Hearst (1999), Kroeze, Matthee & Bothma (2003) use the parameters of novelty and data type to distinguish between information retrieval, standard text mining and intelligent text mining (see Figure 1).

Halliman (2001, p. 7) also hints at a scale of newness of information: “Some text mining discussions stress the importance of ‘discovering new knowledge.’ And the new knowledge is expected to be new to everybody. From a practical point of view, we believe that business text should be ‘mined’ for information that is ‘new enough’ to give a company a competitive edge once the information is analyzed.”

Another issue is the question of when text mining can be regarded as “intelligent”. Intelligent behavior is “the ability to learn from experience and apply knowledge acquired from experience, handle complex situations, solve problems when important information is missing, determine what is important, react quickly and correctly to a new situation, understand visual images, process and manipulate symbols, be creative and imaginative, and use heuristics” (Stair & Reynolds, 2001, p. 421). Intelligent text mining should therefore refer to the interpretation and evaluation of discovered patterns.

FUTURE RESEARCH

Mack and Hehenberger (2002, p. S97) regards the automation of “human-like capabilities for comprehending complicated knowledge structures” as one of the frontiers of “text-based knowledge discovery”. Incorporating more artificial intelligence abilities into text-mining tools will facilitate the transition from
mainly statistical procedures to more intelligent forms of text mining. Fan et al. (2006) consider duo-mining as an important future consideration. This involves the integration of data mining and text mining into a single system and will enable users to consolidate information by analyzing both structured data from databases as well as free text from electronic documents and other sources.

CONCLUSION

Text mining can be regarded as the next frontier in the science of knowledge discovery and creation, enabling businesses to acquire sought-after competitive intelligence, and helping scientists of all academic disciplines to formulate and test new hypotheses. The greatest challenges will be to select and integrate the most appropriate technology for specific problems and to popularise these new technologies so that they become instruments that are generally known, accepted and widely used.

REFERENCES


Discovering Unknown Patterns in Free Text


**KEY TERMS**

**Business Intelligence:** “Any information that reveals threats and opportunities that can motivate a company to take some action” (Halliman, 2001, p. 3).

**Competitive Advantage:** The head start a business has owing to its access to new or unique information and knowledge about the market in which it is operating.

**Hypertext:** A collection of texts containing links to each other to form an interconnected network (Sulli van, 2001, p. 46).

**Information Retrieval:** The searching of a text collection based on a user’s request to find a list of documents organised according to its relevance, as judged by the retrieval engine (Montes-y-Gómez et al., 2004). Information retrieval should be distinguished from text mining.

**Knowledge Creation:** The evaluation and interpretation of patterns, trends or anomalies that have been discovered in a collection of texts (or data in general), as well as the formulation of its implications and consequences, including suggestions concerning reactive business decisions.

**Knowledge Discovery:** The discovery of patterns, trends or anomalies that already exist in a collection of texts (or data in general), but have not yet been identified or described.

**Mark-Up Language:** Tags that are inserted in free text to mark structure, formatting and content. XML tags can be used to mark attributes in free text and to transform free text into an exploitable database (cf. Tseng & Hwung, 2002).

**Metadata:** Information regarding texts, e.g. author, title, publisher, date and place of publication, journal or series, volume, page numbers, key words, etc.

**Natural Language Processing (NLP):** The automatic analysis and/or processing of human language by computer software, “focussed on understanding the contents of human communications”. It can be used to identify relevant data in large collections of free text for a data mining process (Westphal & Blaxton, 1998, p. 116).

**Parsing:** A (NLP) process that analyses linguistic structures and breaks them down into parts, on the morphological, syntactic or semantic level.

**Stemming:** Finding the root form of related words, for example singular and plural nouns, or present and past tense verbs, to be used as key terms for calculating occurrences in texts.

**Text Mining:** The automatic analysis of a large text collection in order to identify previously unknown patterns, trends or anomalies, which can be used to derive business intelligence for competitive advantage or to formulate and test scientific hypotheses.