Discovering Unknown Patterns in Free Text

Jan H. Kroeze
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, these data are 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 has a relatively short history: "Unlike search engines and data mining that have a longer history and are better understood, text mining is an emerging technical area that is relatively unknown to IT professions" (Chen, 2001, p. vi).

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" (Nasukawa & Nagano, 2001, p. 967).

MAIN THRUST

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*" (Rob & Coronel, 2004, p. 597). The same distinction is valid for text mining. Therefore, text-mining tools should also "*initiate* analyses to create knowledge" (Rob & Coronel, 2004, 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). According to Smallheiser (2001, pp. 690-691), text mining approaches can be divided into two main types: "Macro analyses perform data-crunching operations over a large, often global set of papers encompassing one or more fields, in order to identify large-scale trends or to classify and organize the literature…. In contrast, micro analyses pose a sharply focused question, in which one searches for complementary information that links two small, pre-specified fields of inquiry."

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. In the field of academic research, text mining can be used to scan large numbers of publications in order to select the most relevant literature and to propose new links between independent research results. Text mining is also needed "to formulate and assess hypotheses arising in biomedical research … and … for helping make policy decisions regarding technical innovation" (Smallheiser, 2001, p. 690). Another application of text mining is in medical science, to discover gene interactions, functions and relations, or to build and structure medical knowledge bases, and to find undiscovered relations between diseases and medications (De Bruijn & Martin, 2002, p. 8).

Types of Text Mining

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 predefined 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, 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 used by businesses, for example, to categorise customers' e-mails automatically and suggest the appropriate reply templates (Weng & Liu, 2004).

Similarity Detection

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, for example, 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 is found in the work of Fattori et al. (2003), whose text-mining tool processes patent documents into dynamic clusters to discover patenting trends, which constitutes information that can be used as competitive intelligence.

Link Analysis

"Link analysis is the process of building up networks of interconnected objects through relationships in order to expose patterns and trends" (Westphal & Blaxton, 1998, p. 202). In text databases, link analysis is the finding of meaningful, high levels of correlations between text entities. 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). Yoon & Park (2004) use link analysis 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" (p. 49). Text mining is also used to identify experts by finding and evaluating links between persons and areas of expertise (Ibrahim, 2004).

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). Sequence analysis is the discovery