Discovering Unknown Patterns in Free Text

Jan H Kroeze
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 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, 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).