Chapter 3.4
A Survey of Selected Software Technologies for Text Mining

Richard S. Segall
Arkansas State University, USA

Qingyu Zhang
Arkansas State University, USA

ABSTRACT

This chapter presents background on text mining, and comparisons and summaries of seven selected software for text mining. The text mining software selected for discussion and comparison in this chapter are: Compare Suite by AKS-Labs, SAS Text Miner, Megaputer Text Analyst, Visual Text by Text Analysis International, Inc. (TextAI), Magaputer PolyAnalyst, WordStat by Provalis Research, and SPSS Clementine. This chapter not only discusses unique features of these text mining software packages but also compares the features offered by each in the following key steps in analyzing unstructured qualitative data: data preparation, data analysis, and result reporting. A brief discussion of Web mining and its software are also presented, as well as conclusions and future trends.

INTRODUCTION

The growing accessibility of textual knowledge applications and online textual sources has caused a boost in text mining and Web mining research. This chapter presents comparisons and summaries of selected software for text mining. This chapter reviews features offered by each package in the following key steps in analyzing unstructured qualitative data: data preparation including importing, parsing, and cleaning; data analysis including association and clustering; and result presenting/reporting including plots and graphs.

BACKGROUND OF TEXT MINING

Hearst (2003) defines text mining (TM) as "the discovery of new, previously unknown information, by automatically extracting information from

different written sources." Simply put, text mining is the discovery of useful and previously unknown "gems" of information from textual document repositories. Also Hearst (2003) distinguishes text mining from data mining by noting that with "text mining the patterns are extracted from natural language rather than from structured database of facts." A more technical definition of text mining is given by Woodfield (2004) author of SAS Notes for Text Miner, as a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.

Text mining (TM) or text data mining (TDM) has been discussed by numerous investigators that include Hearst (1999), Cerrito (2003) for the application to coded information, Hayes et al. (2005) for software engineering, Leon (2007) for identifying drug, compound, and disease literature, and McCallum (1998) for statistical language modeling. Firestone (2005) emphasizes the importance of text mining in the future knowledge work. Romero and Ventura (2007) survey text mining applications in the educational setting. Kloptchenko et al. (2004) use data and text mining techniques for analyzing financial reports. Mack et al. (2004) describe the value of text analysis in biomedical research for life science. Baker and Witte (2006) discuss the mutation mining to support activities of protein engineers.

Uramoto et al (2004) utilized a text-mining system adopted from that developed by IBM and named TAKMI (Text Analysis and Knowledge Mining) for use with very large text biomedical text documents. In fact the extension of TAKMI was named MedTAKMI and was capable of mining the entire MEDLINE of 11 million biomedical journal abstracts. The TAKMI system allows extracting deeper relationships among biomedical concepts by the use of natural language techniques. Scherf et al. (2005) discuss the applications of text mining in literature search to improve accuracy and relevance. Kostoff et al. (2001) combine data mining and citation mining to identify user community, and its characteristics by categorizing articles.

There is a Text Mining Research Group (TMRG) (2002) at the University of Waikato in New Zealand that maintains a Web page of related publications, links, and software. Similarly there is National Centre for Text Mining (NaCTeM) at the University of Manchester in United Kingdom (UK). The Aims and Objectives of NACTeM is described in article by Ananiandou et al (2005) in which it extensively discusses a need for text mining in biology. According to their Web site of 2002, "text mining uses recall and precision (borrowed from the information retrieval research community) to measure the effectiveness of different information extraction techniques, allowing quantitative comparisons to be made." A Text Mining Workshop was held in 2007 in conjunction with the Seventh Society of Industrial and Applied Mathematics (SIAM) Conference on Data Mining (SDM 2007). Textbooks in text mining have included applications to biology and biomedicine by Ananiadou and McNaught (2006).

Figure 1 of this paper from Liang (2003) shows the text mining process from text preprocessing to analyzing results. Saravanan et al. (2003) discuss how to automatically clean data, i.e., summarizing domain-specific information tailored to user's needs, by discovering classes of similar items that can be grouped into prescribed domains. Hersh (2005) evaluates different text-mining systems for information retrieval. Turmo et al. (2006) describe and compare different approaches to adaptive information extraction from textual documents and different machine language techniques. Amir et al. (2005) describe a new tool called maximal associations which allows the discovering of interesting associations often lost by regular association rules. Spasic et al. (2005) discuss ontologies and text mining to automatically extract information and facts, discover hidden associations and generate hypotheses germane to user needs. Ontologies specify the interpretations of terms, echo the structure of the domain, and thus can be used to