Chapter XX
Mining Text with the Prototype-Matching Method

A. Durfee
Appalachian State University, USA

A. Visa
Tampere University of Technology, Finland

H. Vanharanta
Tampere University of Technology, Finland

S. Schneberger
Appalachian State University, USA

B. Back
Åbo Akademi University, Finland

ABSTRACT

Text documents are the most common means for exchanging formal knowledge among people. Text is a rich medium that can contain a vast range of information, but text can be difficult to decipher automatically. Many organizations have vast repositories of textual data but with few means of automatically mining that text. Text mining methods seek to use an understanding of natural language text to extract information relevant to user needs. This article evaluates a new text mining methodology: prototype-matching for text clustering, developed by the authors' research group. The methodology was applied to four applications: clustering documents based on their abstracts, analyzing financial data, distinguishing authorship, and evaluating multiple translation similarity. The results are discussed in terms of common business applications and possible future research.

INTRODUCTION

It can be argued that computers are now used more for storing and retrieving data than computing data. Organizational computer systems are used for maintaining inventory, production, marketing, financial, sales, accounting, personnel, customer, and other types of data. With enterprise systems, vast amounts of corporate data can be stored digitally and made available to employees when and where needed. Data mining software is often used to further glean information from corporate databases.

A lot of transactional corporate data is numeric but not all of it. Indeed, it's often stated that about 80% of corporate information is textual or unstructured information (for example, see Chen, 2001, and Robb, 2004). An entire information systems specialty—knowledge management—includes collecting, storing, organizing, evaluating, and using textual data such as prevalent with consulting agencies in vast repositories of written reports.

The World Wide Web provides access to planetary-wide databases of textual data for corporate users. Just one of hundreds of online article databases (Education Resources Information Center, or ERIC) has more than 1.2 million citations and 110,000 full text articles. Another, HighWire Press, has more than 1.3 million full text articles. Internal and external data sources offer extensive decision support for managers in dynamic, complex, and demanding business environments. But how can managers, decision makers, and knowledge workers find appropriate textual content among billions of words in internal and external document repositories when it's virtually impossible to do so manually? Seventy-five percent of managers spend more than an hour per day just sorting their e-mails, according to a Gartner Group survey (Marino, 2001).

Compounding the problem is that text, by its very nature, can have multiple meanings and interpretations. The structure of text is not only complex but also not always directly obvious. Even the author of a text might not know the extent of what might be interpreted from the text. These features of text make it a very rich medium for conveying a wide range of meanings but also very difficult to manage, analyze, and mine using computers (Nasukawa & Nagano, 2001). Therein lies the conundrum: There is too much internal and external text to mine manually, but it's problematic for computer software to correctly interpret let alone create knowledge from text.

Text mining (TM) looks for a remedy of that problem. TM seeks to extract high-level knowledge and useful patterns from low-level textual data. Text mining tools seek to analyze and learn the meaning of implicitly structured information automatically (Dorre, Gerstl, & Seiffert, 1999). There are two broad categories of textual mining: *text categorization* and *text clustering.*

Text categorization analyzes text using pre-determined structures or words (i.e., keywords). It is a framework-driven approach, usually based on earlier analysis or expectations. Authors, readers, and librarians may introduce and use keywords, indexes, or mark-ups to outline the main ideas, concepts and themes within a text to make textual searches easier for computers (Anderson, 1999; Chieng, 1997; Lahtinen, 2000; Salton, 1989; Weiss, White, Apte, & Damerau, 2000). However, authors and textual information users can assign different keywords to the same text, or even ascribe different meanings to the same keywords—possibly defeating the speed and accuracy of computer-based textual keyword searches. Readers need only consider their own wayward searches using keyword-based online search engines to understand the depth and breadth of the problem.

Text clustering, on the other hand, differs from keywords or pre-determined structural searches. Text clustering discovers latent groupings of text, where the textual similarities within a group are maximized while similarities among groups are minimized. Effective text clustering uses the characteristics of textual meanings, structure,