Chapter LIV
Text Mining

Antonina Durfee
Appalachian State University, USA

ABSTRACT

Massive quantities of information continue accumulating at about 1.5 billion gigabytes per year in numerous repositories held at news agencies, at libraries, on corporate intranets, on personal computers, and on the Web. A large portion of all available information exists in the form of text. Researchers, analysts, editors, venture capitalists, lawyers, help desk specialists, and even students are faced with text analysis challenges. Text mining tools aim at discovering knowledge from textual databases by isolating key bits of information from large amounts of text, identifying relationships among documents. Text mining technology is used for plagiarism and authorship attribution, text summarization and retrieval, and deception detection.

INTRODUCTION

The proliferation of computers, storage devices, and the World Wide Web makes access to various data sources very convenient. The availability and accessibility of many up-to-date data sources offer an extensive support for ordinary people and decision makers in the dynamic, complex, and demanding environment of today. With technological advances, our technological abilities to collect, generate, distribute, and store data have outgrown our ability to process and understand them. Business and governmental units, collecting and storing information by the click of a mouse button, now want to understand the trends lying behind this information quickly. Understanding those trends allows optimizing the processes of decision making and reaching customers more effectively. Although technology devices can deliver vital data for decision-making purposes anywhere, anytime, the utilization of these advances is mediocre. Technological devices such as computers, Internet-enabled mobile phones, laptops, and personal digital assistants contribute to data multiplication, leading to information overload. Information overload creates data tombs resulting in unavoidable losses and missed opportunities. This environment dictates the strong need for intelligent solutions for data analysis and exploration.

A substantial portion of the available information is stored in text or document databases. This information resides in a company's internal and external documents, technical and financial reports, customer feedback on products, market analyses and overviews, electronic mail and notice-board messages, advertisements, managerial notes, business-related publications, business plans, correspondence with partners and creditors, competitor releases, news articles, research papers, books, digital libraries, and various company-related Web pages. 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 analyze manually, but it is problematic for computer software to correctly interpret, let alone create, knowledge from text.

Text mining (TM) looks for a remedy for that problem. TM seeks to extract high-level knowledge and useful patterns from textual data. Text mining tools seek to analyze and learn the meaning of implicitly structured information automatically (Dörre, Gerstl, & Seiffert, 1999). TM or data mining (DM) from textual databases is an essential part of discovering previously unknown patterns useful for particular purposes from textual databases (Dörre et al.; Hearst, 1999). While users of numeric data can explore a new, previously unknown pattern in numeric data stored in a database, the users of natural-language text can only associate the discovery of previously unknown patterns in textual data with rediscovery or with a new interpretation of what the author of a text had already written. It is very argumentative to state that an author of a text might not know himself or herself what is stated in a text. A new interpretation of all the facts stated in the text can only appear because different readers understand the same text differently based on their backgrounds. Witten, Bray, Mahoui, and Teahan (1998) argue that TM has potential because one does not have to understand the text in order to extract useful information from it.

BACKGROUND

TM has its roots in computational linguistics, natural-language processing, text analysis, cognitive psychology, information retrieval (IR), machine learning, statistics, and information and library sciences. The confluence of multiple disciplines in the area of TM is presented in Figure 1. Some of the parental disciplines, for instance, statistics, artificial intelligence, and information science, are the same for TM and DM. TM can be seen as a subpart of DM that deals with one specific format of data, namely, text.

Working with Textual Data

Since text is the most popular and convenient way of transferring meaning from authors to readers, the amount of digitally available text is mounting. A recent study indicates that 80% of a company's information is contained in text documents (Tan, 1999). Managers and knowledge workers spend a lot of time dealing with textual-information overload looking for useful points in it. For instance, a Gartner Group survey reveals that 75% of managers spend more than an hour per day sorting out and answering their e-mails (Marino, 2001). The dynamic business environment does not allow managers the luxury to devote enough time to read and analyze all available documents that might contain information that might impact managerial decisions.

Being the most common vehicle for written communication, text has a complicated and am-