INTRODUCTION

Intensive usage and growth of the World Wide Web and the daily increasing amount of text information in electronic form have resulted in a growing need for computer-supported ways of dealing with text data. One of the most popular problems addressed with text mining methods is document categorization. Document categorization aims to classify documents into pre-defined categories, based on their content. Other important problems addressed in text mining include document search, based on the content, automatic document summarization, automatic document clustering and construction of document hierarchies, document authorship detection, identification of plagiarism of documents, topic identification and tracking, information extraction, hypertext analysis, and user profiling. If we agree on text mining being a fairly broad area dealing with computer-supported analysis of text, then the list of problems that can be addressed is rather long and open. Here we adopt this fairly open view but concentrate on the parts related to automatic data analysis and data mining.

This article tries to put text mining into a broader research context, with the emphasis on machine learning perspective, and gives some ideas of possible future trends. We provide a brief description of the most popular methods only, avoiding technical details and concentrating on example of problems that can be addressed using text-mining methods.

BACKGROUND

Text mining is an interdisciplinary area that involves at least the following key research fields:

- **Machine Learning and Data Mining** (Hand, et al., 2001; Mitchell, 1997; Witten & Frank, 1999): Provides techniques for data analysis with varying knowledge representations and large amounts of data.

- **Statistics and Statistical Learning** (Hastie, et al., 2001): Contributes data analysis in general in the context of text mining (Duda et al., 2000).

- **Information Retrieval** (Rijsberg, 1979): Provides techniques for text manipulation and retrieval mechanisms.

- **Natural Language Processing** (Manning & Schutze, 2001): Provides techniques for analyzing natural language. Some aspects of text mining involve the development of models for reasoning about new text documents, based on words, phrases, linguistics, and grammatical properties of the text, as well as extracting information and knowledge from large amounts of text documents.

The rest of this article briefly describes the most popular methods used in text mining and provides some ideas for the future trends in the area.

MAIN THRUST

Text mining usually involves some preprocessing of the data, such as removing punctuations from text, identifying word and/or sentence boundaries, and removing words that are not very informative for the problem on hand. After preprocessing, the next step is to impose some representation on the text that will enable application of the desired text-mining methods. One of the simplest and most frequently used representations of text is word-vector representation (also referred to as bag-of-words representation). The idea is fairly simple: words from the text document are taken, ignoring their ordering and any structure of the text. For each word, the word-vector contains some weight proportional to the number of its occurrences in the text.

We all agree that there is additional information in the text that could be used (e.g., information about structure of the sentences, word type and role, position of the words or neighboring words). However, depending on the problem at hand, this additional information may or may not be helpful and definitely requires additional efforts and more sophisticated methods. There is some evidence for document retrieval of long documents, considering information additional to the bag-of-words is not worth the effort and that for document categorization, using natural language information does not improve the categorization results (Dumais et al., 1998). There is also...
Related Content

Approximate Range Queries by Histograms in OLAP
www.igi-global.com/chapter/approximate-range-queries-histograms-olap/10564?camid=4v1a

A Multidimensional Pattern Based Approach for the Design of Data Marts
www.igi-global.com/chapter/multidimensional-pattern-based-approach-design/28167?camid=4v1a

The Issue of Missing Values in Data Mining
www.igi-global.com/chapter/issue-missing-values-data-mining/10959?camid=4v1a

Mining Data with Group Theoretical Means
www.igi-global.com/chapter/mining-data-group-theoretical-means/10699?camid=4v1a