Text Mining–Machine Learning on Documents

DunjaMladenić
Jozef Stefan Institute, Slovenia

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 some work on document categorization that extends the bag-of-words representation by using word sequences instead of single words (Mladenic & Grobelnik, 2003). This work suggests that the usage of single words and word pairs in the bag-of-words representation improves the results of short documents categorization.

The rest of this section gives a brief description of the most important problems addressed by text-mining methods.

Text Document Categorization is used when a set of pre-defined content categories, such as arts, business, computers, games, health, recreation, science, and sport, is provided, as well as a set of documents labeled with those categories. The task is to classify previously unseen text documents by assigning each document one or more of the predefined categories. This usually is performed by representing documents as word-vectors and using documents that already have been assigned the categories to generate a model for assigning content categories to new documents (Jackson & Moulinier, 2002; Sebastiani, 2002). The categories can be organized into an ontology (e.g., the MeSH ontology for medical subject headings or the DMoz hierarchy of Web documents).

Document Clustering (Steinbach et al., 2000) is based on an arbitrary data clustering algorithm adopted for text data by representing each document as a word vector. The similarity of two documents is commonly measured by the cosine-similarity between the word vectors representing the documents. The same similarity also is used in document categorization for finding a set of the most similar documents.

Visualization of text data is a method used to obtain early measures of data quality, content, and distribution (Fayyad et al., 2001). For instance, by applying document visualization, it is possible get an overview of the Web site content or document collection. One form of text visualization is based on document clustering (Grobelnik & Mladenic, 2002) by first representing the documents as word vectors and by performing K-means clustering algorithms on the set of word vectors. The obtained clusters then are represented as nodes in a graph, where each node in the graph is described by the set of most characteristic words in the cluster. Similar nodes, as measured by the cosine-similarity of their word vectors, are connected by an edge in the graph. When such a graph is drawn, it provides a visual representation of the document set.

Text Summarization often is applied as a second stage of document retrieval in order to help the user getting an idea about content of the retrieved documents. Research in information retrieval has a long tradition of addressing the problem of text summarization, with the first reported attempts in the 1950s and 1960s, that were exploiting properties such as frequency of words in the text. When dealing with text, especially in different natural languages, properties of the language can be a valuable source of information. This brings in text summarization of the late 1970s the methods from research in natural language processing. As humans are good at making summaries, we can consider using examples of human-generated summaries to find something about the underlying process by applying machine learning and data-mining methods, a popular problem in 1990s. There are several ways to provide text summary (Mani & Maybury, 1999). The simplest but also very effective way is providing keywords that help to capture the main topics of the text, either for human understanding or for further processing, such as indexing and grouping of documents, books, pictures, and so forth. As the text is usually composed of sentences, we can talk about summarization by highlighting or extracting the most important sentences, a way of summarization that is frequently found in human-generated summaries. A more sophisticated way of summarization is by generating new sentences based on the whole text, as used, for instance, by humans in writing book reviews.

User Profiling is used to provide the information that is potentially interesting for the user (e.g., in the context of personalized search engines, browsing the Web, or shopping on the Web). It can be based on the content (of the documents) that the user has visited or the behavior of other users accessing the same data. In the context of text mining, when using the content, the system searches for text documents that are similar to those the user liked (e.g., observing the user browsing the Web and providing help by highlighting potentially interesting hyperlinks on the requested Web pages) (Mladenic, 2002). Content-based document filtering has its foundation in information retrieval research. One of the main problems with this approach is that it tends to specialize the search for the documents similar to the document already seen by the user.

FUTURE TRENDS

There is a number of researchers intensively working in the area of text data mining, mainly guided by the need of developing new methods capable of handling interesting real-world problems. One such problem recognized in the past few years is on reducing the amount of manual work needed for hand labeling the data. Namely, most of the approaches for automatic document filtering, categorization, user profiling, information extraction, and text tagging requires a set of labeled (pre-categorized) data describing the addressed concepts. Using unlabeled data and bootstrapping learning are two directions giving research results that enable important reduction in the needed amount of hand labeling.