Text Mining

D

Thomas Mandl
Universität Hildesheim, Germany

INTRODUCTION

Text Mining is the extraction of knowledge from many texts. The extracted knowledge is new and created during the extraction process. It cannot be found in any of the single processed texts. Text mining combines several technologies and is applied in diverse application areas. Knowledge derived from text mining comes in the form of distribution patterns and the frequency of words in texts and their parts. Based on such knowledge, the user can explore and determine how topics are dealt with in many text documents, which attitudes on specific topics are expressed, and how these matters evolve with time. Text mining requires algorithms as well as information work of users: "Text Mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools" (Feldman & Sanger 2007, p. 1).

BACKGROUND

The basic algorithms for Text Mining deal with natural language as present in texts. Language technology dealing with the unstructured and vague nature of language needs to create numeric representations. Subsequently, the extracted structural knowledge is further processed by numerical algorithms as they are employed in data mining. Particularly, clustering and classification are applied. For instance, text classification can assign texts to different classes based on their topic.

Text Mining systems allow their users to interact with text collections and to extract some useful information in the form of overall patterns. In order to support the information work of the users, optimized user interfaces are necessary and need to be designed in a user centered way. The extraction and identification of patterns are often easier for the user if data is adequately visualized. Hence, visualization is an inherent part of Text Mining.

The term Web Mining is closely related to Text Mining and means the application of machine learning techniques to data from the Internet: "treat the information in the web as a large knowledge base from which we can extract new, never-before encountered information" (Hearst, 1999). Most information of the Web is stored in text, so that there is a high overlap between the two areas. Machine learning is the computational core of Text Mining. Algorithms try to adapt and improve their output over time. In supervised learning, a teaching input leads the program to a new and better solution for the same input. Data Mining is often referred to as the process in which machine learning is applied and which integrates data preparation and presentation of results.

BASIC COMPONENTS OF TEXT MINING

Text Mining typically begins with the processing of natural language. Initially, the creation of numerical representations for further processing is necessary. Natural language processing tasks are identical for many text mining applications.

Lexical Operations

Texts contain words in many different forms. The words need to be identified and separated, a difficult task for languages without blanks between words (e.g. some East Asian languages). In the case of most European languages, punctuation marks and hyphens need to be regarded.

The following step is grouping words which have a common basic form. These forms could be e.g. different grammatical forms of a verb. Their meaning is basically identical and only their morphology changes.

DOI: 10.4018/978-1-4666-5888-2.ch185

In languages with many cases for nouns and many temporal forms for verbs (e.g. Finnish), this task can be challenging. Identical stemming operations are carried in Information Retrieval.

An example would be the word forms "run," "runs" and "running." They should be all mapped to the same stem "run."

The remaining words are counted and their frequency in each text and in the entire collection is determined. Based on the frequencies, weights are calculated expressing the importance of a word or term for a text document. These weights show the topicality or "aboutness" of a document. This information can be stored in a document-term matrix where a vector contains the weights for all terms regarding one document. Each column shows the distribution of a term over all documents in a collection. (Manning et al., 2008)

Concepts as Collections of Words

For Text Mining, the occurrence of a single word in one document is not as much of importance as it is for Information Retrieval. Text Mining is not about finding one relevant document but rather about the presence of groups of words in documents and often the way the concept represented by these words is dealt with.

Such a concept could be, for instance, "Chinese Government" which would typically comprise phrases like "Chinese Leader" and "Chinese prime Minister" and also the names of ministers. Another concept could be a collection of words with positive meanings. A concept in Text Mining can be understood as a set of words. Concepts are usually created manually or extracted from ontologies or thesauri. There are also different methods for their semi-automatic creation. The success of applications will to some extent depend on the quality of the concept definitions.

The frequency of a concept can be determined by adding the frequency of all terms in this concept, a process which will open opportunities for analysis. For example, the frequency of a concept of "corruption" could be determined in a news corpus. In a further step, the frequency of the occurrence of a concept related to a particular political party can be determined. With this information, the frequency of a party name in the context of the concept "corruption" will be found.

By incorporating the dimension of time, a trend analysis can be carried out. For the example mentioned above, the association of the concept "corruption" and a party name can be recorded for a long period of time and be presented for short time steps like months or years. Thus, the temporal evolution of the phenomenon constituting corruption and a party can be observed. It is important to relate the frequency of a concept to the normal frequency. Concerning the example above, it could be the case that a certain party name appears very rarely. Consequently, it would also appear rarely in the context of corruption. However, it could occur more often in the vicinity of corruption than expected or more often than other party names.

Concept Discovery

The common appearance of terms can be an indicator for their semantic similarity. It could also be a hint that together they belong to a same concept. When terms are similar in the vector space model mentioned above, they often appear together in documents. In other words, they exhibit a similar distribution pattern over documents. In a document-term matrix, they can be identified by searching for similar term vectors. Such associated terms or words should occur more often together than their individual frequency suggests. The frequencies of words differ by several orders of magnitudes. Therefore, it is necessary to calculate the divergence from randomness for joint occurrence. Such calculations can be interpreted as statistical tests for the significance of common occurrence. Similar models are well established in Information Retrieval under the language model (Song & Croft 1999).

Frequently used association measures are mutual information, log-likelihood or Chi-square (Manning et al. 2008). Due to the high number of words, not all pairs can be compared in a typical application, therefore; there is a high demand for efficient methods and optimization approaches.

Classification

Classification algorithms sort objects into predefined classes. In order to do that, these supervised learning algorithms require the presence of positive and negative example objects for each class as shown in Figure 1. The algorithm extracts knowledge from the objects for which the class assignment is known. Their features and their respective values relevant