Chapter 7
Document Classification

ABSTRACT

Keywords can be used as attributes for mining rules or as a basis for measuring the similarity of new (unclassified) documents with existing (classified) ones. The focus is on the problem of extracting keywords from document collection in order to use them as attributes for document classification. Document classification is a hot topic in machine learning. Typical approaches extract "features," generally words, from document, and use the feature vectors as input to a machine learning scheme that learns how to classify documents. This "bag of keywords" model neglects keyword order and contextual effects.

DOCUMENT CLASSIFICATION: ROLE OF KEYWORDS

Survey shows that text classification is a typical scholarly activity in literary study, and automatic text classification methods can be used in three scenarios.

1. The first is **Information Organization** - A classifier can learn the target category concepts (e.g. news article about trade, acquisition, etc.) from the training documents, and then assign new documents into these predefined categories.
2. The second purpose is **Knowledge Discovery** - A successful classifier can provide insights to understand a target concept by revealing the correlations between the features and the concept.

DOI: 10.4018/978-1-7998-3772-5.ch007

3. The third purpose is **Example-based Retrieval** - A classifier might be able to learn a concept from a small number of training documents with the help of semi-supervised learning or active learning methods, and then retrieve more documents similar to the training examples from a large collection.

Automatic document classification methods are tremendously used in research domain of textual documents due to,

- The numerous and important domains of application
- The indispensability in many applications
- The implausibility of manual alternative
- The productivity of machine learning
- Large amount of digital form of documents

NEED FOR DOCUMENT CLASSIFICATION

The goal of text classification is to (semi) automate the categorization process. It is also useful for reducing cost and improving performance (including accuracy and consistency) of text processing.

Scope of Text Classification

Text Classification could be considered as the application of (semi) automatic methods in order to choose, from a set of predefined classification codes, the appropriate one (category / class) for a given new document. Various studies have focused on the construction of a model (Rule or Tree) related to the existence of keywords in order to assign the class of the unclassified document. The text collection is divided into documents and each document is characterized by a number of keywords (Yang, Y., and Pedersen, J.O, 1997). In automatic document classification, for example, for classifying newspaper articles into predefined categories such as politics and sports, the crucial step is how to select appropriate keywords. With traditional classification methods based on the vector space model, frequent words are emphasized and therefore low- frequency words tend to be disregarded. However, there often exist low-frequency words that are effective for classification. For instance,