Abstract

Automatic text categorization (also known as text classification or topic spotting) is the activity of labeling natural language texts with thematic categories from a predefined set. For the purpose of classifying the text documents, there is a need for a set of features and a weighting model which gives relevance value to the features. Currently, bag of words (BOW) is found to be the most widely accepted text representation method. This representation has two major drawbacks. First, the amount of features is very large; second, there is no relatedness between the words. Topic Detection (TD) technique helps the BOW to handle the two drawbacks by detection features very relevant to the document in the space of concept. However, existing TD techniques were not designed for text categorization and often involve huge computational complexity and cost. This paper proposes a topic detection technique for relevant feature extraction. The TD technique extracts topics along with relevant features for each text document. It then finds relatedness between features for each topic. The features extracted for each topic are tightly related to the topic and accordingly the category label. The term frequency measure selects the appropriate features by finding frequency count for each extracted feature for each category label. Thus, the TD technique extracts the relevant features for the classifiers for classification. To evaluate the TD technique, a query categorization system is designed and proposed. The experiments were performed on three datasets (
The experimental results for TD technique show that the topics, along with the set of keywords, detected for documents are indeed relevant. Also, the query categorization system showed satisfactory performance in categorizing the queries using the TD technique.

References
- David D. Lewis "Reuters 21578" dataset.
- Xiao-Bing Xue, Zhi-Hua Zhou, Distributional Features for Text Categorization, IEEE Transactions on Knowledge and Data Engineering, Volume 21 Issue 3, March 2009, Pages 428-442.
- Moulinier and Ganascia J. , Applying and an existing machine learning algorithms to text categorization; in Connectionist, statistical, and symbolic approaches to learning for NLP, springer verlag, 1996.

Index Terms

Computer Science
Software Engineering

Keywords

Concept Derivation  Topic detection  BOW representation  Performance Evaluation