Abstract

Predefined categories can be assigned to the natural language text using for text classification. It is a “bag-of-word” representation, previous documents have a word with values, it represents how frequently this word appears in the document or not. But large documents may face many problems because they have irrelevant or abundant information is there. This paper explores the effect of other types of values, which express the distribution of a word in the document. These values are called distributional features. All features are calculated by tfidf style equation and these features are combined with machine learning techniques. Term frequency is one of the major factor for distributional features it holds weighted item set. When the need is to minimize a certain score function, discovering rare data correlations is more interesting than mining frequent ones. This paper tackles the issue of discovering rare and weighted item sets, i.e., the infrequent weighted item set mining problem. The classifier which gives the more accurate result is selected for categorization. Experiments show that the distributional features are useful for text categorization.
References

8. T. G. Dietterich, “Machine learning research: Four current directions,”

Index Terms

Computer Science

Artificial Intelligence

Keywords

Text Classification, Text Mining, Machine Learning, Compactness, tfidi, Weighted database.