Abstract

Feature selection is one of the well known solution to high dimensionality problem of text categorization. In text categorization, selection of good features (terms) plays a very important role. Feature selection is a strategy that can be used to improve categorization accuracy, effectiveness and computational efficiency. This paper presents an empirical study of most widely used feature selection methods viz. Term Frequency-Inverse Document Frequency (tf idf), Information Gain (IG), Mutual Information(MI), CHI-Square ( 2), Ambiguity Measure (AM), Term Strength (TS), Term Frequency-Relevance Frequency (tf rf ) and Symbolic Feature Selection (SFS) with five different classifiers (Nave Bayes, KNearest Neighbor, Centroid Based Classifier, Support Vector Machine and Symbolic Classifier). Experimentations are carried out on standard bench mark datasets like Reuters-21578, 20-Newsgroups and 4 University dataset.

References

2. 4 universities. [Link to sources]
3. Reuters-21578. [Link to sources]
38. Shang Wenqian, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang.


**Index Terms**

Computer Science

Information Sciences

**Keywords**

High Dimensionality, Feature Selection, Classifiers, Text Categorization