Abstract

The dynamic nature of web and with the increase of the number of web pages, it is very difficult to search required web pages easily and quickly out of thousands of web pages retrieved by a search engine. The solution to this problem is to classify the web pages according to their genre. Automatic genre identification of web pages has become an important area in web page classification, because it can be used to improve the quality of web search results and also to reduce the search time. In this paper, a Combined Stemming Approach (CSA) is proposed to extract genre relevant words and to classify web pages by genre (non-topical) based on word level and linguistic features. Experiments were performed on 7-genre corpus. In order to improve the accuracy of the results, we applied combined stemming and stop word elimination techniques. The proposed approach of extracting features discriminates web pages by genre. The classification results obtained using Random Forest classifier was compared with the results of other researchers, who worked on the same corpus. It is shown that the method proposed is superior in performance in terms of accuracy.

References


Santini M. Characterizing Genres of Web Pages: Genre Hybridism and Individualization Proceedings of the 40th Hawaii International Conference on System Sciences – 2007


Index Terms

Computer Science  Information Sciences

Keywords

Web page classification  Genre  Corpus  Feature Extraction  Combined Stemming Approach