Abstract

Dimensionality refers to number of terms in a web page. While classifying web pages high dimensionality of web pages causes problem. The main objective of reducing dimensionality of web pages is improving the performance of classifier. Processing time and accuracy are two parameters which influence the performance of a classifier. To reduce the processing time, less informative and redundant terms have to be removed from web pages. This research describes hybrid approach for dimensionality reduction in web page classification using a rough set and naïve Bayesian method. Feature selection and dimensionality reduction methods are used for reducing the dimensionality. Information gain method is used as feature selection method. Rough set based Quick Reduct algorithm is used for dimensionality reduction. Naïve Bayesian method is used for classifying web pages to optimal predefined categories. Assignment of web pages to category is based on maximum posterior probability. Words remaining after the process of feature selection and dimensionality reduction will be given to the classifier. Finally the classifier will assign most optimal predefined category to web pages.
Approach for Dimensionality Reduction in Web Page Classification

- Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd Ed., Han, Kamber & Pei, University of Illinois at Urbana-Champaign & Simon Fraser University, 2011
- Ming Mao, Yefei Peng, Michael Spring, "Ontology Mapping: As a Binary Classification Problem", IEEE Fourth international conference on Semantics, Knowledge and grid, 2008
- Sang-Bum Kim, Kyong-soo Han, Hae-Chang Rim, Sung Hyon Myaeng "Some Effective techniques for Naïve Bayes Text Classification", IEEE Transactions on Knowledge and Data Engineering -2006
- Dino Isa, Lam Hong Lee, V. P. Kallimani, and R. RajKumar, "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector
Approach for Dimensionality Reduction in Web Page Classification

Machine"; IEEE transactions on knowledge and data engineering, vol. 20, no. 9, September 2008

**Index Terms**

Computer Science

Information Sciences

**Keywords**

Dimensionality Reduction  Feature Selection  Information gain  Naïve Bayes  Rough Set

Web Page Classification.