A New Text Mining Approach Based on HMM-SVM for Web News Classification

International Journal of Computer Applications
Number 19 - Article 18
Year of Publication: 2010

Authors:
Krishnalal G
S Babu Rengarajan
K G Srinivasagan

10.5120/395-589

Abstract

Since the emergence of WWW, it is essential to handle a very large amount of electronic data of which the majority is in the form of text. This scenario can be effectively handled by various Data Mining techniques. This paper proposes an intelligent system for online news classification based on Hidden Markov Model (HMM) and Support Vector Machine (SVM). An intelligent system is designed to extract the keywords from the online news paper content and classify it according to the pre defined categories. Three different stages are designed to classify the content of online newspapers such as (1) Text pre-processing (2) HMM based Feature Extraction and (3) Classification using SVM. Data have been collected for experimentation from The Hindu, The New Indian Express, Times of India, Business Line, and The Economic Times. The experimental results are based on the news categories such as sports, finance and politics and their accuracies in percentage are 92.45, 96.34 and 90.76 respectively. These results are very good compared to that of other text classification methods.

Reference
[4] Lei Tang, Huan Liu. Bias Analysis in Text Classification for Highly Skewed Data,
A New Text Mining Approach Based on HMM-SVM for Web News Classification

Proceedings of Fifth IEEE International Conference on Data Mining (ICDM'05), 2005, pp. 781-784.


[22] Mladenic, D., Grobelnic, M.: Feature Selection for Unbalanced Class Distribution and


[31] Sang-Bum Kim; Kyoung-Soo Han; Hae-Chang Rim; Sung Hyon Myaeng; Some Effective Techniques for Naive Bayes Text Classification, IEEE Transactions on Knowledge and Data Engineering, Volume 18, Issue 11, Nov,2006.


[40] Jun-Peng Bao Jun-Yi Shen Xiao-Dong Liu Qin-Bao Song , A new text feature extraction model and its application in document copy detection, 1 International Conference on
[41] Harriman, Feature selection and feature extraction for text categorization, Proceedings of the workshop on Speech and Natural Language, Human Language Technology Conference archive.

Index Terms

Computer Science
Text Mining

Key words

Feature Extraction
HMM
kNN
POS
SVM