Abstract

Web content extraction and classification can be viewed as combination of different methods. Nowadays web page contains lot of information including main contents. Contents extraction which are of user’s interest is main task. Text mining is the technique that helps users to find useful information from a large amount of digital text documents on the Web or databases. It is
therefore crucial that a good text mining model should retrieve the information that meets user’s needs within a relatively efficient time frame. A first step toward any Web-based text mining effort would be to collect a significant number of Web mentions of a subject. Thus, the challenge becomes not only to find all the subject occurrences, but also to filter out just those that have the desired meaning. The system described in this paper is capable of extracting main content and classify it. Vector space model method is used for classification.

Reference

- Weiguo Fan1, Linda Wallace, Stephanie Rich, Zhongju Zhang “Tapping into the Power of Text Mining”.
- Neha Gupta, Dr.saba Hilal "A Heuristic Approach for Web content extraction" International Journal of Computer Applications (0975-8887) volume 15-No.5 Feb 2011
- Yves Weissig, Thomas Gottron. “Combinations of Content Extraction Algorithms”.

Index Terms

Computer Science
Information Retrieval

Key words

MVC Architecture VSM model Text Mining Extraction
Classification