Applying Web Analysis in Web Page Filtering

Michael Chau
School of Business
Faculty of Business and Economics
The University of Hong Kong
Pokfulam, Hong Kong
+852 2859-1014
mchau@business.hku.hk

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering, search process.*

General Terms
Design, Experimentation.

Keywords
Information retrieval, Web page classification, Web page filtering, vertical search engines, Web analysis, neural networks, support vector machines, machine learning.

1. INTRODUCTION
Vertical search engines provide Web users with an alternative way to search for information on the Web by providing customized searching in particular domains. However, two issues need to be addressed when developing these search engines: how to locate relevant documents on the Web and how to filter out irrelevant documents from a set of documents collected from the Web. This paper reports the research in addressing the second issue. Traditional approaches, such as a manual approach or a keyword-based approach have their shortcomings. A more promising approach is by using text classifiers, but a major problem is that most text classifiers rely on a large number of testing data and do not effectively incorporate Web characteristics into their models. In this research a machine-learning-based approach that combines Web content analysis and Web structure analysis is proposed. The following research questions are investigated in this research: First, can Web structure analysis techniques be used to help create a vertical search engine? Second, can domain knowledge be used to enhance Web page filtering for a vertical search engine? Lastly, can Web page classification be applied to a large collection with only a small number of training examples?

2. PROPOSED APPROACH
Instead of representing each document as a bag of words, each Web page is represented by a limited number of content and link features. This reduces the dimensionality of the classifier and thus the number of training examples needed. The characteristics of Web structure also can be incorporated into these features easily.

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.
ACM 1-58113-832-6/04/0006.

Based on review of the literature, it is determined that, in general, the relevance and quality of a Web page can be reflected in the following aspects: (1) the content of the page itself, (2) the content of the page's neighbor documents, and (3) the page's link information. A set of 4 to 6 features, calculated based on metrics such as TFIDF, PageRank and HITS, are defined for each aspect. A total of 14 features are defined and used as the input values to machine learning classifiers. A feedforward-backpropagation neural network (NN) [3] and a support vector machine (SVM) [2] are adopted as the classifiers.

3. EVALUATION
An experiment was conducted to compare the proposed approach with two traditional approaches, namely a TFIDF approach and a keyword-based text classifier approach, in the medical domain. A set of 1,000 documents were randomly selected from a medical testbed [1]. A 50-fold cross validation testing was adopted. In general, the experimental results showed that the proposed approach performed better than the traditional approaches in both accuracy and F-measure ($p < 0.005$), especially when the number of training documents is small. When comparing the two proposed methods, it was found that the NN classifier performed better than the SVM classifier ($p < 0.05$). In terms of efficiency, the proposed approach also performed better than the traditional keyword-based approach.

The experimental results are encouraging and show that the proposed approach can be used for Web page filtering by effectively applying Web content and link analysis. The proposed approach also is useful for vertical search engine development, as well as other Web applications.

REFERENCES
[1] Chau, M. and Chen, H. (2003). "Comparison of Three Vertical Search Spiders," *IEEE Computer*, 36(5), pp. 56-62.

[2] Joachims, T. (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings of the European Conference on Machine Learning*, Berlin, 1998, pp. 137-142.

[3] Lippmann, R. P. (1987). "An Introduction to Computing with Neural Networks," *IEEE Acoustics Speech and Signal Processing Magazine*, 4(2), pp. 4-22.

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM