Using Content-Based and Link-Based Analysis in Building Vertical Search Engines

Michael Chau[1] and Hsinchun Chen[2]

[1] School of Business, The University of Hong Kong, Pokfulam, Hong Kong
mchau@business.hku.hk
[2] Department of Management Information Systems,
The University of Arizona, Tucson, Arizona 85721, USA
hchen@bpa.arizona.edu

Abstract. This paper reports our research in the Web page filtering process in specialized search engine development. We propose a machine-learning-based approach that combines Web content analysis and Web structure analysis. Instead of a bag of words, each Web page is represented by a set of content-based and link-based features, which can be used as the input for various machine learning algorithms. The proposed approach was implemented using both a feedforward/backpropagation neural network and a support vector machine. An evaluation study was conducted and showed that the proposed approaches performed better than the benchmark approaches.

1 Introduction

The number of indexable pages on the Web has exceeded three billion and it has become increasingly difficult for search engines to keep an up-to-date and comprehensive search index. Users often find it difficult to search for useful and high-quality information on the Web using general-purpose search engines, especially when searching for specific information on a given topic. Many vertical search engines, or domain-specific search engines, have been built to alleviate his problem to some extent by providing more precise results and more customized features in particular domains. However, these search engines are not easy to build. There are two major challenges for vertical search engine developers: (1) How to locate relevant documents on the Web? (2) How to filter irrelevant documents from a collection? This study addresses the second issue, the Web page filtering problem, and proposes new approaches.

2 Related Work

Web page filtering is important in the process of vertical search engine development. In general, the filtering techniques can be classified as follows: (1) domain experts manually determine the relevance of each Web page (e.g., Yahoo); (2) the relevance

Z. Chen et al. (Eds.): ICADL 2004, LNCS 3334, pp. 515–518, 2004.

of a Web page is determined by the occurrences of particular keywords (e.g., *computer*) [6]; (3) TFIDF (term frequency * inverse document frequency) is calculated based on a lexicon created by domain-experts and Web pages are then with a high similarity score to the lexicon are considered relevant [1]; and (4) text classification techniques such as the Naive Bayesian classifier are applied [3, 9]. Among these, text classification is the most promising approach. Techniques such as Naïve Bayesian model, neural networks, and support vector machines have been widely used in text classification. It has been shown that SVM achieved the best performance among different classifiers on the Reuters-21578 data set [7, 10].

When applied to Web page filtering, few of the text classifiers, however, have made use of the special characteristics of the Web, such as its unique hyperlink structure which has been increasingly used in other Web applications. For example, the PageRank score, computed by weighting each in-link to a page proportionally to the quality of the page containing the in-link, has been applied in the search engine Google for search result ranking [2]. In addition, since Web pages are mostly semi-structured documents like HTML, useful information often can be used to derive some important features of a Web document. Such metrics and other characteristics of Web pages could possibly be applied to improve the performance of traditionally text classifiers in Web page filtering.

3 Proposed Approach

One of the major problems of traditional text classifiers is the large number of features, which result in long processing time. To address this problem, we propose to represent each Web page by a limited number of content and link features rather than as a vector of words. 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.

Based on our review of the literature, we 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 (similarity of the document's title and body text to a domain lexicon); (2) the content of the page's neighbor documents (including parents, children, and siblings); and (3) the page's link characteristics (including PageRank, HITS, number of in-links, and anchor text information). A set of 4 to 6 features , calculated based on metrics such as those mentioned above, are defined for each aspect. A total of 14 features are defined and used as the input values to machine learning classifiers. We adopt a feedforward-backpropagation neural network (NN) [8] and a support vector machine (SVM) [7] as our classifiers.

4 Evaluation

An experiment was conducted to compare the proposed approach with two traditional approaches: a TFIDF approach (Benchmark 1); and a keyword-based text classifier approach using SVM (Benchmark 2), in the medical domain. The proposed Web-feature-based approaches are codenamed Approach 1 (for the neural network

classifier) and Approach 2 (for the SVM classifier) in our experiment. A set of 1,000 documents were randomly selected from a medical testbed created in our previous research [4, 5]. A medical lexicon, created based on the metathesaurus of the Unified Medical Language System (UMLS), was also used in our experiment. A 50-fold cross validation testing was adopted. Testing was performed for 50 iterations, in each of which 49 portions of the data (980 documents) were used for training and the remaining portion (20 documents) was used for testing.

Accuracy and F-measure were to measure the effectiveness of the proposed approaches. The macro-averages for each approach are shown in Table 1. In general, the experimental results showed that the proposed approaches performed significantly 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, we found that the NN classifier performed significantly better than the SVM classifier ($p < 0.05$). In terms of efficiency, the proposed approaches also performed better than the traditional keyword-based approach.

Table 1. Experiment results

	Accuracy	**F-measure**
Benchmark 1	80.80%	0.6005
Benchmark 2	87.80%	0.6646
Approach 1	89.40%	0.7614
Approach 2	87.30%	0.7049

To study the efficiencies of the different approaches, we also recorded the time needed for each system to perform the 50-fold cross validation, including both training and testing time. We found that Benchmark 2 (the keyword-based SVM) took the longest time (382.6 minutes). The reason is that each document was represented as a large vector of keywords, which created a high dimensionality for the classifier. The classifier had to learn the relationships between all these attributes and the class attribute, thus requiring more time. Benchmark 1 (TFIDF) used the least time, as it only had to calculate the TFIDF score for each document and determine the threshold, both of which did not require complex processing. Approach 1 (103.5 minutes) and Approach 2 (37.6 minutes) are in the middle. Approach 1 required a longer time than Approach 2 because the neural network had to be trained in multiple epochs, i.e., in each iteration the training data set had to be presented to the network thousands of times in order to improve the network's performance.

5 Conclusion

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. We believe that the proposed approach is useful for vertical search engine development, as well as other Web applications. We also plan to apply the techniques to Web page filtering in other languages, such as Chinese or Japanese.

Acknowledgement

This project has been supported in part by the following grants: NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management" (IIS-9817473, Apr 1999-Mar 2002), NIH/NLM Grant (PI: H. Chen), "UMLS Enhanced Dynamic Agents to Manage Medical Knowledge" (1 R01 LM06919-1A1, Feb 2001-Jan 2004), and HKU Seed Funding for Basic Research, "Using Content and Link Analysis in Developing Domain-specific Web Search Engines: A Machine Learning Approach" (Feb 2004-Jul 2005). We also thank the medical experts who participated in the user studies.

References

1. Baujard, O., Baujard, V., Aurel, S., Boyer, C., and Appel, R. D.: Trends in Medical Information Retrieval on the Internet. Computers in Biology and Medicine 28 (1998) 589–601.
2. Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia (1998).
3. Chakrabarti, S., Dom, B., and Indyk, P.: Enhanced Hypertext Categorization Using Hyperlink. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Seattle, Washington, USA (1998).
4. Chau, M. and Chen, H.: Comparison of Three Vertical Search Spiders. IEEE Computer 36(5) (2003) 56–62.
5. Chen, H., Lally, A. M., Zhu, B., and Chau, M.: HelpfulMed: Intelligent Searching for Medical Information over the Internet. Journal of the American Society for Information Science and Technology, 54(7) (2003) 683–694.
6. Cho, J., Garcia-Molina, H., and Page, L.: Efficient Crawling through URL Ordering. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia (1998).
7. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the European Conference on Machine Learning, Berlin (1998)137–142.
8. Lippmann, R. P.: An Introduction to Computing with Neural Networks. IEEE Acoustics Speech and Signal Processing Magazine 4(2) (1987) 4–22.
9. McCallum, A., Nigam, K., Rennie, J., and Seymore, K.: A Machine Learning Approach to Building Domain-specific Search Engines. In: Proceedings of the International Joint Conference on Artificial Intelligence (1999) 662–667.
10. Yang, Y. and Liu, X.: A Re-examination of Text Categorization Methods. In: Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (1999) 42–49.