Available online at www.sciencedirect.com

ScienceDirect

Decision Support Systems 44 (2008) 482–494

Decision Support Systems

www.elsevier.com/locate/dss

A machine learning approach to web page filtering using content and structure analysis

Michael Chau [a,*], Hsinchun Chen [b]

[a] School of Business, The University of Hong Kong, Pokfulam, Hong Kong
[b] Department of Management Information Systems, The University of Arizona, Tucson, Arizona 85721, USA

Received 15 January 2006; received in revised form 10 February 2007; accepted 13 June 2007
Available online 22 June 2007

Abstract

As the Web continues to grow, it has become increasingly difficult to search for relevant information using traditional search engines. Topic-specific search engines provide an alternative way to support efficient information retrieval on the Web by providing more precise and customized searching in various domains. However, developers of topic-specific search engines need to address two issues: how to locate relevant documents (URLs) on the Web and how to filter out irrelevant documents from a set of documents collected from the Web. This paper reports our research in addressing the second issue. We propose a machine-learning-based approach that combines Web content analysis and Web structure analysis. We represent each Web page 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. Two experiments were designed and conducted to compare the proposed Web-feature approach with two existing Web page filtering methods — a keyword-based approach and a lexicon-based approach. The experimental results showed that the proposed approach in general performed better than the benchmark approaches, especially when the number of training documents was small. The proposed approaches can be applied in topic-specific search engine development and other Web applications such as Web content management.

Keywords: Web page classification; Link analysis; Machine learning; Web mining

1. Introduction

The most popular way to look for information on the Web is to use Web search engines such as Google (www.google.com) and AltaVista (www.altavista.com). Many users begin their Web activities by submitting a query to a search engine. However, as the size of the Web is still growing and the number of indexable pages on the Web has exceeded eight billion, it has become more 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 topic-specific search engines, have been built to facilitate more efficient searching in various domains. These search engines alleviate the information overload problem to some extent by providing more precise results and more customized features [11]. For example, *LawCrawler*

* Corresponding author.
E-mail addresses: mchau@business.hku.hk (M. Chau), hchen@eller.arizona.edu (H. Chen).

0167-9236/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2007.06.002

(www.lawcrawler.com) allows users to search for legal information and provides links to lawyers and legal information and to relevant government Web sites. *BuildingOnline* (www.buildingonline.com) is a specialized search engine for the building industry, where users can search by manufacturers, architects, associations, contractors, etc. *BioView.com* (www.bioview.com) and *SciSeek* (www.sciseek.com) are two other examples that focus on scientific domains.

Although they provide a promising alternative for users, these vertical search engines are not easy to build. There are two major challenges to building vertical search engines: (1) How to locate relevant documents on the Web? (2) How to filter irrelevant documents from a collection? This study tries to address the second issue and to propose new approaches. The remainder of the paper is structured as follows. Section 2 reviews existing work on vertical search engine development, text classification, and Web content and structure analysis. In Section 3 we discuss some problems with existing Web page filtering approaches and pose our research questions. Section 4 describes in detail our proposed approach. Section 5 describes an experiment designed to evaluate our approach and presents experimental results. In Section 6, we conclude our paper with some discussion and suggestions for future research directions.

2. Research background

2.1. Building vertical search engines

A good vertical search engine should contain as many relevant, high-quality pages and as few irrelevant, low-quality pages as possible. Given the Web's large size and diversity of content, it is not easy to build a comprehensive and relevant collection for a vertical search engine. There are two main problems:

- The search engine needs to locate the URLs that point to relevant Web pages. To improve efficiency, it is necessary for the page collection system to predict which URL is the most likely to point to relevant material and thus should be fetched first.
- After the Web pages have been collected, the search engine system needs to determine the content and quality of the collection in order to avoid irrelevant or low-quality pages.

Search engines usually use spiders (also referred to as Web robots, crawlers, worms, or wanderers) as the software to retrieve pages from the Web by recursively following URL links in pages using standard HTTP protocols [9,13,15]. These spiders use different algorithms to control their search. To address the first problem mentioned above, the following methods have been used to locate Web pages relevant to a particular domain:

- The spiders can be restricted to staying in particular Web domains, because many Web domains have specialized contents [36,50]. For example, most Web pages within the domain www.toyota.com will be relevant to automobiles.
- Some spiders are restricted to collecting only pages at most a fixed number of links away from the starting URLs or starting domains [36,45]. Assuming that nearer pages have higher chances of being relevant, this method prevents spiders from going too "far away" from the starting domains.
- More sophisticated spiders use more advanced graph search algorithms that analyze Web pages and hyperlinks to decide what documents should be downloaded. Cho et al. [16] proposed a best-first search spider that used PageRank as the ranking heuristic; URLs with a higher PageRank scores will be visited first by the spider. The spider developed by McCallum et al. [38] used reinforcement learning to guide their spiders to retrieve research papers from university Web sites. Focused Crawler locates Web pages relevant to a pre-defined set of topics based on example pages provided by the user. In addition, it also analyzes the link structures among the Web pages collected [7]. Context Focused Crawler uses a Naïve Bayesian classifier to guide the search process [19]. A Hopfield Net spider based on spreading activation also has been proposed [8,10]. Page content scores and link analysis scores are combined to determine which URL should be visited next by the spider. The spider was compared with a breadth-first search spider and a best-first search spider using PageRank as the heuristics, and the evaluation results showed that the Hopfield Net spider performed better than the other two.

While these methods have different levels of performance in efficiency and effectiveness, in most cases the resulting collection is still noisy and needs further processing. Filtering programs are needed to eliminate irrelevant and low-quality pages from the collection to be used in a vertical search engine. The filtering techniques used can be classified into the following four categories:

- Domain experts manually determine the relevance of each Web page (e.g., Yahoo) [30].
- In the simplest automatic procedure, the relevance of a Web page can be determined by the occurrences of