Computers and Mathematics with Applications 57 (2009) 1057–1072

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A cross-language focused crawling algorithm based on multiple relevance prediction strategies

Zhumin Chen [a,*], Jun Ma [a], Jingsheng Lei [b], Bo Yuan [c], Li Lian [a], Ling Song [a,d]

[a] School of Computer Science and Technology, Shandong University, Jinan 250061, China
[b] College of Information Science and Technology, Hainan University, Haikou 570228, China
[c] Department of Computer Science, University of Southern California Los Angeles, CA 90088, USA
[d] School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China

ARTICLE INFO

Keywords:
Focused crawling
Multiple relevance prediction strategies
Topic taxonomy
Cross-language
Shark-search algorithm

ABSTRACT

Focused crawling is increasingly seen as a solution to address the scalability limitations of existing general-purpose search engines, by traversing the Web to only gather pages that are relevant to a specific topic. How to predict the relevance of the unvisited pages pointed to by candidate URLs in the crawling frontier to a given topic is a key issue in the design of focused crawlers. In this paper, we propose a novel approach based on multiple relevance prediction strategies to address this problem. For cross-language crawling, we first introduce a hierarchical taxonomy to describe topics in both English and Chinese. We then present a formal description of the relevance predicting process and discuss four strategies that make use of page contents, anchor texts, URL addresses and link types of Web pages, respectively, to evaluate the relevance more accurately, in which we propose a particular strategy using Chinese URL addresses to estimate the relevance of cross-language Web pages. Finally, we get a new focused crawling algorithm (FCMRPS, Focused Crawling based on Multiple Relevance Prediction Strategies) based on the combination of these strategies and Shark-Search, which is a classic focused crawling algorithm. Experiments show that the FCMRPS is more effective than the traditional algorithms, namely Breadth-First, Best-First and Shark-Search, in terms of precision and sum of information.

1. Introduction

Due to the limited bandwidth, storage, computational resources and rapid growth of the World Wide Web, unprecedented scaling challenges have been posed for search engines. Although search engine technology has scaled dramatically to keep up with the growth of the Web, these general-purpose crawlers and search engines have presented some serious limitations as follows:

(1) It is impossible for them to index and analyze all pages and maintain comprehensive, up-to-date search indexes.
(2) They may return hundreds or more links to a user's query, however since they lack the understanding of the query the pages pointed to by these links may not closely relevant to the user's query.
(3) They cannot satisfy the query requests of different background, purpose and period.

* Corresponding author.
E-mail addresses: chenzhumin@mail.sdu.edu.cn (Z. Chen), majun@sdu.edu.cn (J. Ma), jshlei@hainu.edu.cn (J. Lei), boyuan@usc.edu (B. Yuan), lianli@sdu.edu.cn (L. Lian), song_ling@sdjzu.edu.cn (L. Song).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.09.021

(4) Dynamic contents, such as news and financial data, on the Web are growing and changed frequently. Many search engines may take up to one month for refreshing their indices on the full Web. Therefore, the query results may be not valid at the time that the query is issued.

Therefore, fast crawling technology is needed to gather the Web pages with high relevance and quality and keep them up to date. It is also necessary to add capabilities to search engines that respond to the particular information needs expressed by topics or interest profiles. So focused crawling is regard as a potential solution to overcome these limitations.

Focused crawlers traverse a subset of the Web to only gather pages that are relevant to a specific topic. An important assumption implicit in focused crawling is that the pages with respect to related topics tend to be neighbors of each other, i.e. topic locality on the Web [1,2]. Thus, the objective of the crawlers is to stay focused, that is, remaining within the neighborhood in which topic-specific pages have been identified. Focused crawlers work like general-purpose spiders, traversing the Web according to an appropriate traversal priority, instead of the Breadth-First or Depth-First ordering. The ideal focused crawlers retrieve the maximal set of relevant pages while simultaneously traversing the minimal number of irrelevant pages on the Web. This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date.

The basic idea of a focused crawler is to optimize the visit priority of the candidate URLs in a crawling frontier that consists of URLs whose corresponding pages have yet to be fetched by the crawler. A URL should get a higher priority if the page pointed to by it has a higher relevant degree. In this paper, we introduce an innovative approach that combines four strategies to predict the relevance more effectively.

The main contributions of this paper are as follows.

(1) A cross-language hierarchical taxonomy is suggested to represent the topics. Based on the taxonomy users can select their interested topics in English or Chinese, and then the crawler can collect high relevant pages in English, Chinese or both. In addition, the topic context is used to weight a given topic and its contextual topics according to their relative hierarchies in the taxonomy.
(2) A formal description of the process of predicting the relevance of the uncrawled pages to a given topic is discussed.
(3) Four relevance predicting strategies based on page contents, anchor texts, URL addresses and link types of Web pages are introduced, respectively, to improve the relevance computation, in which, a special strategy evaluating the relevance based on the unique characteristic of the Chinese URLs is firstly proposed.
(4) A new focused crawling algorithm, named FCMRPS, is presented based on the combination of the above strategies and the Shark-Search algorithm [3], which estimates the relevance mainly based on the page content and anchor text.

Experiments were carried out on the Web for 30 topics in both English and Chinese. The experiments show that the FCMRPS can obtain significantly higher efficiency than these conventional crawling algorithms, i.e. Breadth-First, Best-First and Shark-Search.

The rest of this paper is organized as follows: Section 2 provides an overview of the focused crawling and the Shark-Search algorithm. Section 3 first introduces a cross-language taxonomy for topic description and then presents a formal description of the relevance predicting process and four relevance prediction strategies. Section 4 describes the details of FCMRPS. Section 5 shows some experimental results and discussions on focused search. Section 6 draws some conclusions and our future work.

2. Related work

Shark-Search [3] is a refined version of Fish-Search which is the first dynamic focused crawling algorithm. Fish-Search [4] is based on the schools of fish metaphor: A school of fish moves in the direction of food. Each URL corresponds to a fish whose survivability is dependent on visited page relevance and remote server speed. Page relevance is estimated based on the page textual content using a binary classification (the page can only be relevant or irrelevant). [5] presents an improved Fish-Search. It points out that the random of search range of original Fish-Search would lead to repeated search. Different fishes moving in different directions can be regarded as different directed graphs. A "distance" parameter that is the distance between the centers of two directed graphs is used to control the search direction. The distance is calculated in graph theory. So, by adjusting "distance" to be a reasonable value between different fishes adaptively, the repeated search problem can be solved. [3] extends Fish-Search into Shark-Search. The given topics are described in keywords. Two improvements are made to the original Fish-Search algorithm to overcome some limitations. One immediate improvement is that relevance between page content and topic is calculated by vector space model and can be any real number between 0 and 1. Another significant improvement is that candidate URLs to be downloaded are prioritized by taking into account a linear combination of page content and anchor text relevance on the source page. Experiments show that the Shark-Search performs between 1.5 and 3 times better than its ancestor does. In [6], a link analysis technology is used to improve the Shark-Search. Some literatures [7–13] make use of PageRank [14] as link analysis algorithm to evaluate the importance of candidate URLs. Although PageRank is effective to rank the results of search engines, they are not suitable for focused crawling for the reason that its process is computationally expensive and based on the overall Web graph [7,10–13]. Therefore, link type analysis is utilized to estimate the relevance of candidate URLs. Links are divided into five groups according to the relative position of the candidate URL to its parent in the Web graph. Then, five heuristic rules are presented to infer the topical relation of a page to its parent page