Data & Knowledge Engineering 88 (2013) 75–93

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

A topic-specific crawling strategy based on semantics similarity

YaJun Du *, QiangQiang Pen, ZhaoQiong Gao

School of Mathematical and Computers Science, Xihua University, Chengdu, 610039 Sichuan, PR China

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form 7 September 2013
Accepted 7 September 2013
Available online 18 October 2013

Keywords:
Search engine
Focused crawling
Formal concept analysis
Web crawler
Concept context graph
Web information systems
Information retrieval

ABSTRACT

With the Internet growing exponentially, search engines are encountering unprecedented challenges. A focused search engine selectively seeks out web pages that are relevant to user topics. Determining the best strategy to utilize a focused search is a crucial and popular research topic. At present, the rank values of unvisited web pages are computed by considering the hyperlinks (as in the PageRank algorithm), a Vector Space Model and a combination of them, and not by considering the semantic relations between the user topic and unvisited web pages. In this paper, we propose a concept context graph to store the knowledge context based on the user's history of clicked web pages and to guide a focused crawler for the next crawling. The concept context graph provides a novel semantic ranking to guide the web crawler in order to retrieve highly relevant web pages on the user's topic. By computing the concept distance and concept similarity among the concepts of the concept context graph and by matching unvisited web pages with the concept context graph, we compute the rank values of the unvisited web pages to pick out the relevant hyperlinks. Additionally, we constitute the focused crawling system, and we retrieve the precision, recall, average harvest rate, and *F-measure* of our proposed approach, using Breadth First, Cosine Similarity, the Link Context Graph and the Relevancy Context Graph. The results show that our proposed method outperforms other methods.

1. Introduction

The search engine is an important tool when searching for information on the web. To satisfy the search requirements of various users, a universal search engine collects as many URLs as possible and even seeks to cover the entire web. It does not matter much whether the URLs of the search results are the most relevant to the user query. Because of the exponential increase in URLs, it is impossible to cover the entire web. Google, for example, only covered 30–40% of Internet URLs in 2005; 60% [1] in 2006. On the other hand, the majority of web pages returned by universal search engines are useless for the user. The reasons are as follows:

- Users' interests in web pages mainly concern one or several topics. An excessive number of web pages is irrelevant to their interests [2].
- Web crawlers cost too much time to crawl the web for each search. The updating period of web pages indexed by search engines is long, and many of the web pages and web sites have become out-of-date. Information that is out-of-date cannot reflect the dynamic changes of the web pages [3].

To address these issues, focused search engines have been developed and have increasingly become popular. In contrast to universal search engines, a focused search engine has a special web crawler. The special web crawler traverses the web to collect

* Corresponding author. Tel.: +86 28 87720554, 13980097252.
E-mail address: dyjdoc2003@aliyun.com (Y. Du).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.09.003

as many relevant web pages as possible. During the crawling, the special web crawler first judges the web pages' relevancy degree to the current topics and weighs them according to their priorities; then, it downloads web pages according to their priority levels in a real-time fashion. The relevancy degree is usually computed by a vector space model; the current topics and unvisited web pages are vectored into two vectors by using the term frequency of the TF–IDF method, and different similarity measuring models are used to compute their relevancy degrees. To some extent, a focused search engine has a better retrieval precision and recall than a universal search engine. In fact, the semantic meanings between the current topic and an unvisited web page are not considered for computing the similarity. In this paper, we discuss the concept of the current topic and an unvisited web page, and we propose a novel semantic ranking method to guide the web crawler when retrieving web pages that are highly relevant to the user topic. Furthermore, we merge the semantic relations among terms into a semantic ranking for an unvisited web page.

Our research has its foundations in previous studies [3–5]. We proposed the topic-specific web crawler with the concept context graph [4]. The concept context graph is only a prototype and basic idea of representing the history crawling of web crawler and extracting its context knowledge. In this paper, we divide the prototype into three parts (mining the core concept; constructing context graph, finding an appropriate position) and develop the three algorithms of implementing the topic-specific web crawler with concept context graph. In reference [3], we discussed the definitions and described the methods of the adding concept and deleting concept on concept context graph. Based on these definitions and methods, in this paper, we develop the two algorithms how to add a new concept to the concept context graph, delete an old concept from the concept context graph, in detail. We proposed the similarity concept context graph (*SCCG*) to guide the focused web crawler [5]. The intent of the core concept includes all user query terms in *SCCG*. We adopted the core concept to constructing the core similarity graph which it is the set of two-tuples based given formal context and wordnet. The innermost layer of *SCCG* is the core similarity graph, is not the core concept. The concepts of the other layers depend on their similarities with concepts of core similarity graph. Obviously, the core concept represents accurately the user query intention, it improves the precision and decreases the recall of focused crawler. On the other hand, the core similarity graph increases the complexities of computing the similarity among the concepts, and constructing the *SCCG*.

A brief summary of the contributions of this paper is as follows:

(1) We propose a novel method for focused crawling with a Concept Context Graph (*CCG*). The context graphs are used as knowledge context and are applied to guide a focused crawler in the next crawling. The context graphs are constructed by using history-clicked web pages in a Link Context Graph (*LCG*) [6] and a Relevancy Context Graph (*RCG*) [1]. The hyperlink relations and the term-based similarities among the history-clicked web pages are very important factors for constructing the context graph in *LCG* and *RCG*, respectively. In our proposed Concept Context Graph, the user chooses and clicks web pages from which submitted user topics are retrieved by Google APIs. By using formal concept analysis, concepts are extracted from these web pages. Similarities between concepts are computed. In our focused crawling approach, we construct the context graph with the similarity between concepts to guide the focused crawling.
(2) We systematically developed the concept context graph method and its algorithm. This method includes three parts: the construct concept context graph, for which there are two algorithms (the mining core concept and constructing *CCG*); finding a proper concept in the *CCG* for the visiting URL; and incrementally updating *CCG* for focused web crawling, for which there are two algorithms (add update *CCG* and delete update *CCG*).
(3) On the same data set, we developed our experiments for comparing our proposed *CCG* with different crawling strategies, such as breath-first web crawling (*BFC*), cosine similarity web crawling (*CosFC*), link context graph web crawling (*LCG*) and relevancy context graph web crawling (*RCG*). The experiments prove that our proposed crawling strategy outperforms *BFC*, *CosFC*, *LCG* and *RCG*.

2. Related work

In this section, we divide the focused crawler into topic-induced and history context-induced focused crawler from the perspective of the implementation method. On the other hand, the concept context graph is constructed by using formal concept analysis (FCA), we recall some researches about FCA.

2.1. Topic-induced focused crawler

The focused web search engine appears along with the search engine. Based on the matching web page's content with the search topic, DeBra P., Houben G. [7] first proposed a crawling strategy called *FishSearch*. The main idea behind *FishSearch* stems from the concept of schools of fish. If a school of fish finds food, then the fish reproduces and continues looking for more food. To search for URLs that have a greater importance first, Cho J. and others [8,9] brought up a new focused web crawling strategy. Here, "importance" represents a web page's similarity to the user's search requirements. The in-degree, out-degree, *PageRank* [10], and URL locations are considered for computing the similarity. Their research shows a crawler that calculates similarity by using the in-degree of the URL and that performs almost as well as *DepthFirst* crawling. Both of these methods are prone to search the URLs in a cluster first, and the selection of different initial URL sets can yield results that are quite different. However, the crawler using the PageRank algorithm does not appear to have this type of problem. It combines the advantages of both *BreadthFirst* crawling and *DepthFirst* crawling. Pant G., Srinivasan P. [11] found a special dependence between link contexts and crawling and applied the context of a hyperlink or link context to the analysis of the hyperlinked structure of the Web to predict a ranked value for the