Applied Soft Computing 36 (2015) 392–407

Contents lists available at ScienceDirect

Applied Soft Computing

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

An improved focused crawler based on Semantic Similarity Vector Space Model

Yajun Du[a,*], Wenjun Liu[b], Xianjing Lv[b], Guoli Peng[b]

[a] School of Computer and Software Engineering, Xihua University, Chengdu 610039, China
[b] Xihua University Library, Chengdu 610039, China

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 27 March 2015
Accepted 15 July 2015
Available online 1 August 2015

Keywords:
Focused crawler
Semantic similarity
VSM
SSRM

ABSTRACT

A focused crawler is topic-specific and aims selectively to collect web pages that are relevant to a given topic from the Internet. In many studies, the Vector Space Model (VSM) and Semantic Similarity Retrieval Model (SSRM) take advantage of cosine similarity and semantic similarity to compute similarities between web pages and the given topic. However, if there are no common terms between a web page and the given topic, the VSM will not obtain the proper topical similarity of the web page. In addition, if all of the terms between them are synonyms, then the SSRM will also not obtain the proper topical similarity. To address these problems, this paper proposes an improved retrieval model, the Semantic Similarity Vector Space Model (SSVSM), which integrates the TF*IDF values of the terms and the semantic similarities among the terms to construct topic and document semantic vectors that are mapped to the same double-term set, and computes the cosine similarities between these semantic vectors as topic-relevant similarities of documents, including the full texts and anchor texts of unvisited hyperlinks. Next, the proposed model predicts the priorities of the unvisited hyperlinks by integrating the full text and anchor text topic-relevant similarities. The experimental results demonstrate that this approach improves the performance of the focused crawlers and outperforms other focused crawlers based on Breadth-First, VSM and SSRM. In conclusion, this method is significant and effective for focused crawlers.

1. Introduction

Web crawlers are programs that collect information from web pages on the World Wide Web (WWW). In the crawling process, the web crawlers first traverse the seed URLs, and then, parse the hyperlinks in the newly downloaded web pages, add them to an unvisited list, select a URL from the unvisited list to traverse, and continue the above operations repeatedly. Web crawlers can be divided into general-purpose and special-purpose web crawlers [1–3]. General-purpose web crawlers retrieve enormous numbers of web pages in all fields from the huge Internet. To find and store these web pages, general-purpose web crawlers must have long running times and immense hard-disk space. However, special-purpose web crawlers acquire many web pages that are relevant to a given topic. Compared with general-purpose web crawlers, special-purpose web crawlers obviously need a smaller amount of runtime and hardware resources. With the sharp growth of web pages in the WWW, special-purpose web crawlers, known as focused crawlers, have become increasingly important in gathering information from web pages for finite resources and in better satisfying the needs of the topic search for the large variety of user interests [4–6].

The most critical task of the focused crawler is selecting the URLs. All of the new URLs are extracted from downloaded web pages and are consecutively traversed by the focused crawlers. The selection of URLs ensures that the crawler acquires more web pages that are relevant to a special topic [7–10]. The selected URLs are classified into two types: the seed URLs from the Internet and the updated URLs from the unvisited list [11]. The seed URLs that are related to a given topic can be selected from the gathered results [12], which are retrieved by inputting topic queries to the general search engine. Moreover, the focused crawler itself can also obtain the seed URLs by providing topic keywords to the focused crawler [13]. The updated URLs are selected from the unvisited list, where the URLs are ranked in descending order based on the weights that are relevant to the given topic. In [14–16], the texts of retrieved web pages were employed to assign topic-relevant priorities to rank the unvisited hyperlinks, and each priority for a hyperlink was computed by averaging the concept values of the web pages that contain the hyperlink.

* Corresponding author.
E-mail address: dyjdoc2003@aliyun.com (Y. Du).

http://dx.doi.org/10.1016/j.asoc.2015.07.026
1568-4946/© 2015 Elsevier B.V. All rights reserved.

Most focused crawlers take advantage of the web page text to calculate the topic-relevant similarities for unvisited hyperlinks to rank them. First, focused crawlers must determine the text type of each unvisited hyperlink to compute their traversed priorities. In [17–19], the documents of the hyperlinks consist of the full texts and anchor texts of the web pages. Specifically, the downloaded priorities for the hyperlinks are computed by linearly integrating the correlations between the full texts, anchor texts and given topic. The correlations are calculated based on information retrieval models such as the Vector Space Model (VSM) [20]. The VSM builds the document vector and query vector that are composed by using the Term Frequency Inverse Document Frequency (TF*IDF) of the terms. The web page must have common terms with the topic to compute its topical similarity. Otherwise, there is no similarity between the web page and the topic. Classic information retrieval models, such as the VSM, achieve topic-relevant degrees based on the terms that are common between the documents and the topic. If there are not any of the same terms between a document and the topic, the document will be topic-irrelevant. However, two lexically different terms does not mean that they must be irrelevant because they might be semantically similar [21]. Consequently, the classic methods are not able to collect conceptually similar documents. To solve the problem, the Semantic Similarity Retrieve Model (SSRM) was proposed to cause focused crawlers to retrieve web pages that have semantically similar terms [22,23]. The SSRM indicates that the document correlation is computed by multiplying the term frequencies and the semantic similarity of the two terms, summing these product values and finally normalizing the accumulated value.

These focused crawlers can reliably predict the priorities of all of the unvisited hyperlinks to rank those hyperlinks. However, there are several problems that are challenges to these focused crawlers, as follows:

(1) Certain focused crawlers cannot take semantic similarity into consideration. For the focused crawlers that are based on the VSM, there must be common terms between the documents of the hyperlinks and the given topic by using the cosine similarity to compute the topical similarities of the documents, to predict the priorities of the hyperlinks. In other words, if there are no common terms between a document of a hyperlink and the given topic, then the topical similarity of the document is zero, i.e., the document is irrelevant to the given topic. However, in the VSM, a document of a hyperlink and the given topic could have similar semantic terms and even the same semantic terms. Then, although there are no common terms between the document and the given topic, the topical similarity of the document is still equal to zero based on the VSM. In fact, due to the existing similar semantic terms, the document should be relevant to the given topic. Therefore, the VSM cannot combine the semantic similarity, and the focused crawlers based on it might not retrieve semantically similar web pages from the Internet.

(2) A number of focused crawlers cannot take the cosine similarity into consideration. For a focused crawler based on the SSRM, although there are no common terms between the documents of the hyperlinks and the given topic, the topical similarities of the documents can be computed by using the semantic similarity and will not be equal to zero, as it would be for a focused crawler based on the VSM. However, in the two term sets of a document of a hyperlink and the given topic, all of the terms in the two sets can be synonyms, and all of the semantic similarities between the two arbitrary terms are equal to the identical value of one. Then, even though the term frequencies of the document are largely different from the term frequencies for the given topic, the topical similarity of the document becomes one by using the SSRM. In fact, due to the existing, largely different term frequencies between the document and the given topic, the topical similarity of the document cannot be equal to one. In contrast, the VSM can obtain the proper value. Therefore, the SSRM cannot combine the cosine similarity, and the focused crawlers based on it could incorrectly gather web pages.

To solve these problems, this paper proposes an improved approach, the Semantic Similarity Vector Space Model (SSVSM). This model combines cosine similarity and semantic similarity and uses the full text and anchor text of a hyperlink as its documents. The SSVSM first computes the TF*IDF values of the terms and semantic similarities among the terms, and these terms are extracted from the full texts, anchor texts and given topic. Specifically, the semantic similarities between the terms are calculated based on the ontology and the relationships in that ontology. In the experiment, the WordNet ontology was selected to compute the semantic similarities between the terms because it is the most popular natural language ontology. Then, the SSVSM establishes the document and topic semantic vectors, mapped to the same double-term set for each unvisited hyperlink. In addition, the document semantic vectors for each unvisited hyperlink include two full text and anchor text semantic vectors. Simultaneously, the SSVSM computes the cosine similarity between the full text and topic semantic vectors as the full text topic-relevant similarity and the cosine similarity between the anchor text and topic semantic vectors as the anchor text topic-relevant similarity for each unvisited hyperlink. Finally, the priority of each unvisited hyperlink is computed by linearly integrating its full text and anchor text topic-relevant similarities. The SSVSM can more accurately predict the priorities of an unvisited hyperlink that are related to the given topic and guide a focused crawler to continuously download a larger quantity of higher quality topic-relevant web pages from the Internet. The experimental results demonstrate that the proposed method improves the performance of the focused crawlers and outperforms the Breadth-First Crawler, VSM Crawler and SSRM Crawler. In conclusion, the SSVSM method is significant and effective for focused crawlers.

The contributions of this paper are as follows:

(1) The SSVSM is proposed. It integrates the cosine similarity and the semantic similarity to predict the priorities of the unvisited hyperlinks. Additionally, relevant definitions are given in this paper, and they primarily include the definitions of the term space, term vector, semantic space and semantic vector.

(2) Four focused crawlers based on Breadth-First, VSM, SSRM and SSVSM have been implemented and evaluated. The performance of the four focused crawlers was evaluated by using the harvest rate, average similarity, the average error.

The remainder of this paper is constructed as follows: Section 2 introduces two typical focused crawlers and related advanced techniques. In Section 3, the improved focused crawler based on the SSVSM is proposed, and the experimental results are analyzed in Section 4. Finally, Section 5 presents the conclusions and outlines further research.

2. Related works

Focused crawlers must predict the priorities for unvisited hyperlinks, to guide themselves to retrieve web pages that are related to a given topic from the Internet. The priorities for the unvisited hyperlinks are affected by two main factors, including topical similarities of the full texts and the anchor texts of those hyperlinks [12,17]. In other words, these priorities are computed by linearly integrating