Neurocomputing 123 (2014) 266–280

Contents lists available at ScienceDirect

Neurocomputing

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

A novel focused crawler based on cell-like membrane computing optimization algorithm

WenJun Liu, YaJun Du *

School of Mathematics and Computer Engineering, Xihua University, Chengdu 610039, China

ARTICLE INFO

Article history:
Received 5 November 2012
Received in revised form
5 June 2013
Accepted 27 June 2013
Communicated by: Xiaofei He
Available online 20 August 2013

Keywords:
Focused crawler
Membrane computing
Optimization algorithm
VSM

ABSTRACT

In many research works, topical priorities of unvisited hyperlinks are computed based on linearly integrating topic-relevant similarities of various texts and corresponding weighted factors. However, these weighted factors are determined based on the personal experience, so that these values may make topical priorities of unvisited hyperlinks serious deviations directly. To solve this problem, this paper proposes a novel focused crawler applying the cell-like membrane computing optimization algorithm (CMCFC). The CMCFC regards all weighted factors corresponding to contribution degrees of similarities of various texts as one object, and utilizes evolution regulars and communication regulars in membranes to achieve the optimal object corresponding to the optimal weighted factors, which make the root measure square error (RMS) of priorities of hyperlinks achieve the minimum. Then, it linearly integrates optimal weighted factors and corresponding topical similarities of various texts, which are computed by using a Vector Space Model (VSM), to compute priorities of unvisited hyperlinks. The CMCFC obtains more accurate unvisited URLs' priorities to guide crawlers to collect higher quality web pages. The experimental results indicate that the proposed method improves the performance of focused crawlers by intelligently determining weighted factors. In conclusion, the mentioned approach is effective and significant for focused crawlers.

1. Introduction

Web crawlers are applications to retrieve as many web pages as possible from the Internet. The relationships among these pages were formed based on their hyperlinks, and pages as nodes and hyperlinks as edges could construct the enormous and complex directed graph [1–3]. The web crawlers usually traversed pages in the above directed graph based on breadth-first algorithm until all pages were collected or there was no vacant storage space. But computer resources, such as running time, memory space and network bandwidth, are very finite, and web pages in the Internet still increase explosively, so that it is impossible to retrieve all web pages from the World Wide Web (WWW). To solve the problem, the focused crawler was proposed to only gather topic-relevant web pages [4–6]. Comparison with the general web crawlers, focused crawlers obviously reduce massive time and space resources and satisfies user needs better. The focused crawlers can be guided to retrieve topic-relevant web pages based on the text contents and the link structures of them.

Focused crawlers mostly pay attention to the creeping performance. The target for focused crawlers was to consume less time and space resources to retrieve more quantity and higher quality web pages related to the given topic [7–9]. Generally, the crawling process of focused crawlers can be divided into two major phases: determining the initial URLs seed and selecting the better unvisited URLs. The initial seed can be classified into the topical seed and the generic seed, and they are obtained based on two different methods. The topical seed URLs were selected from retrieved results by means of inputting the topic-relevant keywords into the general search engine, such as Google [10]. Differently, the generic seed URLs were directly hyperlinks pointing to web directory pages at the top of the hierarchy, which chained to lots of topical web pages at the lower level [11]. The above two ways have respective advantages and disadvantages, and the approach acquiring the initial seed affects the performance of focused crawlers. The better unvisited hyperlinks possessed higher similarities related to the given topic, which were considered as priorities of these hyperlinks [12–14]. Before selecting unvisited URLs to consecutively download pages for focused crawlers, all unvisited URLs are firstly extracted from crawled pages, and various texts of each unvisited hyperlink are subsequently acquired. Secondly, similarities between these texts and the topic are measured by using some information retrieval models, such as the Vector Space Model (VSM) [15]. Thirdly, the priority of each unvisited hyperlink is obtained by integrating these topic-relevant similarities of different texts of hyperlinks. These priorities of

* Corresponding author. Tel.: +86 1 398 0097 252.
E-mail address: dyjdoc2003@aliyun.com (Y. Du).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.06.039

hyperlinks were used to detect the order traversing these unvisited URLs [16–18]. Finally, focused crawlers always select better unvisited hyperlinks with higher priorities to consecutively gather web pages from the Internet.

Focused crawlers need to confirm the document types of unvisited URLs to assign topic-relevant priorities of these URLs. In [19], the full texts of pages were considered as the documents of hyperlinks contained by these pages, and priorities of unvisited hyperlinks were directly topical similarities of web pages including these hyperlinks. In addition, the Semantic Similarity Retrieve Model (SSRM) was put forward to make focused crawlers retrieve pages with semantically similar terms [20,21]. The SSRM also applied full texts to compute priorities of unvisited hyperlinks. The full text can globally indicate the topical relevance of the hyperlink, but it is not obviously neglectful that the anchor text can directly describe the topical information of the page chained by the hyperlink. Both full texts of pages and anchor texts of an unvisited hyperlink were considered as the documents of the hyperlink included by these pages [22,23]. Correspondingly, the priority of the hyperlink was acquired by linearly combining the topical similarity of anchor texts of the hyperlink and topical similarities of full texts of pages containing the hyperlink. Differently, the documents of a hyperlink were composed of anchor texts of a hyperlink and title texts of pages including the hyperlink [24]. Moreover, Ahmed Patel et al. proposed that the priority of a hyperlink was contributed by many structural fragments of pages [25]. Besides the anchor text, these structural fragments contained titles of pages including the hyperlink, heading of sections, surroundings of paragraphs, even table captions and image descriptions containing the hyperlink.

These mentioned focused crawlers can largely retrieve many web pages related to the given topic from the Internet. However, there are some problems for these focused crawlers as follows:

(1) The considerations of priorities of unvisited URLs are not comprehensive. Most of the above literatures about focused crawlers only take full texts of pages and anchor texts of an unvisited hyperlink as the documents of the hyperlink included by these pages in the experiment. The priority of the hyperlink is computed by linearly integrating topic-relevant similarities of these full texts and anchor texts. In these focused crawlers, some other different texts of pages, such as title texts of pages, may also relate to the topic information of the page chained by the hyperlink, but these texts are not still considered as the documents of the hyperlinks. Therefore, these focused crawlers cannot acquire very accurate priorities of hyperlinks, and may be incorrectly guided to retrieve massive irrelevant web pages to the given topic.
(2) It is casual to determine weighted factors for computing priorities of unvisited URLs. The priority of an unvisited hyperlink is calculated by linearly merging topical similarities of various texts and coincident weighted factors. In the most focused crawlers, these weighted factors are acquired through the personal experience, and these values are subjective and permissive. Because these focused crawlers lack self-organizing, self-adaptive and self-learning abilities, these weighted factors are not able to objectively and really display the contribution degrees of different texts to the priority of the hyperlink. Therefore, priorities of unvisited hyperlinks may have very great deviations and these inaccurate values will incorrectly guide focused crawlers to gather massive irrelevant web pages and obviously reduce the performance of focused crawlers.

To solve aforesaid problems, this paper proposes a novel focused crawler based on cell-like membrane computing optimization algorithm (CMCOA), called as CMCFC. In the CMCFC, the documents of an unvisited hyperlink are composed of full texts of pages, anchor texts of the hyperlinks, title texts of pages and surrounding texts of paragraphs containing the hyperlinks. For computing priorities of hyperlinks, there are four topical similarities and four weighted factors. The CMCFC regards four weighted factors corresponding to contribution degrees of similarities of four texts as one object, and utilizes evolution regulars and communication regulars in membranes to achieve the optimal object corresponding to the optimal weighted factors, which make the root measure square error (RMS) of priorities of hyperlinks achieve the minimum. Then, the CMCFC linearly integrates optimal four weighted factors and corresponding topical similarities of four texts, which are computed based on VSM, to compute topical priorities of unvisited hyperlinks. The experimental results indicate that the proposed algorithm improves the performance of focused crawlers. The mentioned method is effective and promising for focused crawlers.

The contributions of this paper are as follows:

(1) This paper proposes a novel focused crawler, the CMCFC, which applies CMCOA to intelligently acquire optimal four weighted factors. The CMCOA utilizes evolution regulars similar to Generic Algorithm (GA), and evolution regulars in CMCOA are modified by using the adaptive method of the population entropy. Compared with GA, CMCOA achieved better optimal values and had faster convergence [35].
(2) Four focused crawlers including the CMCFC and other three crawlers respectively based on Breadth-First, VSM, and SSRM have been implemented and evaluated. The performance of the four focused crawlers is evaluated by using the harvest rate, the average relevance of topic-relevant pages, the number of topic-relevant pages and the average errors.

The remainder of this paper is organized as follows: Section 2 introduces two representative focused crawlers; in Section 3 the novel focused crawler CMCFC is proposed, and the CMCFC specially utilizes the theory of Membrane Computing to acquire optimal four weighted factors; experimental results are displayed and analyzed in Section 4; Finally, Section 5 puts forward the paper conclusions and further research works.

2. Related works

Focused crawlers compute priorities of all unvisited hyperlinks to induce themselves to collect many topic-relevant web pages. In fact, the priority of each unvisited hyperlink is considered as the predictive topical similarity of the unvisited hyperlink. Topical similarities of various texts of web pages including an unvisited hyperlink may influence the topical similarity (the priority) of the unvisited hyperlink. Computing topical similarities of various texts have two representative approaches: VSM and SSRM. The VSM considers the inner product between document and topic vectors as the topical similarity of the document, and the SSRM computes the topical similarity of the document by associating term frequencies and term semantic similarities and accumulating these products. Correspondingly, there are two focused crawlers: the VSM crawler and the SSRM crawler. These two focused crawlers are detailed in the following passage.

2.1. VSM crawler

The VSM crawler computes priorities by using the cosine similarity. In the VSM crawler, full texts and anchor texts of pages including an unvisited hyperlink were considered as documents of the unvisited hyperlink [22,23]. The topical similarities of documents of unvisited hyperlinks are acquired by calculating inner