Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200731

Focused Crawler Strategy Based on Improved Energy Landscape Paving Algorithm

Jingfa Liu[a,b], Wei Zhang[a,b1], Zhihe Yang[a,b], Ziang Liu[c]

[a] *Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong, University of Foreign Studies, Guangzhou 510006, China*
[b] *School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China*
[c] *Faculty of Science, University of Alberta, Edmonton T6G2H6, Canada*

Abstract. The traditional crawlers have difficulty in implementing semantic analysis. Therefore, the focused crawler technologies with topic preference characteristics have received many attentions in the recent years. To increase the precision of focused crawlers and prevent “topic drifting”, this paper adopts the comprehensive relevancy evaluation (CRE) of hyperlinks based on the combination of web content and link structure. In addition, the improved version of the energy landscape paving (ELP) algorithm that is a class of metropolis-sampling-based global optimization method is proposed to avoid the focused crawler falling into local optima. By incorporating the CRE strategy into the improved ELP, a novel focused crawler strategy denoted by IELP is proposed. The experimental results on rainstorm disasters domain show that the precision of the proposed focused crawler is obviously promoted compared to other focused crawlers in literature, illustrating the ability of the IELP to retrieve topic-related web pages.

Keywords. focused crawler, energy landscape paving, rainstorm disasters, link evaluation

1. Introduction

Due to the huge data resources, the Internet has become an important channel for obtaining specific domain knowledge. At present, the scale of web pages on the Internet is massive and growing, while the content of web pages is highly dynamic and complex. Web pages related to specific domain knowledge are very sparse and have the characteristics of big data. Traditional search engines (such as Google, Baidu) and web crawlers (such as Scrapy, Pyspider) face very challenges in accuracy rate of information retrieval. Unlike these methods, focused crawlers filter web pages based on the topic-relevant values of web pages on the Internet. For specific search topic of users, the results returned by focused crawlers are more streamlined and more accurate. Therefore, the development of the focused crawler has become an inevitable

[1] Corresponding Author, Wei Zhang, Guangdong University of Foreign Studies, Guangzhou 510006, China; E-mail: westrive@qq.com.

trend and attracted attentions of an increasing number of scholars.

Generally, focused crawlers are different in topic-relevancy calculation and search strategies. Among the existing crawlers, topic-relevancy calculation methods are mainly divided into link-based evaluation methods and content-based evaluation methods. The link-based evaluation methods focus on the authority of the structure itself. This type of methods ignores the relevancy between the web pages content and the topic, which is likely to lead to “topic drifting” [1] and obtain massive web pages of irrelevant topics by the crawlers, such as the hyperlink-induced topic search (HITS) algorithm [2] and the PageRank[3]. On the other hand, the evaluation methods based on web content analysis evaluate links by calculating and analyzing the relevancy of the web content, the anchor text, the anchor text context, such as the shark-search algorithm [4], probability model based on Bayes algorithm [5] and topic-focused crawlers based on semantic similarity [6]. In searching strategies, the breadth first search (BFS)[7] and the optimal priority search (OPS)[8] are employed mostly. This sort of methods has better performance for crawling a small number of web pages and searching near the relevant ones. However, it ignores some structural features existing between pages themselves, and the design of crawler system in this way will lead to a lack of loyed mostly. BFS is an omission-free search algorithm, which is widely used in traditional web crawlers. Nevertheless, when used in the focused crawler, the accuracy is not high. With the increase of crawling web pages, a large number of irrelevant web pages will be downloaded, which increase the waste of system resources. For example, the design of the MySQL-based news search engine in [9]. OPS gives priority to visit web pages with high-valued links and extend the best web pages according to a certain priority strategy. Obviously, OPS is a sort of greedy algorithm and is prone to fall into local optima. In order to improve global optimization ability of focused crawlers, in recent years some researchers introduce some global optimization algorithms into the focused crawler, such as genetic algorithm [10], ant colony algorithm [11]. Regrettably, due to the incomplete application of some operation operators (such as crossover and mutation) in this type of methods, they only repeat the selection operation and do not generate new links, so the effects are often not good enough.

The rest of this article is organized as follows. The second section gives the comprehensive topic-relevancy evaluation of the unvisited hyperlinks based on the combination of web page content and link structure. The third section proposes a focused crawler strategy based on the improved energy landscape paving algorithm. The proposed crawler allows sub-optimal hyperlinks to be visited so as to discover web pages with greater potential value and to improve the global search performance of the crawler. The fourth section conducted a controlled experiment, and the fifth section summarizes the work of this article.

2. Comprehensive Topic-Relevancy Calculation

Based on the semantic similarity analysis, the calculation of the topic relevancy of the web page text and the anchor text are first described. Then, the comprehensive topic-relevancy evaluation of the unvisited hyperlink is given to direct the search of the focused crawler.

2.1. Topic-Relevancy Calculation of Web Pages

This paper takes "rainstorm disaster" as the topic and describes it in the form of a topic word set. According to the experience of experts in this field, we set the topic-relevant term set as TK={rainstorm, disaster, rainfall, weather, meteorology} in experiments, and assign corresponding semantic weight $W_{TK}(0.8,0.5,0.3,0.1,0.1)$ to this group of topic words. Most of the web pages on the Internet are represented in a form of HTML files. Different topic words extracted from different labels have different influences on the topic-relevancy of the whole web page. According to an analysis of label structures in large amounts of web page texts, this paper divides the main labels into five groups, as shown in Table 1, and assigns different weights $(w_1, w_2, \ldots, w_5)=(2, 1.5, 1.2, 1.0, 0.2)$ to different labels, respectively.

After segmenting the web page and counting the term frequency (TF) of each topic word, we map a web page text into a feature vector $\{tf_1, tf_2, ..., tf_n\}$, where tf_i denotes the TF of the i-th topic word in the web page, and n is the number of topic words. Considering that the topic words occurring in different positions within the web page text hold different weights, the feature vector of a web text can be expressed as a TF vector $D_{TF}=\{(tf_{11}, \ldots, tf_{1j}, ..., tf_{1n}), ..., (tf_{i1}, ..., tf_{ij}, ..., tf_{in}), ..., (tf_{J1}, \ldots, tf_{Jj}, ..., tf_{Jn})\}$. Here J=5 represents the group number into which labels are divided, and tf_{ij} represents the term frequency of the j-th topic word in the i-th group of the labels within the web page. To calculate the weight w_{dk_i} of the i-th topic word in the web page feature set $DK=\{dk_1, dk_2, \ldots, dk_n\}$, this paper uses the improved TF-IDF model [12]:

$$w_{dk_i}=\sum_{j=1}^{J}\left(\frac{tf_{i,j}}{max(tf_{i,j})}\times w_j\right) \tag{1}$$

Here $\max(tf_{i,j})$ represents the maximum TF of the i-th topic word occurring at all positions within the web page text; w_j represents the weight of the j-th web page label. This paper uses the vector space model (VSM) [13] to calculate the topic relevancy of the web page text:

$$R(P)=\frac{W_{TK}\times W_{DK}}{\| W_{TK} \| \times \| W_{DK} \|} \tag{2}$$

Here $R(P)$ represents the topic relevancy of the web page P; W_{TK} represents the semantic weight vector of topic words; W_{DK} represents the feature weight vector of the web page text. $0 < R(P) < 1$ and the topic relevancy is higher if $R(P)$ is closer to 1.

Table 1. Division of labels and their weights

Groups	Labels	Meanings	Weights(wj)
Group1	<title>, <keyword>, <h1>	title, keyword, first-level headline	2
Group2	<h2>, <h3>	second-level headline, third-level headline	1.5
Group3	<h4>, <h5>, <strong>	fourth-level headline, fifth-level headline, Bold text	1.2
Group4	<p>, <td>, <li>	body information	1.0
Group5	Other labels	non-body information	0.2

2.2. Comprehensive Relevancy Calculation of Links

By filtering out links with low relevancy, the crawler is enabled to select high-quality web pages. The three indicators to measure the relevancy of links are the topic relevancy of the anchor text, the topic relevancy of the web page where the link is located, and PageRank value of the page to which the link points.

Hyperlinks enable interconnection between the web pages. The anchor text information of the hyperlink is often one of the important basis for people to judge whether the web pages will be visited. The short anchor text information is like the title of an article, which often points directly to the topic. The topic relevancy of the anchor text is therefore an important basis to identify whether the link is relevant to the topic in the design of focused crawler system. An anchor text is short in most cases, so this paper adopts an improved Term Frequency × Inverse Document Frequency algorithm to calculate the feature weight of the anchor text AR_i:

$$w_{AR_i} = \frac{f_i}{\sum_{j=1}^{n} f_j} \times \log_a\left(\frac{N}{N_i} + 0.01\right) \tag{3}$$

Here f_i represents the term frequency of the i-the topic word in the anchor text; N represents the total number of the crawled web pages; N_i represents the number of the crawled web pages containing the i-th topic word and $a > 1$. After calculating the feature weight vector $W_{AR} = \{ w_{AR_1}, \ldots, w_{AR_i}, \ldots, w_{AR_n} \}$ of the anchor text, we calculate the cosine of the weight vector W_{TK} of topic words and the feature weight vector W_{AR} to obtain the topic relevancy $R(AR_i)$ of the anchor text AR_i of hyperlink $link_i$ by equation (2).

PageRank (PR) value is generally used to evaluate the importance of a web pages, and web pages with high PageRank value are often more reliable. The core idea of PageRank algorithm is to determine the importance of links through cross-reference relations among pages. The PR value of a web page generally depends on the number of its in-links and their average PageRank value. Although it reflects the reliability of a web page, it cannot tell whether the link is relevant to the topic, so the crawler may visit some pages with high PR value but low topic relevancy. Different from traditional PageRank algorithm, this paper gives an improved calculation method of PR value by integrating the topic relevancy of the in-link anchor text into the calculating process of PageRank algorithm. Suppose that P_{next} is the web page pointed to by the hyperlink $link_i$. The PR value of P_{next} is computed by equation (4).

$$PR(P_{next}) = (1-d) + d \times \sum_{i=1}^{U} \left[\frac{PR(P_i)}{S(P_i)} \times (1 + \lambda \times R(AR_i)) \right] \tag{4}$$

Here d represents the damping coefficient, and λ is the adjustment factor to adjust the influence of anchor text's topic relevancy on PR of the web page P_i. U is the total number of all in-links of P_{next} in the crawled web pages. $PR(P_i)$ is the PR value of the i-th in-link web page of the web page P_{next}. $S(P_i)$ denotes the total number of out-links of the web page P_i. $R(AR_i)$ represents the topic relevancy of the anchor text AR_i of the i-th in-link of P_{next}.

According to the above analysis, a comprehensive relevancy evaluation function of unvisited hyperlink $link_i$ is given as follows:

$$E(link_i) = \alpha \times R(AR_i) + \beta \times R(P_i) + \chi \times PR(P_{next}) \tag{5}$$

Here α, β, and χ represent weight factors. Considering the fact that an anchor text can characterize the target contents from links well, this paper sets α to 0.7, β to 0.2 and χ to 0.1. $R(AR_i)$ denotes the topic relevancy of the anchor text AR_i of the hyperlink $link_i$. $R(P_i)$ denotes the topic relevancy of the web page where the hyperlink $link_i$ is located. $PR(P_{next})$ denotes the PageRank value of the web page pointed to by the hyperlink $link_i$. To avoid visiting irrelevant web pages as much as possible when the focused crawler is crawling, this paper sets a threshold of the comprehensive relevancy of links. Links below this threshold are discarded and those above it are inserted into the waiting queue according to the comprehensive relevancy, thus ensuring that the comprehensive relevancy of links in the queue waiting for access is arranged in descending order.

3. Focused Crawler Based on Improved Energy Landscape Paving Strategy

In this section, we first introduce the energy landscape paving (ELP) algorithm, then design an improved energy landscape paving algorithm, and finally propose a focused crawler strategy based on the improved energy landscape paving algorithm.

3.1. Energy Landscape Paving algorithm

The ELP algorithm [14] is a class of Monte-Carlo-based optimization algorithm put forward by Hansmann etc. in 2002, which combines energy landscape deforming algorithm and Taboo Search. In the solution process, the ELP algorithm usually starts from an initial state X. If X is sampled at time t, its corresponding energy function $E(X, t)$ will be modified by the formula $\tilde{E}(X,t) = E(X,t) - k \times H(X,t)$, which leads to lower probability for state X to be sampled next time. Here k is a constant and the penalty term $H(X, t)$ denotes the frequency histogram function that represents the frequency of being sampled in this state. The frequency histogram will be updated during each Monte-Carlo step. The statistical weight for a state X is defined as $exp\ (\tilde{E}(X,t)/(k_b \times T))$, where k_bT is the thermal energy at the temperature T, and k_b is Boltzmann constant. During the ELP's search, as the number that the ELP visits to certain a local minimum increases, the penalty term $H(X, t)$ also increases, which leads to the modified energy value of the state X reduces. As a result, the recently visited repeatedly region will be avoided during searching, which is conducive for the ELP to jump out of the local minimum, search a wider space, and finally may obtain the global optimum state of the problem.

In the focused crawler designing, we use the comprehensive relevancy of hyperlink as its energy value, and the number of times the web page is downloaded as its frequency value of the histogram function. The focused crawler based on the ELP is prevented from repeatedly visiting the web pages that locate at local minima via the influence of punitive frequency histogram. On the other hand, the method of Metropolis sampling is adopted to make it possible for the sub-optimal link to be accessed in advance. This "non-greedy" link selection strategy improves the network coverage of the focused crawler, thus effectively avoiding local optimization of the crawler.

3.2. Focused Crawler Based on IELP

In the original ELP method, there is a technical defect that after the current state X_1 tries to visit the new state X_2, X_2 is accepted only if $random[0,1] < exp\ ((\tilde{E}(X_2,t) - \tilde{E}(X_1,t))/(k_b \times T))$ is satisfied. Due to the existence of a punitive frequency histogram function, it may miss the higher energy around X_1. Liu et al. [15] made an improvement on the original ELP algorithm: when $E(X_1,t) < E(X_2,t)$, accepted X_2 unconditionally, and when $E(X_1,\mathrm{t}) \geq E(X_2,\mathrm{t})$, if X_2 satisfies $random[0,1] < exp\ ((\tilde{E}(X_2,t) - \tilde{E}(X_1,t))/(k_b \times T))$, receives X_2, otherwise does not receive X_2. In the original ELP algorithm and the improved version [15], the frequency histogram function $H(X, t)$ is updated in each Monte Carlo process. That is to say, for a new link, although it is not accepted, its frequency histogram function will be updated. Due to the increase of the frequency histogram function, these links may be difficult to be accepted in subsequent simulations, and some high-quality links will be missed. In response to this defect, we propose a new frequency histogram update strategy based on the improved version [15]: the frequency histogram is updated only when the newly generated link is accepted. On the other hand, the original ELP adds all sub-links to the waiting queue, while the improved ELP (IELP) filters the sub-links. Only keep sub-links whose comprehensive relevance is higher the threshold of links' comprehensive relevancy "r_2", thereby effectively reducing access to low-quality web pages. This strategy is more helpful for crawlers to retrieve high-quality web pages.

The focused crawler based on IELP algorithm starts from the initial seeds. It downloads the corresponding web pages, extracts all the links whose comprehensive relevancy is higher than the set threshold and adds them into the waiting queue, and then selects the next link in the waiting queue by using the IELP algorithm. The specific process of the focused crawler based on the IELP is as follows:

Step 1: Initialize the waiting queue of links, the threshold of web page relevancy "r_1" and the threshold of link relevancy "r_2". Initialize k and let t=1.

Step 2: Rank the waiting queue of links in descending order according to their comprehensive relevancy, and pick out the head link, signed as the current link.

Step 3: Analyze the corresponding web page Phead of the link "head". Calculate the topic relevancy $R(Phead)$ of this page by formula (2), and save this page into a downloaded web page set termed PageDown. If $R(Phead)> r_1$, save this page into the topic-relevant web pages set termed PageSave.

Step 4: If the number of pages downloaded in the PageDown reaches 15,000, stop the algorithm; otherwise go to Step 5.

Step 5: Keep all the sub-links obtained in the current web page into set SubLink, and calculate the comprehensive relevancy $E(link_i)$ of every sub-link $link_i$ in the SubLink by formula (5). If $E(link_i)>r_2$, it will be inserted into the waiting queue; otherwise it will be discarded.

Step 6: Take the comprehensive relevancy of link "head" at the time as its energy $E(head, t)$, and the number of downloads of the current link's corresponding page as its frequency histogram function $H(head,t)$. Calculate $\tilde{E}(head,t) = E(head,t) - k \times H(head,t)$.

Step 7: Pick out the next link termed "next" from the sub-links with equal probability. Calculate this link's energy value $E(next, t)$, frequency histogram function $H(next, t)$ and the value of $\tilde{E}(next,t)$.

Step 8: If $E(head,t) < E(next,t)$, then accept *next*, i.e., let *head*=*next*, *E*(*head*, *t*)=*E*(*next*, *t*), *t*=*t*+1, and go to Step3; otherwise, go to Step 9.

Step 9: If $random[0,1] < exp\ ((\tilde{E}(next,t) - \tilde{E}(head,t))/(k_b \times T))$,then accept *next*, i.e., let *head*=*next*, *E*(*head*, *t*)=*E*(*next*, *t*), *t*=*t*+1, and go to Step3; otherwise, do not accept *next* and restore *head* as current link, let *t*=*t*+1, and go to step10.

Step 10: If all the sub-links in SubLink have already been extracted, but still not any "next" link is accepted, then empty SubLink and go to Step2; else go to Step 7.

4. Experimental Results and Analysis

We have implemented IELP algorithm on the Intel Core(TM) i7-4710 computer using Java language. Some important parameters such as r_2, r_1, T have great impact on experimental results. In fact, if the threshold of links' comprehensive relevancy r_2 is set too high, few links may meet the requirement and there will be not enough links in the waiting queue. Crawlers will run to a dead end and may miss some topic-relevant web pages. On the other hand, if the threshold is set too low, some web pages that are not relevant to the topic obviously may be crawled, leading to a lower precision. This paper performs simulation experiments on different values between 0.1～0.3 of r_2, respectively. The results show that crawlers work best when r_2 is set to 0.12, so we define r_2=0.12 in this paper. Other parameters are set in a similar way. Here T=2×10^{21}K, and r_1 to 0.62.

4.1. Algorithm Evaluating Indicators

Evaluating indicators for the effectiveness of focused crawlers are generally accuracy and recall. Accuracy refers to the ratio of the crawled relevant web pages number *LG* among the total number *DG* of the crawled web pages; recall refers to the ratio of the crawled relevant web pages number *LG* among the total number *TG* of the relevant web pages within the whole network. Given that it is difficult to count the total number of web pages relevant to a particular topic on the Internet, and that web resources are constantly updated, the recall is hard to calculate scientifically. As a result, this paper adopts precision as the indicator to compare crawlers' performance. On the other hand, this paper uses the average topic relevancy *AR* of web pages downloaded (see formula (6)) to analyze the proposed IELP algorithm's performance.

$$AR = \frac{1}{DG} \times \sum_{i=1}^{DG} R(P_i) \tag{6}$$

Here $R(P_i)$ represents the topic relevancy of web page P_i.

4.2. Experimental Results and Analysis

To test the effectiveness of the proposed focused crawler, we run the IELP algorithm, the BFS[7], the OPS[8], the FCSA[16], and the ELP algorithm, respectively. Table 2 lists the results of LG, accuracy, AR, and computational time (s) by five different algorithms when DG reaches 15,000. It is not hard to see that the proposed IELP crawler overmatches the three algorithms (BFS, OPS, FCSA) in literature and the origin ELP algorithm in terms of LG, accuracy and AR, while the BFS and OPS algorithms have a slightly shorter retrieval time. Figure 1 shows the comparison of results of LG by the above five algorithms where the horizontal axis represents the total number of crawled web pages and the vertical axis represents the number of relevant web pages crawled by each method. As the number of crawled web pages increases, the number of relevant web pages crawled by OPS, FCSA, ELP and IELP algorithms increases rapidly, while that by the BFS algorithm grows at a relatively slow speed. IELP is better than the other four algorithms when the number of crawled web pages is larger than 4,000.

Table 2. Comparison of experimental results obtained by five different algorithms when DG reaches 15,000

algorithm	LG	Accuracy	AR	Computational time/s
BFS	1899	0.1265	0.3951	6896
OPS	9053	0.6035	0.5692	7212
FCSA	10596	0.7064	0.6392	7833
ELP	10026	0.6684	0.6212	7742
IELP	11976	0.7984	0.6812	8579

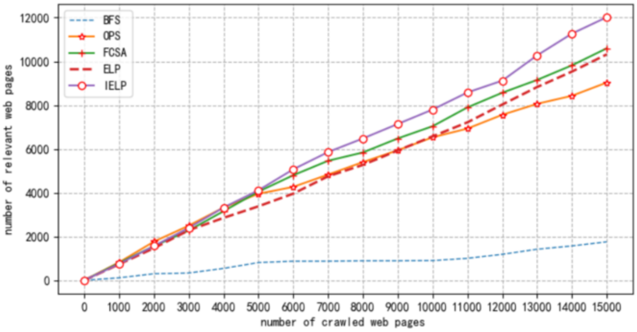

Figure 1. Comparison of the results of the number of crawled relevant web pages for five algorithms

Figure 2 shows the comparison of results on the accuracy of five algorithms where the horizontal axis represents the total number of crawled web pages and the vertical axis represents the precision of each method. As can be seen from Figure 2, the IELP algorithm has the highest precision and tends to be stable when the number of crawled web pages is more than 4,000. Precision of IELP closes to 80%, FCSA 71%, ELP 68%, OPS 60% and BFS around 13% when the number of crawled web pages reaches 15,000. Figure 3 shows the comparison of results on the average topic relevancy of web pages crawled by above five algorithms. The average topic relevancy under the IELP algorithm keeps relatively high values during the whole crawling process and reaches about 0.67, when the number of crawled web pages reaches 15,000. In general, the BFS algorithm has a low precision during the whole crawling process because it does not predict the topic relevancy of web pages during the search process and then visits a large number of irrelevant web pages. The OPS algorithm, which adopts the greedy strategy, downloads the most relevant web pages every time, so its precision is relatively high in the early crawling stage. However, as the number of crawled web pages increases, the greedy strategy adopted by the OPS algorithm leads to the local optimum in the late stage and cannot crawl more high-quality web pages. The FCSA algorithm selects and receives sub-optimal links with a certain probability, which can improve crawlers' shortcoming in falling into the local optimum, and has the advantages of global search. However, this algorithm is not specially designed to solve above problems and it is still likely to repeatedly visit links that have already been searched, which may affect the searching results. The IELP algorithm makes it possible for the sub-optimal links to be visited in advance by means of Metropolis sampling, and on the other hand, it enables focused crawlers to obtain a similar nature to the Taboo Search algorithm by updating frequency histogram, thus avoiding circuitous search by crawlers. This "non-greedy" link selection strategy raises the network coverage rate of focused crawlers, thus effectively avoiding falling into local optimization and improving the crawler's ability to search relevant web pages.

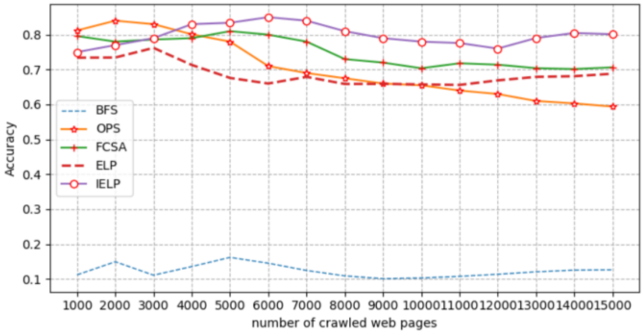

Figure 2. Comparison of the results of accuracy for five algorithms

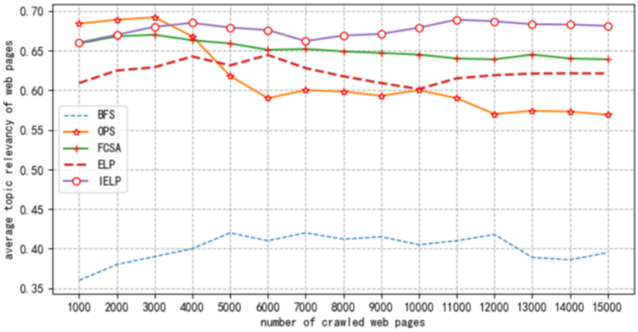

Figure 3. Comparison of the average topic related results of web pages with five algorithms

5. Conclusions

This paper designs a focused crawler based on the improved version of the energy landscape paving (IELP) algorithm that can download relevant web pages on the topic of rainstorm disasters as much as possible. It uses a set of topic words with semantic weight to describe the topic, and designs a method to evaluate the comprehensive relevancy of links in the aspect of the priority of hyperlinks to be visited. Additionally, this paper combines the link search strategy with IELP algorithm to prevent the crawler from falling into local optimization, thus enhancing the precision of focused crawlers. Experiments suggest that the method proposed in this paper is non-greedy and is not limited to current optimum and it also considers future returns, which effectively improves the precision to crawl relevant web pages. Distributed focused crawlers will be studied in the next stage, for improving the speed of crawlers to visit and download web pages in a distributed form.

Acknowledgments

This work is supported by the Special Foundation of Guangzhou Key Laboratory of Multilingual Intelligent Processing, China (Grant No. 201905010008), the Major Program of the National Social Science Foundation of China (Grant No. 16ZDA047), and the Program of Science and Technology of Guangzhou, China (Grant No. 202002030238).

References

[1] Seyfi A, Patel A, Junior JC. Empirical evaluation of the link and content-based focused Treasure-Crawler. Computer standards & interfaces, 2016, 44:54-62.

[2] Liu H, Hong Y, Yao L, et al. HITS-based optimization method for bilingual corpus mining. Journal of Chinese Information Processing, 2017, 31(2): 25-35.

[3] Wang C, Ji XH. Improved PageRank algorithm based on user interest and topic. Computer Science, 2016, 43(3): 275-278.

[4] Prakash J, Kumar R. Web Crawling through Shark-Search using PageRank. Procedia Computer Science, 2015, 48: 210-216.

[5] Zhang W, Chen Y. Bayes topic prediction model for focused crawling of vertical search engine //Proceeding of the 2015 Computing Communications & It Applications Conference. Piscataway: IEEE, 2015: 295-300.

[6] Du Y, Liu W, Lv X, et al. An improved focused crawler based on semantic similarity vector space model. Applied Soft Computing, 2015, 36: 392-407.

[7] Li L, Zhang GY, Li ZW. Research on focused crawling technology based on SVM. Computer Science, 2015, 42 (2): 118-122.

[8] Rawat S, Patil DR. Efficient focused crawling based on best first search //2013 IEEE International Advance Computing Conference. Washington D C: IEEE press, 2013: 908-911.

[9] Chen J. Design and implementation of news search engine based on MySQL. Fudan University, 2013.

[10] Jing WP, Wang YJ, Dong WW. Research on adaptive genetic algorithm in application of focused crawler search strategy. Computer Science, 2016, 43(8):254-257.

[11] Zheng S. Genetic and ant algorithms based focused crawler design// Proceeding of the 2011 International Conference on Innovations in Bio-Inspired Computing & Applications. Piscataway: IEEE, 2011:374-378.

[12] Wu YL, Zhao SL, Li CJ, etc. Text classification method based on TF-IDF and cosine similarity. Journal of Chinese Information processing, 2017, 31(5): 138-145.

[13] Turney PD, Pantel P. From Frequency to Meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 2010, 37(1):141-188.

[14] Hansmann UHE, Wille LT. Global optimization by energy landscape paving. Physical Review Letters, 2002, 88(6): 068105

[15] Liu JF, Li G. Basin filling algorithm for the circular packing problem with equilibrium behavioral constraints. Science China: Information Sciences, 2010, 53(5): 885-895.

[16] Liu JF, Li F, Jiang SY. Focused crawler for rainstorm disaster based on host information and simulated annealing algorithm with comprehensive priority evaluation strategy. Computer Science, 2019, 46(02): 215-222.