CHEMICAL ENGINEERING TRANSACTIONS
VOL. 51, 2016
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Tichun Wang, Hongyang Zhang, Lei Tian

ISBN 978-88-95608-43-3; **ISSN** 2283-9216
DOI: 10.3303/CET1651134

An Improved Shark-Search Algorithm for Agriculture Web Search Engine

Bin Wang[a], Junhao Wang[*a], Xiaohua Sun[b], Na Wang[c]

[a] College of Information Science and Technology,Agricultural University of Hebei,Baoding,China
[b] Department of Digital Media,Hebei Software Institute,Baoding,China
[c.] Department Economics and Management,Baoding Vocational and Technical College, Baoding, China
wb900@126.com

An improved algorithm for agriculture web search engine based on Shark-Search is proposed. Aiming at deficiency of the traditional algorithm including the context of the anchor text and the trapping in local optimum, the new algorithm clusters the links between the different block of the web pages by the k-average algorithm. It randomly selects k texts as the mass of the original class. According to the average value of the text in the class, each text is assigned to the nearest class. By using the tunnel technique, the crawler can complete the search process. The range of the optimality has been expanding. It can get a globally optimal solution from locally optimal solutions. This algorithm is easy to implement and the computing speed can meet the business needs. The experiment indicates that the new algorithm can improved the performance in fetching the pages evidently.

1. Introduction

The research of agriculture search engine is very essential when more and more agricultural information is required by farm workers and the self-employed. As a special crawler category, focused crawler is developed to retrieve as many documents related to a topic of interest as possible while reducing computational resources and the network which was confirmed (Guo et al., 2005; SI, et al., 2010; Ginsberg et al., 2008; Gao et al., 2010; Apte et al., 1994).Traditional focused crawler targets pages that are related to some specific topics. The goal of focused web crawler is getting more pages that are correlative with a certain topic to prepare information for querying which was confirmed Brin et al., 1998; Kleinberg et al., 1999; Chakrabarti et al., 1999). Having been developed for several years, there are many algorithms at present for the focused crawler. The typical examples are Fish-Search algorithm and Shark-Search algorithm which was confirmed (Debra et al., 1994; Herseovici et al., 1998).The advantages of them are the better theoretical foundation and the simple calculation. However, they ignore the relevant information about the linking structure.so there are inadequacies in predicting the value of links. The Shark-Search algorithm has improved a great deal based on the Fish-Search algorithm. But the insufficient of the context of the anchor text and the trapping in local optimum are its obvious defect. For the lack of proposed Shark-Search algorithm, this paper means to ameliorate and demonstrate it with the clustering of link and the tunnelling technique. The comparisons data of the two algorithms show that the improved method can improve the performance in fetching the pages related to the topic evidently.

2. the improved algorithm

The Shark-Search algorithm gains a significant improvement for the Fish-Search method which was confirmed (Menczer et al., 2004). It introduces the vector space model to improve the determination of the correlation. And it makes the most of the correlation of the parent nodes and the anchor text along with the context.
However, there are two drawbacks in current Shark-Search algorithm. One is the insufficient of the context of the anchor text; the other is trapping in local optimum.
This paper improves the algorithm; the method is elaborated as follows.

Please cite this article as: Wang B., Wang J.H., Sun X.H., Wang N., 2016, An improved shark-search algorithm for agriculture web search engine, Chemical Engineering Transactions, 51, 799-804 DOI:10.3303/CET1651134

2.1 the clustering of link

Suppose that a pattern set is expressed as:
The Clustering form can be described below which was confirmed (Sun et al., 2008):

$$U = \{p_1, p_2, ..., p_n\}$$

(1)

The ith pattern is expressed as p_i, and *i= (1, 2, … , n).*
Meanwhile, Suppose that:

$$C_t \subseteq U, t = \{1,2,...,k\} \quad (2)$$

$$C_t = \{p_{t1}, p_{t2}, ..., p_{tw}\} \quad (3)$$

$$proximity(p_{ms}, p_{ir}) \quad (4)$$

The result of clustering is expressed as C_t, and some conditions need to be satisfied: The similarity distance of the pattern is described by the function proximity. Among the formulas, the first subscript of p_{ti} expresses the class, to which the mode belongs; the second subscript expresses the pattern of the class.

(1) $U_{t=1}^{k} C_t = U$. (5)

(2)For $\forall C_m$,

$$C_r \subseteq U, C_m \neq C_r, C_m \cap C_r \neq \Phi. \quad (6)$$

$$MIN_{p_{mu}, p_{rv}}(proximity(p_{mu}, p_{rv})) > MAX_{p_{mx}, p_{my}}(proximity(p_{mx}, p_{my})) \quad (7)$$

There are many kinds of clustering algorithms. The K-means clustering methods are described below.
Step1: randomly select k texts as the mass of the original class.
Step2: repeat the steps below until there are not any more changes.
1) Each text is assigned to the nearest class according to the average value of the text in the class.
2) Update the average value.
The pseudo-code of K-means clustering algorithm is which was confirmed (Gianluigi, et. al, 2002):

```
KMean(X1, X2, … , XN, K)
Initialize the allocation schemes of clustering, the imputation vector are: A [1], A [2],…A [N];
do
{change = false;
For (i=1;i<=N;i++)
{
K = argminkdist(Xi,Ck);
if(A[i]!=K)
{
A[i]= K;
change=true;
}
}
}
while(change==false)
returnA[1],A[2],…,A[N];
}
```

This algorithm is easy to implement and the computing speed can meet the business needs.

2.2 The tunnel technique

The figure1 can describe a case of crossing the irrelevant web page to locate the pages which are really needed.
The Web Community exists in the network, so some nodes have the characteristics of aggregation. And they can form an obviously pages set which was confirmed (Gibson et al., 1998).

The Web Community A and Web Community B are two pages set related to the topic. The web pages corresponding to the seeds set are also related to the topic. The page one that links with the seeds set has the link to the page in pages set A, so the pages related to the topic can be easily get from page one. For the pages set B, the web pages that links with the seeds set has not the link to the page in pages set B, and there is not any page in pages set A directly related to the pages set B. By adopting conventional web crawler method, all the related pages cannot be crawled. By the tunnel technique, the page two can get through the irrelevant pages and reach the pages set B. The page in pages set A can also get through the irrelevant pages and reach the pages set B.

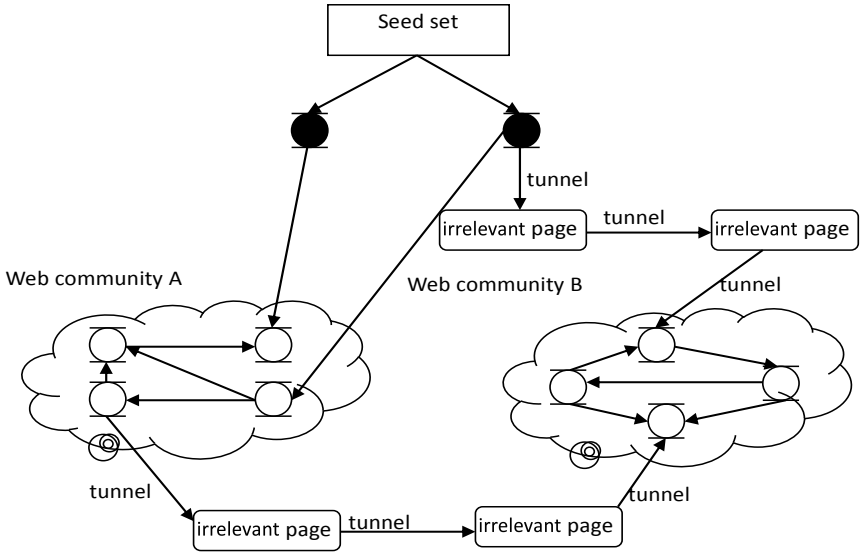

Figure 1: The tunnel technique

By using the tunnel technique, the crawler can complete the Process shown in figure2.

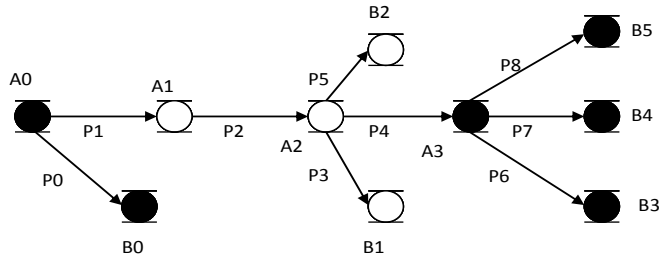

Figure 2: The crawling Process

Traditional focused crawler is targeting web pages that are relevant to some specific topics. But it has many local optimizations. In the figure, the related pages are A0,B0,A3,B3,B4,B5.The irrelevant pages are A1,A2,B1,B2.The search path is P0,P1,P2,P3,P4,P5,P6,P7,P8.Suppose that the focused crawler starts from A0. The B0 page is related to the topic and the A1 page is non-relevant, so the traditional crawler ends its searching when reaches B0 along P0.When using the tunnel technique, the P1 path can also be used for searching. So the path newly adds P1→P2→P4.And the A3 page is related to the topic.it can be added to the queue. Thus, by the tunnel technique, the range of the optimality has been expanded. It can get a globally optimal solution from locally optimal solutions.

2.3 The improved Shark-Search algorithm

Aiming the deficiencies of the Shark-Search algorithm including the insufficient of the context of the anchor text and the trapping in local optimum, the improved algorithm clusters the links between the different block of the web pages by the k-average algorithm. The similarity between the corresponding categories and the topic is figured out. Then the nodes of the tree are numbered by the hierarchy traversal. The numbering of paths corresponding to the link is extracted. The operation is shown in Figure3.

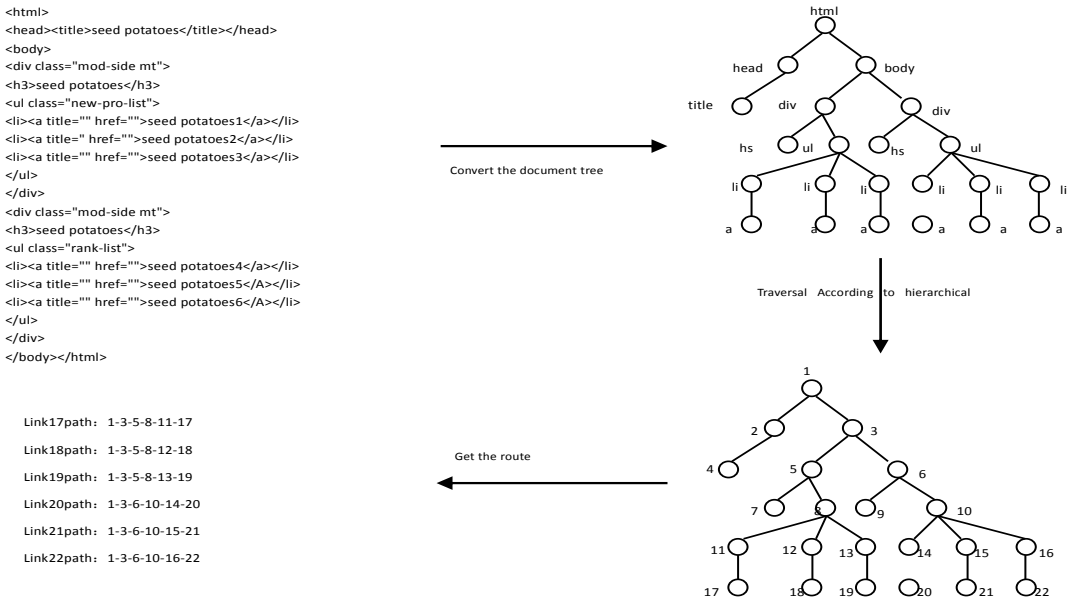

Figure 3: The operation of improved algorithm

In the figure3, the link17,link18 and link19 has the same path 1→3→5→8,and the link20,link21 and link22 has the same path 1→3→6→10.According to the clustering algorithm, the link17, link18 and link19 belongs to the same class and also the link20, link21 and link22 belongs to the same class.

Step1: the web page information is translated into a document object model tree. The context of the anchor text in Shark-Search algorithm is replaced by the similarity to influence the potential_score, as shown in the following steps.

Step2: put the links in web page2 to the queue by the retrieving sequence. Find the matching string which satisfies that the path between any two nodes is greater than or equals to 2.Take all the elements in the string out of the queue and put it into the corresponding class. Repeat the process until all the satisfied links are put it into the class.

Step3: all the links for categorizing are expressed as L, the links that belong to category I are expressed as Gi, the number of current category is expressed as class_num. Define a mark expressed as flag.

1) Initialization. Some settings are listed below.

$$L = \{u_1, u_2, ..., u_n\} \tag{8}$$

$$G_1, G_2, ..., G_n = \Phi \tag{9}$$

$$class\_num = 1 \tag{10}$$

$$flag = 1 \tag{11}$$

2) When the set L is nonempty and *flag=1*, let *flag=0*.

3) Goes through each link u_i in L, if there is the same route as u_i and the value is greater than 1,put u_i in the corresponding Gclass_num and class_num is plused one.set *flag=0*.

4) Repeat 2) until *flag=0* or L is empty.

Step4: according to the previous step, the number of the links that is contained in each class can be figure out as |Gi|.the total of the classes is expressed as cluster_url_num, so:

$$cluster\_url\_num = \max(class\_num) \tag{12}$$

The score of class is expressed as class_score,so:

$$class\_score = \frac{\sum anchor\_score(url)}{cluster\_url\_num} \tag{13}$$

Step4: replace the anchor_context_score with class_score,the new neighborhood_score is:

$$neighborhood\_score(url) = \beta * anchor\_score(url) + (1-\beta) * class\_score(url) \quad (14)$$

Via the above steps, the potential_score(url) of the improved algorithm can be obtained.
Then, the creeper goes through the tunnel by the tunnel technique. The restriction of depth and the number of nodes is lifted. The search time is effectively unconstrained.
The flow of the improved algorithm is as shown in figure4.

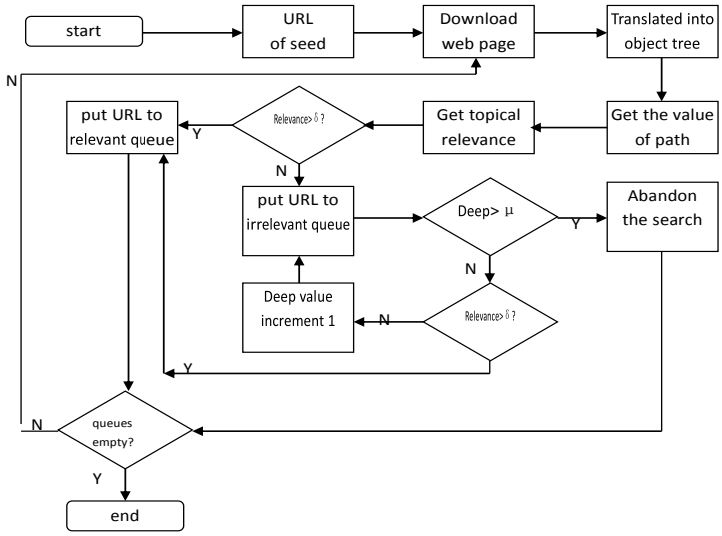

Figure 4: The flow of the improved algorithm

The new algorithm can be summarized as:
(1) The score of class is found out as class_score, and anchor_context_score of the previous algorithm is replaced with it. Then the value of potential_score is figure out.
(2) The URL queue is divided into the related queue and the irrelevant queue. Define δ for the threshold of topical relevance. If the correlation is greater than δ, the URL is put in relevant_Queue,or it is put in irrelevant_Queue.
(3) With the tunnel technique, the nodes in irrelevant_Queue are processed. Set a threshold μto determine whether to continue the crawling.
(4) When the relevant_Queue and irrelevant_Queue are both empty and the crawled pages can meet the demands of users, the process is terminated.

3. experiment results and analyses

The experiment is on a theme of potatoes information searching. The webs for crawling include six agricultural sites, including www.b2cf.cn, www.nong888.cn, www.moa.gov.cn, www.sdny.gov.cn, www.gxny.gov.cn and www.b2cf.cn.The correspondence between URL seeds and the topics is listed in the table. Set the URL seeds in table2 as the initial URL to grab the information of potatoes. The following table shows the results of the Shark-Search and the improved algorithm, respectively. The related pages and the total pages are counted whenever achieves 600 crawled pages. Set the parameter values asα=0.6, β=0.8, γ=0, δ=0.5, μ=3.Under the certain number of crawled pages, the more the number of the relevant pages, the higher the validity.

Table1: Comparison of two algorithms

the total of the pages	the relevant pages		Increase Rate (%)
	Conventional algorithm	The new algorithm	
800	112	201	87.35
1000	283	477	75.76
1600	653	1233	97.59
2500	938	1648	83.17
3200	2190	2666	23.93
4000	2441	2797	16.06

Analysis data of comparison of two algorithms show that the improved algorithm has better performance in fetching the pages Related to the topic. When the total of the pages is 800, the Increase rate can reach 87.35%.however, with an increasing number of the total of the pages, the Increase rate falls a bit. But in general, the examples can prove the efficiency and accuracy of the new algorithm.

4. Conclusions

The two best classical algorithms: Fish-Search and Shark-Search are combined and compared. The Shark-Search algorithm is improved in the disadvantage with the clustering of link and the tunneling technique. The experiment indicates that the new algorithm can improved the performance in fetching the pages related to the topic greatly. As a represent of agriculture web search engine, the various techniques have been used widely and been rewarded. The algorithm proposed in this paper can grab more useful information.

Acknowledgments

This work is supported by 2014 annual plan for scientific research and development of Baoding support project (Grant No.14ZS004) and 2015 annual Science and Engineering Foundation of Hebei Agricultural University, China. (Grant No.LG20150603).

References

Apte D., 1994, Automated learning of decision rules for text categorization [C], ACM Transactionson Informaition System, vol.12 (3): pp 223-225, DOI: 10.1145/183422.183423.

Chakrabarti S., Martin D., Dom B., 1999, Focused crawling: a new approach for topic specific resource discovery [J]. Computer Networks, vol. 31(11): pp 1623-1640, DOI: 10.1016/S1389-1286(99)00052-3.

Debra P., Houben G., Kornatzky Y. and Post R., 1994: Information Retrieval in Distributed Hypertexts[C], In Proeeedings of the 4th RIAO Conference, New York, pp 481-493.

Gao K., Zong B. Q., 2010, Web Information Processing and Extracting [ICMLC], Proceedings of the 9th International Conference on Machine Learning and Cybernetics, vol.5: pp 2350-2355.

Gianluigi G., Sergio G., Ester Z., 2002: A stochastic approach for modeling and computing web communities[C], In Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE 02), pp 43-52.

Gibson D., Kleinberg J., Raghavan P., 1998, Inferring Web Communities from link Topology[C], In Proceedings of 9th ACM Conference on Hypertext and Hypermedia, pp 255-234, DOI: 10.1145/276627.276652.

Ginsberg J., Mohebbi M. H., Patel R. S., et al., 2008, Detecting influenza epidemics using search engine query data [J], Nature, vol.457 (7232): pp 1012-1014., DOI: 10.1038/nature07634.

Guo Q., Guo H., Zhang Z. Q., et al., 2005:Schema Driven and Topic Specific Web Crawling [C], Database Systems for Advanced Applications., Springer Berlin Heidelberg, pp 594-599, DOI: 10.1007/11408079_55.

Herseovici M., Jacov M., Maarek Y. S., 1998, The Shark-Search algorithm an application: Tailored Web Site Mapping [J], Computer Networks and ISDN Systems, vol.30 (1): pp 317-326, DOI: 10.1016/S0169-7552(98)00038-5.

Kleinberg J. M., 1999, Authoritative sources in a hyperlinked environment [J]. Journal of the ACM (JACM), vol.46 (5): pp 604-632. DOI: 10.1145/324133.324140.

Menczer F., Pant G., Srinivasan P., 2004, Topic web crawler: Evaluating adaptive algorithm[C], ACM Transactions on Internet Technology, vol. 4(4): 378-419.

Page L., Brin S., Motwani R., Winograd T., 1998, The PageRank Citation Ranking: Bring Order to the Web[R], Technical Report, Stanford University, CA.

Si X., Liu Z., Sun M., 2010, Modeling Social Annotations via Latent Reason Identification [J], IEEE Intelligent Systems, vol. 25(6): pp 42-49.