International Journal of Database Theory and Application
Vol.8, No.1 (2015), pp.281-288
http://dx.doi.org/10.14257/ijdta.2015.8.1.27

Improved Shark-Search Flash Theme Search Algorithm

Junxiao Liu[1, 2] and Xiangzeng Meng[2]

[1]Library, Shandong Normal University, Jinan, China
[2]School of Communication, Shandong Normal University, Jinan, China
liujx81@163.com

Abstract

Targeting at the web page link features of Flash animation resource, Shark-search theme search algorithm is improved in the search width, link similarity judgment and selection of crawling links, and the search process of "search first and judge later" is employed. According to the experimental result, the improved Shark-Search algorithm may improve the efficiency of Flash theme webpage search, and it is much more applicable for the theme search of Flash resource.

Keywords: *Shark-Search algorithm, theme search, web spider*

1. Introduction

The objective of Flash theme search is related to the theme of search in Web, and it includes the webpage of Flash. Similar to the theme webpage search, the theme spider may crawl to the link in the Seed immediately and directionally according to the appointed themes. At first, it shall take advantage of the heuristic strategy to evaluate the value of crawling link, and then, it may analyze if the crawling link is related to the theme. Different from the theme web search, the web achieved by the Flash theme shall not only guarantee its relation to the theme, but also include Flash resource.

The advantages and disadvantages of theme search algorithm decide the efficiency and accuracy of theme search. At present, many experts and scholars have conducted research work in both theory and practice, and proposed numerous theme search algorithm, including the search algorithm based on the content evaluation with Shark-Search [1] and Best-Fish [2] as the representative, as well as the search algorithm based on the link structure evaluation with PageRank [3] and HITS [4] as the representatives. Some improved theme search algorithms are mainly based on the two algorithms. In algorithm [5], in order to strengthen the adaptive ability of web spider, advantages of algorithm consolidation in predicting the long-term reward is added into the study process of web spider, for predicting the future reward value of the crawling link. In references [6], distance to the target page is estimated through the construction of typical "Web context" based on the reinforcement of learning, so as to enhance the adaptive ability and increment feedback ability of web spider.

The network Flash theme search is similar to the theme search, and both would search the specific theme page in network, but Flash theme search shall also search the Flash resource including specific themes in the webpage. During the search process, web spider shall judge twice, firstly, judge if the webpage includes Flash resource and then judge if the resource is in accordance with the specific theme. Based on the distribution features of Flash in webpage, it can be discovered through the existing theme search algorithm that it is really difficult to apply the theme search algorithm based on "learning reinforcement" and "context drawing" in Flash theme search, and both algorithms divide the search process into training and search, and then construct web classifier according to the training structure and apply it in the judgment of web themes in the search stage. Due to the complexity and diversity of Flash distribution in web page, it is difficult to construct a

ISSN: 2005-4270 IJDTA

classifier with certain universality through training.

Shark-Search algorithm is a classic theme search algorithm [7], targeting at the features of Flash distribution in web page, Shark-search theme search algorithm is improved in the search width, link similarity judgment and selection of crawling links, and the search process of “search first and judge later” is employed. According to the experimental result, the improved Shark-Search algorithm may improve the efficiency of Flash theme webpage search. According to the experimental result, the improved Shark-Search algorithm is much more applicable for the theme search of web Flash resource.

2. Fish-search Algorithm and Shark-search Algorithm

The Fish-search proposed by De Bra *et al.*, [8] was an early dynamic theme web page crawling algorithm. Fish algorithm mainly simulates the Web crawling webpage of theme spider as a feeding process of fish school in the sea, and each fish in the algorithm stands for one URL. When the fish finds the food (discover related web page), its fertility ability may be improved (search width increases), and its descendants have the same life span (invariable search depth); when there is no food discovered (it fails to discover related web page), its fertility maintains unchanged (invariant search width), the life span of its descendants may decrease (search depth -1); when it enters polluted area (the web page does not exists or the reading time is too long), the fish would die (give up the crawling of link). This algorithm mainly determines the priority of crawling URL based on the page contents and theme dependency, as well as the speed in link selection. However, its judgment of correlation is discrete binary judgment, namely relevant / irrelevant. Based on the Fish-search algorithm, Hersovici proposed the Shark-search algorithm [9].

Shark-search algorithm mainly proposes two main improvements based on the Fish-search algorithm. At first, a continuous value function is applied to stand for the correlation, ranging between 0 and 1, rather than the two-value judgment of Fish-search. In addition, the themes of crawling link are impacted by the anchor text, context of anchor text, and parent link correlation inheritance. The links and texts related to the contents of most webpages usually occur continuously, and Shark-Search algorithm just takes advantage of such a feature, and takes the contextual information (including links and texts) of the link as a significant factor for helping decide the correlation of links for visiting to specific themes. However, there may be some limitations, for instance, if the semi-structured characteristic of the webpage is not utilized and only the context of the link is taken as one of the factors, the contexts of links may occur in different blocks. When a link with high correlation degree occurs on the top or bottom of the webpage, it may result that part of the frame links obtain the contextual weight. In addition, since the link texts is usually short, and the range for selecting contexts could not be large, and there must be substantial webpages with the same weight in the same webpage, which may result in the indistinctive differentiation in the correlation of links. In Reference [10], clustering would be conducted for links of different blocks in the web page, and then all similar link anchor text shall be taken as the description text of this category, so as to calculate the correlation with such themes, and replace the impact of anchor text context on the link correlation in Shark-search. In reference [11], similarity of web pages is calculated from three aspects, including the web page, link block and link, and then the three shall be combined according to different weights for further confirming the similarity of the entire page.

3. Improved Shark-search Algorithm

3.1. Web Page Link Blocking

There are two forms of Flash resources in web page, one is the Flash material that can be browsed online, namely, the web embedment, and the other is the Flash material that shall be acquired through downloading, namely in hyperlink form. It can be concluded through the webpage analysis with substantial Flash: web link with Flash usually occurs in link list in the parent page, and the link list is named as “theme group”, and text of “theme group” applied for illustrating the “theme group” is the title, which plays an indicative function for the theme correlation of these links. As for the extraction of titles of the “theme group”, four enlightening rules are extracted, and each “theme group” would be restrained within a pair of table tag, and meanwhile, the internal embedded table tag shall be combined. The rules are shown as follows:

(1) The word size of the text is larger than the surrounding text;
(2) The text is in different color;
(3) The text is usually short in length (generally less than 10);
(4) The text shall be independent.
If any two are satisfied, it shall be deemed as the title of “theme group”.

3.2. Calculation of Similarity in Webpage Content

The web page links would be divided into different “theme groups” according to the table tag, and the correlation degree of the titles of “theme group” shall be taken as the weight of anchor text and URL tag in the crawling link. And in this way, the theme relevancy of crawling link is:

$$\text{Content\_score}(u_i)=\text{score}(\text{block\_title})[\beta*\text{score}(\text{anchor})+(1-\beta)*\text{score}(\text{url})] \qquad (1)$$

In which, score (block_title) is the relevancy between the title of “theme group” of link u_i and the theme, vector space model (VSM) is applied in calculation. In the VSM, all keywords t form the keyword set $T=(t_1,t_2,t_3\cdots t_n)$, each file d in the title file D of “them group” is represented as the vector of paradigm.

$$V_t(d)=(t_1,w_1(d);\cdots t_i,w_i(d);\cdots t_n,w_n(d))$$

In which, $w_i(d)$ is the weight of *ti* in file d, and the calculation of weight employs the TF-IDF word frequency statistics, as shown in equation (2). VSM is employed for calculating the correlation between the title and theme, as shown in equation (3). Score (anchor) and score (url) stands for the correlation between and theme and the anchor text standing for the link u_i and URL address, and Boolean model is employed. β is correlation factor, for adjusting the proportion of anchor text and URL address.

$$w_i(d)=\frac{tf_i\log(\frac{N}{nt_i}+0.01)}{\sqrt{\sum_{i=1}^{n}tf_i\log(\frac{N}{nt_i}+0.01)}} \qquad (2)$$

In which, *tfi* stands for the frequency of keyword *ti* in file *d*; *N* stands for the total files applied for the training texts of characteristics extraction, *tni* stands for the frequency of keyword *ti*.

$$\text{score(block\_title)=sim(d,q)=cos}(\theta)=\frac{\sum_{i=1}^{n}(w_i(d)*w_i(q))}{\sqrt{\sum_{i=1}^{n}w_j^2(d)*\sum_{j=1}^{n}w_j^2(q)}} \qquad (3)$$

In which, $w_i(q)$ is the weight of keyword t_i in the query of q. Generally, it is 1 if it is included in the query, or it is 0. The correlation between the title and theme stands for the cosine of the included angle of two vectors.

3.3. Calculation of Relevancy of Web Page Link

As for the link structure, it is characterized by the "resource adjacency" showed by the theme web page with Flash. The so-called "resource adjacency" refers to a certain part of several parts with Flash resources in this web page. Meanwhile, the theme of Flash resources in the same region may also be the same. The following hypotheses are proposed according to such features:

(1) If a web page is the one with Flash related to the theme, the sub-link of web page may be the one with Flash related to the theme.

(2) If a webpage is the one with Flash related to the theme, the brother link of the parent page may be the one with Flash related to the theme.

Therefore, when calculating the correlation of web page link, the correlation of parent web page and brother web page can be applied to reveal the impact of link structure on the URL link. In order to reflect such impact in real time to each sub-link, a dynamic factor can be introduced, and equation (4) can be applied to stand for the contribution of link structure to the correlation of URL link.

$$Structure\_score(ui) = \frac{\sum_{j=1,u_i \in d_i}^{t} \lambda(d_j)P(d_j)}{t} \qquad (4)$$

In which, *ui* is the crawling link, *t* is the total parent links, *λ(di)* is the dynamic factor, which can be calculated with equation (5); *P(di)* stands for the link correlation inherited from parent link and average link correlation of crawled brother link, which can be applied or measuring the capability of crawling to related themes through the parent link, and it can be calculated in equation (6).

$$\lambda(d_j) = \frac{n' + \theta}{n + \theta} \qquad (5)$$

In which, *n′*is the number of web pages related to the theme of crawled sub-links of the parent link d_j, *n* stands for the total number of crawled sub-link of the parent link d_j, θ is the normalization factor, which is usually 0.5. During the crawling process, *(*d_j*)* would be adjusted constantly.

$$p(d_j) = (1-\sigma)R(d_j) + \frac{\sigma \sum_{k=1,d_k \in d_j}^{N} P(d_k)}{N} \qquad (6)$$

σ is the bias factor, *R(*d_j*)* is the theme correlation of parent link d_j, d_k is the crawled sub-link of d_j. *N* is the total crawled sub-links of d_j, $\sum_{k=1,d_k \in d_j}^{N} P(d_k)/N$ is the average link score of crawled sub-links of the parent link d_j.

4. Flash Theme Search Algorithm based on the Improved Shark-search

The purpose of Flash theme search is to find the web page with Flash, to improve the searching efficiency and speed. During the search process, it shall search first and judge later. It shall search web page with Flash according to the web page link structure, and when it searches the web page with Flash, it can judge if the web page is related to the theme according to the content similarity. The specific algorithm is described as follows: input the seed UB and endow it with large initial value W, search depth D, total crawling pages N, and theme set T.

①Establish two databases: the link database Internal URLs, and Flash database Flash DB; the seed U_B shall be placed into Internal URLs, and its state shall be set as W (Preparation).
② Extract the links with the state W from the link database.
③ Conduct HTML analysis for the link, and extract the sub-link list I_U in the link. Meanwhile, calculate the link structure similarity S_U according to equation ④, and store the link in L_U and corresponding SU into the Internal URLs according to the priority, and set the state of link as W.
④Extract the link I_U with the status W from the Internal URLs, while I_U is not empty and the total crawling web pages <N.
⑤Conduct HTML analysis for link I_U:
If I_U does not include Flash, please turn to step ③;

Else I_U includes Flash

Turn to step ⑥

⑥ Analyze the link I_U, and extract the title *Bu* of “theme group” from I_U, calculate the title *Bu* of “theme group” and S_B similar to theme *T* according to equation ①; extract the link list from “theme group”, and calculate the similarity Cu of each link according to equation ④
If Cu, a given value, stores IU in FlashDB database;
Else

Turn to step②;
⑦End While.

5. Experiment

5.1. Search Width Limit

In the improved algorithm, the search width *W* can be improved. According to the webpage partitioning thought proposed previously, each link shall be endowed with a search width Wblock, with the similarity of title and theme *Score (block_title)* as the coefficient, it is operated according to the following enlightening rules: when encountering a theme-related link, *W* stays the same, or *W* shall be reduced by 1; wen *W* is 0 or the crawling is finished, it shall enter the next “theme group”. *W* is represented by equation (7):

$$W= Score(block\_title)*W_{block}-\rho \qquad (7)$$

In which, when the theme is related, ρ equals 0, or it is 1. Through improving the parameter *W*, the improved algorithm is characterized by increment feedback and self-adaptive features. With the operation of procedures, the system may restudy according to such feedbacks, so as to make the judgment of link correlation accurate.

5.2. Parameter Selection and Evaluation Index

After repeated tests, ρ=0.9, σ=0.7, λ=0.5, W_{black}=10 and D=7 may achieve the best experimental effect. The precision ratio and recall ratio shall be applied for evaluating the effect of the algorithm.

5.3. Simulation Experiment Environment

The improved Shark-search algorithm is realized with Java language. In order to improve the search efficiency, multithreading technology is applied for searching different sites, and the system opens ten threads. The experimental environment is: Windows 7 operating system, Pentium(R) Dual-Core CPU3, 2GHz, 2G internal storage.

5.4. Experimental Result

Ten children's game websites are selected manually as the seeds (six include substantial Flash, four include few or no Flash). The keyword set is the general search algorithm of children's game dictionary (1672 words), and the result is shown in Table 1. Fish algorithm, Shark-search algorithm and improved Shark-search algorithm are taken respectively for operation, and the result is shown in Table 2. The experiment also starts from the relationship between the precision ratio and crawling time, and the three algorithms are tested, and the statistics would be counted every other hour. The result is shown in Table 3.

Table 1. Experimental Result of General Search Algorithm

Number of seeds	Total web pages	Number of efficient web pages	Occupancy of effective web page (%)	Run time(h)	Average crawling speed(min)
10	73925	3059	4.464	30.91	40

Table 2. Comparison of the Experimental Results of Three Algorithms

Algorithm	Number of seeds	Total web pages	Number of efficient web pages	Run time/h	Average crawling speed(min)	Precision ratio (%)	Recall ratio (%)
Standard Fish algorithm	10	36872	1083	17.28	33	2.95	35.22
Shark-search algorithm	10	21691	1863	22.43	14	8.59	60.59
Improved Shark-search algorithm	10	15892	2510	13.28	17	15.79	81.63

Table 3. Comparison of Precision Ration of Three Algorithms with the Changes in Time

Precision ratio			Time(h)
Standard Fish algorithm	Shark-search algorithm	the improved Shark-search algorithm	
2.256	7.485	6.527	1
3.246	12.267	16.463	2
3.568	19.691	21.493	3
3.896	21.572	25.691	4

5.5. Result Analysis

It can be seen from Table 2 that Shark-search algorithm shall calculate the web page similarity and link similarity, and it is slower than the standard Fish algorithm in average speed. The improved Shark-search algorithm only calculates the web page link similarity during the operation process, and it only calculates the web page content similarity when encountering the web page with Flash. Its average speed is higher than the Shark-search algorithm. The improved Shark-search is improved greatly in the precision ratio and recall rate. It can be seen from Table 3 that since the improved Shark-search algorithm "search first and judge later", it may crawl more effective web pages within unit time, and meanwhile, increment feedback and adaptive mechanism is added in the algorithm, so that the result of crawled links can be reflected to the crawling links, so as to improve the precision ratio within unit time.

6. Conclusion

The search width and link similarity calculation in Shark-search algorithm is improved according to the features of web page with Flash. It mainly searches first and judges later, so as to serve the Flash theme search better. The search algorithm has already been running in the Flash resource search system (stand-alone edition and network edition). The network edition has already been running in the internet, but it only supports the login of guest. In the next step, other parameters of the algorithm shall be further optimized, and the search of dynamic web page would be added.

References

[1] Aggarwal C, AL-Garawi F and Yu P. Intelligent crawling on the World Wide Web with arbitrary predicates, Proceedings of 10th International WWW Conference. New York: ACM,**(2001)**

[2] Menczer F. Complementing search engines with online Web mining agents. Decision Support Systems, **(2003)** 35 (2), pp. 195-212.

[3] De Bra P,Houben G, Kornalzky Y,et al.Information retrieval in distributed hypertexts.Proceedings of the 4th RIAO Conference, New York,**(1994)**, pp. 481-491.

[4] Cho J,Garcia-MolinaH, Page L.Efficient crawling through URL ordering. Computer Networks,**(1998)**,30(1-7), pp. 161-172.

[5] Rennie J,MeCallum A. Using reinforcement learning to spider the Web efficiently, Proceedings of the lnternational Conference on Machine Learning(ICML99).San Francisco: Morgan Kaufmann Publishers Inc. **(1999)**, pp. 335-343.

[6] Diligenti M,Coetzee F M,Lawrence S,et al,Focused carwling using context graphs, Proceedings of the International Conference on Very Large Database(VLDB '00),**(2000)**, pp. 527-534.

[7] Hersovici M, Jacovi M, Maarek Y S, et al.The shark-search algorithm-An application：Tailored Web site mapping.Proceeding of the 7th International World-wide Web Conference.Brisbane,Australia：ACM Press **(1998)**, pp. 317-326.

[8] DeBra P, Houben G, Kornatzky Y,et al.Information tetrieval in distributed hypertests ,Intelligent Multimedia, Information Retrieval Systems and Management. New York, USA: ACM Press, **(1994)**, pp. 481-491.

[9] M Hersovici,M Jacovi,YS Maarek,et a1.The shark-search algorithm-An application：Tailored Web site mapping[A].Proceedings of the 7th International World-wide Web Conference.Brisbane,Australia:ACM Press,**(1998),** pp. 317-326.

[10] Su qi,Xiang kun,Sun bin. Shark-Search algorithm based on Clustering Links. Journal of Shandong University(Natural Science). **(2006)**, 41(3), pp. 1-4.

[11] Chen jun,Chen zhumin. Shark-Search algorithm based on segmentation of the web page. Journal of Shandong University(Natural Science). **(2007)**, 42(9), pp. 62-66.

[12] Salton G, Buckley C.Term weighting approaches in automatic text retrieval. Information Processing and Management, **(1988)**, 24, pp. 513-523.

Author

Junxiao Liu, was born in Shandong, China. She received the bachelor and master degree in educational technology in July 2002 and July 2007 from Shandong Normal University. Her current research interests are the content analysis of flash. Besides, she has begun to do some research on the search algorithm of flash. She worked at Shandong Normal University after her graduation. She has participated in some projects about multimedia researches.