2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing

Crawling Strategy of Focused Crawler Based on Niche Genetic Algorithm

Huilian Fan, Guangpu Zeng
School of Mathematics and Computer Science,
Yangtze Normal University
Fuling, Chongqing, China
fhlmx@163.com, feiya_cq@sina.com

Xianli Li
Electronics Information Department,
Chongqing College of Electronic Engineering
Chongqing, China
lisansanchqi@163.com

***Abstract*—In order to improve the search efficiency of focused crawler, we design a new crawling strategy based on the niche genetic algorithm. Rather than colleting and indexing all accessible hypertext documents to be able to answer all possible ad-hoc queries, the new crawling strategy, combined the advantages of hyperlinks structure and web content strategies, uses hyperlink as genetic individual and topic-keywords based VSM is used to evaluate individual fitness, and imports new URLs to implement crossover and mutation, and the URLs that have the same prefix are regarded as niche. Guide the crawl direction by niche genetic algorithm to selectively seek out pages that are likely to be most relevant to a pre-defined set of topics. Compared with the other algorithms, experiments show that the strategy has higher precision and recall in searching the topic pages.**

Keywords—focused crawler; niche genetic algorithm; Vector Space Model; topic relevancy

I. INTRODUCTION

The rapid growth of the World-Wide Web poses unprecedented scaling challenges for general-purpose crawlers and search engines. The first generation of crawlers on which most of the web search engines are based rely heavily on traditional graph algorithms, such as breadth-first or depth-first traversal, to index the web. A core set of URLs are used as a seed set, and the algorithm recursively follows hyperlinks down to other documents. Document content is paid little heed, since the ultimate goal of the crawl is to cover the whole Web [1]. The motivation for focused crawler comes from the poor performance of general-purpose search engines, which depend on the results of generic Web crawlers. So, focused crawler aim to search and retrieve only the subset of the world-wide web that pertains to a specific topic of relevance. Focused crawling search algorithm is a key technology of focused crawler which directly affects the search quality. The ideal focused crawler retrieves the maximal set of relevant pages while simultaneously traversing the minimal number of irrelevant documents on the web. Compare to the standard web search engines, focused crawler yield good recall as well as good precision by restricting themselves to a limited domain [2]. In this paper, we propose an improved genetic algorithm-based crawling strategy (GAFC) that will seek, acquire pages on a specific topic that represent a relatively narrow segment of the web. Based on vector space model, the strategy uses topic-keywords which extracted from the user submitting query expression to evaluate topic relevancy.

II. RESEARCH STATUS OF RELATED WORKS

A. Basic Idea of Niche Genetic Algorithm

Genetic Algorithm (GA) is a global optimization algorithm derived from the mechanism of genetics and evolution. In GA, a new population is obtained by exerting the three operations of selection, crossover and mutation to the individuals of the current population. The optimal solution is included in the final population. But its one drawback is that when dealing with multi-modal functions with peaks of unequal value, simple GA are characterized by converging to the best peak of the space (or to a space zone containing several of the best peaks) and to lose an adequate individual sampling over other peaks in other space zones. This phenomenon is called genetic drift and is not a correct behavior for several kinds of problems in which one may be interested in knowing the location of other function optima[3].

In niche genetic algorithms (NGA), the analogy with nature is straightforward, as in an ecosystem there are different subsystems (niches) that contain many diverse species. The number of individuals in a niche is determined by its resources and by the efficiency of each individual in taking profit of these resources [4]. Using this analogy, it is possible to maintain the population diversity in a GA. Each peak of the multi-modal function can be seen as a niche that supports a number of individuals directly proportional to its "fertility'', which is measured by the fitness of this peak relatively to the fitness of the other peaks of the domain.

B. Research Status of Focusing crawler

The main difference between focus crawlers is how to determine the URLs crawling order. There are two kinds of crawling strategy.

- Web Content based search strategy. It evaluate topic relevancy with Web content, such as URL string, anchor, text information, etc. For example, BestFirst algorithm[5],whose basic idea is that given a priority queue of URLs, according to relevancy estimation criterion the best URL is selected for crawling, where relevancy estimation criterion is based on the relevancy between the topic-keywords describing the specific topic with crawled Web page. Advantage of

978-0-7695-3929-4/09 $26.00 © 2009 Crown Copyright
DOI 10.1109/DASC.2009.49

such algorithm is that have theoretical basis. However, due to neglect of the linked structure between Web pages, it has some shortcomings in forecast the linked pages relevancy.

- Web linked structure based search strategy. It evaluate topic relevancy with Web linked structure, such as Hyperlink Induced Topic Search (HITS) algorithm[6] that is a link analysis algorithm that rates Web pages. It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. Usually, authority page's content have high topic relevancy and hub page is directory for Web pages having topic relevancy. Generally speaking, Good hub pages point to the good authority pages and good authority page are linked by good hub pages. Although taking into account the link structure between Web pages, but those method ignores the relevant page with the specific topic, and in some cases, HITS will occur topic drift problem.

III. CRAWLING STRATEGY BASED ON NICHE GENETIC ALGORITHM

A. Design Idea

Compared to the general crawler, the core problem to be solved of the focused crawler is how to determine whether the Web pages relate to pre-defined topic and how to predict topic relevancy and eliminate irrelevant pages while the crawler search process. We propose a crawling strategy, which has the advantages of strategies based on hyperlinks structure and Web content, use topic-keywords based on improved topic-based Vector Space Model [7] to evaluate topic relevancy and import the niche genetic algorithm to guide the crawling process.

The crawler regard hyperlink as genetic individual. Based on the principle, "most Web pages linked by good pages are good pages" and "most Web pages which contain hyperlinks to link good pages are also good pages" [8], the crawler take some hyperlinks from those Web pages which meet the topic relevancy criteria to achieve crossover. Mutation is taking from some links which point to the page meeting requirement of the specific topic relevancy. Finally, use niche rules to eliminate some URLs, whose topic relevancy is relatively lower in the niche, to obtain next generation. The focused crawler avoids searching irrelevant regions of the web and can take a set of well-selected web pages exemplifying the user interest by crossover operation. Through mutation operation and niche rules, the focused crawler can search for further relevant web pages and recursively explore the linked web pages. The crossover, mutation and niche rule are defined by the following Section C. Fig.1 shows the basic idea of crawling strategy. Following experiment result shows that the strategy not only has higher capability of local optimization, but also improves global search performance of the focused crawler.

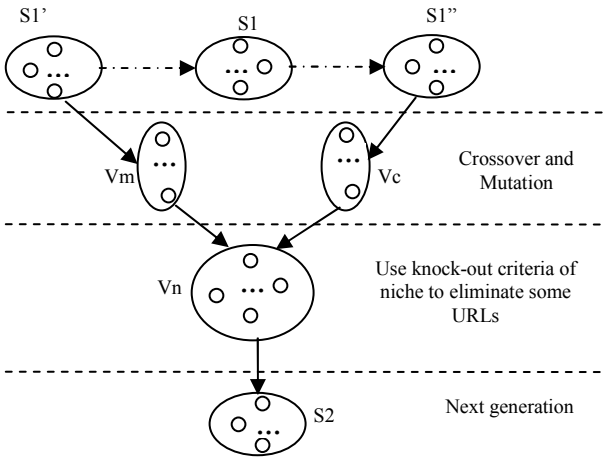

Figure 1. The Basic Idea of Crawling Strategy

Where

S1 is the set of URLs of Web pages meeting the requirement of topic relevancy,

S1' is the set of URLs of Web pages which contain URLs to link S1,

S1" is the set of URLs of Web pages which linked by URLs embedded in S1,

Vc is the result set of URLs of Web pages by crossover operate on S1",

Vm is the result set of URLs of Web pages by mutation operate on S1',

$Vn = Vc \cup Vm$,

S2, which is the result set of URLs of Web pages by using knock-out criteria of niche to eliminate some URLs from Vn, is next generation.

B. Determining Topic Relevancy

Use topic-keywords based VSM to evaluate topic relevancy. The first task is to map the interest topic of user to a set of feature words which called topic-keywords. With the assist of domain expert, those keywords are taken from the user submitting query expression or pre-defining template document used to information retrieval and then convert into n-dimensional vector $\alpha = (q_1, q_2, ..., q_n)$, where q_j is the weight of the j^{th} keyword in the vector α.

Next, Segment the Web page which will be crawled and count the occurring number of the feature words in the Web page. According to the principle, the frequency of the word which has the highest occurring number is *1*, calculate frequency of other feature words. Consider the Web page is a semi-structured text, and feature words which occur in different HTML tags have different contribution for topic[9], so weight of feature words is expressed as $x_i * W(i)$,where x_i is the i^{th} feature word occur frequency and $W(i)$ which is weight coefficient of the i^{th} feature words is given by the following expression:

$$W(i)=\begin{cases}5.0 & \text{the i}^{th}\text{ feature word occur in <title>tag}\\ 3.0 & \text{the i}^{th}\text{ feature word occur in<a>tag}\\ 2.0 & \text{the i}^{th}\text{ feature word occur in<meta>tag}\\ 1.0 & \text{others}\end{cases}$$

Finally, We translate the Web page to n-dimensional vector $\beta = (x_1 * W(1), x_2 * W(2), ... x_n * W(n))$. Then, user

interest topic and the crawling Web pages are mapped to the vector space, respectively, α, β. The topic relevancy between Web pages with user interest topic is converted into vector matching problem. Topic relevancy is defined by the following formula[10]:

$$Similarity(\alpha, \beta) = cos(\alpha, \beta) = \frac{\sum [q_i \times x_i W(i)]}{\sqrt{\sum q_i^2 \times \sum [x_i^2 W(i)^2]}}$$

C. Crawling Strategy

Based on the algorithm of general crawler, the improved strategy of the focused crawler is comprised of four procedures: via selection based on use topic relevancy as fitness to acquire the high topic relevancy pages, via crossover to achieve depth search, via mutation to implement breadth search, via knock-out criteria of niche to eliminate low relevancy pages. The details of the crawling strategy are explained below:

Input: V which is a initial seed set, *pc* which is crossover probability , *pm* which is mutation probability, *pn* which is threshold of knock-out criteria of niche, S_0 which is topic relevancy threshold, *URL_NUMBER* which is threshold used to determine the page is Hub or Authority.

Output: Web pages set P, whose topic relevancy is greater than S_0, and the corresponding queue denoted as *downloaded_url*.

Crawling Strategy:

1. URLs in the initial seed set V are joined the wait process queue, denoted as *wait_queue*.

2. Visited these pages which pointed by URLs in the *wait_queue* and then these URLs are joined visited queue, denoted as *visited_queue*.

3. **Selection** Different from the general genetic algorithm to find the optimal solution, the goal of focused crawler is to find more topic relevancy Web pages as much as possible. So, completed evolution of each generation, all authority pages whose topic relevancy greater than user-set threshold are meeting the requirement of user. The Details of selection are described below:

First, calculate topic relevancy of each page downloaded successfully, denoted as S, check the hyperlink number of the page if $S > S_0$. Then, determine the Web page whose topic relevancy is greater than S_0 is Hub or Authority according to the hyperlink number greater than *URL_NUMBER* or not. If the page is Authority, save the Web page to P and corresponding URL is joined the downloaded queue, denoted as *downloaded_url*, and is joined the set which denoted as *andauthority_urls*. If the page is hub, the corresponding URL is joined the set which denoted as *hub_urls*.

4. Analyzing these new pages that just joined set *hub_urls*, extracts URL from them and eliminate all visited, downloaded and duplicate URL, denoted as V_1. Then, calculate the topic relevancy of Web page which pointed by URL in V_1, denoted as S, and save the result to *temp_hashURL* which is a hash table, whose key value is the URL and corresponding value is node, whose class structure is defined below:

```
class Node
{
    String url;              //corresponding URL of the Web page
    String title;            //title of the Web page
    LinkedList meta_List;    //keywords contained in meta tag of the Web page
    LinkedList url_List;     //urls List contained in the web page
    int url_number;          // number of urls embeded in the web page
    String content;          //topic content of the page
    boolean isHub;           /* if url_number > threshold, the web page is Hub, else
                                is Authority page */
    double s;                //topic relevancy
}
```

5. **Crossover** First, sort the *temp_hashURL* in descending order by topic relevancy S. Then, according to crossover probability *pc*, select first c URLs from *temp_hashURL* as crossover result , denoted as V_c, where c=size of *temp_hashURL* $\times$*pc* .

6. **Mutation** First, collect these URLs embedded in the Web page which point to URLs in *authority_urls*, the result, which eliminated duplication and visited URL, denoted as V_2. Next, according to mutation probability *pm*, random sampling m URL, denoted as V_m, where m=size of $V_2 \times$*pm* .

7. Define $V_n = V_c \cup V_m$.

8. After use knock-out criteria of niche to eliminate some URLs in V_n, the remaining of V_n is joined *wait_queue*, where the criterion is described below:

Because the Web site is organized by content categories, that is the contents under same directory are possible relevant, so according to the feature of URL composition structure (http://<host>:<port>/<path>), we regard these URLs which have the same prefix, such as have same host or host and top-n level directory where n can be defined by user, as niche. The elimination criteria of the niche is the bottom-out system, also known as out at the end of the system. Our elimination criteria is knock out last m or m percent URLs in niche descending sorted based on their topic relevancy, where m can be defined by user.

9. Empty these temp variables: *temp_hashURL*, V_1, V_2, V_c, V_m, V_n, *authority_urls*, *hub_urls*.

10. **Stop Crawling Condition** The strategy is to continue crawling until either *wait_queue* is empty or number of successfully download Web pages is greater than user-set threshold, or the number of generation is over user-set threshold.

IV. Experimental Results and Discussion

In this section, we compare the efficiency of our strategy to BestFirst and HITS algorithm when retrieving"2009 call for papers on information technology" documents.

We design the topic-keywords as "{call for paper, information technology, IT, computer, software, network, database, artificial intelligence, multimedia, 2009}",corresponding n-dimensional vector *α=(5,3,3,3,3,3,3,3,3,1)* record the weight of each topic-keyword. The strategy is implemented by Java. The number of initial seed set V is choose 20 URL from the search result in Google with topic-keywords to search, crossover probability *pc* is 0.6, mutation probability *pc* is 0.4, knock-out criteria is eliminate 10% URL in niche, threshold *URL_NUMBER* to determine the page is Hub or Authority is

15, topic relevancy threshold S_0 is 0.05, condition of stop crawling is that number of downloaded pages, which topic relevancy is greater than S_0, is over 4500. Fig.2 shows the experiment result.

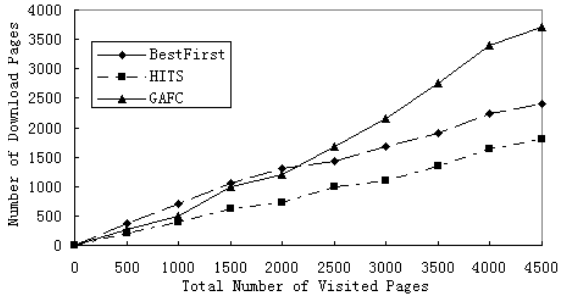

Figure 2. The Number of Visited and Downloaded Pages which are Topic Relevancy.

Superposition ratio can be computed as follows:

$$sratio=\frac{\text{number of superposition pages which downloaded by each algorithm}}{\text{total number of pages which downloaded by all algorithm}}$$

Table I shows the downloaded page's superposition ratio between the three algorithms.

TABLE I. COMPARISON OF SUPERPOSITION RATIO OF THESE ALGORITHM

	GAFC with BF	GAFC with HITS	HITS with BF
Superposition ratio	71%	59%	85%

Our experiments show that the precision of our strategy is lower than precision of Best-First and HITS algorithm in the early crawling, but with continue searching, the performance advantages of our strategy is gradually displayed, the precision of our strategy is significantly higher than other two algorithms. Not only that, but through compare superposition ratio between these algorithms, we find our strategy has large search scope because the superposition ratio of our strategy is lesser than the others, so our strategy has higher recall than the others.

V. CONCLUSIONS

Focused crawler is a new research approach of search engine. Its search algorithm is a key technique which directly affects the search quality. We have proposed the crawling strategy that acquires the topic keywords from specific topic, and then explores the Web, guided by topic relevancy and niche genetic algorithm. In crawling process, our strategy not only considers whether the content of visited Web pages relate to specific topic, but also take account of hyperlinks structure of crawling pages. The experiment result shows that quantity and quality of downloaded pages are more than mere content-based search strategies or link structure-based search strategy.

ACKNOWLEDGMENT

This work is supported by the Natural Science Project of Chongqing Municipal Commission of Education (Project No.KJ091309)

REFERENCES

[1] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, Marco Gori. "Focused Crawling using Context Graphs", Proceedings of the 26th International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt, 2000,pp. 527–534.

[2] Martin Ester, Matthias Groß, Hans-Peter Kriegel, "Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies",Proceedings of the 27th International Conference on Very Large Database,VLDB2001,Roma,Italy,pp.633-637.

[3] Yan Li, "Using Niche Genetic Algorithm to Find Fuzzy Rules", Proceeding of the 2009 International Symposium on Web Information Systems and Applications(WISA'09), Nanchang, China, May,2009, pp. 064-067.

[4] JELASITYM,DOMBIT, "GAS,a concept on modeling species in genetic algorithm", Artificial In telligence,Vol. 99(1),1998,pp.1-19.

[5] CHO J , GARCIA-MOLINA H , PAGE L, "Efficient crawling through URL ordering",Computer Networks,Vol.30(1-7),1998,pp.161-172.

[6] Kleinberg J. "Authoritative Sources in a Hyperlinked Environment", Proc. ACM-SIAM Symposium on Discrete Algorithms, New York,1998,pp.668-677.

[7] Jörg Becker, Dominik Kuropka, "Topic-based Vector Space Model",Proceedings of the 6th International Conference on Business Information Systems,BIS2003, Colorado Springs, USA,2003,pp.337-341.

[8] PAGEL,BRINS,MOTWAN IR, "The PageRank citation ranking:Bringing order to the Web", Stanford,CA:Stanford Digital Libraries Working Paper,1998.

[9] SONG Ju−ping,WANG Yong−cheng,YIN Zhong−hang, "A System for Analyzing Topic—Specific Web Pages", JOURNAL of SHANGHAI JIAOTONG UNIVERSITY,Vol.37(3),2003,pp.401-403.

[10] Pang-Ning Tan,Michael Steinbach,Vipin Kumar(Editor).Introduction to Data Mining.Beijing:POSTS&TELECOM PRESS,2006,pp.54-55.