2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

A New Crawling Method Based on AntNet Genetic and Routing Algorithms

Bahador Saket[1] and Farnaz Behrang[2+]

[1]Urumia University

[2]Amirkabir University of Technology

Abstract. Centralized crawlers are used to gather web pages in a special field. These crawlers are facing challenges such as the problem of local search and how to foresee quality of web pages including their retrieval. The recommended method in this article uses AntNet routing and genetic algorithms to solve these problems.

Keywords: Genetic algorithm, AntNet routing algorithm, centralized crawler

1. Introduction

A search engine may be divided into two general parts of online and offline. Online section takes the queries from users and sends appropriate pages to user from among indexed pages that had been saved before. The duty of offline section is gathering web pages and indexing them. Search engines use crawlers to retrieve and save web pages. A high efficiency crawler must have at least one of the two characteristics that come in the following: first it should use intelligent strategies in its decision makings (such as choosing a link from among the unfollowed links until when the related page is retrieved, and second, its supporting hardware and software structures should have been optimized for a large number of pages at time unit [21, 22]. For this purpose, fault tolerance and considerations related to web server sources must be added to crawler. Research on the first characteristic includes strategies to identify important pages [8 and 15], retrieving documents related to a topic [4, 5, 9, 18] and recrawling to maximize refresh frequency of web archive [6, 7]. But since implementation is difficult, little work has been done about the second characteristic.Today the existing crawlers use local search, this method has three problems [17]:

- Some sites that are related to a similar topic may have no link to each other due to reasons of competition. Among these sites we can refer to commercial sites.
- Links may be one-way, i.e. there may be a link from one web page to the other but there might not be a link from the second page to the first one. Therefore beginning in second page, we cannot reach the first page.

[+] Corresponding author. Tel.: +98-9128030489
E-mail address: fbehrang@aut.ac.ir

- A cluster of web pages may be separated from a cluster of their related pages through a number of pages that are not related to the specific topic. In this case when centralized crawler reaches unrelated pages, stops its work in that route and will not review next pages.

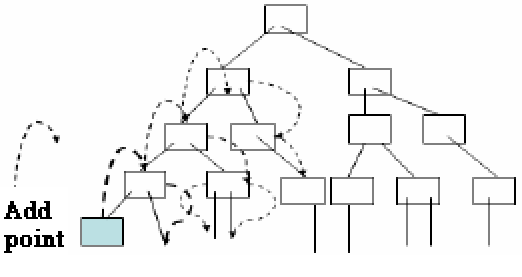
2. Using Genetic Algorithms

Ordinarily when we work with a very large search engine and do not have much information about it, genetic algorithms are appropriate methods of finding optimum solutions. To solve the problems of local search, Qin and Chen presented a method based on genetic algorithms. In this method, address of some pages is given to crawler and their related pages are retrieved and the first generation is created.In selection task, the degree of relationship between the pages and the specific topic is studied and each page is given a special score. Pages whose scores exceed a certain amount are selected and saved and other pages are discarded. In Cross-over task, the links of present generation pages are extracted. Each link is given a special score depending on the pages in which link is located. Then a predetermined number of links are selected randomly (roulette wheel selection method is used and the more the score of a link, the more the likelihood of its selection) and the related pages are retrieved and new generation is created. In mutation, a metasearch is made in known time intervals; maximum four key words are randomly selected and are searched by several regular search engines (such as Google) and a predetermined number of found pages with highest priority are retrieved and added to the new generation of pages. To gain more pages, most of previous tasks are repeated. To avoid the problems associated with local search we use the genetic algorithms of our recommended method.

3. Recommended Method

In regular crawling method, links of each page is independently added to the set of links. After that no relationship is considered among them until the end of work. But in our recommended method which has been inspired by AntNet routing, after a page to which a link refers is retrieved, the scores of other links that exist in the page containing that link are corrected according to quality of that page.

In nature, ants are restlessly gathering foodstuff and repeatedly go forth and back the distance between food sources and their nest. In their route, ants leave a substance named Pheromone, which gradually evaporates after a while. Ants might travel the distance between their nest and a food stuff through various routes, but what is clear is that ants that travel the nearer route will make more travels between nest and food in a specific time interval and so leave more Pheromone in that route. Each ant selects its root randomly but the likelihood of selecting the route that has more Pheromone is more. In this manner, ants could travel optimized routes. Ants algorithm was first presented by Dorigo et.al to solve issues of travelling salesman problems. Today it is used in various problems such as routing in AntNet network, graph painting and the likes.Random selection of links in crawling method is based on the genetic algorithms presented by Qin and Chen, such as selecting route in AntNet routing. The turning point of these two methods is our recommended method. The general idea is that if we retrieve a page that is very much related to our specific subject, namely, if the score which is calculated for it exceeds the foreseen level (the score calculated for the link), the score of the pages that are linked to that page and the pages that are linked to these pages, and up to several higher levels is increased and following that, the scores of the links that come out of those pages will also increase. Of course the more we go farther the page, its influence will diminish.



3.1. Figure 1: The trend of correcting the scores of pages and links after retrieving a new page, which is similar to AntNet routing algorithm

Similarly, we can act for unrelated pages and deduct from the score for the pages to which they are linked and change the score of the links that have not been viewed yet. Doing this scoring the links is made with more care and more appropriate links will be followed.

4. Scores of Links

1.1 How to Calculate Scores of Web Pages

To being crawling, several key words are defined and each is given a special weight and crawling is made based on them. After a web page is retrieved, a score is given for that page per each key word. This is shown in formula 2 with wordscore.

$$1)\ TF_i = \frac{frquency_i}{frequency_{max}}$$

$$2)\ wordscore_i = \frac{frequency_i}{N} + \alpha \times TF_i$$

In formula 1, frequency shows the number of repetitions of key words with index i in web page, and frequency $_{\max}$ is the number of repetitions of the word which is repeated in that page more than other words. In formula 2, N is total number of the words existing in page and α is a fixed amount to determine the extent of influence of TF_i in final amount of word score. Considering the score calculated for each key word and the weight allocated to it, according to formula 3, $pagescore_1$ is calculated for each web page.

$$3)\ pagescore_1 = \sum_{i \in keywords} (wordscore_i \times weight_i)$$

Of course the score from formula 3 could not express quality of a page. Assume that the weight of two key words is equal. If the amount of TF calculated for a word is high and is low for another word, and or both words have normal and similar TF amounts, the calculated score will not differ remarkably, although the score calculated in the first mood must be low. To calculate proper score of pages, the amount of distribution of TF of key words proportionate to the allocated weight must be studied. This must be done using vector space model. In this method, two vectors are used which the number of their dimensions is equal to key words and one of them includes weight of key words and the second includes page $score_1$ calculated per key word. The Cos of the angle between these two vectors is calculated according to formula 4, which is shown as C. Then according to formula 5, the appropriate score for each page is calculated.

$$4)\ C = \frac{\sum_{i \in keywords} (wordscore_i \times weight_i)}{\sqrt{\sum_{i \in keywords} (wordscore_i)^2 \times \sum_{i \in keywords} (weight_i)^2}}$$

$$5)\ pagescore = C \times pagescore_1$$

4.1. How to Calculate the Score of Links

As mentioned before, a method named anchor text was suggested by Pinkerton. We use this method in our work (Of course not as the only method). An anchor window is specified for each link, which includes part of the web page in which that link exists. (for example, the two upper lines of link address in addition to its two lower lines). Then in anchor window, 1 score is given to each key word according to formula 6.

$$6)\ wordscore_i^* = \frac{frequency_i^*}{N^*} + \alpha \times TF_i^*$$

In formula 6, the asterisk amounts have been calculated in anchor window area instead of page area. Like calculation of score for page (formulae 3, 4, 5), for each link, the score that is shown by link score$_l$ is calculated according to word score*. It is clear that the existing links of pages with high scores pint to pages with higher qualities. For this purpose, according to formula 7, the score calculated for each page is used as a basis for the score of its links.

7) $linkscore_2 = pagescore + \beta \times linkscore_1$

In Formula 7, β is a constant amount that determines the degree of influence of linkscore$_1$. The fact that at least several links should be followed from an initial page to retrieve the current page shows the depth of this page. Najork and Wiener showed that pages with lower depth have better quality with regard to other pages. If the depth of a page is low, the depth of pages whose links refer to that is also low. In calculation of the score of links like formula 8, depth of pages containing those links are also considered. In this formula, pagedepth shows the depth of page.

8) $linkscore_3 = linkscore_2 \times (1 + \frac{1}{pagedepth})$

Like what has been expressed in [1, 2], if there are links to a page from several high quality pages, it is foreseen that that page will have a high quality too. If a link is extracted from a current page which has been obtained from the previous pages and which a score has been given to it from before, final score of link is calculated by combining the previous and new score according to formula 8.

9) $linkscore = MAX(linkscore_3, linkscore_{old}) + \gamma \times MIN(linkscore_3, linkscore_{old})$

In formula 9, λ is a constant amount and linkscore$_{old}$ is a saved score which has been calculated for the specific link from previous pages.

4.2. Determining Order of Following Links

Random selection is made both in genetic algorithms and AntNet routing algorithm, and may be considered as a turning point of combination of these two methods. Here from among all links extracted by Roulette-wheel method, address of pages to be retrieved must be selected randomly (in this method, links with higher scores have more chance of selection). Each time a predetermined number of links are selected and the pages related to them are retrieved.

10) $p_L = \frac{linkscore_L}{\sum_{i \in linkset} linkscore_i}$

In formula 10, P$_L$ shows the likelihood of selecting link L and linkset includes all extracted links that the pages referred by them have not been retrieved yet.

4.3. Correcting Score of Links by AntNet Routing Algorithm

If after retrieval of a page, the score that is calculated for it (pagescore) has a difference exceeding a threshold amount with regard to a foreseen score (linkscore), the score of all links is corrected in the paged that were linked to this page. The algorithm required for this task is shown in figure

```
If $(| difference \ l |> threshold)$
 For each page P that has a link to this new downloaded page
   If $((pagescore_P < pagescore_{downloaded\ page}) and\ (difference > 0))$
  or $((pagescore_p > pagescore_{downloaded\ page}) and\ (difference < 0))$
     For each link L in the P
     {
$linkscore_L = \begin{cases} linkscoreL + sign(difference) \times \Delta_{\max} & if\ |\delta \times difference\ l |> \Delta_{\max} \\ linkscoreL + \delta \times difference & otherwise \end{cases}$
}
```

Figure 2: The algorithm for correction of links score based on AntNet routing algorithm

The amount of difference is equivalent to the score of retrieved page (pagescore) minus the score of the link (linkscore) to which it refers. Threshold is the mean difference calculated for previous pages. Δ_{max} shows the maximum changes that could be applied on the links' score. Sign function, with regard to the mark of its input amount, returns +1 and -1 amounts and δ is a constant amount.

Testing the Recommended Method and Providing the Results

To test our method we did not use a fixed set of web pages but the crawler directly connected to the internet and retrieved the pages. In this way we have access to a very large set and since time has no influence on assessment of standards, the results of test are quite valid. We tested our three methods BFS (best first search), genetic algorithm and the recommended method and compared their results. The amounts of α and β were considered 4 and amounts of λ and δ were taken 0.25. In order to implement genetic algorithm mutation, each time two key words were randomly selected and a URL address is made as follows to request web

http://www.google.com/search?hl=en&q=

keyword1+keyword2 &btnG=Google+Search

We get the URL addresses that we need to start crawler by searching key words in Google search engine and selecting several found pages. In the first experiment conducted with three methods, the key words were search, engine and computer. The graphs related to the results are shown in figures 3 and 4. In the 2nd test, the key words artificial and intelligence and computer were shown. The harvest rate standard in each moment shows the percent of retrieved pages in relation with the subject, with regard to total retrieved pages. In Table Ⅰ, the average amounts of harvest rate obtained from three methods are compared. In the first test, our recommended method was 14% better than BSF and 9% better than genetic algorithm. In the 2nd experiment, this improvement is 12% and 10% respectively.

TABLE I COMPARISON OF MEAN HARVEST RATE AMONG THREE METHODS

Recommended method	Genetic algorithm	BFS	Name of method
96%	87%	82%	Test 1
59%	49%	47%	Test 2

TABLE II COMPARING THE MEAN SCORE OF PAGES RETRIEVED IN THREE DIFFERENT METHODS

Recommended method	Genetic algorithm	BFS	Name of method
0.70	0.33	0.41	Test 1
0.61	0.46	0.42	Test 2

Table Ⅱ contains the mean scores of the pages gathered in three methods. As you see in this table, in test 1, the results of our recommended method is respectively 29% and 37% better in comparison with BFS method and genetic algorithm. This improvement is respectively 19% and 15% in test 2.

5. Conclusion

The main objective of this article is to present a method to properly determine the quality of links, or in another words, to properly foresee quality of webpages that have not been retrieved so far but a link is available to them. For this purpose we use an algorithm similar to AntNet routing algorithm. To avoid local search problems, our recommended method is based on genetic algorithms. Moreover, here to calculate the page scores and initial scores of links, different methods have been appropriately combined with each other instead of using a special method. The results of conducted tests show priority of our recommended method with regard to other methods. This method may simply be added to focused crawlers and increase their efficiency to some extent.

6. References

[1] S. Chakrabarti, M.V.D. Berg, B. Dom, “Focused Crawling: a New Approach to Topic-Specific Web Resource Discovery,” *31th Computer Networks Conf*, pp. 1623–1640, 1999.

[2] S. Chakrabarti, M.V.D. Berg, B. Dom, “Distributed Hypertext Resource Discovery through Example,” *25th Int. Conf. on Very Large Data Base*, USA, pp. 375–386, 1997.

[3] J. Cho, H. Garcia-Molina, “The Evolution of the Web and Implications for an Incremental Crawler,” *26th Int. Conf. on Very Large Data Bases*, USA, pp. 200-209, 2000.

[4] J. Cho, H. Garcia-Molina, “Synchronizing a Database to Improve Freshness,” *ACM SIGMOD Int. Conf. on Management of Data*, USA, pp. 117-128, 2000.

[5] J. Cho, H. Garcia-Molina, L. Page, “Efficient Crawling through URL Ordering,” *7th Int. World Wide Web Conference*, Australia, pp. 161-172, 1998.

[6] M. Diligenti, F. Coetzee, S. Lawrence, “Focused Crawling Using Context Graphs,” *26th International Conference on Very Large Databases (VLDB)*, Cairo, Egypt, pp. 527–534, 2000.

[7] M.Najork, J.Wiener, “Breadth-First Search Crawling Yields High-Quality Pages,” *10th Conference on Word Wide Web*, Hong Kong, pp. 114-118, 2001.

[8] J. Qin, H. chen, “Using Genetic Algorithm in Building Domain-Specific Collections: An Experiment in the Nanotechnology Domain,” *38th Annual Hawaii International Conference on System Sciences*, USA, p. 102.2, 2005.

[9] J. Rennie, A. McCallum. “Using Reinforcement Learning to Spider the Web Efficiently”. *16th International Conference on Machine Learning*, USA, pp. 335-343, 1999

[10] V. Shkapenyuk, T. Suel, “Design and Implementation of a High-Performance Distributed Web Crawler,” *18th international conference on data engineering*, USA, pp. 357- 368, 2001.

[11] H. Younes, D. Chabane, “High Performance Crawling System,” *6th ACM SIGMM international workshop on Multimedia information retrieval*, New Yourk, USA, pp. 299-306, 2004.