Performance Evaluation of Web Crawler

Sandhya [1,] M. Q. Rafiq[2]
Department of Computer Engineering,
Aligarh Muslim University, Aligarh(UP) India
[1]sandhya.pundir@gmail.com, [2]mqrafiq@hotmail.com

ABSTRACT : *Extracting information from the web is becoming gradually important and popular. To find Web pages one typically uses search engines that are based on the web crawling framework. A web crawler is a software module that fetches data from various servers. The quality of a crawler directly affects the searching quality. So the time to time performance evaluation of the web crawler is needed.*

This paper proposes a new URL ordering algorithm .It covers major factors that a good ranking algorithm should have. It also overcome limitation of PAGERANK. It uses all three web mining technique to obtain a score with its parameters relevance .It is expected to get better result than PAGERANK, as implementation of it in a web crawler is still under progress .

Keywords : *Web crawler, URL, Web Pages*

1. INTRODUCTION

The World Wide Web (WWW) is a big network where you can get a huge amount of information. The Web is a collection of interconnected documents and other resources, linked by hyperlinks and URLs.[1]Searching within the Web is performed by the special engines, known as Web Search Engines. Among the processing, ranking has been a key technical link in design of search engine, which attracts widespread attention. Different strategies are implemented on this topic

.Some of them is based on classical information retrieval technologies, such as Vector Space Model (VSM) [1], extended Boolean Model [1], probability model [1], BM25 [1]etc.; Others analyze Web link structures, for example, the well-known Page Rank algorithm, which was proposed by Google in 1998 [4], and the hub and authority.

Website ranking is very useful .Generally speaking Web pages from important websites always have higher weights in results ranking. Furthermore, important websites should be crawled first and have higher updating priority when designing spiders [5];

However existing technologies of site ranking are limited to one setting of ranking[9][4], namely ranking based on the link analysis. In the computing, they suppose that it was of equal probability to click all the hyperlinks in one page. But in fact the choosing for next page is of in equable Probability, people tend to select the pages they are interested in. In other words people tend to click the hyperlinks which have higher semantic relevance between the anchor texts and the page contents. Semantic relevance should be considered in the computing of site ranking[5].

For most users they would click the first hyperlink, which is more relevant to the page content. Besides, for site ranking the updating frequency of websites is also important[5].

Obviously if a website rarely updates, even though it has lots of out-links, the site should not be given a high ranking. In this paper, semantic relevance, page popularity from server logs[2] and combine time frequency into the final ranks.

To sum up, main contributions in this paper are:

First, site ranking algorithms, using anchor texts semantic relevance is used in computing rank values. Second, time labels in Web pages are considered in computing the updating frequency of websites. And the updating frequency of websites is further imported in computing of Site Ranking. Thirdly a public popularity score is calculated for the site. At last a Final score based on the importance of these factors is calculated. This algorithm is under implementation on architecture shown as in Figure1.

The rest of the paper is organized as follows. It starts with a brief review of related works in Section 2. Then in Section 3, semantic relevance, public popularity of site, time frequency are discussed and website ranking, algorithm is proposed. The result is discussed in Section 4. Finally, Section 5 concludes this paper and gives directions for future works.

1.1 Proposed architecture and Components of Crawler

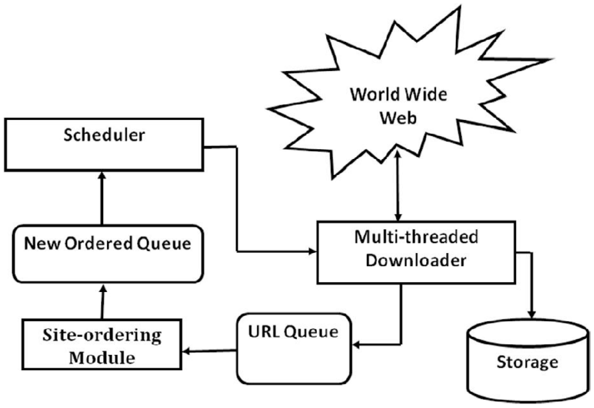

Figure 1

1. **Multi-threaded Downloader:-**It download documents in parallel by various parallel running thread.

2. **Scheduler**:-Selects the next url to be downloaded.

3. **URL Queue**:-A queue having all url of page.

4. **Site-ordering Module**:-It score the site based on various factors as described below and order them based on the score.

5. **New ordered queue**:-Urls sorted based on their score.

6. **World Wide Web**:- Collection of interlinked documents.

7. **Storage**:-to save the downloaded documents.

1.2 Factors Affecting Performance of Web Crawler

A critical look at the available literature [1] [3][5] indicates the following issues that need to be addressed:

Issue 1: Overlapping of web documents

Overlap problem occurs when multiple crawlers running in parallel download the same web document multiple times.

Issue 2: Quality of downloaded web documents

The quality of downloaded documents can be ensured only when web pages of high relevance are downloaded by the crawlers.

Issue 3: Network bandwidth/traffic problem

In order to maintain the quality, the crawling process is carried out using either of the following approaches:

- Crawlers can be generously allowed to communicate among themselves or
- They cannot be allowed to communicate among themselves at all.

Both approaches put extra burden on network traffic.

Issue 4: Change of web documents

Changing of web documents is a continuous process. This change must be reflected at the search engine repository failing which a user may have to access an obsolete web document.

2. Related Work

Lots of previous work has focused on the crawling ordering strategy so far [4] [9].Computing PageRank for whole web graph is both time consuming and costly [5].Recent work has made many modifications in traditional PageRank[3][5][7][10] to improve the performance ,but still these algorithm are computationally expensive.

As an alternative here a new url ordering algorithm is proposed. Major advantage is that it will be relatively inexpensive. Because website can process their logs efficiently.

3. PROPOSED ALGORITHM

To achieve better results for above mentioned parameters of web crawler a url ordering algorithm is proposed. In this algorithm a new site rank is calculated which covers all three types of web mining technique i.e web content mining, web usage mining and web structure mining.

As a result of using all three web mining technique covering all issues it is believed to achieve an efficient site rank algorithm. Algorithm steps are as follows:-

1. Input a url.
2. Extract whole site.
3. Remove the stop word and suffix.

4. Calculate tern weight using TF-IDF.
5. Now calculate content similarity.
6. Calculate public popularity score using server logs.
7. Obtain site updating frequency.
8. Final site score is now obtained in accordance to relevance of above factor i.e .Final Rank=0.223(result at step 6) +0.2387(result at step 7) +0.35(result at step 5) (1)

1.1 Algorithm Explanation

A web crawler's working start with a seed url. Every url is associated with a web page or site. Then content of page are downloaded. We know that all content are not important. To weight the page in accordance to importance its stop word and suffix are removed. By this content to be used for ranking become less in size and more relevant.

TF-IDF(Term Frequency –Inverse Document Frequency)

This scheme is used to calculate weights of the term or words in the document.

In *TF scheme*, the weight of a term t_i in page pj is denoted by f_{ij} . The following approach is applied [3]:

$$t\, f_{ij} = f_{ij} / max\ (f1j, f2j, \ldots\ldots f|V|j) \qquad (2)$$

where f_{ij} is the frequency count of term ti in page *dj* , and

|*V*| is the vocabulary size of the collection.

TF-IDF Scheme: Let *N* be the total number of pages in web database, df_i be the number of pages in which term t_i appears at least once, and f_{ij} be the frequency count of term t_i in page d_j . The inverse document frequency (denoted by idf_i) of term t_i is computed by:

$$idf_i = \log N/df_i \qquad (3)$$

The term weight is computed by:

$$w_{ij}= tf_{ij} \times idf_i \qquad (4)$$

To compute the TF-IDF weight of each term, the improved method to determine how important a term is in a page is used.

$$w_{ij} = \{0.5 + 0.5\times tf_{ij}\}\times \log N/df_{ij} \qquad (5)$$

By Equation 5, we can see that if a term t_i appears in every document, then $N=df_{ij}$ and $w_{ij} = 0$, which means that t_i has no way to any page So, improve the above formula to be:

$$w_{ij} = \{0.5 + 0.5\times tf_{ij}\}\times\log N+1/df_{ij} \qquad (6)$$

The following similarity measure to compute the similarity between pages p_a and p_b

$$Sim(p_a,p_b) = (d_a.d_b)/(||d_a||2+||d_b||2-d_a.d_b)$$
$$= (\Sigma\ w_i p_a \times \Sigma\ w_i p_b) / \Sigma w^2_{ipa}+\Sigma w^2_{ipb}-\Sigma w_{ipa}*w_{ipb} \qquad (7)$$

After this public popularity of the page is calculated

Popularity information is exploited from web logs on a website server. Four different type of access information is extracted from web logs namely, the Total External Count (TEC), the Unique External Count (UEC), the Total Internal Count(TIC),and the unique Internal Count(UIC).External count is request made to url from outside local network and Internal count is local request made .Then weighted score for each url is calculated as follows.

$$Total_{acc}=TEC_{acc}+UEC_{acc}+TIC_{acc}+UIC_{acc} \qquad (8)$$

$$WS=\frac{x*\ TEC_{acc}}{Total_{acc}}+\frac{y*\ UEC_{acc}}{Total_{acc}}+\frac{z*\ TIC_{acc}}{Total_{acc}}+\frac{w*UIC_{acc}}{Total_{acc}} \qquad (9)$$

Where WS=Weighted Score

$TEC_{acc=}$ TEC algorithm accuracy

$UEC_{acc=}$ UEC algorithm accuracy

$TIC_{acc=}$ TIC algorithm accuracy

$UIC_{acc\ =}$ UIC algorithm accuracy

And x,y,z,w are raw external, unique external, internal and unique internal counts for the URL.

Next step is to calculate the updating frequency for each page. $$Freq(\ s)= \frac{x*Na}{D} + (1-x)*\frac{Nna}{D} (10)$$

Where *Na* denotes the count of updated related pages in a website, and *Nna* denotes the count of updated non-topic pages. *D* is the updating time interval for calculating updated pages. x is a damping factor, *0<x<1*, usually set to *0.85*.

Now a final score depending on the importance of different parameters score is calculated by equation 1. These importance score is influenced by the score of a survey [6].

4. EXPERIMENTS

4.1 Data Collection

In this section, experimental studies which will be carried on real data that will be crawled from internet by proposed crawler . Url ordering proposed will be compared with traditional PageRank produced by a freeware. Web Crawler is implemented in Java on windows-XP platform and experiments are continued on Intel core2duo n series CPU with 3GB RAM.

4.2 Evaluation Method

In order to measure the performance of the proposed ranking algorithms, it will be evaluated in two ways. First, top 100 URLs returned by the above mentioned algorithms will be used. Then a count of different urls present will be done .This will be indication for site recommendation. Also, pages of spam sites should be identified. Minimum number of overlapping document, more relevant page ,less traffic consume less bandwidth and most updated page storage are to be considered.

5. RESULTS

As all the three web mining technique are employed in this above algorithm. Using website logs is inexpensive. Semantic relevance choose more accurate probability. According to their relevance a weight factor is multiplied to obtain more accurate site score. Weight factor also plays a important role in obtaining more precious results. How accurate results obtained will be mentioned as early as possible.

This algorithm also considers the problem of traditional PageRank .It is expected that it will give better result. as it is able to fulfill above mentioned issues.

First is less overlapping, to be obtained as different content, page popularity and update frequency give precious score.

Secondly, a good score will help in download a highly relevant page first, so better quality expected.

Thirdly, when sites are carefully prioritized there are chances of less ambiguousness and frequent unnecessary traffic can be avoided.

Finally, change frequency is also taken into consideration which helps to retrieve most updated page.

6. CONCLUSION AND FUTURE WORK

In this paper a new URL ordering algorithm is proposed based on the content similarity, popularity information from web logs and site updating frequency. It is expected to perform well and better than traditional pagerank and also have anti-spamming ability.

It also has a drawback that new pages has not been accessed are penalized and also do not have good updating frequency.

Using last modified date popularity information may address this issue. Finally, the proposed algorithm is under implementation and actual results will be compared as early as possible.

7. REFERENCES

[1] Bhaskar Reddy,Kethi Reddy,” Improving efficiency of web crawler algorithm using parametric variations” Ph.d thesis submitted in June 2010 at Thapar University India.

[2] Arvind chandramouli ,Susan gauch andJosua eno “A popularity-based URL ordering Algorithm for Crawlers”, *Rzesow ,Poland may 13-15 2010 IEEE*

[3] Shaojie Qiao ,Tianni Li, Jiangtao Qiu," SimRank: A Page Rank Approach based on Similarity Measure " 2010 IEEE.

[4] Dilip Kumar Sharma , A.K.Sharma "A Comparative Analysis of Web Page Ranking Algorithms "(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676

[5] Hongzhi Guo, Qingcai Chen, Xiaolong Wang, Zhiyong Wang, Yonghui Wu," STRank: A SiteRank Algorithm using Semantic Relevance and Time Frequency "Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics .

[6] http://palatnikfactor.com/2010/05/18/components-of-googles-ranking-algorithm-in-2010-linking-still-king/

[7] Yi Zhang,Lei Zhang,Yan zhang,Xiaoming Li,"XRrank;Learning More from Web User Behaviors"2006 IEEE

[8] Qiancheng jiang,Yan Zhang,"SiteRank-Based crawling Ordering Strategy for Search Engine" 2007IEEE

[9] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia," Page Ranking Algorithms: A Survey", 2009 IEEE International Advance Computing Conference (IACC 2009).

[10] Apostolos Kritikopoulos, Martha Sideri, Iraklis Varlamis," WORDRANK: A METHOD FOR RANKING WEB PAGES BASED ON CONTENT SIMILARITY "2007 IEEE