International Journal of Research in Advent Technology, Vol.2, No.2, February 2014

E-ISSN: 2321-9637

Web Crawler Using Priority Queue

Ms. Kiran P. Lokhande[1]
A.G.P.C.O.E. Nagpur,
Maharashtra india
klokhande26@gmail.com

Prof. Sonal S. Honale[2]
Assist. Professor A.G.P.C.O.E
Nagpur ,Maharashtra india
sonalhonale@gmail.com

Miss H.N.Gangavane[3]
BDCOE Wardha
Maharashtra india
harsha24_hng@rediffmail.com

Abstract- The World Wide Web (WWW) has billions of documents and these documents are attachedto each other using hyperlinks. Web crawler is a heart of Search engine that gathers these documentsfrom WWW. Maximum documents present on WWW are dynamic and changes periodically.Hence, Crawlerneeds to refresh these documents to update database of search engine. In this paper, we haveproposed a priority based focused web crawling algorithm. The web pages corresponding to URL(Uniform Resource Locator) are downloaded from web and calculated the relativity score ofdownloaded page with focus word. We store URL and its relativity score with focus word in priorityqueue instead of normal queue. So, every time priority queue returns maximum Score URL to crawlnext. The overall performance gain over simple crawler is 87% and over focused crawling is 24%.
Keywords: Priority, focus word, web pages, downloader, search engine.

1. INTRODUCTION

Web search engine is designed to find information which is related to search query specifiedby user from WWW. Search engine stores millions of web pages and their links. These web pagesare needed to be refreshed that makes it more reliable. Search engine uses web crawler for thisvpurpose. Web crawler is a continuous running program which downloads web pages periodicallyfrom WWW. The downloaded pages are indexed and stored in a database as shown in Fig. 1.[1]**Figure 1:** Architecture of Search engineThere are two types of web crawling breadth first crawling and best first crawling [2].

1.1 BREADTH FIRST CRAWLING

Breadth first crawling method is same as breadth first search in a graph. Web crawler startswith initial seed URLs. It downloads web pages for given URLs. Then extract new URLs from thedownloaded pages, add them into queue and pick up URL one by one and repeat same process forspecific count or until queue is empty. The architecture of classical web crawler is shown in fig. 2
This has the following four components:

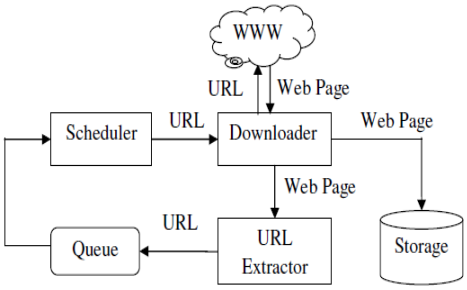

Figure 2: Architecture of usualWeb Crawler

1.1.1Queue
Queue is a data structure used by crawler to store URLs extracted from the downloaded pages for another used in crawling process.

1.1.2 Scheduler
Scheduler selects URL from the queue and sends it to downloader for downloading the web page atgiven URL.

1.1.3 Downloader
Downloader downloads web page at given URL.

1.1.4 Storage
After downloading page from the web, crawler stores page in stable storage.

1.2 Best first crawling
The best first crawling focuses on to download only relevant pages of a particular given topic.A crawler using a best first crawling strategy is known as a focused crawler [2]. In other wordsFocused crawling is a variation of breadth first crawling where web pages related to particular topicor set of topics are downloaded only.

In this paper, we are presenting a priority based focused web crawler that will downloadrelevant pages related to a particular topic or focus string only. We have used priority queue insteadof simple queue to store web pages and their similarity score with focus string. Every time when adelete operation performed on queue will return maximum score web page.The reminder of this paper is organized as follows: section 2 contains related work in this area.Section 3 is the architecture of priority based focused web crawling. Section 4 contains the algorithmof priority based focused web crawling. Section 5 describes the experimental results and Section 6 isthe conclusion and future work.

2. RELATED WORK

The various crawling algorithms has been proposed which are as follows:N. Singhal et. al.[1] have designed a incremental web crawler. The incremental crawler visitsthe internet periodically to update its database. Based upon updation web documents, web documentsare categorized and grouped as very frequently, frequently less frequently. The crawler visits a sitefrequently and the frequency of visits may be adjusted according to the category of the site. Thisarchitecture is more suitable for parallel web crawler.S. Ganesh et. al.[3] have proposed an ontology based web crawler. In this approach, a newmetric called association-metric has been proposed. The association-metric analyzes the semanticcontent of the URL based on the domain dependent ontology. After downloading the page, theassociation metric estimates the relevancy of the links in that page. Finally, reordering of URL isdone based on relevancy of web page.D. Mukhopadhyayet. al.[4] have proposed a domain specific web crawler which crawlsdomain specific Web pages from the World Wide Web(WWW). Crawler uses ontology of a domainfor which web pages has to be crawl.X. Chen et. al.[5] gave the methodology for focused crawling. They have focused on contentof web page to improve page relevance and also used link structure to improve the coverage of aspecific topic. They considered only two factor, content of web page and link structure, to getrelevancy of web page.D. Hatiet. al.[6] have proposed an adaptive focused crawling based on link analysis. In thisapproach, they first calculate the score of unvisited URL based on its anchor text relevancy score,Relevancy score of its parent, its description in Google search engine and calculate the similarityscore of description with topic key words. The major issue of this technique is URL queueoptimization.S. Thenmalaret. al.[7] have proposed an algorithm for focused crawling based on ontology.They are preparing topic as an overall conceptual vector that is obtained by combining conceptvectors of individual pages associated with seed URLs. Here the role of ontology is to obtainconcepts associated with seed page. The next URL to be crawl is based on the conceptual rank of theweb page at that level which is obtained by conceptual matching between conceptual vectors of allweb pages at each level.

3. PRIORITY BASED FOCUSED WEB CRAWLING

The overall crawling process is shown by figure 3.

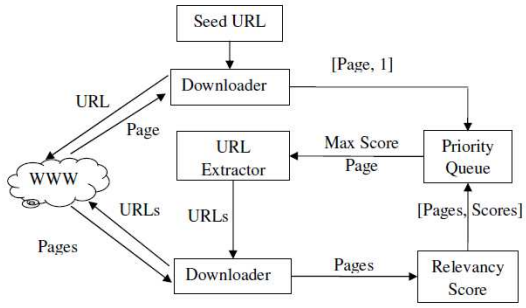

Figure 3: Architecture of Priority based focused crawler.

The crawling process begins with initial seed URL and focus word.

Step 1: Crawler downloads web page corresponding to given URL.

Step 2: Now, it extracts all the URLs present in downloaded page.

Step 3: Again, crawler downloads all the pages corresponding to all new extracted URLs.

Step 4: Now, we calculate cosine similarity between focus word and all downloaded pages that
Works as relativity score.

Step 5: we add page and its relativity score into the priority queue and every time when we delete apage from priority queue, queue will return a maximum similarity page.

Step 6: Now, repeat step 2-5 for either specified number of pages or until queue is empty.

Let, crawler starts crawling with initial seed URL, we assign score 1 to initial seed URL.Web page is downloaded from web for seed URL and new URLs are extracted from downloadedPagei.e. URL 1, URL 2,..., URL N as shown in figure 4.Now, crawler again downloads web pages for every new URL which are Page 1, Page 2,Page N. We calculate similarity score between web pages and focus word. Let Page 1, Page 2,Page N has score 0.7, 0.5,....., 0.8 respectively where 0.8 is maximum score. Downloaded pages

andtheir score are inserted into priority queue. Now a page is deleted from priority queue, page withhighest score is selected. The maximum score URL is URL N, crawler will extract all the URL fromPage N first.

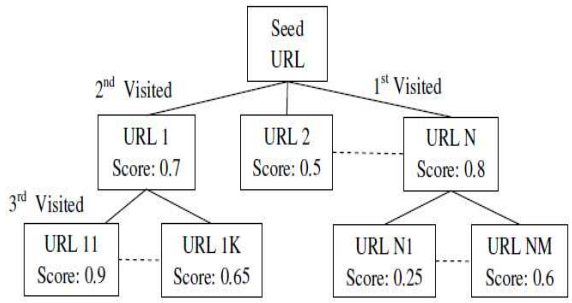

Figure 4: Priority based focused crawling process

Again, it will repeat the same process i.e. extracting URLs (URL N1, URL N2,...., URL NM)and downloading pages (Page N1, Page N2,...., Page NM). Calculate the similarity score betweennewly downloaded pages and focus word. Let Page N1,....., URL NM has score 0.25, ..., 0.6respectively where 0.6 is maximum. Downloaded pages and their score is inserted into priorityqueue. Now a page is deleted from priority queue at this time Page 1 is selected because it hasmaximum score then remaining other pages present in queue. Every time maximum score page isselected for crawling. This will be the advantages of using priority queue over simple queue. This will defiantly improve performance of crawling process over normal crawling process.

4. CRAWLING ALGORITHM

The priority based focused crawling algorithm works as follows:
Input: Initial seed URL, Focus_String and PQueue.
Output: Web_Pages related to Focus_String.
Step 1: Page :=downloadPage(URL);
Step 2: addPQueue(Page, 1);
Step 3: While PQueue is not empty do
Step 4: Page :=dePqueue();
Step 5: newURLs :=extractURL(Page);
Step 6: for each ith URL in newURLs do
Step 7: Page[i] := downloadPage (newURLs[i]);
Step 8: RScore[i] := SimScore(Page[i], Focus_String);
Step 9: addPQueue(Page[i], RScore[i]);
Step 10: end for;
Step 11: end while;

Descriptions of various modules used in algorithm are as follows:

1.1 addPQueue(Page, Score)
addPQueue module add a new downloaded page and its similarity score with Focus_StringintoPriority Queue.

1.2 dePQueue()
dePQueue module returns a page which has maximum score from Priority Queue.

1.2 downloadPage(URL)
downloadPage module downloads web page from WWW corresponding to given URL.

1.3 extractURL(Page)
extractURL module extracts all URLs which are present in given Page.simScore(Page, Focus_String)simScoremodule calculates cosine similarity score between Page and Focus_String.

5. EXPERIMENTAL RESULTS

The performance of Focused crawler is measured by harvest rate. Harvest Rate measures therate at which relevant pages are crawled and how effectively irrelevant pages are filtered off from thecrawl [8].Harvest ratio (1)WhereR: No. of relevant web pages crawled, andN: Total No. of web pages crawledThis harvest ratio must be high. We have used standard data set present on WWW in the formof open directory named "http://www.dmoz.org". We have evaluated and compared the harvest rateof our crawler with simple crawler and focused crawler. The figure 5 shows the harvest rate ofbreadth first crawler, focused crawler [2] and priority based focused crawler. The 1000 web pagescrawled on focus word Computer, Science, Regional and Sports, and average of harvest rate are

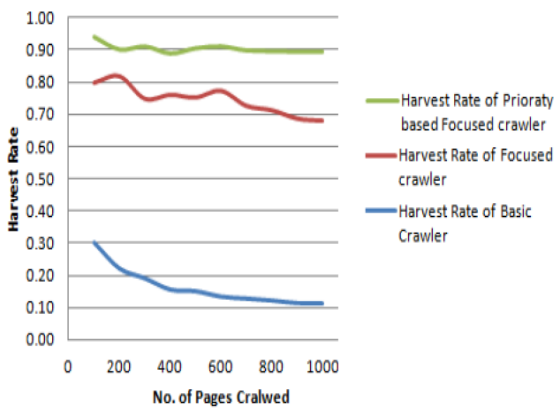

Figure 5: Priority based focused crawling process

6. CONCLUSION

In this paper, we proposed a priority based focused crawler which keeps all URLs to be visitin priority queue along with their relativity score. When we delete URL from priority queue, itreturns maximum score URL. Thus, every time a highest priority URL is return for crawling.Experimental results shows that our crawler gives 24% improved results over focused and 87%improved results over simple crawler. The main problem of this crawling strategy, it is more timeconsuming. In future, we will try to reduce the crawling time by implementing algorithm parallel
.

REFERENCES

[1] N. Singhal, A. Dixit, and A.K. Sharma, Design of a Priority Based Frequency RegulatedIncremental Crawler, International Journal of Computer Applications, 1(1), 2010,.

[2] Ah Chung Tsoi, Daniele Forsali, Marco Gori, Markus Hagenbuchner, and Franco Scarselli, ASimple Focused Crawler, WWW2003 ACM, 2003.

[3] S.Ganesh, M.Jayaraj, V.Kalyan, S. Murthy, and G.Aghila, Ontology-based Web Crawler, Proc.IEEE International Conf. on Information Technology: Coding and Computing (ITCC'04),2004.

[4] D. Mukhopadhyay, A. Biswas, and S. Sinha, A New Approach to Design Domain SpecificOntology Based Web Crawler, Proc. 10th IEEE International Conf. on InformationTechnology, 2010.

[5] X. Chen, and X. Zhang, HAWK: A Focused Crawler with Content and Link Analysis, Proc.IEEE International Conf. on e-Business Engineering, 2008.

[6] D. Hati, B. Sahoo, and A Kumar, Adaptive Focused Crawling Based on Link Analysis, Proc.2nd IEEE International Conf. on Education Technology and Computer (ICETC), 2010.

[7] S.Thenmalar, and T. V. Geetha, Concept based Focused Crawling using Ontology,International Journal of Computer Applications, 26(7), 2011, 29-32.

[8] S. Chakrabarti, M. van den Berg, and B. Dom, Focused crawling: a new approach to topicspecificWeb resource discovery, Proc. 8th International WWW Conf., 1999.

[9] A. Suganthy, G.S.Sumithra, J.Hindusha, A.Gayathri and S.Girija,, "Semantic Web Servicesand its Challenges", International Journal of Computer Engineering & Technology (IJCET),Volume 1, Issue 2, 2010, pp. 26 - 37, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[10] AlameluMangai J, Santhosh Kumar V and Sugumaran V, "Recent Research in Web PageClassification – A Review", International Journal of Computer Engineering & Technology(IJCET), Volume 1, Issue 1, 2010, pp. 112 - 122, ISSN Print: 0976 – 6367, ISSN Online:0976 – 6375.

[11] Houda El Bouhissi, MimounMalki and DjamilaBerramdane, "Applying Semantic Web Services", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 2, 2013, pp. 108 - 113, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.