A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency

S S Vishwakarma
Department of CSE
Radharaman Institute of Technology and Science, Bhopal, M.P, India

A Jain
Department of CSE
Radharaman Institute of Technology and Science, Bhopal, M.P, India

A K Sachan
Department of CSE
Radharaman Institute of Technology and Science, Bhopal, M.P, India

ABSTRACT
The Web crawler is a computer program that downloads data or information from World Wide Web for search engine. Web information is changed or updated rapidly without any information or notice. Web crawler searches the web for updated or new information. Approximate 40 % of web traffic is by web crawler. In this paper a web or network traffic solution has been proposed. The method of web crawling with filter is used. This approach is query based approach. The proposed approach solves the problem of revisiting web pages by crawler.

General Terms
HTTP GET Request

Keywords
Web, Search Engine, Web Crawler, Web Crawling Traffic, HTTP GET request.

1. INTRODUCTION
The World Wide Web (WWW) is the main source of information collection in the contemporary world [14]. A massive amount of users are using WWW daily to access any type of information they want. But access to this information is not straightforward. Software is required through which users can browse the WWW. Such software's are called Web Browser. Web browsers are used for accessing information on the WWW.
But Internet is such a massive collection of web sites that it is very difficult to search for the specific web site that we want unless and until we remember the name of the URL. Since it is impossible to remember URL of millions of web sites the web developers came up with the most revolutionary part of WWW that is Search Engines.

Search engines are used to search for a web sites or pages which are related to the content supplied by the user in the search engine. Search engines can be used by almost all the users for reaching particular page as it is very easy to use. Main component of search engine is web crawler [13]. Web Crawler crawl WWW and tags web pages which contain relevant information matching the user supplied search string. There are almost 5,500 sear operational crawlers currently in the WWW. Search engine needs updated information to generate correct results for user queries. Web crawler revisits the web pages for search of updates. But a main drawback in this process is that even if a single web page is updated in the web site, the whole web site is scanned by the crawler to get the updates in that single page. This generates a huge amount of web traffic [1].

In an interesting fact based on experimental research tells that in 2002 almost 40% of the web traffic was generated by Web Crawlers [1] which rise to 50% in 2010 [2].

Hypertext transfer protocol is a protocol can be used to intermediate between the client and server computing. This is used to enable the communication between the client and server [15]. HTTP defines the following nine methods: Head, Get, Post, Put, Delete, Trace, Option, Connect, and Patch.

In this paper a web or network traffic solution has been proposed. The method of web crawling with filter is used. This approach is query based approach. The proposed approach solves the problem of revisiting web pages by crawler. In our technique the crawler need not to scan the entire web site but only that web page which is updated. This reduces the web traffic by a large amount. It uses one parameter LAST_VISIT, last visit parameter time of last crawling time of particular web crawler. We first discuss the related work in this field and then elaborate our work.

2. LITERATURE REVIEW
A lot of research has been conducted so far on Crawler which mostly emphasize on the structuring of Web Crawlers.
In [3] Google software engineers announced that Google Search had discovered one trillion (10^{12}) unique URLs [a]. On which 109.5 million domains are operated. Out of which 74 % were commercial and other web sites operating in the, **.**com domain [4].

In [5] authors shows, New pages are created at a very high rate in web sites. A researcher shows that 8 % of web pages are created every week. Death of web pages is also very high; Today's 80 % of web pages accessible will not be accessible after one year. After one year only 20 % of pages accessible. Every week new content is generated 5 %. After one year 50 % of new content will be generated on the web.

In [6] authors performed crawling on 55000 web pages every hours for 5 weeks and shows that 34 % of web pages do not change during 5 weeks. 66 % of web pages are changed during that time. Average 66 % of web pages are changed after 123 hours. New pages are changed after 33 hours and websites of industry or trade are changes after 218 hours.

In [7] authors suggest placing web crawler at different geographical areas. A specific Web crawler downloads web pages within its geographical area which makes crawling faster and efficient.

In [1] authors suggest placing active routers at key places in network. Active routers record underling traffic for indexing.

Another idea is elaborated in [8, 9] about usage of mobile crawlers at web servers. Crawler check updates in web site and send updates for indexing to search engine [8, 9].

Unethical crawlers cause problems to network, web servers. Causing denial of service, copying user private information and downloading copyright data [10].

Research is in progress on Crawler ethics. Researchers have derived the formula to calculate the ethicality of web crawlers [2]

The use of dynamic web page to inform the web crawler about the new pages and updates on web site is discussed in [11, 12]. Dynamic web page accepts one parameter LAST_VISIT. This parameter Indicates time and date of last visit of web crawler to website. Parameter value is passed by HTTP GET request.

3. PROPOSED APPROACH

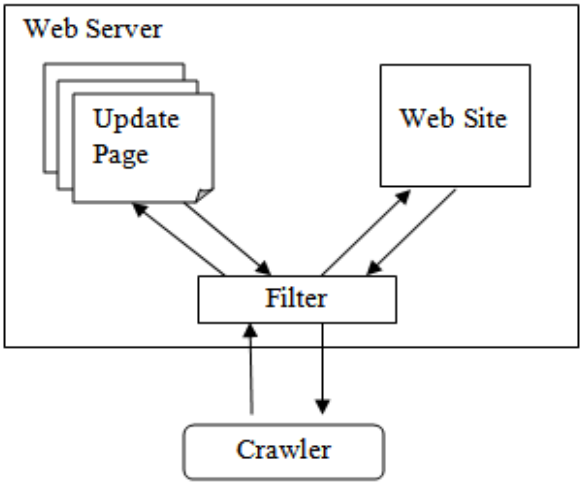

Figure 1: Proposed Architecture

Proposed approach is query based approach. It uses one parameter LAST_VISIT, last visit parameter time of last crawling time of particular web crawler. The author proposed architecture (see Figure 1).

Web crawler sends the HTTP GET request to any web document to web server. Request having one parameter LAST_VISIT indicates the last crawling time of web crawler.

A filter is used to check each request user agent and perform one operation depending upon the value of user agent. User agent act as users in computing:

If request is generated by web crawler and its user agent indicates it, Filter directs it to update web page.

If request is generated by web browser filter bypass the request to requested resources.

If request having parameter visit is equal to true it bypass the request to requested resources.

Dynamic web page receive the HTTP GET request with parameter LAST_VISIT and generates result.

Dynamic web page use list data structure to generate results, having list of URLs of web pages updated after crawler last visit.

Web crawler receives the results having list of URLs of web pages updated after crawler last visit.

Crawler only visits the list of URLs instead of visiting whole web site.

4. ALGORITHM USED

1. Web crawler send HTTP GET request with parameter LAST_VIST
2. Filter receive crawler generated request
3. Filter direct crawler request to update web page
4. Dynamic web page receive HTTP GET request with parameter LAST_VIST
5. Dynamic web page send updated URLs list
6. Web crawler receive updated URLs list
7. Crawler visit updated URLs list

5. EXPERIMENT

A website structure can be used to perform experiment (see Figure 2).

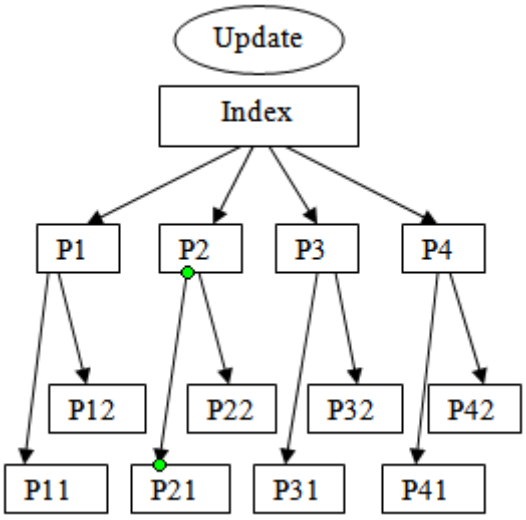

Figure 2: Website Structure

We perform experiment on four scenarios:

5.1 Scenario-I

If we update the index web page then as per existing approach crawler scan all the web pages i.e. 13 web pages and as per proposed idea crawler download index web page and update web page i.e. 2 web pages (See Figure 3).

5.2 Scenario-II

If we update the four web pages i.e., page 1, page 22, page 31 and page 4 then as per existing approach crawler scan all the web pages i.e. 13 web pages and as per proposed idea crawler

download web page 1, page 22, page 31, page 4 and update web page i.e. 5 web pages (See Figure 3).

5.3 Scenario-III

If we update the six web pages i.e., page3, page22, page4, page11, page 32, and page index then as per existing approach crawler scan all the web pages i.e. 13 web pages and as per proposed idea crawler download web page3, page22, page4, page11, page 32, page index and update web page i.e. 7 web pages (See Figure 3).

5.4 Scenario-IV

Last we update the eight web pages i.e., page 11, page 12, page 2, page 21, page 3, page 32, page 41 and page 42 then as per existing approach crawler scan all the web pages i.e. 13 web pages and as per proposed idea crawler download web page 11, page 12, page 2, page 21, page 3, page 32, page 41, page 42 and update web page i.e. 9 web pages (See Figure 3).

6. RESULT AND SIMULATION

Our simulation shows the comparison between existing web crawling and proposed idea. First we shows the comparison between normal web crawling (WC) and scenarios performed in experiment i.e. S-I, S-II, S-III and S-IV (see Figure 3).

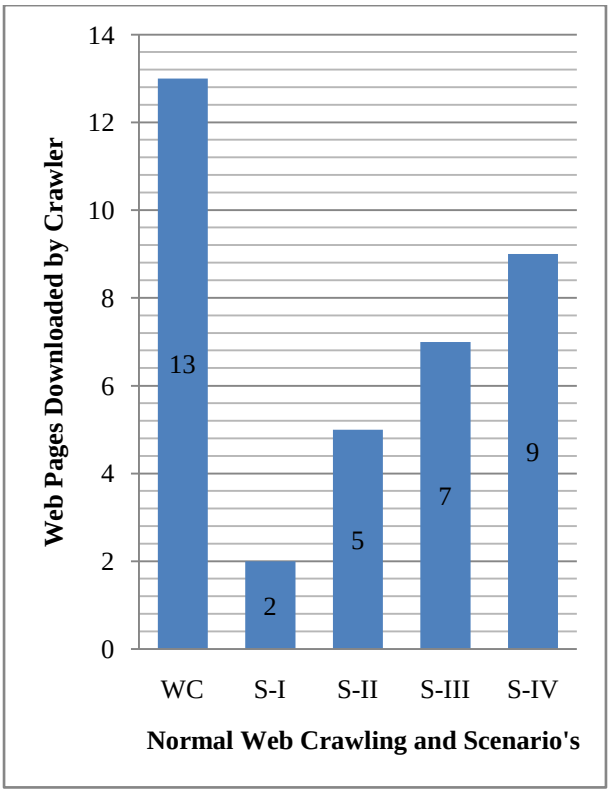

Figure 3: Comparison of Web Crawling and Scenario

Secondly we show how many times performed scenarios are much more efficient than normal web crawling. In this scenario-I is 6.5 times, scenario-II is 2.6 times, scenario-III is 1.85 times and scenario-IV is 1.44 times is more efficient than normal web crawling (see Figure 4).

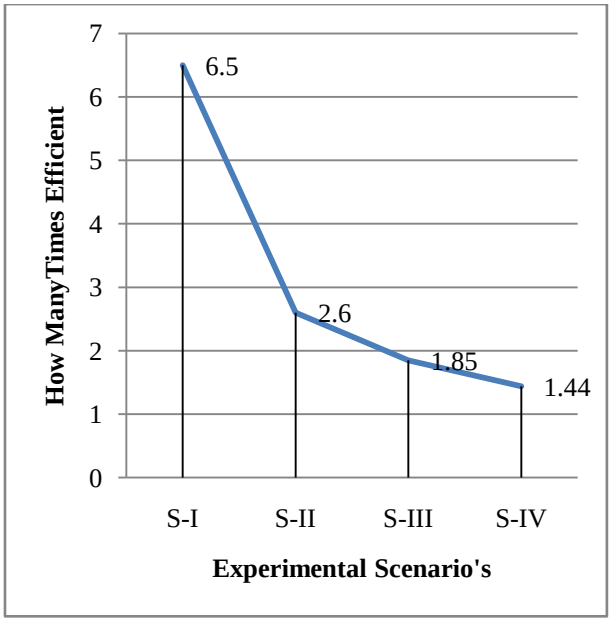

Figure 4: Comparison graph of Web crawling and Scenario

Thirdly we shows the comparison between normal web crawling **(**WC) and scenarios performed in experiment i.e. S-I, S-II, S-III and S-IV (in Bytes) (see Figure 5).

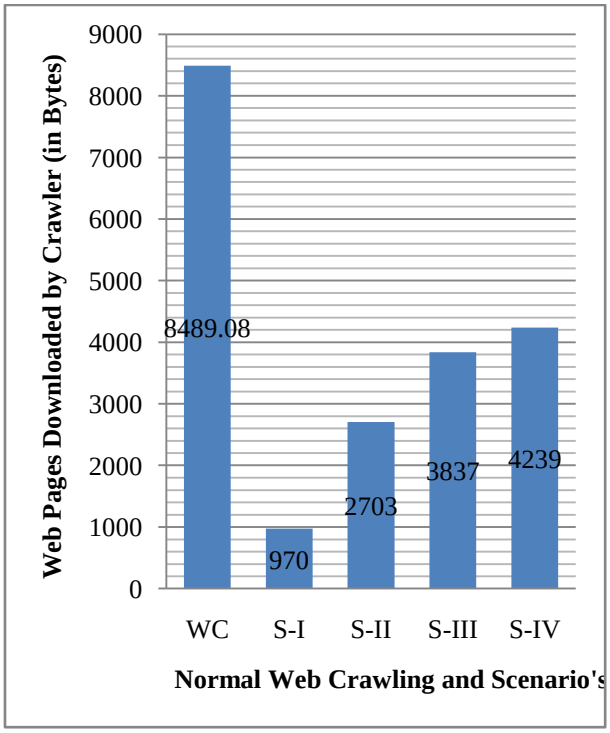

Figure 5: Comparison of Web Crawling and Scenario (In Bytes)

Fourthly we show how many times performed scenarios are much more efficient than normal web crawling as per size (in bytes). In this scenario-I is 8.75 times, scenario-II 3.14 times, scenario-III 2.21 times and scenario-IV 2.00 times more efficient than normal web crawling (see Figure 6).

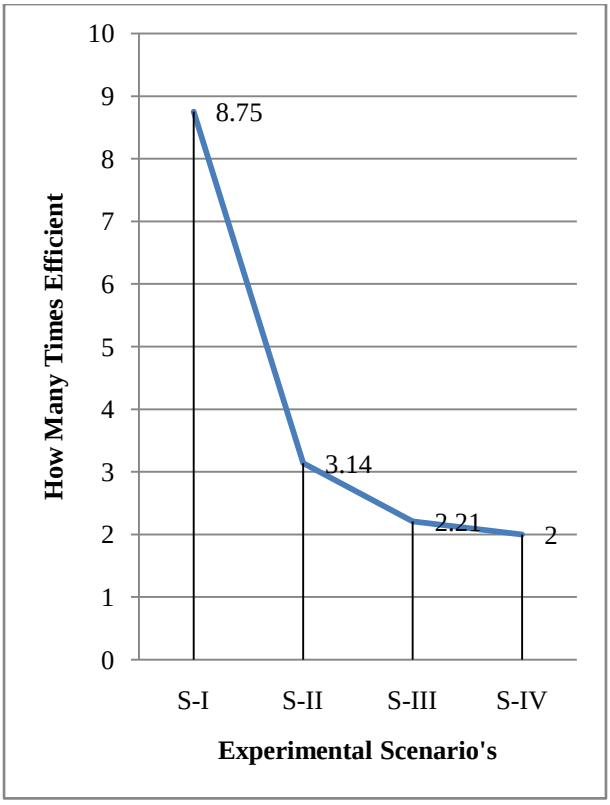

Figure 6: Comparison Graph of Web Crawling and Scenario (in Size)

Our scheme is also much more efficient than existing approach [11]. In existing approach if request is come from index page than direct it to index page but in our approach direct it to update page. In our approach force is used to crawler send request direct it to update page. Overall in our scheme request only direct it to update page. In proposed scheme crawling policies were changed.

7. CONCLUSION OR FUTURE WORK

Scheme can be implemented on existing system with some changes in a policy. Proposed scheme is much more efficient on existing system and unethical crawling. In proposed approach Crawler visits only update page instead of visiting full website. With the help of this approach reduces web crawling or network traffic. Many types of queries can be implemented with the help of this scheme. This scheme Useful in web mining and for what changes at the web page and how much change should report to web crawler. This approach is also helpful to do work on unethical crawling.

8. REFERENCES

[1] Yuan X, H Macgregor and J. Harms, "An efficient scheme to remove crawler traffic from the internet." Proceedings of the 11th International Conference on Computer Communications and Networks, Oct 2002. 14-16, IEEE CS Press, (pp: 90-95).

[2] Sun. Y, Council G. Isaac and Giles C. Lee, "The Ethicality of Web Crawlers", in the proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto Canada august 2010.(pp: 668-675)

[3] Alpert, Jesse; Hajaj, Nissan (July 25, 2008). "We knew the web was big..." The Official Google Blog.

[4] "Domain Counts & Internet Statistics". Name Intelligence. Retrieved May 17, 2009.

[5] Alexandros Ntoulas, Junghoo Cho and Christopher Olston, "What's new on the web ? the evolution of the web from a search engine perspective" WWW2004, may 17-22, 2004, New York, USA, ACM 1-58113-844-X/04/0005.

[6] Etyan Adar, Jaime Teevan, Susan T Durnais and Jonathan L Elsas, "The web changes everything: Understanding the dynamics of web content" WSDM 09, February 9-12-2009, Barcelona, Spam, ACM 978-1-60558-390-7.

[7] Cambazoglu, B.B.; Junqueira, F.; Plachouras, V.; Telloli, L., "On the feasibility of geographically distributed web crawling." (ISBN: 978-963-9799-28-8) In the proceedings of Third International ICST Conference on Scalable Information Systems, ICST, Vico Equense, Italy (2008).

[8] Bal.S and Nath.R,"Filtering the web pages that are not modified at remote site without downloading using mobile crawler". Information Technology journal 9(2)2010 ISSN 1812-5638, Asian Network for Sciencetific information. (pp: 376-380)

[9] Pahal N, Kumar S, Bhardwaj A and Chauhan N," Security Mobile Agent Based Crawler = (SMABC)"= International Journal of Computer Applications 1(14), February 2010. (pp: 5–11)

[10] Thelwall. M and Stuart. D, "Web crawling ethics revisited: Cost, privacy and denial of service". Journal of the American Society for Information Science and Technology. 2006. Volume 57, Issue 13 November 2006. (pp: 1771 - 1779)

[11] Shekhar mishra, anurag jain and A K Sachhan, "A Query based Approach to Reduce the Web Crawler Traffic using HTTP Get Request a Dynamic Web Page". International Journal of Computer Applications (0975 – 8887) Volume 14– No.3, January 2011.

[12] Shekhar mishra, anurag jain and A K Sachhan, "Smart approach to Reduce the Web Crawler Traffic of existing system using HTML based update file at web server". International Journal of Computer Applications 11(7), December 2010 (pp: 34-38).

[13] "Web Crawler", From Wikipedia, http://en.wikipedia.org/wiki/Web_crawler

[14] "World Wide Web", From Wikipedia, http://en.wikipedia.org/wiki/World_Wide_Web

[15] "Hyper Text Transfer Protocol",

http://en.wikipedia.org/wiki/hypertext_Transfer_Protocol