Efficient Crawling Through Dynamic Priority of Web Page in Sitemap

Rahul kumar[1] and Anurag Jain[2]

[1]Department of CSE Radharaman Institute of Technology and Science, Bhopal, M.P, India

ABSTRACT

A web crawler or automatic indexer is used to download updated information from World Wide Web (www) for search engine. It is estimated that current size of Google index is approx $8*10^9$ pages and crawling costs could be around 4 million dollars for a full crawl if only considered network costs. Thus we need to download only most important pages. In order toward, we propose "Efficient crawling through dynamic page priority of web pages in Sitemap" which is query based approach to inform most important pages to web crawler through sitemap protocol in dynamic page priority. Through the page priority web crawler can find most important pages from any website and may just download them. Experimental results reveal our approach has better performance than existing approach.

Keywords:

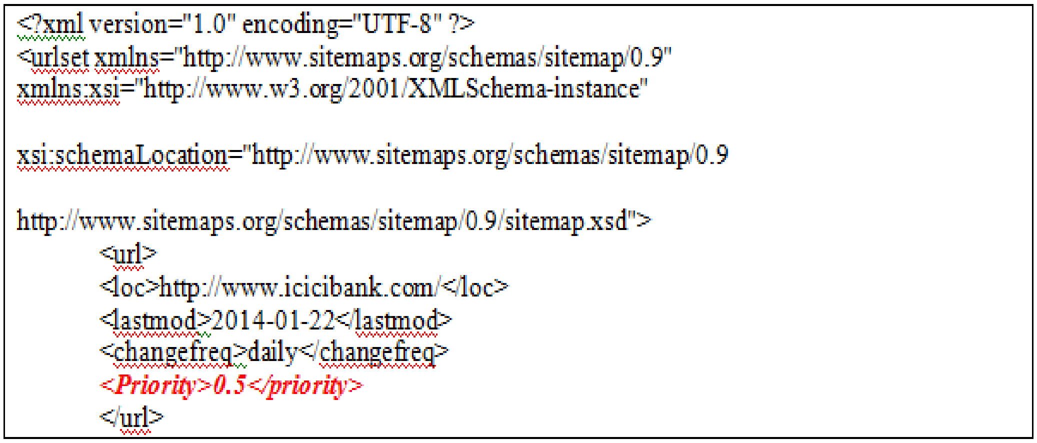
WWW, WebCrawler, Indexer, Search engine, Priority, Sitemap.

1. INTRODUCTION

The web is the main source of retrieving data or information in the universe. A large amount of users use web for retrieving any type information as they want and for this web browser are used for accessing information. A World Wide Web [WWW] is the collection of millions of web pages that can have text audio, video, image and so on [19]. A web server and web browser are used Hyper Text Transfer Protocol [HTTP] to communicate each other [20]. A search engine is used to retrieving web pages which are associated to the string pass by the user in search engine. Search engines typically "crawl" Web pages in advance to build local copies and/or indexes of the pages. This local index is then used later to identify relevant pages and answer users' queries quickly[2].Today's all most every user used search engine for getting information or web pages. An indexer is used for indexing the web pages at web server. One of the main parts of search engine is web crawler. A web crawler is a computer program that browses the WWW in sequencing and automated manner [18]. A crawler which is sometimes referred to spider, bot or agent is software whose purpose it is perform web crawling [14].

This can be used for accessing the Web pages from the web server as per user pass queries commonly for search engine. A web crawler also used sitemap protocol for crawling web pages. Sitemaps file is an XML file that lists a site's URLs, along with additional metadata detailing: when was the page last updated, how frequently does the page change on average and how important it is relative to the other pages in the site [16, 17]. The purpose of Sitemaps is to enable search engines to crawl the site more intelligently [14].

Sitemaps files are XML files with a list of URLs with additional metadata, as shown in the example below



```
<?xml version="1.0" encoding="UTF-8" ?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9

http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
        <url>
        <loc>http://www.icicibank.com/</loc>
        <lastmod>2014-01-22</lastmod>
        <changefreq>daily</changefreq>
        <Priority>0.5</priority>
        </url>
```

In the WWW maximum 40% of web traffic is generated by web crawler & the changes rate of web pages is too high [10].But in web crawling almost 50% of the request are generated by web crawler [10]. Crawling approach is modified and crawler download only updated web pages after last visit. In this crawler only search the updated URL of web pages instead of searching full URL's. There for this will help to decrease the crawling traffic on the web server and the result crawler work fast. But in this priority of URL's are static therefore crawler search the whole URL's based on static priority. This will also generate traffic [13].

The solution to this problem is that the crawler only search the updated URL's based on highest priority and priority of URL's generate dynamically.

2. RELATED WORK

For more efficient crawling research is being carried in different areas such as network level, crawler level, web server and web crawler coordination. In this we represent the surveys of related works and problems identification.

In 2004, Junghoo Cho et al [1], "Impact of Search Engines on Page Popularity" says that top 20% of the web pages with the highest number of incoming links find that 70% of the new links after 7 months, while the bottom 60% of the web pages find virtually no new incoming links during that period .i.e. popular (important) pages are getting more popular while unpopular pages are getting relatively less popular.

In 2005, Ricardo BaezaYates et al [3], "Crawling a Country: Better Strategies than Breadth First for Web Page Ordering" perform experiment in approx 100 million web pages and find that crawling the large sites first scheme has practically most useful then on-line page importance computation . The crawler uses the number of un-crawled pages found so far as the priority for picking a web site, and starts with the sites with the large number of pending pages.

In 2007, C. J. Pilgrim et al [5], "Trends in Sitemap Designs – A Taxonomy and Survey" sitemap is a map, diagram or textual description of the structure or content of a website. By the help of sitemap users understand where they are, where they have been and where they can go, and can guide users to the desired page [4].

O. Jiang et al [6], say that "Full information on the web is not available due to constraint of time, network bandwidth and hardware .So site rank –based strategy has the best performance in discovering high quality pages".

Junghoo Cho et al [7], "Efficient Crawling Through URL Ordering" find that a crawler is to select URLs & to scan from queue of known URLs so as to find more important pages first when it visits earlier URLs that have anchor text which is similar to the driving query or short link distance to a page; that is known to be hot.

In 2009, Uri Schonfeld et al [8], "Sitemaps: Above and Beyond the Crawl of Duty "The sitemaps protocol is a web protocol supported jointly by search engines to help content creators and search engines to unlock this hidden data by making it available to search engines.

In 2009, Eytan Adar et al [9]," The Web Changes Everything: Understanding the Dynamics of Web Content" perform web crawling on 55000 web pages and fine that a large potation of web pages changelings more than hourly.

In 2010, Sun et al [10], "The Ethicality of Web Crawlers" analyzed various log file of different web site. They found that on an average 50% of web request is generated by web crawler. In this paper we can find most valuable web pages so crawler can download these pages for search engine.

In 2011, Shekhar Mishra et al [11], "A Query based Approach to Reduce the Web Crawler Traffic using HTTP Get Request and Dynamic Web Page "authors proposed a query based approach to inform updates on web site by web crawler using Dynamic web page and HTTP GET Request .

In 2012, Dr. Bharat Bhushan et al [12], "Increasing the Efficiency of Crawler Using Customized Sitemap" authors proposed that when a crawler revisiting the websites and find that which web pages have been updated or newly added since last visit, then there is no need to download the complete website every time. With this scheme it will be less time consuming for web crawlers to maintain the freshness of downloaded websites used by search engines.

In 2012, S S Vishwakarma et al [13], "A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency**"** The authors proposed a modify approach for crawling by the use of filter and this is a query based approach. Filter always redirects the updated web pages and crawler downloads all updated web pages after LAST_VISIT.

In 2013, Damien Lefortier Yandex et al [15], "Timely crawling of high-quality ephemeral new content" says a web crawler traditionally fulfills two purposes discovering new pages and refreshing already discovered pages and most ephemeral new pages can be found at a relatively small set of content sources & it is done to periodically re-crawl content sources and crawl newly created pages linked from them, focusing on high-quality (in terms of user interest) content

As per literature survey we will indentified the following problems

- The Rate of changes of web pages is too high.
- Due to changes rate of web pages network traffic is also too high.
- Due to changes rate of web pages high amount of network bandwidth is consumed.
- The crawling cost is high for updated information.

So our motivation is to develop a scheduling policy for downloading web pages from the WWW which guarantees that, even if we do not download all of the known pages, we still download the updated web pages consists of highest priority.

3. PROPOSED APPROACH AND METHODS

As per literature survey the priority of URLs of web pages in a website is static as per sitemap structure. The priority of web pages cannot generate dynamically if the priorities of URLs of web pages generate dynamically than web crawler also helpful to only download the updated URLs of web pages consists of highest priority. Proposed approach is a query based approach. The authors propose the use of dynamic web page to inform the web crawler about the new URLs consists of highest priority and updates on web site. Initially we generate sitemap of any web site.

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

If request is generated by web crawler and its user agent indicates it then the filter directs it to update web pages.

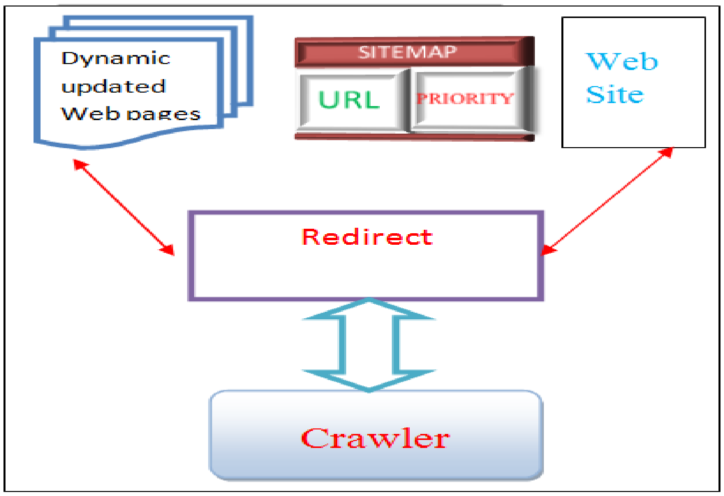


Figure1: proposed approach

Dynamic web page receives the HTTP GET request with parameter LAST_VISIT and searches the list data structure to find the priority based updated URLs of web pages updated after crawler's last visit and also has a component of sitemap.

The URL priority is generated dynamically and number of hit for each URL .The priority of a particular URL is higher which is having highest number of hit. This scheme is efficient for finding the high priority based updated URL in a website.

After searching the list data structure we are having an updated URLs list based on HTTP request with parameter LAST_VISIT and this will sends to the web crawler. The crawler receives the updated URLs list consist of highest priority and download them & shown all above URLs.

Algorithm Used:

- Web crawler sends to HTTP GET request with LAST_VIST parameter to dynamic web page.
- A filter is used to check request.
- Dynamic web page receive HTTP GET request with LAST_VIST parameter.
- Dynamic web page search list Data structure as per received request.
- Dynamic web page sends updated URLs list consists of highest priority to web crawler.
- Web crawler receive updated URLs list and downloads.

4. SIMULATION & RESULT

A general website structure is used to performed different scenarios in simulation & result as shown in figure (2).

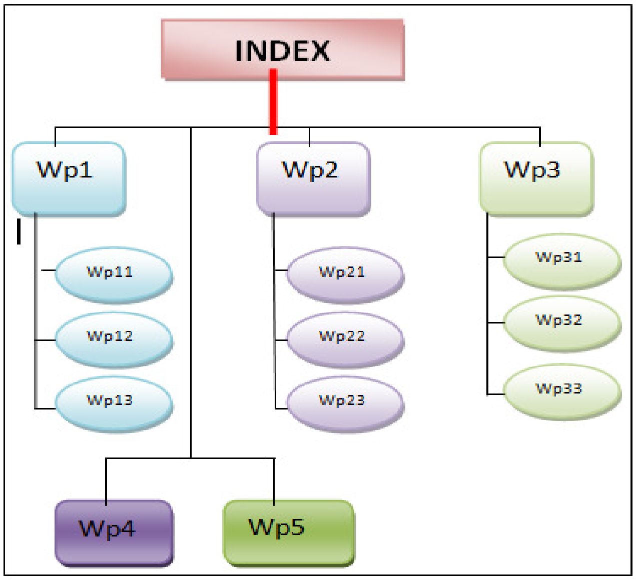


Figure 2: Structure of Website

Experimental Scenarios:

Scenario: I

In this scenario as per performed experiment dynamically generated priorities of web pages are shown in figure3. In this scenario web crawler only downloads no of web pages whose priorities are greater than 0.75

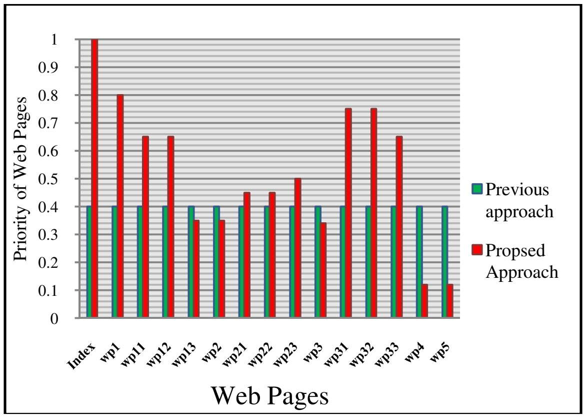


Figure 3: COMPARISON OF WEB PAGES PRIORITY

In scenario-I experiment is performed for four web pages. If we update 4 web pages i.e. Index, Wp1, Wp31, Wp32 and their priorities are generated dynamically i.e. 1.0, 0.8, 0.75, and 0.75. Now as per Normal Web Crawling Approach web crawler downloads all the 15 web pages as per used in experimental website structure, as per Existing Web Crawling Approach web crawler downloads 4 web pages i.e. updated web pages are Index, Wp1, Wp31, Wp32 and as per Proposed Web Crawling Approach web crawler only downloads 2 web pages i.e. updated web pages consist of highest priority, as shown in figure 4.

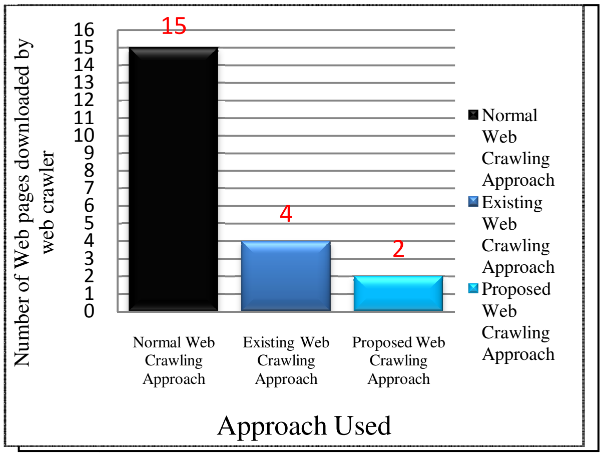


Figure 4: Comparison of Number of Web Pages downloaded by Web Crawler Using Different Approaches.

Scenario: II

In scenario II as per performed experiment dynamically generated priorities of web pages are shown in figure5. In this scenario web crawler only downloads no of web pages whose priorities are greater than 0.80.

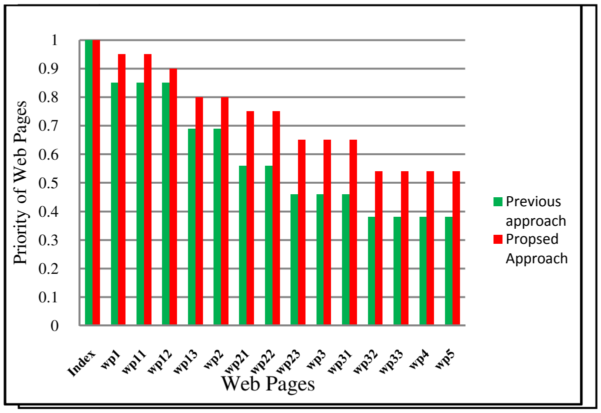


Figure5: COMPARISON OF WEB PAGES PRIORITY

In this scenario experiment is performed for six web pages. If we update six web pages i.e. Index, Wp1, wp11, wp12, wp13, Wp4 and their priorities are 1.0, 0.95, 0.95, 0.90, 0.80 and 0.80. Now as per Normal Web Crawling Approach web crawler downloads all the 15 web pages, as per Existing Web Crawling Approach web crawler download 6 web pages i.e. updated web pages are Index, Wp1, wp11, wp12, wp13, Wp4 and as per Proposed Web Crawling Approach web crawler downloads only 4 web pages i.e. updated web pages consist of highest priority (>0.80) as shown in figure 6.

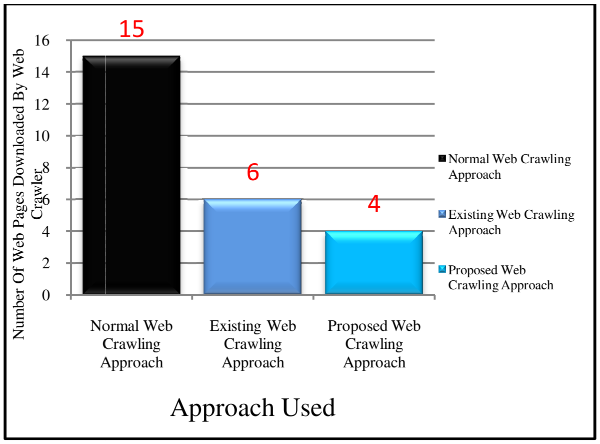


Figure6: Comparison of Number of Web Pages downloaded by Web Crawler Using Different Approaches.

Scenario: III

In scenario III as per performed experiment dynamically generated priorities of web pages are shown in figure7. In this scenario web crawler only downloads no of web pages whose priorities are greater than 0.90.

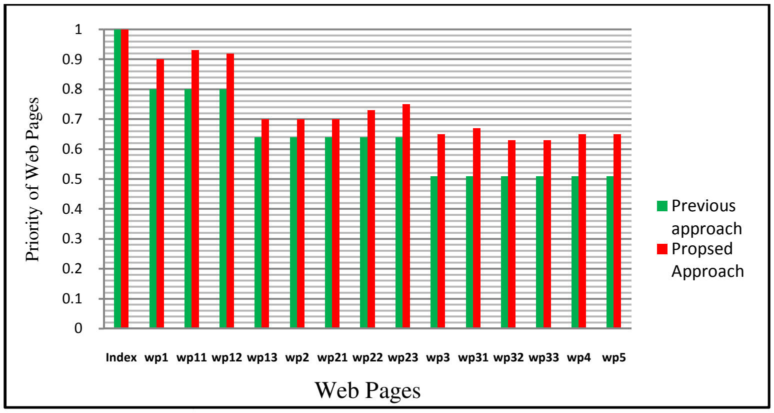


Figure7: Comparison of Web Page Priority

In this scenario experiment is performed for eight web pages. If we update 8 web pages i.e. Index, Wp1, wp11, wp12, Wp2, wp23, Wp4, Wp5 and their priorities are generated dynamically i.e. 1.0, 0.9, 0.93, 0.92, 0.7, 0.75, 0.65, 0.65. Now as per Normal Web Crawling Approach web crawler downloads all the 15 web pages, as per Existing Web Crawling Approach web crawler downloads 8 web pages i.e. updated web pages are. Index, Wp1, wp11, wp12, Wp2, wp23, Wp4, Wp5. And as per Proposed Web Crawling Approach web crawler downloads 3 web pages i.e. updated web pages consist of highest priority(>0.90) as shown in figure (8)

]

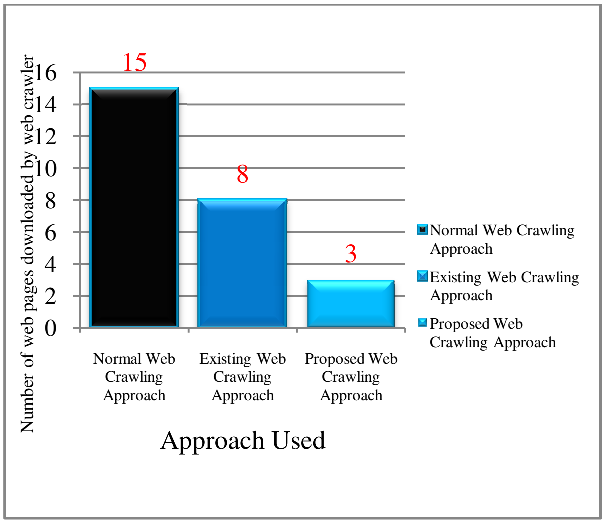


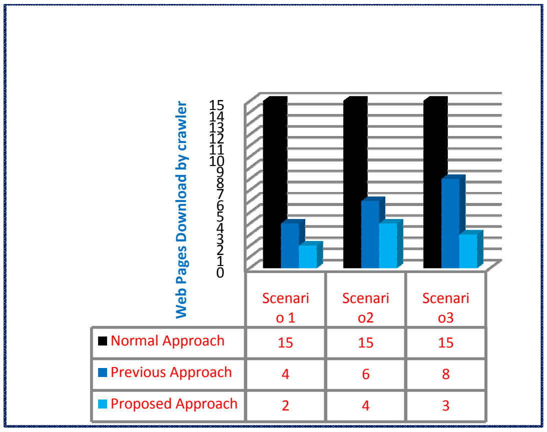
Figure8: Comparison of Number of Web Pages downloaded by Web Crawler Using Different Approaches.

5. COMPRASSION BETWEEN SCENARIOS

The simulation shows comparisons between numbers of web pages downloaded by web crawler using different approaches as shown in table V.I and figure(9).

Scenario	Normal Web Crawling Approach	Existing Web Crawling Approach	Proposed Web Crawling Approach
I	15	4	2
II	15	6	4
III	15	8	3

Table 5.1



	Scenari o 1	Scenari o2	Scenari o3
■ Normal Approach	15	15	15
■ Previous Approach	4	6	8
■ Proposed Approach	2	4	3

Figure 9: Comparison of Number of Web Pages downloaded by Web Crawler Using Different Approaches in all Scenarios.

Experiments show that our proposed approach only downloads two hot updated web pages but previous approach downloads four web pages. Same results are seen in second & third experiment i.e. proposed approach downloads only four web pages but the existing approach downloaded six and eight web pages.

Advantages:

- Web crawler always has new updated information
- Web crawlers download only updated web pages or URLs consists of highest priority.
- Reduce the load on web server or network traffic.
- Using proposed schemes crawling cost is also low.
- Less amount of network bandwidth is consumed

6. CONCLUSION AND FUTURE WORK

Proposed scheme is implemented on existing system with some modification in policy. In this scheme is more effective as per our experiment performed. In this scheme the web crawler only downloads most valuable updated web pages priority based this also helpful for large data base because we can get important valuable pages in very short duration. With the help of this approach we can reduce web crawling and network traffic. This will also help to prioritize new pages with seemingly higher quality found on the same content source at the same time. In feature this scheme is more effective in distributed environment or cloud environment. This scheme is also very useful in web data mining.

REFERENCES

[1] Junghoo Cho, Sourashis Roy" Impact of Search Engines on Page Popularity" WWW2004, May 17–22, 2004, New York, NY USA. ACM xxx.xxx.

[2] Alexandros Ntoulas, Junghoo Cho, 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, New York, USA. ACM 158113844X/ 04/0005.

[3] Ricardo BaezaYates Carlos Castillo Mauricio Marin Andrea Rodriguez," Crawling a Country: Better Strategies than BreadthFirst for Web Page Ordering" International World Wide Web Conference Committee (IW3C2). WWW 2005,, May 10–14, 2005, Chiba, Japan.

[4] Sifer, M. & Liechti, O. 1999, 'Zooming in One Dimension Can be Better Than Two: An Interface for Placing Search Results in Context with a Restricted Sitemap', Proceedings of the IEEE Symposium on Visual Languages: VL'99, Tokyo, Japan, pp. 72-79.

[5] C. J. Pilgrim" Trends in Sitemap Designs – A Taxonomy and Survey" Australian Computer Society, Inc2007.

[6] Qiancheng Jiang, Yan Zhang," SiteRank-Based Crawling Ordering Strategy for Search Engines" State Key Laboratory of Machine Perception Peking University 100871 Beijing, China.

[7] Junghoo Cho , Hector Garcia-Molina ,Lawrence Page ," Efficient Crawling Through URL Ordering" Department of Computer Science Stanford, CA 94305

[8] Uri Schonfeld , Narayanan Shivakumar," Sitemaps: Above and Beyond the Crawl of Duty "International World Wide Web Conference Committee (IW3C2) WWW 2009, April 20–24, 2009, Madrid, Spain. ACM 978-1-60558-487-4/09/04.

[9] Eytan Adar, Jaime Teevan, Susan T. Dumais, Jonathan L. Elsas" The Web Changes Everything:Understanding the Dynamics of Web Content" WSDM'09, February 9-12, 2009, Barcelona, Spain

[10] Yang Sun, Isaac G. Councill, C. Lee Giles," The Ethicality of Web Crawlers" 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

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

[12] Dr. Bharat Bhushan, Meenakshi Gupta, Garima Gupta" INCREASING THE EFFICIENCY OF CRAWLER USING CUSTOMIZED SITEMAP" International Journal of Computing and Business Research (IJCBR) Volume 3 Issue 2 May 2012

[13] S S Vishwakarma, A Jain ,A K Sachan," A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency" International Journal of Computer Applications (0975 – 8887) Volume 46– No.1, May 2012.

[14] Yousse Bassil "A Survey on Information Retrieval, Text Categorization, and Web Crawling" Journal of Computer Science & Research (JCSCR) - ISSN 2227-328X Vol. 1, No. 6, Pages. 1-11, December 2012

[15] Damien Lefortier Yandex, Liudmila Ostroumova Yandex, Egor Samosvat Yandex" Timely crawling of high-quality ephemeral new content" arXiv:1307.6080v2 [cs.IR] 24 Jul 2013
[16] "Sitemaps Generation "From http://xmlsitemapgenerator.org/
[17] "Sitemaps", from Wikipedia,http://en.wikipedia.org/wiki/Sitemaps
[18] "Web crawler", From Wikipedia,http://en.wikipedia.org/wiki/Web_crawler
[19] "World Wide Web", From Wikipedia,http://en.wikipedia.org/wiki/World_Wide_Web
[20] "Hyper Text Transfer Protocol", http://en.wikipedia.org/wiki/hypertext_Transfer_Protocol

Authors

Rahul Kumar has received his Bachelor's Degree in Computer science Engineering from RGPV UniversityBhopal India.