ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Crawling the website deeply: Deep Page crawling

Pooja Tevatia*
Deapartment of Computer Science & Engineering
SET, Sharda University
Greater Noida, Gautam Buddha Nagar (U.P), India
pooja.mtechcse@gmail.com

Vinit Kumar Gunjan
Department of Computer Science & Engineering
SET, Sharda University
Greater Noida, Gautam Buddha Nagar (U.P), India
vinitkumargunjan@gmail.com

Dr (col.) Allam Appa Rao
Vice Chancellor
JNTU University
Kakinada, Andhra Pradesh, India
apparaoallam@gmail.com

Abstract: This paper presents a study of how deep web crawling can be much more efficient than that of a normal crawling. World Wide Web can be divided into two parts: Surface Web and Deep Web. The Surface Web refers to the part of the Web that can be crawled and indexed by general purpose search engines, while the Deep Web refers to the abundant information that is "hidden" behind the query interfaces and not directly accessible to the search engines. Hence, there is need to access the Hidden Web. The normal crawling can retrieve only Surface web pages ignoring the large amounts of high quality information 'hidden' behind search forms that need to be filled manually by the user. The retrieved Hidden web documents are thereof stored in a repository.

Keywords: Deep web crawlers, Search engine crawler, Crawler designing, Google crawler, User centric crawler

I. INTRODUCTION

Search engine is used to access information from World Wide Web. Users provide search queries in the Search engine's interface, In response to which, Search engines fetch the database and produce the result after ranking on the basis of relevance. Search engine builds its database with the help of Web Crawlers, which uses set of algorithms that tracks the Web and collects important information about web documents. In order to do the work significantly & maximize the download rate the Web search engines run multiple crawlers in parallel. As web crawlers play an important role in order to locate URLs, hyperlinks therefore has been an useful tool for gathering timely information from the cyber frontier. Many libraries and databases work like general-purpose Web crawlers, and expose the content only through their own search engines.DP9 is an open source gateway service that allows general search engines, (E.g. Google, Inktomi) to index OAI-compliant archives. DP9 does this by providing consistent URLs for repository records and converting them to OAI queries against the appropriate repository when the URL is requested.

II. DESCRIPTION

A. *SEARCH ENGINES*

The content of the WWW repository is useful to millions of users in millions of Ways. Some simply access it using URL's but it becomes difficult task to access it, if unknown to the URL.Search engine plays an important role in fetching the data according to the query of the user [1].

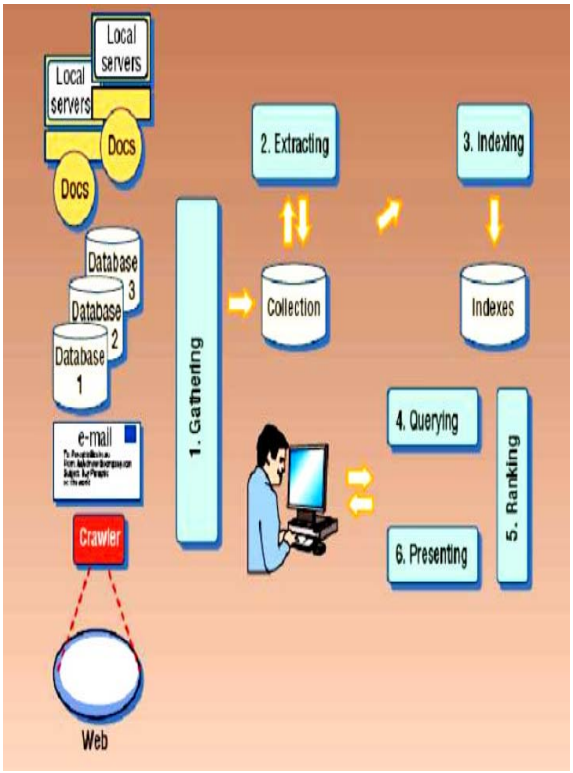

Fig.1: Generic structure of search engine

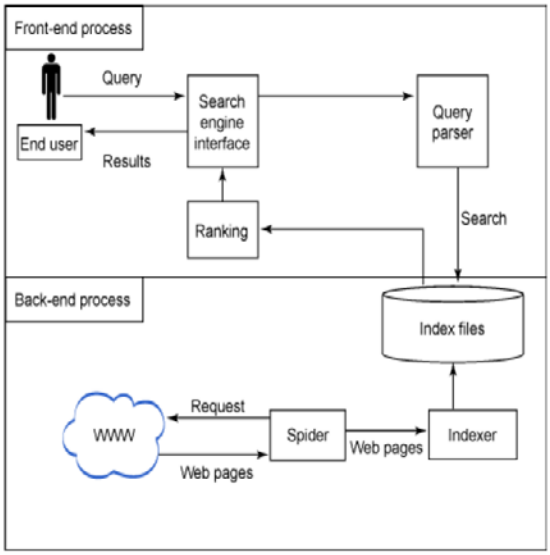

Fig 2: Working steps of a typical search engine

B. COMMON SEARCH ENGINES

1) Google
2) Yahoo
3) MSN
4) AltaVista
5) E-Bay
6) AOL Search
7) Live Search
8) You tube

C. WEB CRAWLERS

These are the automated computer programmes or software agents which are designed to index the results in the web repository and fetch the results from it .e.g., Googlebot is a crawler for Google search engine which helps in indexing of data in web repository. When query is done by the user, the desired results are produced. These crawlers browse the WWW in methodological and automated manner [2].

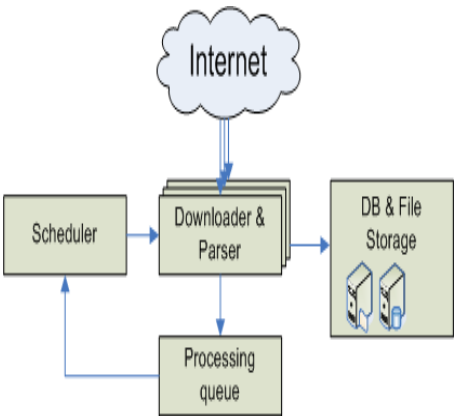

Fig 3: Figure showing general purpose web crawler.

D. WORKING OF BASIC WEB CRAWLER

Starting from the seed page followed by the sublinks associated with it crawler starts working. The process is repeated in frequent interval of time until a higher level objective is reached. Common web crawler implements the following steps [3, 4].

1) Acquiring seed URL of web document from processing queue.
2) Downloading web document.
3) Parsing the content of document to extract URL's associated to other resources and update processing queue.
4) Storing web document for further processing.

E. BASIC STEPS OF WEB CRAWLER WORKING

1) Selection of the seed URL.
2) Addition of seed URL to the frontier
3) Pick the URL from the frontier
4) Fetching of the web-page associated with that URL
5) Parsing the fetched web-page to find new URL links
6) Add found URL's into the frontier
7) Repetition of steps beyond step 2till the frontier gets empty.

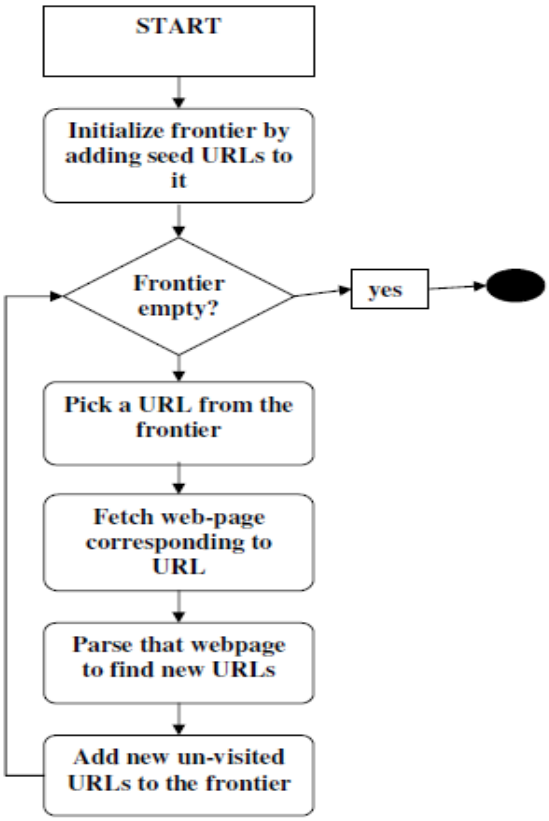

Fig 4: Figure showing sequential flow in a web crawler

F. Deep vs. Surface Web Quality

The issue of quality has been highlighted throughout this study. A quality search result is not a long list of hits, but the right list. Searchers want answers. Providing those answers has always been a problem for the surface Web, and without appropriate technology it will be a problem for the deep Web as well. Effective searches should both identify the relevant information desired and present it in order of potential relevance quality. The searches may be the same for the two sets of requirements, but the answers will have to be different. Meeting those requirements is daunting, and knowing that the deep Web exists only complicates the solution because it often contains

useful information for either kind of search. If useful information is obtainable but excluded from a search, the requirements of either user cannot be met [5].

An intriguing possibility with the deep Web is that individual sites can themselves establish that authority. For example, an archived publication listing from a peer-reviewed journal such as *Nature* or *Science* or user-accepted sources such as the *Wall Street Journal* or *The Economist* carry with them authority based on their editorial and content efforts. The owner of the site vets what content is made available. Professional content suppliers typically have the kinds of database-based sites that make up the deep Web; the static HTML

Pages that typically make up the surface Web are less likely to be from professional content suppliers. By directing queries to deep Web sources, users can choose authoritative sites. By careful selection of searchable sites, users can make their own determinations about quality, even though a solid metric for that value is difficult or impossible to assign universally.[6,12]

G. THE DARK WEB

Whatever technology we use, the waste elements of society come up with the ways to misuse it. There is a dark side of the deep web called as dark web. The notorious elements of the society use the portion of deep web to propagate the objectionable materials.

Freenet, a distributed, decentralized information storage and retrieval system which allows users to surf the internet anonymously. It has come under a lot of flak as it not only allows users to remain anonymous so that their paths are untraceable, but also hides the fact that someone is using Freenet at all. This has given a lot of leeway to cyber criminals to flourish under the guise of anonymity. This guise of anonymity though is good for users in countries where internet is highly censored, where freedom of speech is stifled. But at the same time a large number of web sites on paedophilia, terrorist activities, and virus coding and other cyber-crimes are floating on Freenet. Another case in point is the Russian mafia arm called Russian Business Net-work (RBN) which is synonymous with online cyber-crime. RBN takes advantages of the unused or discarded web addresses that are lying in the Deep Web. They activate these addresses for a couple of minutes, send out millions of spam email and then deactivate the address. This makes them untraceable. They also host web sites whose content relates to child porn, malware, spyware among other such cyber-crimes [7].

Fig 5: Figure differentiating surface web and deep web

H. DEEP WEB CRAWLER

Normal search engines create their index by crawling surface Web pages. To be discovered, the page must be static and linked to other pages. Normal search engines cannot retrieve information in the deep Web as the deep Web is different from the surface Web in a number of ways; Deep Web sources store their information in searchable databases or repository which only produces results dynamically in response to a direct request. The deep Web contains 7,500 terabytes of information compared to 19 terabytes of information in the surface Web. The deep Web contains nearly 550 billion individual documents compared to the one billion of the surface Web. More than 200,000 deep Web sites presently exist; sixty of the largest deep Web sites collectively contain about 750 terabytes of information. On average, deep Web sites receive fifty per cent greater monthly traffic than surface sites and are more highly linked to than surface sites; however, the deep Web site is not well known to the Internet searching in public. Total quality content of the deep Web is 1,000 to 2,000 times greater than that of the surface Web. [8]

I. Deep Web Characteristics

Deep Web information has some significant differences from surface Web or normal web information. Deep Web documents are on average 27% smaller than surface Web documents. Though individual deep Web sites have tremendous diversity in their number of records, ranging from tens or hundreds to hundreds of millions, these sites are on average much larger than surface sites. The mean deep Web site has a Web-expressed (HTML-included basis) database size of 74.4 MB. The deep Web site receives somewhat more than two times the traffic of a surface Web site. Deep Web sites on average are more highly linked to than surface web and are highly popular.

G. Why Deep Web Content cannot be Indexed

The Surface Web use inverted index as a data structure to index the web information and keyword interface to retrieve the information. But the Deep web is more difficult task because the deep web is having the large repository of data where hidden web information is created dynamically by user query.

It is infeasible to issue many hundreds of thousands or millions of direct queries to individual deep Web search databases. It is implausible to repeat this process across tens to hundreds of thousands of deep Web sites. And, of course, because content changes and is dynamic, it is impossible to repeat this task on a reasonable update schedule. For these reasons, the predominant share of the deep Web content will remain below the surface and can only be discovered within the context of a specific information request [9].

J. BASIC CRAWLING TERMINOLOGY

K.

A. *Seed Page:*

By Crawling, we mean to track the web by following the links from a set of URL's or single URL.This URL is the base using which a crawler starts searching procedure. These URL's are referred to as "Seed Page". The selection of a good seed is the most important factor of a crawling process

B. *Frontier Page:*

The crawling process starts with a given URL,Extracting links from it & adding them to the repository of un-visited list of URL's, these unvisited links or URL's are known as "Frontier". It is also called processing queue.

C. *Parser:*

Its job is to parse the fetched webpage to extract list of New URL's from it and returns the new unvisited URL's to the repository known as Frontier [3].

L. DESIGNING ISSUES OF WEB CRAWLER

Web crawler starts off with an initial set of URLs called, *Seed.* Firstly it places *Seed* in a queue, where after retrieving all URLs are kept and prioritized. In some order the crawler gets a URL from this queue and the page associated with it is downloaded. The URLs associated with

the downloaded pages are extracted and the list of URL is placed in the queue [3,7,13].

From this queue, the crawler gets a URL (in some order), downloads the page, extracts any URLs in the downloaded page, and puts the new URLs in the queue. This process is repeated until the crawler decides to stop. Because of the huge and increasing size of the web and its frequent rate of change many issues arise day by day, some of them are following:

1) Revisiting of pages by the crawler
2) How frequent the crawler should revisit the pages as some pages get frequent updating and some pages are not updated for years.
3) How to run parallel crawlers
4) How could be the process be parallelized as there may be indexing of same pages more than once.
5) Minimization of load on the visited pages

When the crawler collects pages from the Web, it consumes resources belonging to other organizations. For example, when the crawler downloads page *p* on site *S*, the site needs to retrieve page *p* from its file system, consuming disk and CPU resource. Also, after this retrieval the page needs to be transferred through the network, 15 which is another resource, shared by multiple organizations. The crawler should minimize its impact on these resources. Otherwise, the administrators of the Website or a particular network may complain and sometimes completely block access by the crawler.

M. COMMON WEB CRAWLERS

1) ASP seek
2) Crawler4j
3) Nutch
4) Pavuk
5) Ubicrawler

N. *OPEN SOURCE WEB CRAWLERS*

1) Datapark Search
2) GNU
3) Heritrix
4) WIRE
5) WebCrawler

O. PAGE IMPORTANCE FACTORS

A. Content Keywords

If the query words are available in a crawled page the page is considered as relevant, the more the query word matches the more page is considered to a relevant.

B. Popularity of link

Number of backlinks to a crawled page is used to see the page relevance.

C. Query similarity

Sometimes the query made by the user is smaller and sometimes larger; Similarity between the crawled page and the query is used as the page relevance score

D. Seed page similarity

The pages associated with the seed URL is used to measure the relevance of the pages being crawled. The seed pages are combined together into a single document and the cosine similarity of this document and a crawled page is used as the page's relevance score [10].

III. METHODOLOGY

A. AlGORITHM'S FOR WEB CRAWLING

The most important part of a web crawler is Scheduler which decides the next URL from the list of un-visited URLs for refreshing the contents. A new URL is selected from the frontier by the crawler, so it is also a major task to select URL from the frontier which entirely depends on the crawling algorithm as of the reason that crawler aims to achieve several goals at the same time and many of them contradict among them. For updating the changes made in a page crawler revisits the page but at the same time it is also desired by the crawler to go through the new pages in order to index them. In order to use the available resources efficiently, it uses the concept of parallelism. Crawler prioritizes its download as it can download limited pages in a limited time. It would also be desirable if the crawler downloads good pages in order to avoid the wastage of resource but without downloading it is a tough task for the crawler to decide whether which page is good or which page is bad [11].

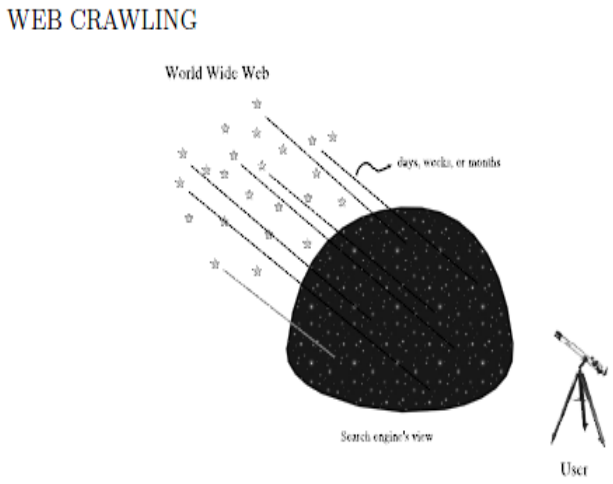

Fig 6 : As the crawling process takes time and the Web is very dynamic, the Search engine's view of the Web represents the state of Web pages at different times. This is similar to watching the sky at night, as the stars we see never existed simultaneously as we see them

B. Steps implemented(GOOGLEBOT ACTIVITY IN 90 DAYS)

After the crawling of the seed page is done, Crawler goes through the sitemap at frequent intervals. It is always advisable to create crawler friendly sitemap. In addition to sitemaps we tried to provide the links of all the pages on the homepage in the form of hidden keywords which is mostly considered as seed by the crawler .We made the hyperlink of all the pages and after taking the observation at frequent intervals we concluded that if all the URL's are provided at the homepage, results in the fastest crawling of the pages. Hyperlinks may also be provided with the word or image contents.

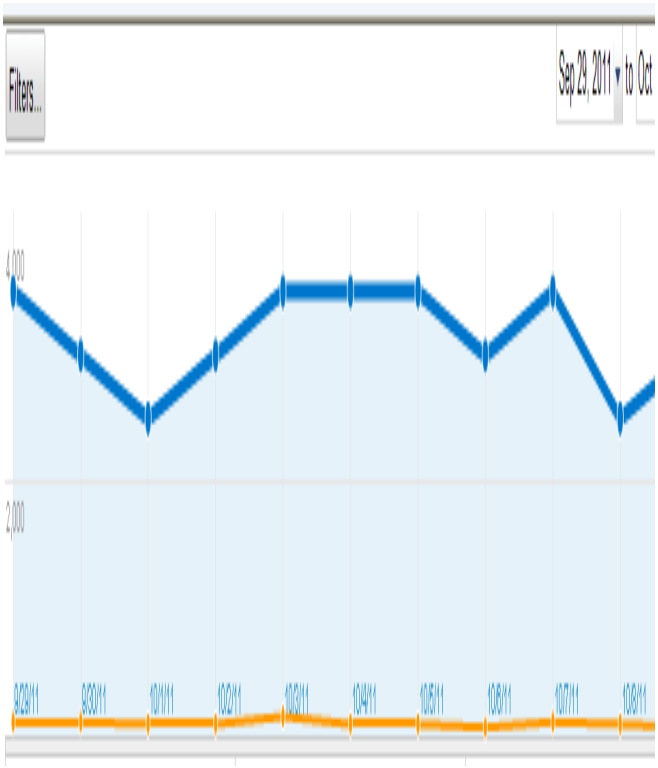

Fig 7. Rate of crawling before hyperlinking the pages at homepage

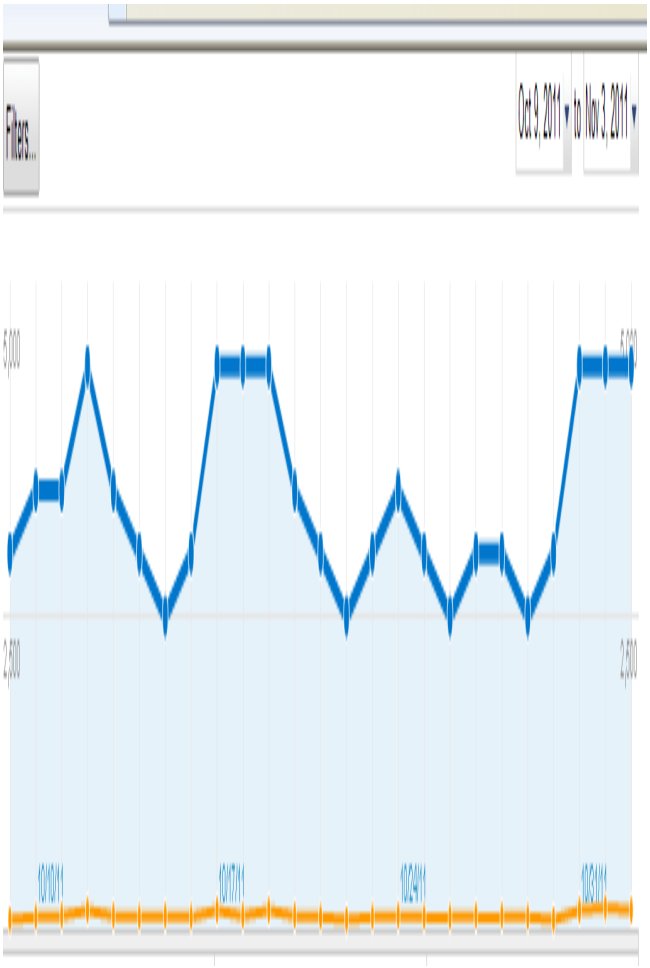

Fig8. Rate of change of crawling after hyperlinking the pages at homepage

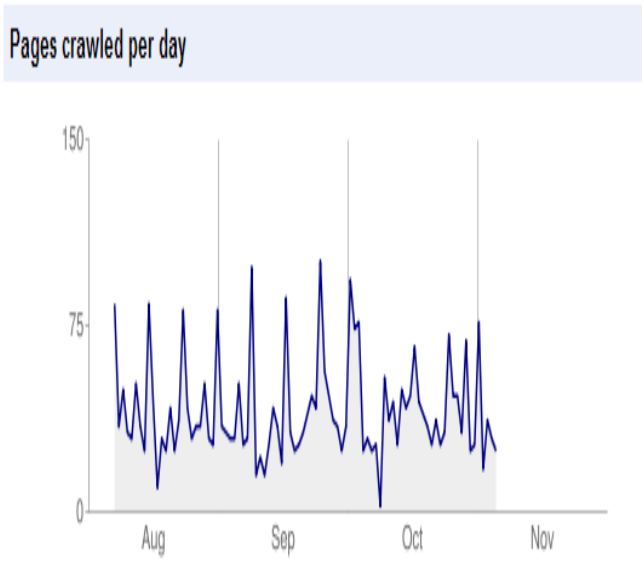

Fig 9. Number of pages crawled by Googlebot /day

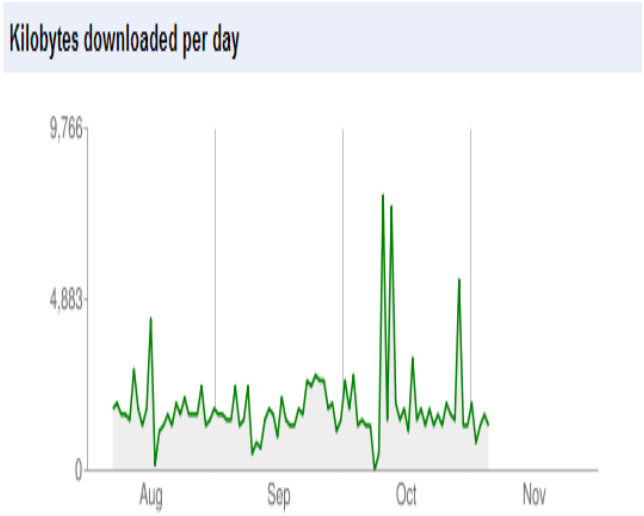

Fig 10. Data downloaded from the website/day

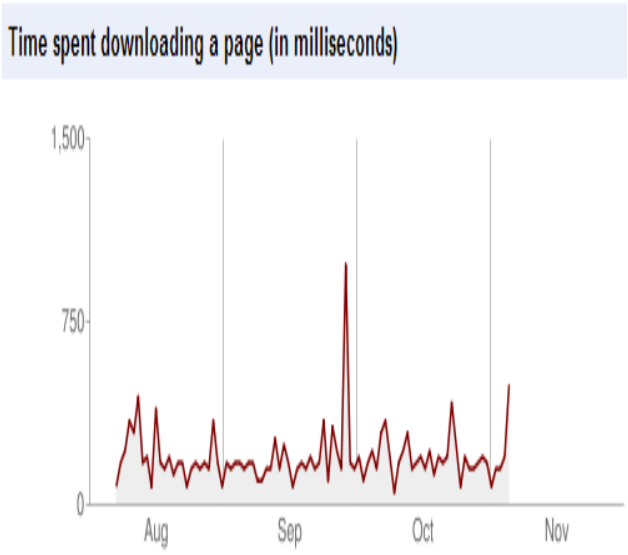

Fig 11. Time spent by an user in downloading a page

Table1.Googlebot activity

	High	Average	Low
Pages Crawled per day	101	42	4
Kilobytes downloaded per day	7958	1885	141
Time spent downloading a page(in milliseconds)	1004	208	50

IV. CONCLUSION AND FUTURE WORK

The importance or quality of deep web information cannot be avoided by serious information seekers, However deep web information are not indexable but they are quite a good repository of total information. This large amount of data stored in the repository of deep web makes it difficult to get indexed.
In future we endeavor to design an architecture which can continuously update and refresh the deep web information and update the repository periodically.

V. ACKNOWLEDGEMENT

We are very thankful to the Research Technology Development Centre team and the course coordinator for their ideas and excellent computing, Research & Development facilities at the University campus. In addition to Sharda University we also pay tribute to Dr IMS lamba for his moral support and the webmaster team of BioAxis DNA Research Centre, India who allowed us access to their website in order to implement our ideas.

VI. REFERENCES

[1] (2012)Berkeley website. [Online]. Available: http://www.lib.berkeley.edu/TeachingLib/Guides/.../SrchEngCriteria.pdf
[2] S.S. Dhenakaran and K. Thirugnana Sambanthan "WEB CRAWLER - AN OVERVIEW"*IJCSC*, Vol. 2, No. 1, January-June 2011, pp. 265-267.
[3] Sandeep Sharma, "Web-Crawling Approaches in Search Engines", M. Eng. thesis, Thapar University, Patiala, India, June. 2008.
[4] Avanish Kumar Singh,Amit Kumar Pathak,"Novel Architecture of Web Crawler for URL Distribution" IJCSt Vol. 2, ISSue 3, September 2011
[5] Umara Noor , Zahid Rashid , Azhar Rauf 'A Survey of Automatic Deep Web Classification Techniques', International Journal of Computer Applications (0975 – 8887) Volume 19– No.6, April 2011
[6] (2012) Scribd website. [Online]. Available: http://www.scribd.com/doc/37664189/IJCSS-V2-I4
[7] (2012) Techquark website. [Online]. Available:http://www.techquark.com/2011/01/crawling-deep-web.html
[8] Bergman, Michael K., "White Paper: The Deep Web: Surfacing Hidden Value", Volume 7, Issue 1, August, 2001Example of a One-Column figure caption.
[9] Anuradha, A K Sharma, "Exploring the hidden web: A Review", Vol.2, No.1, Jaunary-June2011, pp.257-258
[10] (2012) Soft solutions India website [Online]. Available: http://www.softsolutionsindia.net/blog/tags/23/Web-Crawler/SEO Web-crawler.aspx
[11] Ignacio Garc´ıa Dorado, "Focused Crawling: algorithm survey and new approaches with a manual analysis", Master's thesis, Lund University, Lund, Sweden, April 7. 2008.
[12] Brian Pinkerton. "Finding What People Want: Experiences with the WebCrawler". In Proceedings of the Second International World Wide Web Conference, 1994.
[13] J S. Chakrabarti, et al. "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text". WWW Conference, 1998.

VII. BIOGRAPHY

Pooja Tevatia has completed her B.Tech degree in IT from Chaudhary charan singh university in 2009.She is pursuing MTech in computer science from Sharda university ,Greater Noida, Uttar Pradesh, India. Her research interest includes Web technology & Network security**.**

Vinit kumar gunjan has completed his B.Tech degree in computer science in 2009.He is pursuing MTech in computer science from Sharda university, Greater Noida, UttarPradesh, India. His research interest includes Biological (DNA) database development, Network & internet security & cloud Computing.

Dr Allam Appa Rao, Vice-Chancellor of the JNTU Kakinada, is an iconic and towering personality in the field of education and research. His contributions to the field of Computer Engineering have been exemplary and spilled over into numerous other areas of science and technology, making him a pioneer of scientific advancements meant for the benefit of Society. Dr Allam Appa Rao began his career in 1969 and went on to complete his Ph. D in Computer engineering from Andhra University in the year 1984. This is the first Ph. D in Computer Enengineering from Andhra University.