International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347- 5552, Volume -1, Issue-1, September-2013

Deep Web Crawler: A Review

Smita Agrawal, Kriti Agrawal

Abstract- **In today's scenario, there is an ample amount of data on the internet that can be accessed by everyone. This is the data that can be indexed by search engines. There are softwares named Web Crawlers that explore the WWW in an efficient manner. But there is also a large amount of data that is still out of reach from the access of the conventional search engines. This is known as Deep Web or Invisible Web. Web pages that are hidden created dynamically as a result of queries send to particular web databases. For traditional web crawlers, it is almost impossible to access the content of deep web due to its structure. To retrieve the contents of deep web is a challenge in itself. This paper discusses the methods and tools of crawling the web that is hidden beneath the surface.**

Index Terms- **Deep web, surface web, search engines, crawling, information retrieval**

I. INTRODUCTION

In present scenario, WWW is very important and most common place for the users of the Internet as now days it has become the intrinsic part of the most society persons. Web is searched for information retrieval by millions of people in their daily routine and no doubtly, web has become the fast information retrieval system for web searchers, by using some fast retrieval techniques and softwares. Among them, web crawler is the essential and important software.

Web crawler is the software that explores the World Wide Web in an efficient, organized and methodical manner [1]. Its main purpose is to index the contents of the website using various methods and techniques. It finds and downloads the web pages that are relevant to the user search in very limited period of time. A little part of the website server should be used by the appropriate web crawler. A reasonable web crawler should fetch few pages in very less time.

Web crawler may also be known as web robot, web spider, automatic indexer, internet bots, www robots or software agents. [1]

Web crawler searches the web document or page, collects all information of that page, save them in indices for later processing by the search engine that can perform indexing of downloaded pages to search in a very fast manner.

Smita Agrawal, Assistant Professor, Department of Information Technology, International Institute For Special Education, Lucknow, INDIA (sony.smita2003@gmail.com).
Kriti Agrawal, Former Assistant Professor, Department of Computer Science, INDIA (kritiagrawal2010@gmail.com)

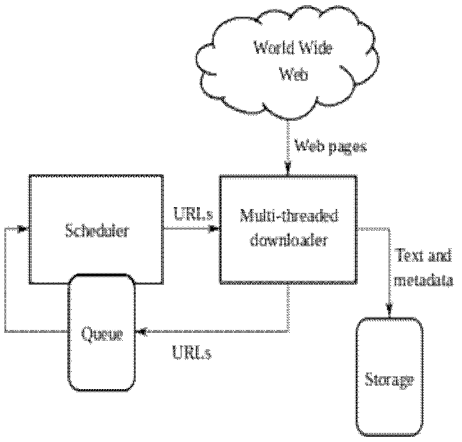

Fig. 1: Architecture of a Standard Web Crawler

The page or set of pages that is to be visited by web crawler are called as *seed URL* [13]. Crawler visits all those pages which start from the first page and store all links associated with that page in the list of visited links, known as frontier. These new pages are visited and their links are also saved. Crawlers perform this process until it finds all the pages related to that particular search [13].

In short, Web crawler starts from first page, visits several web pages and prepares a list of websites by using the links between them within a given time. The whole process of finding and downloading web pages for indexing them is called as web crawling.

There are many general purpose crawlers, some of them are listed below- [1]:

- ***Yahoo Slurp*** is the Yahoo's! Search crawler.
- ***Googlebot*** is the Google's crawler.
- ***Bingbot*** is the Microsoft's bing web crawler**.**
- ***RBSE*** was the first published web crawler**.**
- ***WebFountain*** is a distributed modular crawler.
- ***Scooter*** is the AltaVista Web crawler.

II. DEEP WEB V/S SURFACE WEB

Web searchers don't get much relevant information as traditional search engines cannot find such specific and depth content which is demanded by the user. Traditional search engines only searches static pages and those pages which are linked to one another .Even with the help of web crawler ,they cannot find dynamic pages which are generated dynamically or generated only by user specific query and closed or get invisible after the user query is stopped. But with the help of deep web crawling technique, most search engines can now search in depth

or even those dynamic pages which were invisible while searching by using web crawling techniques.

Deep web is also known as 'invisible web' as these pages are out of reach from traditional search engines. It is also called as hidden web, deepnet or undernet [1]. In 1994, Dr. Jill Ellsworth "first coined the phrase "invisible Web" to refer to information content that was "invisible" to conventional search engines [3]. In 2001, Bergman coined it as "Deep Web".

In brief, deep web, part of Internet, can be defined as the information concealed behind Hyper Text Markup Language (HTML).

It is also believed that deep web is a vast source of methodized content on the World Wide Web and to access contents of deep web has been a challenge in the web community.

Surface web is also called as open web, visible web or indexable net which can be accessed by the traditional search engines. Search engines access those web pages and websites which are linked to one another and are generally static and accessible to all web users. Approximately 80% of the information on the Web belongs to the 'invisible web' [5]. The concept used behind surface web is to perform indexing of web pages and store them for future processing also.

Fig.2: Surface v/s Deep web[2]

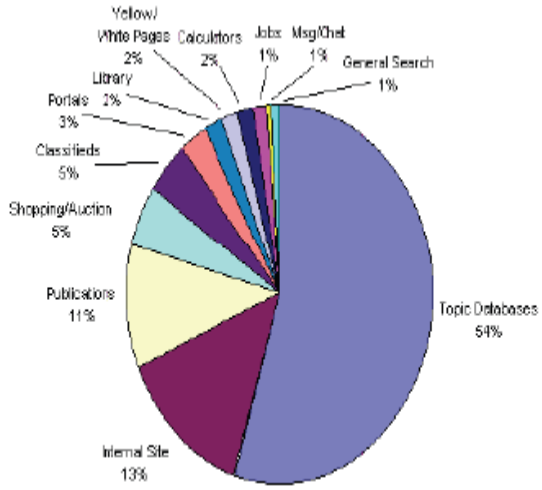

Fig. 3: Distribution of Deep websites based on the content[12]

Table 1: Bergman (2001) contrasts these two parts of the Web: [4]

Surface Web	**Deep Web**
Millions of web pages	Over 200,000 databases
1 billion documents	550 billion documents
19 terabytes	7,750 terabytes
Broad shallow coverage	Deep vertical coverage
Results contain ads	Results contain no ads
Content unevaluated	Content evaluated by experts

III. EXISTENCE OF DEEP WEB: REASON

Web that contains Dynamic contents, unlinked pages, private web, contextual web, etc. is basically termed as Deep Web.

Conventional search engines typically do not index the pages that have following features and all these form Deep Web:

a. Pages which have dynamic content.
b. Websites that require registration and login.
c. Content that is unlinked (not linked to by other web pages).
d. Pages that have scripted content and can only be accessed by links produced by scripting languages.
e. Content that has limited access.
f. Pages that have Non-HTML content or have text content in the form of multimedia content.
h. Pages not hosted on http or https

Fig. 3: The Deep Web [13]

IV. DEEP WEB CRAWLING: METHODS

Traditional Web crawlers can efficiently crawl the surface or visible web but they cannot crawl the deep or hidden web so efficiently. Deep web crawling means accessing the web pages that are not indexed. Deep crawling is the process of efficiently crawling individual websites for very specific structured and unstructured content.

Crawling deep web can be seen as the process of collecting data from search interfaces by generating queries. One of the major challenges in crawling deep web is the selection of the queries so that most of the data can be retrieved at a low cost.

The concept of crawling the deep web is closely related to the federated search [6] is a technology that allows the simultaneous search of multiple searchable resources. A user makes a single query request which is distributed to the search engines participating in the federation. The federated search then aggregates the results that are received from the search engines for presentation to the user [1].

This is the process of combining search results from different sources in some helpful way.

Deep web crawling considers structured as well as unstructured query interfaces. Deep web crawling can be understood by following steps:

a) Locating form interfaces that lead to the content related to deep web.
b) Selecting the resources that are relevant for deep web crawling.
c) Extracting the content lying beneath the form interface of selected resource content.

Strategic approaches may be taken to target deep web content. With a technique called screen scraping, specialized software may be customized to automatically and repeatedly query a given Web form with the intention of aggregating the resulting data. [1]

V. DEEP WEB CRAWLERS: TOOLS

There are many Deep Web search tools that are designed specifically and can be used to retrieve information from the great treasure of deep web. “The first commercial Deep Web tool (although they referred to it as the “Invisible Web”) was @1, announced December 12th, 1996 in partnership with large content providers. According to a December 12th, 1996 press release, @1 started with 5.7 terabytes of content which was estimated to be 30 times the size of the nascent World Wide Web. (“America Online to Place AT1 from PLS in Internet Search Area: New AT1 Service Allows AOL Members to Search “The Invisible Web”).

DWIM (Deep Web Intelligent Miner) [7] is open source software with respect to Deep Web crawling which can be employed as an experimental platform for researches and a crawler engine for developers. DWIM is a task-specific, configurable, open source platform for Deep Web crawling. Task-specific means DWIM can retrieve the Deep Web content whose query from is provided; Configurable means that DWIM allows users to configure his crawler. DWIM implements alternative methods for the same function.

Mozenda (http://www.mozenda.com/) is a Software as a Service (SaaS) company that enables users of all types to easily and affordably extract and manage web data. With Mozenda, users can set up agents that routinely extract data, store data, and publish data to multiple destinations. Once information is in the Mozenda systems users can format, repurpose, and mashup the data to be used in other online/offline applications or as intelligence.

Artirix (http://www.artirix.com)Artirix is a Search Application Platform which is run as a cloud based software service for clients who view search as a core component of their business, application or portal.

There are several algorithms and crawlers that have been proposed over the time for crawling the deep web such as Deep iCrawl: An Intelligent Vision Based Deep Web Cralwer [8], HiWE [9] etc.

Raghavan and H. Garcia-Molina proposed architecture of the deep web crawler in the paper Crawling the hidden web (2001) [9]. In this paper, they gave a prototype for a hidden web crawler named HiWE. This is based on the concept of extraction of task specific information. The main advantage of this strategy is that it minimizes the extraction of relevant information. The limitations of HiWE include its inability to support partially filled forms and to identify dependence between certain elements of form.

L. Barbosa and J. Freire proposed a crawling strategy to crawl the deep i.e. hidden web in the paper Searching for hidden web databases (2005) [18]. In this paper, they proposed a focused crawler based on the concept of automatic form filling. This technique has the main advantage of saving resources and time.

R. Anita et al. proposed an intelligent deep web crawler Deep iCrwal, in the paper Deep iCrawl: An Intelligent Vision Based Deep Web Cralwer (2011) [8], that is purely vision based. In this, they have tried to use the visual similarities of the data elements. There are three phases in

Deep iCrawl – creation of integrated interface, query analysis and raction of information from hidden web pages. [8]

Dilip Sharma and A.K.Sharma proposed a architecture for deep web crawler in their paper A Novel Architecture for Deep Web Crawler (2011) [20]. To minimize limitations of existing deep Web crawlers, a novel architecture was proposed based on QIIIEP specification. The proposed architecture is cost effective and has features of privatized search and general search for deep Web data hidden behind html forms. [20]

Namrata and Dr. Subhash also proposed the architecture of a deep web crawler, in their paper Deep Web Crawl for Deep Web Extraction (2013) [19] for extracting information from deep web, which is based on the concept of multiple HTTP connections to WWW. This system is designed to be used on the client side [19].

Some search engines are specially designed to search the deep or hidden web. Following are the name of some search engines that can be used to search the DeepWeb:

A. Clusty is a meta search engine, meaning it combines results from a variety of different sources, filtering out duplicates and giving the best content. [10]

B. SurfWax gives the option to grab results from multiple search engines at the same time. [10]

C. Scirus is a science search engine dedicated to only searching science-specific content. At the time of this writing, Scirus searches over 370 million science-specific web pages, including scientific journals, scientists' dedicated homepages, courseware, pre-print server material, patents, and much more. [10]

D. Humbul (http://www.humbul.ac.uk/) [11] Covers languages and literatures, philosophy, history, religion, archaeology, American studies and other related areas. It is maintained by the University of Oxford.

E. DeepPeep aims to enter the Invisible Web through forms that query databases and web services for information. Typed queries open up dynamic but short lived results which cannot be indexed by normal search engines. By indexing databases, DeepPeep hopes to track 45,000 forms across 7 domains. The domains covered by DeepPeep (Beta) are Auto, Airfare, Biology, Book, Hotel, Job, and Rental. Being a beta service, there are occasional glitches as some results don't load in the browser. [15].

F. TechXtra (http://www.techxtra.ac.uk/) concentrates on engineering, mathematics and computing. It gives industry news, job announcements, technical reports, technical data, full text eprints, teaching and learning resources along with articles and relevant website information. [15].

G. Infomine has been built by a pool of libraries in the United States. Some of them are University of California, Wake Forest University, California State University, and the University of Detroit. Infomine "˜mines' information from databases, electronic journals, electronic books, bulletin boards, mailing lists, online library card catalogs, articles, directories of researchers, and many other resources. [15].

H. IncyWincy (http://www.incywincy.com/) is an Invisible Web search engine and it behaves as a meta-search engine by tapping into other search engines and filtering the results. It searches the web, directory, forms, and images. With a free registration, search results can be tracked with alerts. [15].

I. Nationsoline (http://www.nationsonline.org/oneworld/) [11] A wealth of up-to-date information on individual countries, cultural, political, social and economic information.

There are many more websites that are used to retrieve relevant information from the websites that are not shown in the search of a conventional search engines.

VI. CONCLUSION

Deep web has plentiful information contained in it. It is a repository of very useful contents that are important for researchers at many levels. To use these resources, there is need of an efficient method to get the relevant and desired content which is lying beneath the surface web i.e. deep web. Although some very useful algorithms and softwares are designed to explore the hidden web, yet there is much scope of finding new methods of crawling the so called deep web that can be cost and time effective.

REFERENCES

[1] http://en.wikipedia.org/wiki/
[2] http://www.brandpowder.com/how-deep-is-your-web/
[3] http://papergirls.wordpress.com/2008/10/07/timeline-deep-web/
[4] Bergman, Michael. "The Deep Web: Surfacing Hidden Value." Journal of Electronic Publishing. 7:1. 2001
[5] List-Handley, C. J. Information literacy and technology. 4th ed. Dubuque, Iowa: Kendall/Hunt, p. 36, 2008
[6] J. Callan, "Distributed information retrieval," in Advances in Information Retrieval, (W. B. Croft, ed.), pp. 127–150, Kluwer Academic Publishers, 2000
[7] http://code.google.com/p/deep-web-intelligent-miner/
[8] R.Anita, V.Ganga Bharani, N.Nityanandam, Pradeep Kumar Sahoo, Deep iCrawl: An Intelligent Vision Based Deep Web Cralwer, World Academy of Science, Engineering and Technology, Volume 50, 2011
[9] Raghavan and H. Garcia-Molina. Crawling the Hidden Web. In Proc. of VLDB, pages 129–138, 2001
[10]http://websearch.about.com/od/invisibleweb/tp/deep-web-search-engines.htm
[11] http://people.hws.edu/hunter/deepwebgate03.htm
[12] http://www.brightplanet.com/
[13] Margaret H. Dunham, Data mining –Introductory and Advanced Topics
[14] http://www.emeraldinsight.com/
[15] http://www.makeuseof.com/
[16] http://techdeepweb.com/
[17] http://www.completeplanet.com
[18]L. Barbosa and J. Freire., Searching for Hidden-Web Databases. In Proceedings of WebDB, pages 1--6, 2005
[19] Namrata Bhalerao, Dr. Subhash K. Shinde, Deep Web Crawl For Deep Web Extraction, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 3, March – 2013
[20] Dilip Kumar Sharma, A.K. Sharma, A Novel Architecture for Deep Web Crawler, International Journal of Information Technology and Web Engineering, Volume 6 Issue 1, January 2011, Pages 25-48