Timestamp based Recrawling Technique (TSBCT)

Babita Ahuja
Assistant Professor
MRCE, College
Faridabad

Neelu Chaudhary
Assistant Professor
MRCE, College
Faridabad

ABSTRACT
In this era of digital tsunami of information on the web, everyone is completely dependent on the WWW for information retrieval. Most of the information is hidden behind the query interface. In the query interface the user types the keyword to access the web pages. These pages are known as the Hidden web, Invisible Web or Dark Web. Such kind of web pages cannot be indexed by the Search Engines. As these are not indexed by the search engines these pages cannot be returned and displayed to the users. This paper discusses the various reasons due of which they are not indexed by the search engines and the possible solutions for these reasons.

General Terms
Web Crawler, Indexing, Hidden Web.

Keywords
Hidden Web, Search Engine, Surface WebQuery Interface.

1. INTRODUCTION
Search Engines uses the crawlers that crawl the World Wide Web and they collect Web pages by starting with a Uniform Resource Locator, URL, downloading the Web page at that location, and recursively retrieving all the pages pointed to by the hyperlinks on the page. There are thousands of search engines on the Internet, each with their own crawler, continuously collecting pages. The part of the web that is in the reach of our search engine is called the surface web or the visible web. The part of the Web that is out of reach of our search engines is called the Hidden Web. The surface web contributes 1% only as the part of the total web space whereas the hidden web makes a share of 99% of the total web space as shown in Figure 1. The content provided by many Hidden Web sites is often of very high quality and can be extremely valuable to the users.

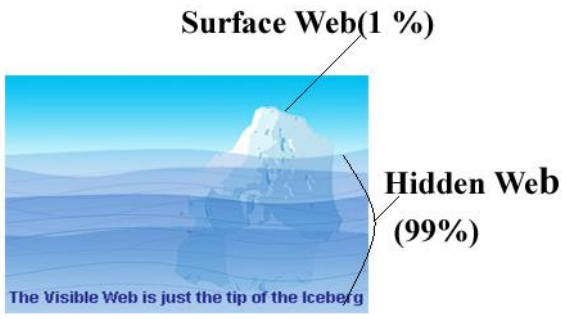

Figure1. The hidden web and the surface web

2. TYPES OF INVISIBILITY IN WWW
There are certain kinds of material on the web which cannot be accessed via the internet as they are not indexed by the search engines. WWW has four types of hidden web or Invisible content as shown in Figure 2. The four types of invisibility are:

- The Opaque Web
- The Secret Web
- The Proprietary Web
- The Proper Invisible Web

2.1 The Opaque Web
This web forms that part of the web that can be indexed by our web crawlers but they are not indexed

Web Crawler is not able to index these pages because of the following reasons:

2.1.1 Depth of Crawl:
It costs more to the search engines if they increase their depth of crawling and crawl each and every page of all websites. If they index every web page then the indexing cost also increases. Limiting the depth of crawl also reduced the cost of indexing a particular Web site.

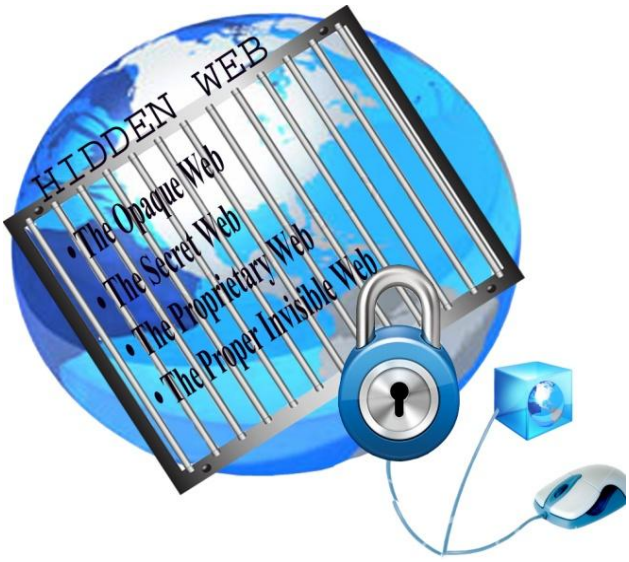

Figure2. Types of Hidden Web

2.1.2 Disconnected URL's:

Some URL's are disconnected as they don't have links pointing to them from other Web pages.

2.1.1 Frequency of Crawling:

Most search Engines Crawl the web at a lesser rate (after a month or more). As the web is very dynamic, so new web pages are added and removed at very frequent rate. These pages make a major the part of invisible web.

2.2 The Secret Web

These web pages can be technically indexed but they are not indexed by the search engines as the Web Administrator doesn't wish this to happen.

Web Crawler is not able to index these pages because of the following reasons:

2.2.1 Robots Exclusion Protocol:

The web Administrator can prohibit a web page to be indexed by search engine by placing the links of those pages in a robot.txt file. The robots.txt file gives instructions about their site to web robots; this is called The Robots Exclusion Protocol. If a crawler wants to visit a URL, http://www.mywebsite.com/welcome.html then before it does so, it firsts checks for robot.txt file shown in Figure 3.

The "User-agent: *" means this section applies to all robots. The "Disallow: /" tells the robot that it should not visit any pages on the site.

2.2.2 Meta Tag- Noindex:

The Web Administrator uses the "noindex" Meta tag. When the crawler reads this Meta tag in the head section then they cannot read further and index the body as shown in Figure 4.

2.2.3 Authentication:

These are Password protected web pages.

2.3 The Proprietary Web

This web has a proprietor and it is controlled as a property. If users desire to view such pages then they must register first and only then they can view such kind of pages. The registration procedure can be paid or it can be free of cost. Web Crawler is not able to index these pages because of the following reasons:

Figure3. The Robot.txt file

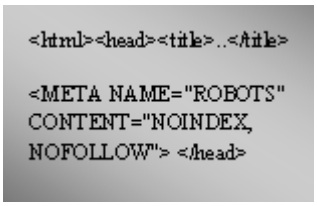

Figure4. A Web page using the "Noindex Tag"

2.3.1 Registration Procedure:

The web Crawler cannot index such kind of web pages as the procedure of registration cannot be accomplished by the crawlers.

2.4 The Proper Invisible Web

These web pages are not indexed by search engines because of the following reasons:

Web Crawler is not able to index these pages because of the following reasons:

2.4.1 Format of Files:

There are some file formats which cannot be handled by today's search engines. Ex:

- Audio Files
- Video Files
- PDF Files
- Flash Files
- Executable files

2.4.2 Frequently changing web Pages:

There are some dynamically generated Web pages. These pages are generated by some server side scripts on the fly. So these pages can be indexed but they are not by the Search Engines. Such kinds of pages are changed within minutes, so if they are indexed then they will provide the stale information. So they are not indexed. Ex: Share market values change very frequently so stale information will be very harmful there so they are not indexed.

2.4.3 Pages behind Query Interface:

Certain kind of information can only be accessed by the help of the query interfaces. When user fills the query interface and submits the form information from the databases is extracted and displayed to the users. Crawlers aren't programmed to understand either the database structure or the command language used to extract information

3. RELATED WORK

In the recent years, the importance of hidden web has become very renowned and considered as a very important part of WWW. The studies are conducted that are extending current-day crawlers to build repositories that include pages from the "hidden Web", the portion of the Web behind searchable HTML forms [1]. To determine the freshness of the web pages is another problem. The work has been conducted that gives a new technique to continuously update/refresh the Hidden Web repository [2]. The query-based database crawling has been modeled to fetch the data from the hidden web [3]. Jian Qiu [4] also presents new index structures for querying the hidden web. But this study again considered only single attribute. Here, clustering of data is done to compress the index. But this technique is inefficient for storing the multi-attribute based hidden web data. HiddenSeek [5] uses a keyword based indexing and searching technique for single-attribute hidden web sites. This approach uses the inverted index for indexing and searching method the hidden web data. HiddenSeek takes a term frequency of keyword as a factor for ranking the results i.e, whether the keyword appears in the URL of a page. The MetaQuerier [6] was designed by Chang, 2005. The goal of MetaQuerier is twofold– First, to make the deep Web systematically accessible, it will help users find online databases useful for their queries. Second, to make the deep Web uniformly usable, it will help users query online

databases. This system focuses on the query interface processing and the processing of the query results is not involved in detail.

4. PROPOSED WORK

There are some dynamically generated Web pages. These pages are generated by any server side programming like servlet, asp, jsp, php etc. on the fly. So these pages can be indexed but they are not indexed by the Search Engines. If they are indexed then they will provide the stale information. So they are not indexed. The problem with the current crawlers is that they recrawl the web after a very long span of time. During this period the already indexed web pages can be updated on the website or new web pages can be added

We propose a multithreaded technique to overcome this problem. This technique has two processes one the updated traditional crawler (UTC) and the Timestamp based recrawl technique (TSBCT) running at the same time. The UTC will index all the pages of a website in a traditional manner except that it will store the time of modification of the web pages in the index table too as shown in Figure 5. At the same time TSBCT, will start recrawling backward and will refetch the updated or fresh web pages at a regular schedule as shown in Figure 6. The quickly updated pages are the part of hidden web. The proposed solution of Timestamp Based Recrawling Technique (TSBCT) will crawl very frequently and will give the users the fresh information and the will bring light on this part of the Dark Web. At the same time the UTC module will crawl the rest of the web.

Algorithm Steps:

1. UTC will be provided with the seed URL. All the URL's in the seed link will be indexed and stored in the search engine repository. The indexing will be done as shown in Table 1.

2. The URL's indexed in step 1 are then checked and the links present in them are also retrieved, indexed and stored in search engine repository. TSBCT will also start.

Table 1. The structure of the index table

URL's	Keywords	Last Date of modification
http://snapdeal.com	Selling, purchase,Online store	28 May 2012

2.1 TSBCT will start revisiting the links from the index table while indexing the new links. TSBCT will check the freshness of the page. By the last time of

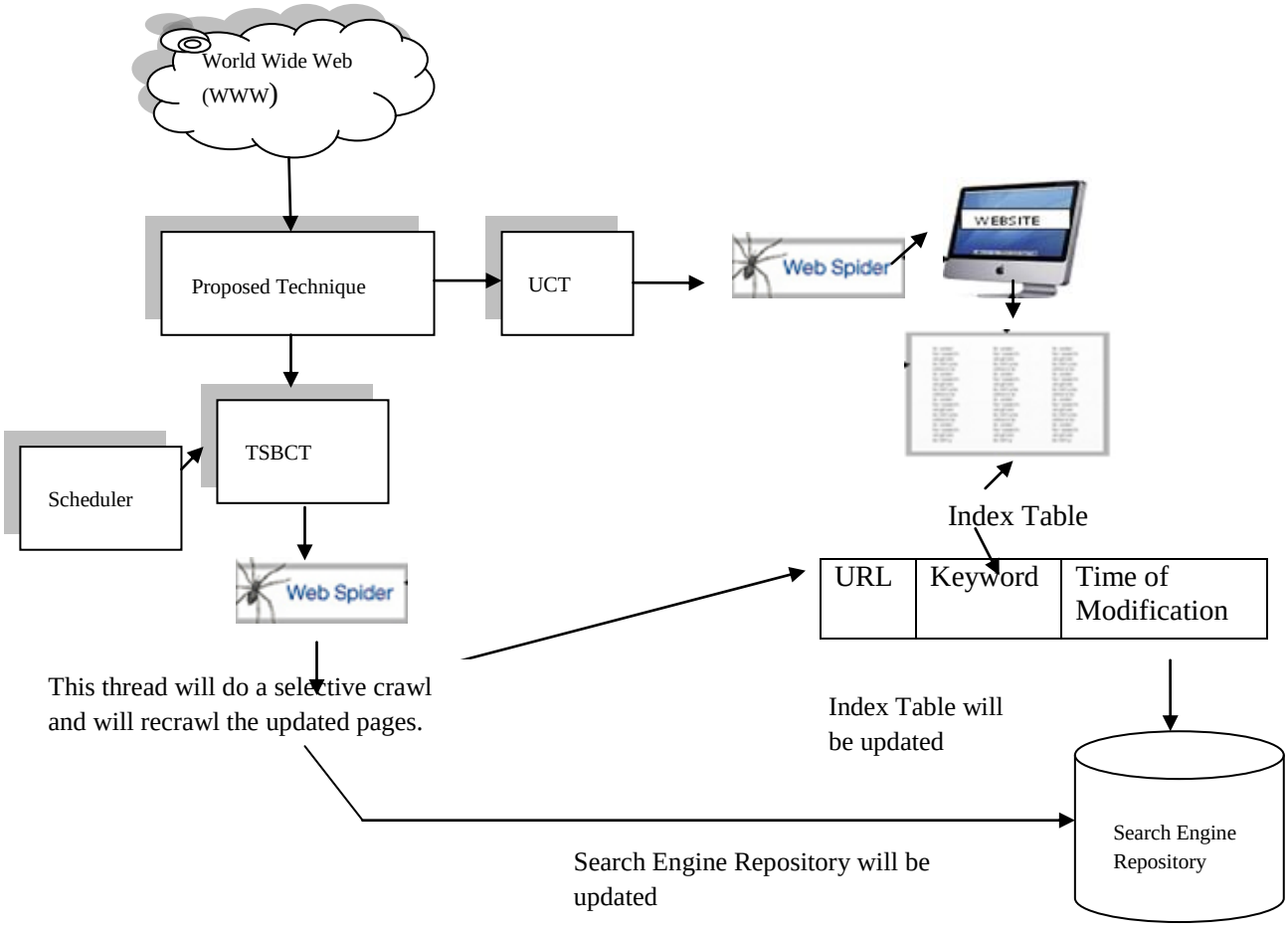

Figure5. Proposed Technique

modification of the page the freshness will be dertermined.

2.2 If the last time of modification of the page stored on web server comes out to be different from the indexed one then the page will be reindexed and updated in search engine repository.

This procedure will go on recursively.

Advantage of Proposed Solution:

- The crawling is done only for the pages that are updated often so it will not incur more expenditure.
- The dark web will be uncovered without increasing the cost of crawling.
- The non-updated pages need not to be recrawled unnecessarily.

5. CONCLUSION

The proposed system of TimeStamp Based Recrawl technique (TSBCT) will crawl the WWW very frequently and will drag the frequently updated pages from the darkness of Hidden Web. TSBCT will only check the date of modification and if page is found fresh then it will go through the metadata and the contents of the page. So it will be more fast. The TSBCT will after a month will add only the new born pages and the updated pages. So by TSBCT the non-updated pages will not recrawl again and again as done by the traditional crawler.

6. REFRENCES

[1] Sriram Raghavan Hector Garcia-Molina Computer Science Department Stanford University Stanford, CA 94305, USA, "Crawling the HiddenWeb"

[2] Rosy Madaan / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 753-758, "A Framework for Incremental Hidden Web Crawler"

[3] Ping Wu Ji-Rong Wen, Huan Liu, Wei-Ying Ma "Query Selection Techniques for Efficient Crawling of Structured Web Sources"

[4] Jian Qiu, Feng Shao, Misha Zatsman, Jayavel Index Structures for Querying the Deep Web, Workshop on the Web and Databases (WebDB), 2003, 79-86

[5] Ntoulas, A., Zerfos, P., Cho, J. Downloading Textual Hidden Web Content Through Keyword Queries. In Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL05). 2005.

[6] Chang, K; He, B; Zhang, Z. (2005). Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web.*CIDR*, pp44-55