IJCSC International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 257-258

EXPLORING THE HIDDEN WEB: A REVIEW

Anuradha and A.K Sharma

Department of Computer Engineering, YMCA UST, Faridabad, INDIA, *E-mail: anuangra@yahoo.com, ashokkale2@rediffmail.com*

ABSTRACT

World Wide Web (WWW) is broadly divided into two categories. First is Surface web that contains 1% of information content of the web. Search engine crawl along this web to extract and index text from HTML documents on the websites, then make this information searchable through keywords. Second is Hidden web that contains 99% of information content of the web. Most of this information is contained in the backend databases and is not indexed by search engines. Thus users are searching only through 1% of Web. However, these hidden web pages are dynamically created through search query interfaces. Hence, Indexing and retrieving this content poses several challenges. This paper expounds the reasons why Hidden web is hidden and how can we retrieve hidden web content.

1. INTRODUCTION

The term "Hidden web" refers to the large repository of information that search engines don't have direct access to, like databases. Hidden web data, stored in structured or unstructured databases [1], is inherently hidden behind search forms. It is qualitatively and quantitatively different from the Surface Web. The quality content of the hidden Web is 1,000 to 2,000 times greater than that of the Surface Web whereas overall the hidden web contains approximately 7,500 terabytes of data and 550 billion individual documents in contrast to the Surface Web, which is reported to about 167 terabytes [4]. Since the hidden web is the biggest source for structured data and is not publically indexed yet, accessing the same is a challenging task especially when the pages are created dynamically through search interfaces. On average, Hidden websites receive fifty per cent greater monthly traffic than surface websites and are more highly linked to than surface sites. In recent years, we have witnessed the rapid growth of databases on the Web, or the so called "Hidden Web". A July 2000 survey [1] estimated that 96,000 "search sites" and 550 billion content pages in this hidden web. [7] estimated 450,000 online databases. With the virtually unlimited amount of information sources, the Hidden Web is clearly an important frontier for data integration.

2. WHY HIDDEN WEB IS HIDDEN AND WHY IS IT IMPORTANT

Several online databases provide dynamic query-based data access through their query interfaces, instead of static URL links. This Query interface is considered as an entrance to Hidden Web, as the tremendous amount of information is hidden behind these search forms in web pages and traditional crawler cannot replicate the query submission carried out by human beings. A Web search interface for e-commerce typically contains some HTML form control elements such as textbox (i.e., a single line text input), radio button, checkbox and selection list (i.e., a pull-down menu) that allow a user to enter search information.

Search engines only search a very small portion of the web and this fact makes the Hidden Web a very tempting resource. There is a lot more information out there than we could ever imagine. Think about the Web as a library and we find book of our interest on front table. But we cannot! We have to search it. This is where search engines will not necessarily help us, and the Hidden Web will.

3. WHY HIDDEN WEB CANNOT BE INDEXED

The traditional search engines use inverted index as a data structure to index the web data and keyword interface to retrieve the data. But, surfacing the Hidden Web is more difficult task in many respects. First, the index structures for the deep web deal with the structured data as well as the large volume of data. Second, the Search query interfaces often have more than one attribute and require their respective values to be submitted. Since hidden web pages are created dynamically by firing user query, they cannot be indexed by search engines. There are many reasons behind this. Some of these are:

1. Some websites have millions of pages but just a small percentage are indexed by search engines because of the depth of the site.
2. Pages are protected by the webmaster: Files like "robots.text" and robots "noindex" or "nocache" meta tags in the HTML code of a web page can prevent search engines accessing the content.
3. Pages that are the result of a submitted query and consequently do not have a static URL can be impossible for search engines to find, since the spiders cannot replicate the query submission carried out by human beings.

4. Many websites limit access to some pages of the websites or require a password. Pages that can only be accessed after entering a password cannot be reached by search engine spiders.
5. Pages which are not linked to by other pages, which may prevent crawlers from accessing the content.
6. Pages with only JavaScript or Flash-based content cannot be easily indexed.
7. Hidden Web databases contains 100,000 unique records.100,000 direct queries would have to be issued to that database to obtain all records from that database. Multiply this by tens of thousands of hidden web databases and suddenly we have millions of queries that need to be issued to hidden web databases and then listed as links on static web pages to be searchable on surface web.

4. STRATEGIES TO EXTRACT THE HIDDEN CONTENT

4.1 Search Directiories:

A directory offers a hierarchical representation of hyperlinks to web pages and presentations broken down into topics and subtopics. Some directories offer a gateway to hidden web. By using these gateways, we can find relevant databases and then use the database specific search tool to extract the information we want. Some of them are www.completeplanet.com, Invisible web directory, www.directsearch.com etc.

4.2 Meta Search Engine

A metasearch engine or all-in-one search engine performs a search by calling on more than one search engine to do the actual work. It does not maintain its own database of information. By submitting searches to other search engines, it queries the databases of other search engines. Some of them are Dogpile, MetaCrawler, Turbo10, Profusion. Advantage of this type of search engine is that we can access a number of different search engines with a single query. The disadvantage is that we will often have a lot of matches that will not be of our interest.

4.3 Specialty Search Engine

Specialty Search engine exist for a multitude of topics, including shopping, news, travel etc. Many web presentations exist that provide search tools to find specialty search engines. These will provide subject guides as well. This search engine is another way to tap the hidden web.

4.4 Developing Hidden Web Search Engine

In order to extract information from Hidden web, It is necessary to develop a special search engine that retrieve the hidden web content with a view to give high quality information to the web user in integrated form. This search engine will automatically discover domain specific hidden web databases from the Web. Then crawl and integrate the hidden web content by querying search interface forms. Upon user querying, search engine will search from this integrated database by forwarding the search to right direction of domain. Several recent research projects, e.g., MetaQuerier [7] and WISE-Integrator [8], are exploring this exciting direction.

5. CONCLUSION

After the above discussion we have come to the following conclusion. First, the hidden web is not really hidden, but because searchable databases are not indexable or queryable by today's search engines, they appear hidden to the average Internet user. Second, it is difficult to index the hidden web content by the same technique used by conventional search engines as the hidden web pages are created dynamically by submitting different values to different fields of query interfaces. At the end, we can say that this is just the tip of the iceberg. As time goes on, the Hidden Web will only get bigger, and that's why it's a good idea to learn how to use it now.

REFERENCES

[1] BrightPlanet.com. The deep web: Surfacing Hidden Value. Accessible at http://brightplanet.com, July 2000.

[2] Thanaa M. Ghanem, and Walid G. Aref. "Databases Deepen the Web". *IEEE Computer*, **73 (1)**, 116-117, 2004.

[3] Steve Lawrence, and C. Lee Giles. "Accessibility of Information on the Web". *Nature*, **400 (6740)** 107-109, 1999.

[4] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. "A Large-scale Study of the Evolution of Web Pages". *In Proceedings of the 12th International World Wide Web Conference*, pp. 669-678, 2004.

[5] Ed O'Neill, Brian Lavoie, and Rick Bennett. Web Characterization. Accessible at "http://wcp.oclc.org".

[6] GNU. wget. "Accessible at http://www.gnu.org/software/wget/wget.html".

[7] Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. "Toward Large Scale Integration: Building a Meta-querier Over Databases on the Web". *In CIDR 2005 Conference*, 2005.

[8] Hai He, Weiyi Meng, Clement Yu, and Zonghuan Wu. "Wise-integrator: An Automatic Integrator of Web Search Interfaces for E-commerce". *In Proceedings of the 29th VLDB Conference*, 2003.

[9] C. Sherman, and G. Price, "The Invisible Web: Uncovering Information Sources Search Engines Can't Reach", *Information Today*, Inc., 2001, pp. 71-75.

[10] Y.Y. Chen, "Study of Internet Invisible Information Resources and its Effective Obtaining Strategies", *Information Research*, 2007. 2, pp. 51.

[11] S.B. Yuan, "Study of Invisible Web and its Strategies", *Library Tribune*, 2005. 5, pp. 191.

[12] H.J. Liu and S. Li, "Extraction and Integration of Invisible Web Resources", *Information and Documentation Services*, 2007. 1, pp. 68. The 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC'08).