DESIGN OF A HIDDEN WEB CRAWLER BASED SEARCH ENGINE

ABSTRACT

The World Wide Web is a global information medium of interlinked hypertext documents accessed via computers connected to the internet. Most of the users rely on traditional search engines to search the information on the web. These search engines deal with the Surface Web which is a set of Web pages directly accessible through hyperlinks and ignores a large part of the Web called Hidden Web which is hidden to present-day search engines. It lies behind search forms and this part of the web containing an almost endless amount of sources providing high quality information stored in specialized databases, can be found in the depths of the WWW.

World Wide Web (WWW) is broadly divided into two categories:

- The surface web contains 1% of information content of the web. Search engine crawl along the web to extract and index text from HTML documents on the websites, then make this information searchable through keywords.
- The hidden web contains 99% of information content of the web. Most of this information is contained in the databases and is not indexed by search engines.

This means if we are searching for information from surface web only, we search through only 1% of WWW and miss 99% of it whereas 95% of hidden web is free publicly accessible information.

As the Hidden web information that is hidden behind the search query forms can only be accessed by interacting with these forms, development of automated system that interacts with the search forms and extracts the hidden web content would be of great value to human users.

Today, the web is crowded with home-pages and sites that sell various types of products. Since the companies selling same type of products are not at all interested to publish the products of their competitors on their site, it would be nice if there is a free web service
which collaborate the marketing of the products of the competitor web sites. Infact all the information is available on the internet but buried behind search interfaces and stored inside the databases. Therefore, the required web service has to dig, gather, normalize and store the information to make it searchable for the users. In this thesis, the Hidden Web is studied in detail. Various challenges and their possible solutions are discovered to extract the Hidden Web data.

To accomplish both the tasks stated above, in this work a “Design of a Hidden Web Crawler Based Search Engine (HWCBSE)” is being proposed that works in four steps. The first step detects and extracts the search interfaces from predefined set of Hidden websites. After extracting the interfaces, it fills out web search forms automatically and extracts the result pages. Since the Hidden Web sites commonly return information in the form of lists and tables, a novel method to extract and integrate the data from these tables into one main repository is proposed in the second step which is based on structure of the web page. Since the data stored in main repository is the relational information and this information could be found by writing and firing SQL queries, users should have knowledge of database schema and SQL (Structured Query Language). To remove this limitation i.e to make the search based on keywords only (like traditional search engine searches), an inverted index database is developed in third step of this work that stores the attributes of main repository table against their respective values. The last step processes the user query by first splitting it in the form of keywords and then processes these keywords to construct SQL query that is fired on main repository which in turn returns the relevant results.

In this thesis, design and implementation of “Hidden Web Crawler Based Search Engine (HWCBSE)” has been reported. This prototype is capable of extracting, collecting and integrating the result records of various Hidden web databases efficiently and accurately.