Chapter 1

INTRODUCTION

1.1 GENERAL

The Internet and the World Wide Web (WWW) have changed the computer and communications world like nothing before. The objective behind the Web was to create a common information space for communication and information sharing. No other information system has taken such a meaningful place in our life in such a short time than is world-wide largest data network called the World Wide Web. However, when searching for information in the data network, the user is constantly exposed to an ever-growing flood of information.

Gathering information from the Web and identifying relevant resources which match a query is studied in the field of Information Retrieval (IR)[17]. Today, it is impossible to imagine the Web without search engines. They unite several sophisticated retrieval techniques and have become essential tools for navigating the Web. However, in addition to advantages of the current search engines, there are technical limitations too. The source of one such limitation is the nature of data retrieval and indexing. As a result, traditional search engines are only capable of answering queries based on the collections of web documents that are pre-processed and gathered by crawlers. This technique is efficient for the static web pages, which remain same for longer periods of time.

While most of the information is written in HTML and designed as static web pages, a large amount of pages behind search interfaces are dynamically generated using server side scripts on contents of back-end databases in response to form based queries are also presented to user. Contents in these dynamically generated pages are changed according to the user’s choice or some constraint such as time constraint. For example, ebay.com, amazon.com etc. are the Hidden websites where information is changed with user input and also they are updated very frequently. The pages are built up using templates and are dynamically filled
with information from back-end databases. In this way, the desired and more concise information is delivered to user in very short time.

However, today content of many resources in the Web remain hidden behind search interfaces. This part of the web is termed as Hidden Web. These are also known as dynamic web pages that can be hardly crawled by web crawlers and if so, they corrupt the search result with outdated information. Such pages are not helpful in case of virtual marketplaces or real-time information systems [9].

While many users in search of the information swim at the surface of the Web via search engines, an enormous amount of high quality information often stored in databases can be found in the depths of the Hidden Web. Bergman [6] states that Hidden Web is 500 times larger than the Surface Web and there are about 96000 Hidden Web sites and an estimate of 7,500 terabytes of data resides inside the Hidden Web. Another study [9, 10, 11] shows that the Hidden Web consists of about 91,000 terabytes and another survey published by Madhavan [12] based on a small fraction of 25 million randomly selected web pages taken from Google’s index identified over 647,000 Hidden web resources. In contrast, the Surface Web contains static web pages, easily reachable by traditional search engines. PIW, also known as publically indexable web, only contain 167 terabytes of data. Therefore, traditional search engines are only able to access 20% of the Web [6], ignoring the fact that the structured, high quality data lies hidden behind query interfaces.

Precisely, the Web can be partitioned into the two parts: Surface Web indexed by conventional search engines and the Hidden Web, which consist of structured data hidden behind search forms and published within dynamic web documents.

There are some false assumptions about web that leads to need of Hidden Web Search Engine. These are:

1. *Everything is available on the Web and if it is not there, it is not worth finding it.*
   
   To the user, information on the web seems better than from other sources because it is easy to access. However, the vast information lies inside the backend databases.
2. *Google can search the whole web.* This is the wrong impression created in minds of users by flooding number of results upon submitted even a small query. Bergman [6] stated that the Hidden Web to be about 500 times more than the size of the Surface Web. The Surface Web contains only 1% of information from WWW. So, we are getting results from 1% only and missing 99% of WWW.

3. *The desired result lies in first 10 results.* The searched results are ranked by ranking algorithms. This ranking may be based on frequency of keyword in the web page or on more sophisticated methods of evaluating linking, as is the case for Google. However, there is no guarantee that the first ten results are the best i.e only the user can judge. To increase the ranking of a web page, web designer can use various optimizing methods to ensure page’s placement in first ten results.

4. *Searching is easy.* It is well said “searching is easy; finding is difficult”. Today’s search engine is made for searching the information. We can easily find the results according to our query. But this easiness depends upon how appropriate query is? or expertise of user. The successful research requires time to make it finding.

5. *Everything is truthful and accurate.* We cannot expect the information on the web to be accurate, and truthful. Many websites contain replicated and mature information. Students and other researchers need to develop evaluation skills to cope with this kind of web environment.

If the Hidden Web comprises all information available on the World Wide Web that cannot be found by traditional search engines, then clearly the existence of the Hidden Web needs exploration. Hidden Web presents many problems for the information world, but the most important is how to express effectively its content and importance.
1.2 TRADITIONAL SEARCH ENGINE:

Search engine is a program that searches documents for specified keywords and returns a list of the documents where the keywords are found. Although search engine is really a general class of programs, the term is often used to specifically describe systems like Google, Alta Vista and Excite that enable users to search for documents on the World Wide Web.

Search engines are automated robots or spiders that systematically comb the web for servers and web pages. Once a page is found, the robot reads the words on the web pages and adds them to its database for later recovery when queried by a user [14]. Doing a search with a search engine is simply querying its database of words. The search engine does not scan the web on user’s behalf when he/she types the words to be searched. The search engine merely checks its database of words found by the “bot” on the web. It returns to user, URLs containing those words. This process is all well and good until we realize that search engines suffer from some limitations. Some of these limitations are:

- Search engines are more likely to index sites that have more links to them (more “popular” sites).
- Search sites are more likely to index commercial sites than educational sites.
- Indexing of new or modified pages by just one of the major search engines can take months.
- Search engines are designed to read flat web pages.
- Search engines are not designed to see the database-driven, dynamically constructed web pages.

And even if the search engine crawlers could get into the back-end databases used by dynamically generated web sites, some dynamically created web pages have variable URLs and some have same URL for all queries. Thus, a search engine can not rely on the URL found to be accurate on the next search. Therefore, it is not possible to index the Hidden web pages using traditional approach and hence there is a need to build the “Hidden Web Search Engine”.

4
1.3 MOTIVATION OF THE THESIS:

1.3.1 Example in Car Domain:

Let’s suppose a user wants to buy a used car. He fills the search box of search engine with “used car”. Search engine will return the result page containing result indexes as shown in figure 1.1.

Fig 1.1: Google results for query “used car”

User has to open the link one by one, result of which is shown in figure 1.2 (a) & 1.2(b). He has to fill the search form of each and can get the result. But after getting these result pages, he has to analyze the page for his requirement.
Consider another query “Honda city used car”. Again search engine will respond with the index pages as shown in figure 1.3.

After opening the links, it is seen that again it is returned with the search forms, and irrelevant data as shown in figure 1.4. Again user has to fill the forms to find the result.
But scenario is very different when online back-end databases of these sites are explored. The figure 1.5 shows that data actually resides inside the databases but it is not surfaced by the search engine. This means the result pages are not crawled by the web crawler.

A careful look at examples shown indicates that the data remains hidden behind the search interfaces. On one hand, these interfaces provide web users with an online access to Hidden Web databases. At the same time, these search forms are terrible barriers for the web
crawlers. Unlike human beings, the web crawlers have great difficulties in filling out forms and retrieving information from returned pages.

So, this is the first motivation behind designing the proposed system which automatically interacts with the search interfaces and extracts the data behind them. Again looking at the figure 1.5, we can see that there is some similarity in the presentation style of result data that pushes this research work to move in the field of integration. This motivation is discussed in next section.

1.3.2 Keyword Searching and Integration:

Search engines answer the user’s query by returning result pages based on some ranking algorithms. Here, the search is keyword based but it is very far away from true integration. Instead of linking the information, the search engines return the pages and the user has to traverse the pages manually to find the information. Search engines can not answer knowledge-rich queries because the Surface web documents are mostly unstructured. On other hand, the Hidden web which is many times larger than the Surface Web, offers a better platform for information processing and knowledge-rich query processing. This is because Hidden web information is stored in some form of structured database, and the information is produced dynamically on demand when a query is submitted through a web form. Mostly, this information is produced as tables or in some structured form like lists.

Since, search engine works on keyword based searching and cannot understand the knowledge rich data, it cannot display the results in summarized way or we can say it cannot present data in structured form. However, there are some similarities in presentation of result pages and if the crawler could be constructed to understand this regularity then it would be easy to extract and integrate the data at one place. For example, if the data from different Hidden web sites can be extracted and collected at one place then it would be easy to present the summarized data to the user.

So, there should be an information service that must be capable of accessing and processing heterogeneous collections of highly structured data delivered in HTML format and this becomes the second motivation behind the proposed system.
1.4 PROBLEM DESCRIPTION:

The two motivations discussed above show that there is a lot of data buried in Hidden Web sites and there exists no such system that can handle wide variety of such problems that require integration of different heterogeneous sources. So, the main objective of this research is to develop a robust and easy-to-use system that can handle the following tasks. These are:

- Understand the search query interfaces, extracts them, automatically submits the form, and retrieve the result pages.
- Extracts table records/ list records from table/list displayed in result pages and collects them in main repository to be used for later searching.
- Provides a user friendly interface where user can fill the query and find the result in integrated form i.e in tabular form which is easy to analyze and understand.

Thus, there is a need to “Design a Hidden Web Search Engine”.

1.5 CHALLENGES:

The system proposed in the last section has several challenges. These are:

- The first challenge is the automatic detection of search interfaces which is the entry to the web site.
- The second challenge is to recognize field labels as there are large number of web coding practices and, thus, no common rules regarding the name and position of field label in the page code.
- The web pages in the Hidden Web are dynamic pages that are constructed by embedding database records into predefined templates. The content of these pages usually have implicit structures or schemas that are different from others. Different companies choose different number of product features in interfaces as well as in the result pages. So, automatic schema matching is another challenge.
- Automatic filling of forms is another complicated task. If the queries using all possible combinations of attributes are fired then there are millions of queries and in
return millions of result pages. So, an efficient method should be there to fill the form which should not be time-consuming and can return large number of results.

- If the web site does not change and if all the websites have same structure of their web pages then it will be very easy to extract the information from table of back-end database by generating the extraction rule. But, the World Wide Web has been dominated by the HTML based on a browsing standard which is designed for human user and not for the data extraction programs it is therefore difficult to extract information by same HTML parsing methods. This is again the major challenge.

To overcome these challenges, in this work a “Hidden Web Crawler Based Search Engine (HWCBSE)” has been developed.

1.6 ORGANIZATION OF THESIS:

This thesis studies the Hidden Web at several different levels. Several contributions are made to extract the data from Hidden Web sites. The following is an outline of the contents of this thesis:

- This chapter introduced the concept of Hidden web, presented the motivation of thesis, challenges to fulfill the main objective, and summarized the contributions of this thesis.
- Chapter 2 explores the unseen facts about the Hidden Web. It also reviews various publications related to the search interfaces.
- Chapter 3 discusses related work done in the area of data extraction from Hidden Web sources and also presents the comparative study of the existing techniques.
- Chapter 4 presents the architecture of the proposed work “Design of Hidden Web Crawler Based Search Engine (HWCBSE)”. The proposed system has four components: Hidden Web Crawler, Data Extractor, Indexer and Query Processor. This chapter also describes briefly all these components.
- Chapter 5 describes in detail various components of HWCBSE.
- Chapter 6 explains the implementation and results of the proposed system.
• **Chapter 7** summarizes the contributions of the proposed work and provides some guidelines for future work in this area.

• In **Appendix A**, Document Object Model (DOM) of an HTML webpage is presented which discusses various functions and properties that can be used to extract the result from web pages.

Finally, bibliography includes over 150 references to publications in this area. The next chapter is a review of the most important facts and publications in the area of Hidden Web and search query interfaces.