Chapter 7

CONCLUSION AND FUTURE SCOPE

7.1 CONCLUSION:

Hidden Web data integration is a major challenge nowadays. Because of autonomous and heterogeneous nature of hidden web content, traditional search engine has now become an ineffective way to search this kind of data. They can neither integrate the data nor they can query the hidden web sites. Hidden Web data needs syntactic and semantic matching to achieve fully automatic integration.

In this thesis, fully automatic and domain dependent prototype system is proposed that extracts and integrates the data lying behind the search forms. The main contributions of this thesis are:

- **Search interface detection & extraction** - A search query interface allows a user to search the desired data by selecting the options that describes the items of his/her interest. This phase of proposed prototype automatically detects the domain specific search interfaces by looking at domain ontology (consist of domain word, attributes, their synonym and some keywords like submit, go etc.) for particular domain in source code and then extracts the interface by fetching the code that lies between `<FORM>` and `<FORM>`.  

- **Attribute-Value repository formation** - This phase extracts the value for each attribute of an interface and then store these attribute values in a repository which is used for filling the search form at later end.

- **Form submission analysis and new submission plan** - This phase analyzes the form submission plan. There can be many ways to submit the form with the queries. All the possibilities are tried to submit the form and results are analyzed. After analyzing the result, new form submission plan has been proposed that extracts the whole database by submitting values of mandatory field.
• **Study of indexing technique:** After extracting results by submitting forms, the result pages are indexed using traditional and database approaches. Traditional approach destroys the beauty of the relational data (tables) by parsing the page into words. So, the traditional approach is not suitable for Hidden Web data indexing. On other hand, database approach indexes the data by storing the URLs of the pages against each query. Here, the query processor matches the user query with the fields of the index database and return the result URLs. Study of this technique shows that it is again not considered as an efficient way of indexing. The first reason is that many sites return the same resultant URLs in response to different queries. So, the same results (URLs of the page) can not be stored against different combination of fields (queries). Second reason is that at the end of this phase user is again provided with large number of results which he/she has to analyze. So, this study shows that Hidden web data cannot be indexed by traditional approach.

• **Effective data extraction technique based on web structure** – Since, HTML is a building block for the web documents, this phase of thesis extracts the individual data units from table data using HTML DOM tree structure of the web page.

• **Integration of data records extracted from result page** – After collecting table for each website, a large repository has been maintained which contains data from various Hidden web sources of same domain.

• **Indexing** - Indexer receives the data from repository and build an inverted index to support keyword based searching (like a traditional search engine search keywords in various documents).

• **Query Processing** – To support keyword based searching, query processing method has been proposed in this phase that splits the user query into keywords then searches these keywords in inverted index for attribute name. As soon as attribute name corresponding to keyword is found, SQL query is automatically generated using template for that attribute and respective keyword (value).

• **User friendly search interface** – Finally the Graphical user interface is designed for user interaction where the user can fill the query in the form of keywords and find the desired result in integrated form.
The prototype “Hidden Web Crawler Based Search Engine (HWCBSE)” has been implemented in ASP.NET technology and SQL server is used as the database. The main objective of designing such a system is fulfilled as the results from various Hidden websites are accurately extracted and integrated. Moreover, user can get the desired results on the same result page in summarized form by filling only keywords in the search box of search engine and he/she does not have to learn about the database schema and SQL queries.

7.2 FUTURE SCOPE:

In this research work, various challenges in the area of Hidden web data extraction and their possible solutions have been discussed. Although this system extracts, collects and integrates the data from various hidden websites successfully, this work could be extended in near future. In this work, a search engine shell has been created which was tested on a particular domain. This work could be extended for other domains by integrating this work with the unified search interface [140].