Chapter 6

CONCLUSION & FUTURE SCOPE

6.1 CONCLUSION

In this dissertation, an effective technique to collect the data from the Hidden Web has been developed. More specifically, the main challenges involved in developing a crawler for the Hidden Web have been addressed and resolved.

During this work, Hidden Web crawling at many different levels has been studied. The main objective was to study crawling strategies for Hidden Web and to build a Hidden Web crawler by resolving the following challenges:

- **Identification of *Entry Points* for Hidden Web.** *Since the search interfaces acts as the entry points for the hidden web contents, a mechanism that automatically extracts domain-specific search interfaces has been proposed and implemented in this work.*

- **Large number of Domains.** *A domain-assisted approach for crawling has been developed. It helps in identifying the domain and in finding the relevant information thereof for the Hidden web.*

- **Integration of search forms of the same domain.** A novel framework for interface integration is being proposed. The proposed Interface integration involves two steps: Interface Matching and Interface Merging. The "Interface Matching" employs an "extensible domain-specific Library" for quickly identifying regions in the interface repository comprising of important mappings while "Interface Merging" merges all the interfaces into a Unified Search Interface (USI) based upon the semantic mapping identified in first step of Interface Integration by merging two interfaces at a time.

- **Automatic Query Generation.** To fill the USI automatically, a Domain-specific Data Repository has been employed. It contains labels and their corresponding values. In this

work, Data Extractor Engine and Search Interface Parser are used to create Domain-specific Data Repository which is further used for automatic query generation.

The Domain-specific Hidden Web Crawler (DSHWC) has been implemented using .NET technology and SQL Server 5.0. High values of *Precision, Recall* and *F-measure* for various tests conducted on Domain-specific Hidden Web Crawler (DSHWC) indicate that it efficiently crawls the hidden web pages. The classification of the proposed work into different phases not only improves the performance of each phase but also renders the crawling a modular and extensive framework with the expectation that new functionality can be added by third parties according to their requirements.

6.2 FUTURE SCOPE

In this thesis, the problems related to Crawling the Hidden Web has been explored extensively. Some of the possible extensions and issues that could be further explored or extended in the near future are as follows:

- **Designing of a Search Engine based on DSHWC.** As the size of the hidden web contents are very large and it would continue to grow with the time. Therefore, work can be done to design a search engine that could be able to crawl, extract and index the content of these hidden databases.

- **Updating the Hidden-Web pages.** The information on the Web today is constantly evolving. Once DSHWC has downloaded the information from the Hidden Web, it needs to periodically refresh its local copy in order to enable users to search for up-to-date information. Therefore, work can be done in this direction to crawl the Hidden Web information incrementally.

- **Indexing the Hidden Web Pages.** Search engines typically allows to search the information to the users by maintaining large-scale *inverted indexes* which are replicated dozens of times for scalability and which are then pruned in order to reduce the cost of

operation. Thus, future work can be done towards obtaining efficiency in maintaining the index for Hidden Web in general and searching the required information in the indexed documents in particular.