CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

As the size of the web increases and its contents become more diverse, the role of a Web crawler becomes even more important. In this thesis

- The architecture for Effective Migrating Parallel Web Crawling approach with domain specific and incremental crawling is proposed. Domain specific crawling yields high quality pages. The incremental crawling yields the fresh page. The crawler is implemented in Java. The implemented model supports all features of real time III tier architecture.

- In this thesis a neural network based change detection method in migrating parallel web crawler is implemented. The method for Effective Migrating Parallel Web Crawling approach detects changes in the content and structure using neural network.

- Crawlers can be generously allowed to communicate among themselves or they cannot be allowed to communicate among themselves at all, both approaches put extra burden on network traffic. Here a fuzzy logic based system is proposed and is implemented using MATLAB fuzzy logic tool box which predict the load at particular node and route of network traffic.

- The architecture of migrating parallel web crawler is validated using finite state machine. Test cases are generated for the validation of the architecture. The approach for generating the test cases through FSM is very reliable and efficient and does not support for the invalid test cases. Empirical validation is performed with SPSS tool to assess the performance of the proposed approach.
In chapter 1 World Wide Web, Internet, Search engines and Web crawlers are discussed. The evolution of WWW and how the web has become the source of information sought by users is discussed. In chapter 2, hyper text systems are discussed. A discussion on indexing and querying web pages is given. Web characterization is also discussed. Distributed systems are discussed. A discussion of web crawlers and their types is also discussed. Web crawling issues are discussed. Based on the literature reviewed change detection techniques have been identified, that provided the basis for the work to be carried out. In chapter 3 the outline of complete model is discussed. In chapter 4 the architecture for Effective Migrating Parallel Web Crawling approach with domain specific and incremental crawling strategy that makes web crawling system more effective and efficient is proposed. Domain specific crawling will yield high quality pages. The crawling process will migrate to host or server with specific domain and start downloading pages within specific domain. Incremental crawling will keep the pages in local database fresh thus increasing the quality of downloaded pages. The crawler is implemented as a client server system based on JAVA, Apache TOMCAT as server and MySQL as database has been designed. The implemented model supports all features of the real time III tier architecture. The Net Beans 7.4 is used as Development Platform. In Chapter 5 a neural network based change detection method in migrating parallel web crawler is implemented. The method for Effective Migrating Parallel Web Crawling approach detects changes in the content and structure using neural network. Neural network based change detection method in migrating parallel web crawler will yield high quality pages and detect for changes, will always download fresh pages. In chapter 6 the crawling process is discussed to be carried out using either of the following approaches: Crawlers can be generously allowed to communicate among themselves or they cannot be allowed to communicate among themselves at all, both approaches put extra burden on network traffic. Here a fuzzy logic based algorithm is proposed and it is implemented in MATLAB using fuzzy logic tool box which predict the load at particular node and route of network traffic. In Chapter 7, the architecture of migrating parallel web crawler is validated using finite state machine. Test cases are generated for the validation of the architecture. The approach for generating the test cases through FSM is very reliable and efficient and does not support for the invalid test
cases. Valid input strings are generated as test cases. Empirical validation is performed to assess the performance of the proposed approach.

8.2 Future work

The potential areas for future work are discussed in this section. The assumption is made that all Web pages can be reached by following the link structure of the Web. As more and more pages are dynamically generated, however, some pages are “hidden” behind a query interface. That is, some pages are reachable only when the user issues keyword queries to a query interface. For these pages, the crawler cannot simply follow links but has to figure out the keywords to be issued. While this task is clearly challenging, the crawler may get some help from the “context” of the pages. The research directions in migrating parallel crawler include:

- Security could be introduced in migrating parallel crawlers
- Migrating parallel crawler could be made polite
- Location awareness could be introduced in migrating parallel crawlers

This future work will deal with the problem of quick searching and downloading the data. The data will be collected and analyzed with the help of tables and graphs.