CHAPTER 8
CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

The Web has become a very significant resource for accessing a variety of information. Currently, billions of webpages are available on the Web, much more are created and hosted on the Web daily. Therefore, the use of SE is becoming a primary activity to search the information on the Web. The quality of SE results depends upon the crawlers, which are responsible to keep the page repository up-to-date. Although SE has adopted many efficient techniques to constantly index more and more Web pages.

An extensive literature survey of existing work in the domain of Web crawlers and search systems have been carried out. The existing architectures, algorithms, optimization techniques and Web crawlers are critically analyzed which helps to improve the understanding of the search system in depth in general and provides the motivation for doing the present work. Many problems in the existing work have been identified – (i) redundancy of results (ii) wastage of network resources (iii) inefficient Web page change detection technique (iv) poor quality of search results and (v) inefficient text summarization techniques.

Firstly, the problem that has been found in existing Web crawling is the way to determine the freshness of a webpage at the RS. Currently, it is determined at the SE end only which results in wastage of resources by indexing the webpages that already have been indexed. To address this problem, the authors have proposed a Parallel Domain Focused Crawler (PDFC) based on page change behavior. Page change is detected by comparing the old and new ASCII count value of the webpage. The PDFC has been implemented in Java and the results are compared with the existing crawler PMCS in terms of Web page change behavior, load on the network, time consumption and network utilization. The results have shown
that the PDFC has reduced the average network load by one fourth as compared to PMCS with compression and it is one third when they are used without compression. Further, the experimental results have shown that the PDFC has improved the performance of network utilization significantly as compared to the PMCS.

Secondly, the problems that have been found in the existing MSEs are – (i) they are not able to present the result in an effective manner (ii) they use positional ranking and count function which fails to deal with the relevant search results (iii) They provide the name of the folders based on the highest frequency of words, but the name assigned to these folders may not be relevant to the keywords searched by the user. To address these problems, an MSE has been proposed based on clustering of results. The proposed MSE has been implemented in C# and the relevancy of search results against the given query has been evaluated. The results are compared with existing MSEs – MetaCrawler, WebCrawler, Excite, Dogpile and Genome. The experimental results have shown that the proposed MSE provides the more relevant results, better ranking, better naming of clusters and removes the duplicate links as compared to the existing MSEs.

Thirdly, the problems that have been found in existing ATS are – (i) use of less number of HTML tags used in summary generation (ii) importance of a Web page is measured by position of text which fails to provide the relevant results (iii) presentation of preprocessed data which diminishes the meaning of up-to-date data (iv) redundancy of results and (v) wastage of network resources. To address these problems a multi-document ATS based on feature keyword has been proposed and implemented in ASP.NET. The proposed summarizer has been compared with the following ATS - Free Summarizer, Auto Summarizer, Tools4noobs, MEAD and Comparative Summarizer in order to test it. The experimental results have shown that the proposed ATS provide better summary for single as well as for multiple documents as compared to existing
summarizers. Results further shows that the proposed ATS is efficient in time, saves network utilization and is capable of summary generation for ever changing environment. Its applications can be found in mass media, news broadcasting, medical science and snippets generation etc.

Fourthly, the problems that have been found in existing SE evaluation techniques are – (i) they are not effective in ever changing environment and (ii) most of the time the evaluation has been done manually which is a very time consuming process. To address these problems an SE evaluation framework based on TLS has been proposed. TLS evaluates the proposed framework on relevancy of the results, user efforts and precision of the results. The proposed framework has been implemented in C# and the queries set of three domains - technical, medical and mixed have been given to it. The experimental results have shown that for technical and medical domain query set Google provides better results whereas for mixed domain query set AltaVista provides most relevant results than others. When the comparison of relevancy, user efforts and precision of the results are considered collectively then it has been found that Google wins the race in all the three domains.

### 8.2 Future Scope

In this thesis, the problems related to the existing search systems have been explored. However, there are still many open problems and issues that need to be explored before providing access to the variety of Web search resources to the user. Some of these problems are discussed below and their solutions can further improve the existing search systems.

In this work, the Parallel Domain Focused Crawler has been proposed which works on textual data only. This work can be extended to work on images and videos.

The multiple document ATS proposed in this thesis summarizes the textual data only. This work can be extended to develop the methods that can summarize the images, audio and video as well.