CHAPTER 7

CONCLUSION

7.1 INTRODUCTION

This chapter reviews the significance of the proposed web crawling algorithms, and the results obtained during the present work and offers suggestions for future research. Before proceeding with the review of the work done, the objectives of the thesis stated earlier in the introductory chapter are recalled.

The primary objective of this research is to develop effective and efficient web crawling algorithms that account for the retrieval of the more relevant pages at the earlier stage of crawling. Based on this objective it is proposed to develop new approaches that are derived from specific web crawling paradigms such as keywords, logarithmic method, probability method etc.

7.2 HIGHLIGHTS OF THE WORK DONE

The thesis offers fresh motivation for fundamental algorithms that are finding new applications in the context of web crawling and emphasizes the fact that the models proposed depend on efficient retrieval of relevant pages with few levels of crawling. In this dissertation four effective web crawlers are developed and implemented, which can discover and identify the important pages earlier, retrieve the pages promptly, and maintain the retrieved pages fresh. The crawling is expressed using different measures and approaches and various crawling models have been developed.
To make the strategies more effective and to obtain better results in web crawling, new concepts on the integration of link, text content, logarithmic, probability measures of web crawling, topic specific and two-level approaches for relevant retrieval of pages are introduced. The proposed techniques are experimented in a web environment and the performance are evaluated and compared with the performance of the two traditional crawling schemes, such as Integrated Page Rank Algorithm (IPRA) and Improved Weight Algorithm (IWA). The results reveal that the proposed schemes outperform the existing schemes in terms of threshold, precision percentage, recall and accuracy percentage. The number of levels of crawl is greatly reduced and the percentage of the precision, recall and accuracy is increased, and consequently the crawling performance is improved.

Analysis has been carried out to verify the correctness of the algorithms. Based on the investigations made and results obtained, the following conclusions are arrived at:

i. The Logarithmic and Probability Measure Algorithm (LPMA) identifies and retrieves the most valuable pages fetched by the crawler. This method provides more number of relevant pages. Different anchor texts are given to retrieve the web pages and the effectiveness of the algorithm is evaluated by comparing with the Integrated Page Ranking Algorithm (IPRA) and the Improved Weight Algorithm (IWA). The advantage of the algorithm is that it makes use of the similar and also the dissimilar keywords with the probability measure for improving the efficiency of crawling. This search algorithm shows the total number of retrieved pages, within few levels of crawl.
ii. The Topic-Specific Algorithm (TSA) for web crawling using probability measure has fetched more number of relevant pages based on the topic specified by the user. Evaluation is made on different topics, threshold values and it is found that more relevant pages are retrieved. This algorithm reduces the levels of the crawler and fetches the best web pages of user interests at the earlier stages of crawling compared to the Improved Weight Algorithm (IWA). The computation of relevancy score for each and every web page in the URL queue using link weight, Levenshtein distance and probability measure and placing the outgoing URL when relevancy score is greater than a stipulated threshold value has reduced the number of crawling iterations. Efficiency of the method is tested varying the co-efficient factors of the relevancy score computed for the different URLs. There is reduction in the crawling levels for the relevant retrieved pages by the proposed method.

iii. In the Integrated Four Measures Algorithm (IFMA), a comparative study has been performed on four different measures individually and the integrated measure, that is the combination of the four measures for different anchor texts. The computation of relevancy score for each and every web page in the URL queue using the four measures link weight, Levenshtein distance, logarithmic distance and probability measure individually and the relevancy of the integrated measure has reduced the number of crawling levels and fetches the best web pages of user interests and outperforms the Improved Weight Algorithm (IWA).

iv. The Two-Level Approach (TLA) identifies the top URLs that exist for the given seed URL. The computation of relevancy score for each and every web page in the URL queue using title of document, snippet and parent URL at the first level discards the irrelevant pages in the initial and subsequent crawling and skews the search to more relevant pages. The
computation of relevancy score for each URL in URL queue using page content of document at the second level, the most relevant pages are retrieved effectively. The URLs having most relevant pages are considered for next iteration to give more effective result compared to the Improved Weight Algorithm (IWA). Hence, more numbers of pages of most relevance are retrieved by the proposed crawling at an earlier stage.

7.3 FUTURE ENHANCEMENTS

The new crawling techniques introduced in this research deal with users’ opinions. As the web grows larger and its contents become more diverse, the role of a web crawler becomes even more important. It is difficult to maintain a page up-to-date if the pages change too often. As more and more pages are dynamically generated, frequent changes may become a problem.

The main objective of this research is to develop a model for web crawling, to study the crawling strategies and to build a web crawler implementing them. The implementation is done on a portion of the web to retrieve the web pages. The performance of these algorithms is to be tested with a large volume of web pages. In the algorithms used in this work for testing, some of the relevant pages were not retrieved. However, they were improved by indexing the web pages based on keywords, topics and topic related query in proposed strategies. Further extensions can be done on this work by analyzing and proposing semantic queries. The need for interpreting the inherent meaning of the query and indexing based on that will be an interesting research direction. The work can be extended to have a periodic crawler for performing better indexing.