Chapter 8

CONCLUSION AND FUTURE SCOPE

8.1 CONCLUSION

The Web has become a vital resource for accessing a variety of information. There are billions of Web pages and, everyday, new content is produced and hosted on the web. Therefore, the use of search engines is becoming a primary Internet activity. Although search engines have developed many efficient techniques to constantly gather more and more information of user's interest. Nevertheless, the literature review carried out in dissertation identified many open research areas of tremendous interest where the quality of search results can be improved.

The objective of this dissertation was to build a unified framework of search engine with which the crawling, indexing and the query processing tasks could be efficiently mechanized so as to improve the overall quality of search results.

To fulfill the above said objective, in this dissertation, a novel P & P framework of search engine has been developed, which is a blend of a set of proposed Pre-mining and Post-mining techniques. The framework addresses the main challenges involved in improving the relevancy of search results.

During the whole work, various technologies involved in the working of a search engine have been studied. A review of present web mining techniques in the context of search engines have been carried out and the major challenges in building cost effective search engines have been identified. Following are the contributions made by the work carried out in this dissertation:

- A unified framework of search engine: In this work, a unified framework of search engine called “P&P (Pre-Mining and Post-Mining based) Framework” has been proposed, which simultaneously can optimize crawling, indexing and query
processing tasks carried out by search engines. A couple of Pre-mining and Post-mining techniques have been developed for optimizing various processes of a search engine. The proposed techniques aim to provide relevant document retrieval.

- **Efficient data structures**: Efficient data structures have been designed for the sake of better interpretability. These data structures have been made to represent valuable information related to relevant document retrieval. The proposed techniques have been made to utilize the information contained in these data structures.

- **Optimization of Crawling and Indexing process**: An optimization technique (Pre-Mining technique) to utilize usage data in the crawling and indexing process has been proposed so as to make crawlers adaptive in nature and download important pages, which were previously not crawled and indexed. The index has also been optimized such that it refers to current information about important pages and their accessing behavior.

- **Semantic Indexing over Keyword based Indexing**: A Semantic Search technique (Pre-Mining technique) has been developed without affecting the keyword based indexing employed by existing search engines. The approach relies on the online web resources such as dictionary based sites to retrieve possible semantics of the query keywords to automatically generate meaningful queries, which are in turn searched for in the keyword based indexes of existing search engines to return relevant documents.

- **Efficient Index structures for XML documents**: An incremental content-centric approach (Pre-Mining technique) of XML indexing has been developed in this dissertation that uses frequently used paths from historical query logs to improve the query performance over XML indexes. The XML index has been made to be updated incrementally according to the changes in query workload and thus minimizing the overhead of reconstruction.
Quality of Search Results: A Duplicate page detection and elimination technique (Post-Mining technique) has been proposed wherein duplicate detection is performed offline only for the historical queries submitted by users, while elimination is done at the query processing time to return quality results to the user.

Efficient Result Organization Scheme: An efficient technique (Post-Mining technique) to organize the search results has been developed which produces clusters of results with ranked pages within. A novel ranking mechanism based on page-query similarity has also been proposed for ranking the pages within each cluster.

Ranking Algorithm based on Page Accessing Behavior: A novel page ranking technique (Post-Mining technique) based on Web usage and structure mining has been proposed that takes number of user visits on links of pages into account to assign a rank score to the page.

Optimization of Rank Scores and better Query Recommendation: A rank optimization and query recommendation technique (Post-Mining technique) has been proposed in the final stage, which improves the score assigned to a page based on the sequential patterns of page accesses discovered from the query logs. The technique also proposes an efficient query recommendation approach based on mining the query logs.

The Pre-mining and Post-mining techniques have been implemented and tested over the sample web graph and query logs. The result analysis indicates that the proposed techniques result in relevant pages in an easy browsable representation. The classification of the proposed framework into two phases not only improves the performance of each phase but also renders the provision of adding new functionalities.

8.2 FUTURE LINES OF RESEARCH

In this work, the problems related to existing search engines have been explored. However, there are still many open questions and issues that are worth exploring before we can provide
users with sophisticated, seamless, and transparent access to the large number and variety of web sources by the use of search engines. Below is a description of some of these problems, which range from improving systems that already exist, to dealing with sources that are largely ignored by current search engines:

- The search engines must have a mining mechanism to mine hidden web resources such as relational tables, non HTML pages and image repositories so as to better optimize the quality and relevancy of search results.
- The semantic search system proposed in this dissertation can be extended to serve complex user queries, besides serving topical and informational queries. This will require building efficient query analyzers so as to direct the user search towards the right direction.
- Crawlers of the search engines can be made to utilize the information regarding duplicate pages found by the proposed approach, to enhance the crawler efficiency also.
- More advanced mining mechanisms using AI and NLP techniques may be devised which can provide more comprehensive information about relevancy of the query terms and thus allow identifying related documents.
- The query logs can further be mined for identifying object associations at three levels: associations between accessed pages, associations between user queries and the associations between query keywords. The mining done at these three levels can be in query expansion mechanisms employed by query processor resulting in a better web search.


[40] HTML Tutorial. www.w3schools.com/html/default.asp


