Chapter 7
CONCLUSION

Seen are beautiful, Unseen is more beautiful. Web Mining, though challenging, poses wonderful opportunities before the researchers. This chapter summarizes the research findings and discusses possible enhancements to the proposed methods and some interesting future directions

7.1 Introduction

As people become more dependent on World Wide Web, various intelligent web services have turned to be an immediate necessity. Though many new techniques have already proposed and implemented, highly complex nature of Web Data and varying desire of Web Users urge serious research interventions in this domain. Mining of Web Access Pattern is the very fundamental and the most tedious job in learning users' interests on the web. In this thesis, concepts and techniques of Web Access Pattern Mining have studied in detail and some efficient and scalable methods for mining Web Access Patterns are proposed. Using benchmark datasets, all the proposed methods are experimentally proved to be efficient and scalable. This chapter first summarizes the research work and then discusses some interesting future directions.

7.2 Research Summary

World Wide Web is a rich information repository which people can depend on for information storage and retrieval. But the reliability and quality of information provided by the Net for a particular user depends on how intelligent the system is. The system should be able to learn the users' interests and exhibits the web as he/she desires. Web log mining is an effective tool for learning the users' interests on the web. Frequent patterns of user access to web pages can be mined from Web logs.

Sequential Pattern Mining, an important data mining technique with broad applications, is an important tool in Web Access Pattern Mining.

In this thesis, an in depth study of sequential pattern mining is carried out. Various approaches to Sequential Pattern Mining, its possibilities and drawbacks are well discussed. WAP-Tree is an efficient tree structure introduced by Pei *et al*. which provides an easy way of holding access sequences and facilitates tedious support counting. WAP-Mine, the corresponding mining algorithm, outperformed all earlier pattern mining method. In this thesis a detailed discussion on WAP-Tree and WAP-Mine is provided. Prominent research proposals based on WAP-Tree for improving the efficiency of Access Pattern Mining are also systematically studied and compared.

Though the WAP-Tree structure is efficient in holding access sequences and in facilitating support counting, it has got some limitations. All WAP-Tree based methods involve two database scan, first to find out the frequent items and second to create tree using frequent sub-sequences. This introduces a time delay in the mining process. Generally, tree structure is very compact, but the Aggregate tree used in WAP-Tree based methods, node structure requires more space. Moreover, WAP-Tree based methods either use complicated linkages of tree nodes or construct intermediate trees for projection databases for finding out the first occurrences.

We have proposed, ***FOL-Mine***, an efficient pattern growth method for Web Access Pattern Mining. The proposed method involve only one database scan and access sequences are stored in using a very efficient linked structure. The method also uses a linked list structure FOL that elegantly manages the First Occurrence positions of frequent events, which is crucial in improving the performance of access pattern mining. Performance and scalability are systematically proved experimentally.

The huge number of patterns generated during frequent pattern mining process is a real concern in terms of the space occupied. Maximal Pattern mining is one of the techniques for handling this issue. Web Log Database itself is huge in size and generally the access patterns are very lengthy too. Lengthy patterns contain exponential number of smaller access patterns. So, space is a major constraint in Web Access Pattern Mining as well. Concept of Maximal Pattern can be adapted to Web Mining as a solution. A pattern α is a maximal frequent pattern in set *D,* if α is frequent in *D* and there exists no super frequent pattern of α *exists in D.* In this thesis a very efficient maximal pattern mining algorithm, ***FOLMax-Mine***, is proposed for mining the maximal access pattern. The method uses the modified version of structures proposed in

FOL-Mine. The efficacy of proposed method over the ordinary access pattern mining is systematically studied and established through experiments.

Weighted sequential pattern mining is an approach that treats various items in the sequences with varying weights. This field of research has emerged from the fact that all items in a sequence do not have the same importance. Many attribute can be attached to items to show their relative importance. Thus, weighted method models the real life sequence database in a better manner. Weighted sequential pattern mining can be adapted to mine web access patterns more efficiently from web log data.

The basics of weighted pattern mining are discussed in detail. A new weighted access pattern mining algorithm, ***FWAP Mine***, to mine all Weighted Access Patterns in a web log database is proposed. The new method uses frequency of user visit to give weights to web pages during the mining process. Through extensive experimental evaluation the algorithm is proved to be promising.

7.3 Enhancements and Future Research Directions

7.3.1 Enhancements

- Web log grow rapidly and incrementally and so is the case of Web Access Sequence Database. It is undesirable to mine access patterns from scratch each time when a small set of access sequences are added to the database. Incremental mining has added advantage in this case. Data structure used in our access pattern mining algorithm supports incremental mining. It does not delete any item from the access sequences from WASD. So new access sequences can be added and frequent items can be found without dropping already loaded sequences from the data structure. So our algorithm can be easily enhanced to incremental mining.

- Maximal pattern mining loses the support information which is essential in making the information usable in other applications. Closed pattern mining is another aspect where we can have the support information. So our maximal pattern mining method can be modified to Closed Pattern Mining.

- In the Weighted Access Pattern Mining algorithm we have considered only the frequency of user access to attach weight to sequence. Other attributes like number of back access and time spent on each page can also be considered for attaching weights and improve the performance.

7.3.2 Future Research Directions

Intelligent response by the Web is an important research concern as the fast and personalized response is a primary concern of web users. Web Access Pattern Mining is the fundamental step in many techniques for improving the surfing experience a user. In this thesis we have proposed, implemented and analyzed three efficient methods for improving access pattern mining. These methods can be extended to improve the intelligent response of the web.

Web Recommendation: **-** Knowledge regarding the user path is a primary requirement for Web Recommendation. After completing the mining of web access patterns, they can be plotted on a pattern tree structure. While a user interacts with the web, current access path can be matched against the patterns in the tree and recommendation rules can be generated [Zhou, Hui, Chang, 2004].

Web Caching: **-** Web caching becomes an attractive solution because it represents an effective means for reducing bandwidth demands, improving web server availability, and reducing network latencies. Access pattern information along with support provides rich information regarding the possibility of next user access. So using more efficient methods for access pattern mining becomes an important concern.

Prefetching:- Though the web performance is improved by caching, benefits of caching is limited due to the fast changing nature of web data. Page Prefetching is evolved as a solution to this problem by reducing the retrieval latency. Prefetching reduces user access time, but at the same time, it requires more bandwidth and increases traffic. An important factor of a prefetching algorithm is its ability to decide the data to be fetched in advance. Web Access Pattern information provides a better way of identifying the possibilities of next user access.

Personalization: **-** Web personalization aims at customizing the presentation of web contents to Web users according to their specific tastes or preferences. Information regarding access pattern can be used to identify the user preferences or user tastes.

Web User Clustering: - Web clustering is an important research field in Intelligent Web Mining. Clustering web user facilitates the modelling of user profiles. User Clustering based on access pattern is a fertile field of research.

7.4 Conclusion

As a result of the research and development, quality of web related services has improved significantly. But as a result of the unprecedented growth of the Web and abundance of information, Web Users are not able to locate what they actually want within a reasonable amount time. Web Access Pattern Mining has been identified as a key step in improving the efficiency of web services. This thesis has proposed three efficient approaches towards an intelligent web. But due to the fast development in the technology and explosion in the number of users, Web Mining area still gives lots of research opportunities.