ABSTRACT

Generating accurate and actionable knowledge from a huge amount of data is a real challenge. Data mining has emerged as a research direction to meet this challenge. World Wide Web is the repository of all kinds of data that people from various fields, with varying interests and having variant levels of expertise depend for their day to day activities. Making the Web respond to its users intelligently, providing them with most relevant and accurate data in minimum time is a real challenge. The unstructured, uncontrolled, dynamic, ever-changing nature of Web Data makes the task more complex. Web Intelligence (WI) has been recognized as a vital field for scientific research.

Web Data Mining is a process that discovers the intrinsic relationships among Web Data, which are expressed in the forms of textual, linkage or usage information, via analyzing the features of the Web and web-based data using data mining techniques. Techniques like Web Recommendation, Web Caching, and Web Personalization are used for improving the web surfing experience of user. Efficient mining of Web Access Patterns is crucial in all the techniques used for making the Web more intelligent. Thus, Web Access Pattern Mining is a thrust area of research.

Sequential pattern mining that mines frequent patterns from sequence data can be adapted to mine web access patterns. In this work an in-depth study of Sequential Pattern Mining is carried out. Various approaches to Sequential Pattern Mining are discussed in detail. Theoretical background of Web Access Pattern Mining is also discussed.

Web Access Pattern Tree (WAP-Tree) and the WAP-Tree based mining algorithm WAP-Mine proposed by Pei et al. is an important contribution in the field of Web Access Pattern Mining. WAP-Tree holds the Access Sequences compactly and facilitates the efficient counting of support. WAP-Mine algorithm was proved to outperform the earlier apriori based Access Pattern Mining algorithms. Many proposals are put forward by the researchers to improve the performance of WAP-Mine in terms of time and space.

Though the WAP-Tree structure is efficient in holding access sequences and facilitating support counting, it has got some limitations too. All WAP-Tree based methods involve
two data base scans, 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.

Moreover, though tree structure is very compact in nature, node structure is very complex in the Aggregate tree used in WAP-Tree based methods and requires more space. Further, in order to find the first occurrences of a frequent symbol, WAP-Tree based methods either use complicated linkages of tree nodes or construct intermediate trees for projection data bases. This thesis proposes an efficient web access pattern mining algorithm, FOL-Mine. The FOL-mine algorithm generate access patterns by suffix building in the projected data base of each frequent event and eliminates the need for construction of pattern tree. First Occurrence List is introduced in the proposed algorithm for efficient handling of suffix building. Rebuilding of projection databases is completely eliminated in the new method. Experimental analysis of the algorithms reveals significant performance gain over other web access pattern mining algorithms.

Huge number of patterns generated during mining is a major concern in pattern mining due to the space requirement. Mining of maximal patterns, patterns having no super frequent pattern, is a solution. Motivated by the concept of Maximal pattern mining, in this thesis we have proposed an algorithm, FOLMAX-Mine, for mining maximal access patterns. It is a top-down method that uses the concept of first occurrence to reduce search space and thereby improving the performance.

Weighted sequential pattern mining field of research has emerged from the fact that all items in a sequence do not have the same importance. It is identified that Weighted Sequential Pattern Mining can be adapted to mine web access patterns more efficiently from web log data. 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.

Web Access Pattern Mining is the most crucial and time consuming part of techniques like Web Recommendation, Web Caching, Page Prefetching, towards an Intelligent Web. Thus making Web Access Pattern Mining more effective and efficient is a first step towards the intelligent Web. Three algorithms for improving the efficiency and effectiveness of Web Access Pattern Mining are proposed in this thesis. The extensive performance analysis using benchmark datasets shows that all suggested approaches meet the objective in an efficient and scalable way.
Proposed Web Access Pattern Mining method FOL-Mine can be enhanced to Incremental Mining of access patterns. Incremental Mining is more dynamic in nature and takes recent user accesses also into consideration. FOLMax-Mine, the new method presented for Maximal Access Patterns can be enhanced to Closed Access Pattern mining easily. Similarly, the proposed Weighted Access Pattern Mining method FWAP-Mine can be enhanced by considering other factors like access time and number of back access. The proposed Access Pattern Mining methods can be used for implementing new methods for Web Recommendation, Web Caching, Page Prefetching and User Clustering.