CHAPTER 7

CONCLUSION AND FUTUREWORK

7.1 CONCLUSION

The increasing popularity of the web has greatly attracted the web mining technology. A vital research area in web mining is web usage mining which mainly focuses on the discovery of patterns in the browsing and navigation data of web users. WUM has been a potential technology for understanding behavior of the user on the web. There are several techniques proposed by different researchers for the web usage mining. This thesis discussed about various techniques available for web usage mining. This thesis mainly discusses about three vital steps in WUM such as preprocessing, pattern discovery and pattern analysis. It is obvious that enhanced cluster recovery provides highly accurate guessing of a web user’s future visit if the user’s cluster can be exactly determined.

Data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found. The data cleaning phase implemented in this thesis will helps in determining only the relevant logs that the user is interested in.
The problem of web users clustering is to use web access log files to partition a set of users into clusters such that the users within a cluster are more similar to each other than users from different clusters are solved by using modified Fuzzy Possibilistic C-Means algorithm and it is compared with FCM algorithm. An important feature of the algorithm is that it preprocesses the data and divides the web users into clusters that can be used to classify future web users. Thus, the proposed FPCM approach is best suited for the web users clustering applications effectively.

As the outcome, the proposed approach accomplishes improved presentation in web log mining. It can be observed that the proposed FPCM offers better clustering outcome when match up with FCM. In addition the processing time of the expected FPCM is extremely low. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique and it reveals the number of records in web mining.

A decision theoretic rough set model can be used as the web mining component for a web based system. This extended model allows the component to provide additional decision support to the users. The proposed RFPCM is a hybrid methodology, which comprises a judicious integration of the principles of rough and fuzzy sets. The concept of crisp lower bound and fuzzy boundary of a class, which is introduced in the RFPCM, enables efficient selection of cluster prototypes. At last, the effectiveness of the proposed algorithm is established, along the with a comparison of FPCM algorithm on a set of synthetic data sets.
7.2 FUTURE WORK

Web usage mining is a powerful technique used to extract the information from past behavior of users. Data preprocessing plays an important role in this mining activity. The data preparation process is often the most time consuming and computationally intensive step in the web usage mining process, and often requires the use of special algorithms and heuristics not commonly employed in other domains. The future work involves developing autonomous agents that analyze the discovered rules to provide meaningful courses of action or suggestions to users. Future scope of web mining includes predicting user needs in order to improve the usability, scalability, user retention, and framing an efficient framework for web personalization through efficient use of web log file.

As a conclusion of the attainment of the study, more areas of research can be taking these structural concepts of web mining. This presents the fuzzy clustering and fuzzy possibilistic algorithm based on the web log. In order to recover the correctness and to decrease difficulty, fault and execution time in this research requires some upcoming improvement. Thus, a focus on web custom study techniques and architectures for more useful mixing and mining of substance, usage, and structure data from dissimilar sources is likely to direct to the next invention of more useful and more gifted applications, and more difficult tools for web usage mining that can obtain brainpower from user dealings on the web.