Summary and Conclusion

5. SUMMARY AND CONCLUSION

The rapid e-commerce growth has made both business community and customers face a new situation. Due to intense competition on one hand and the customer's option to choose from several alternatives business community has realized the necessity for intelligent marketing strategies and relationship management.

Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the WWW. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help to understand the user behavior and the web structure. From the business and applications point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-business, eservices, e-education, on-line communities and so on. Accurate Web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and find the most effective logical structure for their Web space. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure, and page content. However as the size and complexity of the data increases, the statistics provided by existing Web log file analysis tools may prove inadequate and more intelligent mining techniques will be necessary.

An important knowledge that can be obtained from web log files is the user's navigation pattern. The challenge in obtaining such knowledge is the users are constantly shifting their focus and different users have different navigational behaviour with different needs associated with them. The navigation pattern knowledge can be used to help users from getting lost in the cyberspace by predicting their future request.

In the present research a novel method to discover user's navigation process and predict future request is proposed. The system consists of an offline phase and an online phase. The offline phase performs web log file preprocessing, which transforms the raw log entries into a form which can be used by the subsequent stages. The preprocessed files are then clustered to find the navigation pattern. This knowledge is used as prior information in the second phase, (i.e.,) the online phase. The clustering method used in an ant-based clustering, where artificial ants act as agents, which do not communicate with each other but influence themselves through the configuration of objects on the floor. Thus, the agents construct groups of similar objects or construct clusters. In the online phase, a classification algorithm based on Longest Common Sequence algorithm is used. The main aim of this algorithm is to use the knowledge from offline stage and predict the users' next request.

Several experiments were conducted with weblog data collected from http://www.samplesite.com. The preprocessing of these raw data proved that around 88% of the raw data comprises of irrelevant entries, like image requests, JavaScript or css requests, secured transmissions, etc. Unique user identification can be achieved only when combining the user identification result with session identification. It was also found that approximately 40-50% of user entries were repeated without session identification. This might be due to the fact that the users spending more time in a web site are recorded in multiple entries in the web log file. Moreover, since user identification is performed through IP address values.

Clustering using ant-based algorithm was very efficient. From the clustered results, several important statistics like number of visits made to a webpage, the most popular web page among different users, common navigation pattern between users can be identified. Classification using LCS algorithm is proved to be effective in terms of discovering user navigation pattern and online prediction of future request. The performance of classification depends on the performance of the clustering technique as

classification works on the data which is built by the clustering algorithm. The accuracy of the prediction was calculated and it was found that around 80% of the times, the system was able to predict the future user request correctly.

FUTURE RESEARCH DIRECTIONS

The following directions can be used to improve the proposed system as it will help researches to find new kinds of WWW community patterns.

- The present work has been tested only with data obtained from one web server. Weblog files from different web servers can be obtained and analyzed.
- Future research can also try to combine clustering and association rules to discover more knowledge from the clustered data.
- Different clustering algorithms can also be investigated to improve the trend analysis and knowledge discovery.