CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENT

7.1. CONCLUSION

This research work focused on designing and developing web prediction algorithms to improve the browsing experience of users. The proposed algorithm was built in three steps namely preprocessing, clustering and classification. Users browsing details are stored in a special file called “weblog”, whose entries are auto-generated by the server. These files consist of “hidden asset” called “knowledge” that has become the most valuable resource to both web designers and online traders. Web Usage Mining (WUM) is a domain area that focuses on analyzing and predicting user’s experience/preferences using data mining techniques.

Analysis of user’s navigation pattern from weblog data is mainly used to discover the users’ access patterns and usage trends. Users behavioral pattern is used to identify the real world problems like developing websites, knowledge about the visitor’s behavior, product recommendation etc. The prediction of User Navigation Pattern was analyzed by various techniques like preprocessing, potential user identification, clustering and classification.
In Phase 1, Weblog file typically stores number of cumbersome activity. The preprocessing technique consists of cleaning the weblog data, path filling, user and session identification and clustering session. Cleaning of data removes the irrelevant entries in weblog files by discarding all images, video, audio entries, unsuccessful HTTP entry codes (Status code ≠ 200), and retaining only entries having GET and POST method.

The User Identification is performed using IP address. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a website. The unique user is identified in the session by accessing the website for more than 30 minutes. Session identification classifies the interested user in the website. Path completion is a process of appending the accessed page that are not contained in the weblog but those which be actually occurred.

The path filling method performs the path completion using an optimized 2-Way hash structure. This hash structure is used to represent user accessed page sequence using an Access History List (AHL). The 2-way hash structure is optimized to fill the path of only those users who are actually interested in the website. The final step of preprocessing is the clustering of the sessions using rough set clustering technique. The clustering algorithm is applied on the most probable user sessions to find the group of users accessing the web page of same interest. This preprocessing converts the usage, content, and structure information contained in the data sources into pattern discovery. In Phase 2, the potential users in the website are identified by using the rule based and interested measure algorithm which reduces the time spent on analyzing the robot entries.
The decision tree classifier namely improved C4.5 algorithm is used to identify the interested users. In a site, all users have different visited page based on their session. Each user have different browsing goals while accessing each page of a site and the users interest may vary from one page to another in each session. It is necessary to include an interest measure to identify the potential user in a session. Interest is calculated as the ratio of total amount of time spent on a page to the number of times a page was visited during a session. Potential users are those who have interest measure greater than threshold are identified as interested users by Rule based algorithm.

In Phase 3, the prediction system is composed of two steps namely clustering and classification. It ensembles three clustering algorithms (Ant based clustering algorithm, Improved Pairwise Nearest Neighbour algorithm and Graph partitioning algorithm) and three classification algorithms (Maximum Likelihood Classification Algorithm, Longest Common Sequence Classification Algorithm and Markov Model based Classification Algorithm). The majority voting algorithm is used in both clustering and classification techniques to select the navigation pattern by combining them to ensemble model.

This research work focuses on discovery of user navigation pattern and predicts the user’s future request in weblog mining environment. The experimental result has been carried out through Java programming language and SQL server database that permits to mine the knowledge from weblog data in an accurate and time efficient manner for improving the browsing experience of users. The experimental result evaluates the performance of proposed approach.
The experimental evaluation has been carried out through the dataset collected from the site www.microsoft.com. The proposed work is validated by the performance metrics like Accuracy, F1 measures, Coverage rate, and Memory allocations.

The accuracy of proposed preprocessing CD-PF-US-RST is 16.27% higher when compared to existing preprocessing algorithm. While considering the storage space saved, the preprocessed weblog data showed 1.35mb/sec in proposed preprocessing when compared with the existing system. In potential user identification, the accuracy of hybrid model produces a memory gain of 19.12% with respect to accuracy when compared to the existing one.

The ECLUCLA prediction algorithm showed 11.43% and 10.38% accuracy efficiency gain over ECLULCS and GPECLA respectively. The percentage of accuracy in the proposed method of ensemble clustering based ensemble prediction algorithm is 94.64%. The experimental results proved that the proposed amalgamation techniques are efficient both in terms of clustering and classification. The proposed ECLUCLA framework improved the web server performance, reduced the memory complexity, and predicted future user requests.

7.2. FUTURE ENHANCEMENT

In e-commerce, many types of users can access the website. But, this work focuses only on identification of potential users. In future, three
types of users can be identified from weblog files such as frequent users, Synthetic user and Potential user. It helps to improve the browsing behavior of user and predict their interest. Heuristic based user and session algorithm can be used to optimize the preprocessing procedure and produce the optimized and cleaned weblog. This improved technique can be implemented in online phase of the website.

WUM has become a well established field of research, where WUM processes can be applied to integrate the semantics within website design and aim to improve the results of WUM applications. Efforts in this direction seem to be a productive way to create much more effective WUM based systems that are consistent with the emergence and proliferation of the semantic web.

Furthermore, the web mining seems to be adequate for applying other mining techniques in parallel with the current algorithms. The system extensions for augmentative and detailed analysis could be the use of queries in one-dimensional representations, persistent data structures and multidimensional data structures in order to apply queries based on time.