CHAPTER 8

SUMMARY AND CONCLUSION

World Wide Web is the most useful media in today's information explosion era, for gathering, sharing and distributing information. E-Commerce is one of the areas where web-based technology is being adopted successfully. It provides various advantages like the support of huge number of activities that arise during a purchase or sale, advanced facilities that ensure successful communication and transaction, support different styles of purchase (sales), flexibility of time and place, ability to handle large number of customers easily and sharing and re-use of resources.

Today, almost all e-commerce websites aim to improve users' experience during a transaction, so as to bring satisfaction with their purchase (sale). The reason behind this is that a well-satisfied user uses the facility repeatedly, which consecutively will improve the business and profit. Users' experience involves a person's behavior, attitudes, and emotions about using a particular product, system or service. User Experience includes the practical, experiential, affective, meaningful and valuable aspects of human–computer interaction and product ownership. Additionally, it includes a person's perceptions of system aspects such as utility, ease of use and efficiency.

This research work analyzes these web log files to improve the browsing experience of the user, by predicting future page requests of the user. Usage of the proposed next web page prediction system will help the users to quickly and efficiently find their purchase interest and complete their business efficiently.

The proposed performs next web page prediction system consist of three main steps, namely, preprocessing, potential user identification and prediction of next page. The research work builds the prediction system by enhancing the

working of each of these steps in order to improve the overall system in terms of accuracy and speed. The research methodology is designed to treat each of the steps of the proposed system as a separate phase.

Phase I of the study focuses on preprocessing techniques to convert the raw web log data into a form that is more suitable for analysis. The techniques involved in this phase are cleaning of web log data, user identification and session identification.

Data cleaning is usually site-specific, and performs tasks like, removing unwanted references to embedded objects that are not important for the purpose of analysis. The research work removes all entries that contain (i) unwanted and redundant data (requests for images, javascripts, flash animations, video, etc), (ii) non-human accesses (robot entries) (iii) erroneous references (failed page requests).

The task of user identification is to identify the users of a web site and the set of pages accessed by the users. Unique users are identified from their IP address in this research work. The User' identity is not essential for web usage mining. However, it is mandatory to distinguish among different users.

The server log records may contain multiple entries for each user, since a user may visit a site more than once. The term user activity record is used to refer to the sequence of logged activities belonging to the same user. Session construction is the process of segmenting the user activity record of each user into sessions. Each entry represents a single visit to the site. The goal of a session construction is to re-construct, from the web log data, the actual sequence of actions performed by one user during one visit to the site. For this purpose, the usage of acyclic graphs is proposed. Four types of acyclic graph construction methods are analyzed. They are Directed Acyclic Graph (DAG), Hierarchical Directed Acyclic Graph-based (HDAG), Partial Ancestral Graph-based (PAG) and

Mixed Ancestral Graph-based (MAG). Using these methods, the access patterns of each user are identified.

Phase II of the research work, in order to reduce the size of web log data file, identifies potential users and prunes non-potential users. This task was performed in two steps, namely, attribute extraction and classification. Three types of attribute groups were extracted. They are temporal attributes, Page Attributes and Communication Attributes.

Temporal attributes extracted are access time, total session time and statistics such as the time a visitor accesses the site, the total time a visitor stays at the site and the different amounts of time a visitor stays on various pages. Three page attributes, namely, total number of accessed pages in a session, width and depth, were treated as page attributes. The third category selects access request attributes which have access methods as GET or POST. The selected attributes are discretized, using which the training dataset was created. Finally, to identify potential and non-potential users, the TSVM (Transductive Support Vector Machine) semi-supervised classifier was used. All the identified non-potential user details are pruned from the dataset, thus reducing the size of the dataset.

Predicting web pages purely based on user similarity is not very efficient as the browsing characteristics of the user may vary according to their emotions and behavior. To accommodate this in the proposed next web page prediction system, the final phase of the study proposes the use semantic web technology. The proposed prediction system, first uses an associative rule based semantic web usage mining algorithm that integrates human emotions and behaviors through self-reporting and behavioral tracking, for generating periodic access patterns.

The algorithm first creates a personal web usage lattice using user's web access activities and ontology. In this stage eight temporal concepts (Early Morning, Late Morning, Noon, Early Afternoon, Late Afternoon, Evening, Night,

and Late Night) as periodic attributes were used. The semantic web log was modified to include the timestamp along with the URL address. Using the same concepts, a global web usage (for all users) lattice was also generated. The web usage ontology was generated by combining the personal web usage lattice of a user with the global web usage lattice by using the concept of instance mapping. In this stage, instead of using all periodic sessions, to save time and to improve classification accuracy, a Particle Swarm Optimization algorithm, was used to select optimal session intervals. Finally, using fuzzy associative rules, usage pattern sequence for each user is generated. An associative classifier is then used to predict the future web page requests for the user.

The experiments were conducted in three stages to evaluate the performance of the Preprocessing, potential user identification and prediction tasks respectively. During experimentation it was found that the application of preprocessing steps significantly reduces the web log file size both in terms of number of transactions and file size. Comparison between the four session construction algorithms, the DAG algorithm worked efficiently in terms of all the performance metrics selected.

The experiments were conducted in three stages to evaluate the performance of the Preprocessing, potential user identification and prediction tasks respectively. The web log was obtained from Welcare Equipments (P) Limited, for a period of one year. During experimentation, to analyze the scalability property of the proposed algorithms, this web log dataset was grouped as small sized dataset (15 days, 1 month and 2 months data), medium sized dataset (3, 4 and 5 months data) and large sized dataset (6 and 12 months data).

During experimentation, the preprocessing algorithms were evaluated using performance metrics like size of log data before and after applying preprocessing steps, memory size of log data before and after applying preprocessing steps.

Analysis of the four selected acyclic graphs for session construction was performed using the number of session identified and speed of session construction. The potential user identification algorithm was analyzed using performance metrics like precision, recall, f-measure, accuracy and speed of identification. The same performance metrics was also used to evaluate the prediction of next web page algorithm.

The following facts were found from the experimental results analyzing the performance of the various steps of next web page prediction system.

- The application of preprocessing steps significantly reduces the web log file size both in terms of number of transactions (41.58% average efficiency gain) and file size (21.58% average efficiency gain).

- On average, the DAG algorithm performed session construction in ≤2.23 seconds, while HDAG, PAG and MAG algorithms took ≤2.40 seconds, ≤2.40 seconds and ≤3.49 seconds respectively.

- Comparison between the four session construction algorithms revealed that the performance of the DAG algorithm is efficient in terms of all the performance metrics selected, when compared HDAG, PAG and MAG algorithms.

- The usage of TSVM algorithm to identify potential users was also successful both in terms of accuracy and speed.

- On average, the TSVM algorithm, when compared with SVM algorithm, showed an efficiency gain of

    - 5.34% with SWL, 5.16% with MSL and 4.92% with LWL while considering Precision.

- 5.48% with SWL, 5.91% with MSL and 5.80% with LWL while considering Recall.
- 5.41% with SWL, 5.54% with MSL and 5.36% with LWL while considering F-Measure.
- 3.35% with SWL, 3.47% with MSL and 2.98% with LWL while considering Accuracy.

- On average, the proposed ontology and associative rule based prediction system showed an average efficiency gain, as given below, over existing algorithm
  - Precision - 5.15% (SWL), 5.32% (MSL) and 5.83% (LWL).
  - Recall - 5.10% (SWL), 5.40% (MSL) and 5.45% (LWL).
  - F-Measure- 5.13% (SWL), 5.36% (MSL) and 5.65% (LWL).
  - Accuracy – 5.05% (SWL), 5.24% (MSL) and 5.71% (LWL).

- The experimental results also showed that the inclusion of Particle Swarm Optimization algorithm improved the process of next page prediction.

It can be concluded that all the enhancement operations proposed for the development of next web page prediction system was successful and the positive results obtained shows that the proposed system can be used web masters and administrators for improving customer loyalty and customer relationship management.

8.1 FUTURE RESEARCH DIRECTIONS

The following can be considered in future to improve the algorithms proposed in this research work.

- TSVM algorithm for web log classification may be further enhanced to motivate the customers of e-commerce site increase their buying habit. A method may be proposed to automate the selection of the best set of training samples so that the system is trained well to predict labels for all possible combination of attribute values.

- For future work in OSIPSO, the gap between the systematic representation of preferences and customers' actual preferences may be decreased. A new method of preference selection from customers is projected for mapping the customer requirements and customer preference ontologies.