6 Conclusion

6.1 Summary

Rapid development of Web applications has led to vast amount of information available on the Internet. As a result, the WWW has become a very popular medium to disseminate, retrieve and analyse information. Consequently, users currently face the problem of information overload and in particular, finding the desired and exact data. The emergence of the Web has therefore put forth a host of challenges to Web researchers for information retrieval and management. Web mining offers a probable solution to the problems mentioned above by utilizing data mining methods to extract useful information from Web. Web mining, in turn, could be classified into three categories: Web content mining, Web structure mining and Web usage mining. This research work has concentrated on Web usage mining as an alternate method for not only deriving the usage profiles, but also, for making suitable tailor-made Web recommendations according to the users’ preferences.

The various stages of the research work can be summarized as follows:

- A Web server log typically maintains a history of page requests maintained in CLF or ELF. Information about the request, includes client IP address, request date/time, page requested, HTTP code, bytes served, user agent and referrer. These data can be combined into a single file or separated into distinct logs, such as an access log, error log, or referrer log. Since irrelevant items such as embedded images are stored in the Web server log and the identity of users and sessions are not apparent, preprocessing of the log files, involving various phases like data cleaning, user and session identification using heuristics, is mandatory. This results in the formation of a user session file. In some cases,
after preprocessing, the data available in the user session file may still not be in a suitable format on which various types of mining tasks need to be performed. Hence, in addition to preprocessing the log file, formatting of the data is essential. This formatting phase has been experimented using the data set available in http://www/cs.depaul.edu which consists of a sample of users visiting the site with details about the pages visited along with referrer page in each session. In this study as discussed in Chapter 2, the session file is formatted as a sequence of pages visited in each session by different users eliminating the need for the referrer page for the mining tasks.

In addition, processing of user session data such as data transformation tasks can be performed that would influence the quality of the discovered patterns resulting from the mining algorithms. The Web log files of msnbc.com Web site have been used for this research. This data set consists of sessions, each comprising page visits of anonymous users. Each session may contain the same page repeated consecutively. In this work, repetition of the same page visits consecutively has been suppressed with the assumption that the user might have kept the page open for a long time.

In this research work, Web usage data in the session file is characterized in a matrix model which reflects the Web user’s relative navigational preference/interest in each session. This framework makes it possible to methodically perform analysis on the collected Web usage data. As a result, this framework provides a mathematical base for discovering Web usage patterns and making Web recommendations. The transformation of user sessions into a usage data model referred to as session-pageview matrix has been performed. The session-pageview matrix has been normalized between [0,1], referred to as weighted session-pageview matrix, which signifies the relative interest/significance on the pageviews visited across each session. This serves as the input to the Pattern
Discovery phase.

- The Pattern Discovery phase deals with capturing sessions of “similar” browsing behaviour hidden in the usage data. This phase, as discussed in Chapter 4, has employed the probabilistic-based clustering approach such as EM algorithm to obtain transaction clusters from which the centroid of each cluster has been calculated. The centroids represent an aggregate view of the usage interest in the pageviews in each cluster, referred to as aggregate usage profile. Generally, the usage patterns are explicitly captured at the user transaction level or pageview level. They, however, do not reveal the underlying characteristics of user navigational activities for the purpose of analysis of clusters.

In our approach, a simple and effective method for cluster analysis/interpretation has been attempted by examining the centroids of the discovered profiles to determine the significance of each cluster. The centroids provide the ability to distinguish among pageviews in terms of the significance within the profile. The centroids also enable us to determine the cluster/profile that contains the maximum usage interest for each pageview.

With the help of the centroids, the discovered clusters can be evaluated using measures such as Precision, Recall and Purity. The **Precision** measure of a cluster with respect to a pageview is calculated by determining the ratio of the usage interest of each pageview in a cluster to the total usage interest of all pageviews in the corresponding cluster.

The **Recall** or modified Recall measure of a cluster with respect to a pageview finds out the ratio of the usage interest of a pageview in a cluster to the total usage interest of the corresponding pageview in all clusters. Since the relative usage interest is considered, it has been referred to as modified recall measure.

Hence, these measures represent the relative significance of a pageview within
a cluster and between clusters. In particular, in this study, the maximum usage interest of each pageview among all clusters is determined showing that the clusters are more focused in those pages having maximum usage interest, henceforth referred to as significant pages, than in the other pages. The quality of the clusters has also been assessed using the Purity measure. This measure identifies the pageview that has maximum usage interest in a cluster. On the other hand, purity value of each cluster signifies the pages that are more focused than the other pages in the corresponding cluster. The average purity value determines the quality of the clusters, higher the average purity, better the clustering result.

Weka tool has been used for determining clusters using the EM algorithm. The discovered clusters/patterns have been used for one of the various Web usage applications such as recommendations of Web pages.

- Having determined the clusters and the significant pages in each cluster in the offline phase, one of the important applications of Web usage mining is the identification of current interests of an online user and recommendation of the appropriate pages. This work has utilised the previously discovered aggregate usage profiles obtained using clustering in the offline phase. The online phase consists of obtaining matching profile(s) and recommendation of Web pages. As discussed in Chapter 5, the new active session consisting of the most recent ‘n’ pages visited by the user, termed as the sliding window, is taken into consideration. For each set of pages in the sliding window, the relative weights of the pageviews are determined automatically which signifies the usage interest and are compared to the already existing usage profiles using the well-known cosine similarity measure to obtain matching cluster(s). Clusters greater than the threshold value of 0.045 is taken into account. As the users browse through various Web pages, the window also slides and, correspondingly the appropriate matching clusters are determined. This is important and is taken care of in
the study since users’ interests are dynamic in nature. Therefore, in this study, matching the new users to the appropriate profiles is performed based on short-term, previous navigational visits in the current session rather at the end of the session. Also, it has been found that there is a similarity in the results obtained for matching clusters using the cosine similarity measure in Chapter 5 and the significant pages derived from the aggregate usage profiles in Chapter 4. This can be observed graphically which is represented in figure 5.2a, figure 5.2b and figure 5.2c and in comparison with the aggregate usage profiles as shown in table 4.3.

Once the matching profiles are determined, recommendation of Web pages need to be performed. Previous researchers have computed a score for each page in the matching profiles, called recommendation score, and pages whose recommendation score is above a predefined recommendation threshold have been chosen for recommendation. On the other hand, in this research, instead of computing the recommendation score, a new maximization function is used, as in equation 5.3, based on the aggregate usage profiles. The pages representing the maximum values, referred to in Chapter 5 as significant pages, can be selected for recommendation of Web pages. This has enabled the system in the immediate recommendation of unvisited pageviews which are likely to draw the interest of the user in the current session. Suitable measures such as Precision, Recall and F-measure have been employed to evaluate the recommendation set. A comparison in the evaluation measures obtained through the recommendation score and maximization function has been made. It has been shown graphically that recommendation accuracy is better when using the maximization function than the recommendation score in figure 5.3a, figure 5.3b and figure 5.3c.

Experiments for the entire work have been tested on publicly available UCI repository msnbc.com data set. The experimental results conducted on the real world datasets have
verified the effectiveness of finding better quality clusters and have an improvement in recommendation accuracy.

To conclude, it has been proved that through adequate processing of Web log files and user session files, it is possible to perform Web usage mining tasks resulting in a reasonable optimization of the clustering results and recommendation accuracy.

6.2 Possible Future Work

In this dissertation, we have concentrated on the research of Web usage mining for Web recommendation using data mining techniques. The theoretical and experimental studies have shown the effectiveness and applicability of the proposed models and approaches.

The future work can be continued along the following directions:

- Integration of semantic knowledge of Web pages using OWL (Web Ontology Language) into Web recommendation: The current research is mainly based on analysis of Web usage knowledge without considering other Web data sources. With the development of semantic Web and ontology research, it is believed that ontology knowledge of Web pages can provide deeper understanding of Web pages as a result of conveying the conceptual information. Ontology knowledge could be viewed as a high-level knowledge representation over the content knowledge. Hence, integrating the ontology knowledge with the usage knowledge will substantially improve the accuracy and efficiency of Web recommendations.

- Expanding the scope of current research to other associated areas: With the development of new applications over the Internet, especially Web 2.0 technology, many new types of Web data, such as email traffic, Web-blog and
Wiki pages are available. These data types have produced a large amount of new knowledge resources, which leads to new research directions, for example, social network analysis.