Abstract

With the Internet usage gaining popularity and the steady growth of users, the World Wide Web (WWW) has become a huge repository of data. The wealth of information, in turn, has lured users to seek and retrieve information from the Internet. In addition to it, Web users are interested in the extraction of potential, useful knowledge from the large amounts of Web data. Due to the nature of Web data, information retrieval becomes a complex task and Web users are generally drowned by the “tsunami of information” hitting them while accessing data on the Internet. The primary objective of this research is to mitigate this overload by providing ‘tailor-made’ responses. One of the approaches to achieve this, is the discovery of similarity in interests of Web users during their browsing sessions. Most often, it is also necessary to identify and capture the interests of new users which may vary dynamically during a session.

The fast pace of today and sheer magnitude of data available online make it imperative to use automated data recovery techniques. Web mining is a process that discovers the intrinsic relationships among Web data, which are expressed in the form of textual, linkage or usage information. Generally, Web users may exhibit various types of behaviours associated with their information needs and intended tasks when they are traversing the Web. Their behaviours are explicitly characterized by sequences of clicks on different Web items performed by them. Web usage mining is the application of data mining techniques to discover usage patterns from Web data generated by the user’s interaction with the Web.

To summarize, usage data that is collected when a user browses a specific Web site, represent the interaction between the user and the Web site. The complete process of Web usage mining broadly consists of two components: offline component and online component. The offline component is further divided into three phases: Data preprocessing and data transformation, Pattern discovery and Pattern analysis. The
data stored in the Web log files raise a great challenge for the users since it is available in an unsuitable format to extract useful information from the WWW. The preparation of a suitable data set is a pre-requisite and significant task for the mining of Web data. Web usage mining provides an approach to the collection and preprocessing of usage data resulting in the creation of a user session file wherein, each session consists of a sequence of user’s pageview requests. The session file is further formatted to conform to a data model representing the behaviour and the interests of users. The entire set of sessions can be represented as a usage matrix referred to as session-pageview matrix where the $i^{th}$ row represents the pages visited by the user in the session and the $j^{th}$ column represents the sessions in which the page $j$ has been visited. The cell value $(i, j)$ in the session-pageview matrix corresponds to the frequency of the pageview $j$ in session $i$.

The ability to extract useful knowledge hidden in the Web data and to act on that knowledge is an important strategic asset in today’s competitive world. Thus, the emergence of the Web has put forth a number of challenges for knowledge discovery and analysis. Web usage mining can be considered as a special case of Knowledge Discovery in Databases. Capturing clickstream data will be helpful to model and analyze the users’ browsing behaviour. This analysis requires the automatic discovery of meaningful patterns and relationships from a large collection of data stored in the user session file. Most Web usage mining systems use different data mining techniques in the pattern discovery phase, such as clustering, association rule mining and sequential pattern mining to extract usage patterns from navigational data. Nevertheless, the difficulty of assigning class labels to large quantities of data in supervised learning methods has resulted in the adoption of clustering methods. Most clustering algorithms are basically classified into Hierarchical clustering, Partitional clustering and Probabilistic clustering. With reference to the clustering algorithms, the probabilistic clustering algorithms have drawn more attention to research. In theory,
Probabilistic clustering approaches are able to determine the optimal number of clusters automatically. Expectation Maximization (EM) is a well-known algorithm used for clustering for finding the Maximum Likelihood Estimate (MLE) of the parameters of a mixture model. Therefore, EM clustering algorithm is applied to the transformed usage data to partition user sessions into transaction/session clusters.

Each transaction cluster may itself contain a large number of transactions, each consisting of several pageviews. However, these transaction clusters do not represent an aggregate view of the users’ interests on the Web pages. The final goal in clustering transactions is to provide the ability to analyze each cluster and generate an aggregate view so as to be helpful for various applications. Thus, the resulting usage patterns are transformed into aggregate usage profiles that provide an aggregate representation of the similar activities or interests of groups of users.

Moreover, the discovered patterns should be suitable for discovering useful information or knowledge. As a result, analysis of the discovered patterns through various techniques/tools is necessary to filter out uninteresting patterns. Interpretation and visualization of the discovered patterns is a major phase in Web usage mining process so that potentially useful information can be extracted and analysed.

The outcome of the offline component is the derivation of aggregate usage profiles using mining techniques. Experiments are performed on real-world data set to discover user access patterns and evaluate the clustering technique in terms of quality of clustering.

Nevertheless, the discovery and analysis of usage patterns obtained from Web data would not be useful unless they can be employed in various applications. The ultimate goal of any Web-based system is to cater to the needs or interests of a particular user or a segment of users. Hence, in the context of the Web, one of the applications of Web usage mining includes automatic recommendation/personalization. Specifically,
the goal of recommendation based on Web usage mining is to recommend a set of pageviews to the current, active user customized to the user’s perceived interests as determined by the matching usage patterns discovered through Web usage mining. Identification of the current interests of the user based on the short-term navigational patterns, instead of explicit user information, has proved to be one of the potential sources for recommendation of pages which may be of interest to the user.

The online component, consisting of two sub-phases, is responsible for matching the current user’s session to the aggregate usage profiles obtained through mining tasks and provide suitable page recommendations which may be of interest to the user. The aggregate usage patterns/profiles discovered through data mining techniques namely clustering provide input to the recommendation engine that recommends appropriate pages based on the intelligence gained from the usage profiles.

To determine the pageviews to be recommended, previous researchers have computed a recommendation score for each pageview based on the matching aggregate usage profiles. In this research work, a novel approach has been employed to identify significant pages which represent the user’s (session) interest in the cluster. The significant pages could be used for the purpose of recommendation of pageviews to the current user. To evaluate the effectiveness and efficiency of the recommendations of pages, experiments are conducted in terms of the proposed recommendation accuracy metric. It has been found that the proposed approach shows an improvement in the correctness and completeness of the recommendation set in comparison to the recommendation score approach.