Abstract

ABSTRACT

Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining techniques. It mines the secondary data (web logs) derived from the users' interaction with the web pages during certain period of Web sessions. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. In the present research work, a hybrid method is proposed, which uses the clustering and classification methods to find and predict user's navigation behaviour. The proposed system works in two phases, (i) the offline phase and (ii) the online phase. The offline phase takes care of preprocessing and clustering, while classification and prediction is performed during the online phase.

Preprocessing is the step which transforms the raw log file into a form that is more suitable for mining. Four steps are used during preprocessing, they are, (i) data cleaning (ii) user identification (iii) session identification and (iv) formatting.

The clustering method used is 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, Longest Common Sequence classification 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. Clustering of web data using ant-based algorithm proved to be very effective and several web usage statistics like most frequently viewed webpage, number of visits made to each page, which time of a day has the most traffic, common usage pattern, could be easily inferred.

Classification using LCS algorithm also proved to be effective in terms of discovering user navigation patterns and online prediction of future request.

The accuracy of the online phase on prediction was also calculated and it was found that around 80% of the times, the system was able to predict the future user request correctly. The results thus prove that the application of clustering and classification have positive impact during user navigation pattern discovery