ABSTRACT

The last decade has witnessed rapid growth and adaptation of the Internet. As a large community of users depends on the Internet for day-to-day services, an efficient and effective mechanism is required for information search and retrieval. The Internet contains a wide range of information and serves a range of communities of users. Keyword based search mechanism is mainly used to gather information. The major challenges are information mismatch, uselessness of the information and overload of information. This has a tremendous impact on the relevance of the information search and prompts for gathering knowledge and capturing information needs of end users. Concept based web personalization has played a major role to alleviate the problems. To improve the effectiveness of information personalization, the generation of user profile must be more efficient and accurate.

Many concept based user profiling approaches were proposed to capture users’ positive and negative preferences. Ontology based user profiling is a semantic approach to derive richer concept based user profiles. This research focuses on topic ontology construction to derive user profiles with user positive and negative preferences more independently and efficiently. Topic ontology is constructed based on user search queries or keywords. Interest score is assigned to topics in topic ontology and maintained in the user profile using Spreading Activation algorithm. User
interests are changed over a period of time. As per the change of user interest, the user profile will also be changed over time.

This work focuses on generation of profile based on two types. The first one is based on search activity, which takes search queries and accessed pages. According to the profile changes, the search engine is personalized by assigning interest scores to the topics. It gives weight to the topics. Session is introduced to capture the user browsing activities. User profile refers to the user interest in a specific search session. Therefore, interest score is activated to related topics and maintained in a same search session. The second is the generation of the personalized profile based on a wide range of Internet usage such as search queries, accessed pages, Wikipedia information, blogs, social-networking sites and local directory documents. These documents are collected to generate richer user profiles. Midstream was used to group all user information in order to perform the evaluation of search engine personalization. Top results are returned to users when a query is submitted to the midstream. Users click on the results which they find are relevant to their queries.

An experimental setup is designed to evaluate the performance of the proposed topic ontological profiling approach. User queries are taken as training set, and for each query in the training set, topic ontology construction process automatically organizes the topics based on their frequency. Then it is categorized into top $n$ relevant documents list, and the topic ontology hierarchy is developed. User queries or topic weights are updated in the training set using the activation scores. The test data consists of user
information from the Web and local directory. The user accessed information from the Web was collected as data set for over the period of three months.

The performance of the proposed topic ontological profiling approach is compared with the performance of the existing concept based profiling approaches. Recall and Precision are the measures that are used to summarize and match up to the user search results. The average precision and recall values are 0.9864 and 0.9632 respectively when the topic ontological user profile is used. These values are compared with average precision and recall values of \( P_{\text{Click} + \text{Joachims-C}}, P_{\text{Click} + \text{mJoachims-C}}, P_{\text{Click} + \text{SpyNB-C}} \) methods, and it is found that even though those methods captured both the positive and negative preferences of user, they gave slightly low precision and recall.

In order to improve the search quality and to facilitate efficient information and retrieval techniques, classification of clicked documents is needed. In this work, SVM classifier is proposed to classify clicked documents. Classification is performed by SVM classifier based on the semantic hierarchy in constructed topic ontology. An experiment is conducted and compared with Query Clustering method. The finest results with 97.06% of overall accuracy are obtained with Support Vector Machine whereas in the case of Query Clustering only 87% is obtained. It shows that the accuracy of classification using Support Vector Machine approach is better than the Query clustering method.