Journal of Computer Science 6 (7): 785-793, 2010
ISSN 1549-3636

An Application of Session Based Clustering to Analyze Web Pages of User Interest from Web Log Files

[1]C.P. Sumathi, [2]R. Padmaja Valli and [3]T. Santhanam
[1]Department of Computer Science, SDNB Vaishnav College, Chennai, Tamil Nadu, India
[2]Department of Computer Science, Mother Teresa Women's University Kodaikanal, Tamil Nadu, India
[3]Department of Computer Science, DG Vaishnav College, Chennai, Tamil Nadu, India

Abstract: Problem statement: With the continued growth and proliferation of e-commerce, Web services and Web-based information systems, the volumes of click-stream and user data collected by Web-based organizations in their daily operations have reached astronomical proportions. Analyzing such data can help these organizations optimize the functionality of web-based applications and provide more personalized content to visitors. This type of analysis involved the automatic discovery of usage interest on the web pages which are often stored in web and applications server access logs. **Approach:** The usage interest on the web pages in various sessions was partitioned into clusters such that sessions with "similar" interest were placed in the same cluster using expectation maximization clustering technique as discussed in this study. **Results:** The approach results in the generation of usage profiles and automatic identification of user interest in each profile. **Conclusion:** The significance of the results will be helpful for organizations for web site improvement based on their navigational interest and provide recommendations for page(s) not yet visited by the user.

Key words: Web usage mining, expectation maximization, usage profile, web page interest

INTRODUCTION

With the continual growth of Web-based information systems, click-stream (sequential series of pageview requests) data and user data are collected by Web-based organizations in their daily operations. The necessity to understand the large amount of data is inevitable in all fields of business, science and engineering. The ability to extract useful knowledge hidden in these data and to act on that knowledge is an important strategic asset in today's competitive world. Capturing click-stream data will be helpful to model and analyze the users' browsing behavior. This type of analysis requires the automatic discovery of meaningful relationships from a large collection of semi-structured data stored in the application server logs. The discovery of usage profiles automatically from click-stream data will enable organizations to improve their web site management and Web personalization by providing dynamic recommendations.

Web mining can be considered as a special case of the Knowledge Discovery in Databases (KDD) (Cooley *et al*., 1997; Srivatsava *et al*., 2000; Agarwal and Srikant, 1994; Mobasher *et al*., 2000b). Web usage mining deals with the automatic discovery and analysis of "interesting" patterns from clickstream and associated data collected during the interactions with Web server on one or more Web sites. To discover knowledge from the raw data, it is necessary to perform the following three steps:

- Data collection and preprocessing
- Pattern discovery
- Pattern analysis

This study focuses on the discovery of "similar" interests of groups of sessions based on the navigation behavior. To achieve this:

- First, a usage model is developed based on the browsing behavior in various sessions
- Using this model, we learn "similar" interests of groups of sessions. This is called as the aggregate usage profile
- Each usage profile consists of pages of varying user interest/significance
- In the proposed method, the significance of the pages in each profile is determined. These profiles would later help in various applications of web usage mining such as web site improvement, assigning a new user to the appropriate cluster and

Corresponding Author: C.P. Sumathi, Department of Computer Science, SDNB Vaishnav College, Chennai, Tamil Nadu, India

recommend pages of interest not yet visited by the user to provide personalized web content

Related work: A lot of research is being done in the area of Web Usage Mining. Based on the goals of the analyst and applications, various algorithms can be applied for cluster analysis. On the whole, clustering is the process of grouping the samples into clusters such that samples within a cluster have high similarity compared to each other but dissimilar to samples in other clusters. As mentioned in (Mobasher *et al*., 2000a), two types of clustering can be performed on usage data: Transactions Clusters and Page Clusters. Each type of clustering is helpful in different applications such as personalization and recommendation, system improvement, web site structure, business intelligence and user behavior. Similarity measures form the core component for every clustering algorithm.

For several years, focus on cluster analysis has been mainly distance-based cluster analysis. There are various distance-based similarity measures such as the Euclidean distance measure, Manhattan distance, Minkowski, Mutual Neighbor Distance (MND), Simple Matching Coefficient, Jaccard Coefficient and Rao's coefficient. The usage of these similarity measures depends on the features of the samples. However, in (Chaofeng, 2009), a Sequence Alignment Method has been used for measuring similarities between web pages by considering the URL and the viewing time of the URL. The proposed algorithm for Web Session Clustering Based on Increase of Similarities (WSCBIS) has been implemented along with k-means clustering and Robust Clustering using linKs (ROCK) proving the decrease in time and space complexity. Research regarding clustering of URLs using Sequence Alignment Method has also been done in (Hay *et al*., 2004). In this study, Hay *et al.* (2004) have clustered web users using two different similarity measures: SAM (non-Euclidean distance-based measure) and Association measure (Euclidean distance-based measure). The sequential order of pages is taken into consideration and not the position of the pages. Such sequences are called open sequences. As mentioned in (Hay *et al*., 2004), sequences with the same elements occurring in the same order and irrelevant of the positions of the elements are called open sequences. For example, the open sequence (1, 3, 5) occurs in the sequences (4, 1, 2, 3, 6, 5), (1, 2, 3, 4, 2, 5) and (3, 1, 3, 5, 2). Unlike most research, where users are grouped into cluster with similar pages, it was proved that SAM retrieves sequences not only with similar pages but the order of pages is also considered compared to the associative measure which is Euclidean-distance based. Hence users are clustered based on their sequential order of web navigation.

Stochastic methods have been proposed for clustering user transactions for the purpose of user modeling. Since each user may reveal different types of navigation behavior, the patterns should also capture the overlapping interests of these users.

Mixture models are able to capture complex, dynamic user behavior (Cooley *et al*., 1997). To determine the user behavior in web usage mining systems, (Mustapha *et al*., 2009) deals with model-based clustering method using Expectation-Maximization (EM) algorithm which is an extension of k-means algorithm. EM algorithm is used for finding the parameter estimates in probabilistic models. The EM algorithm has been compared with the k-means algorithm and the results showed an improvement in the accuracy of the algorithm. A variant of the Model-Based Clustering has been done in (Pallis *et al*., 2005) in which the interpretation and visualization of model-based clustering schemes using the concept of Correspondence Analysis (CO-AN) has been done. User sessions are clustered using the first-order Markov model using the EM algorithm. CO-AN is a multi-variate statistical analysis method to interpret and visualize Web users' navigation patterns. This is helpful for commercial Web sites to understand the customer behavior and provides scope for site improvement.

A similar research on model-based clustering has been done in (Igor *et al*., 2003) based on first-order Markov model and a using a visualization tool, Web Canvas. The results have shown that learning time scales linearly with sample size using model-based clustering compared to agglomerative distance-based methods in which the learning time scales quadratic ally with sample size. In addition to the discovery of navigation patterns, prediction of future navigation behavior has been included in (Borges and Levene, 2008). Different scoring metrics such as the hit and miss score, the mean absolute error and the ignorance score have been employed to determine the quality of prediction. In (Lee and Fu, 2008a), two levels of prediction of users' browsing behavior have been proposed. Using Markov Model, browsing behavior is predicted at the category level and using Bayes Theorem, prediction is done at the web page level. A combination of Markov model and Bayes theorem results in a two-level prediction of user's browsing behavior. The results proved that the hit ratio is effective and accurate in both the levels. An extension of (Lee and Fu, 2008a) has been dealt with in (Lee and Fu, 2008b) in which the overlapping or heterogeneous

nature of user's behavior and improvement in hit ratio has been considered.

Fuzzy Relational Clustering Algorithms have been applied for web usage mining (Krishnapuram *et al*., 2001; Labroche, 2007). Clustering of relational data using fuzzy approaches has been implemented using Fuzzy C-Medoids (FCMdd) and Robust Fuzzy C-Medoids (RFCMdd) in (Krishnapuram *et al*., 2001). These algorithms have been applied in Web Usage Mining for discovering user profiles. Similar research has been performed in (Labroche, 2007) for discovery of user profiles using Ant clustering algorithm and a linear version of Fuzzy c-Medoids.

Another approach to observing path traversal and clustering based on that data is advanced by Shahabi *et al.* (1997). The basic approach there is to define a path similarity measure for a given Web site. Then, the logged data about a user's paths is clustered using a simple K-means algorithm to aggregate users into groups. However, it is not clear how the similarity metric is devised and whether it can produce meaningful clusters. Approaches, essentially based on the association rule ideas (Agarwal and Srikant, 1994; Etzioni, 1996), have been proposed in (Cooley *et al*., 1997). However, these approaches assume that logs contain user IDs, which is not common in the real world except in the rare case that the ident protocol is used and the clients are agreeable to release the user names. A related topic that has been recently gaining momentum is the idea that we can learn much about users and customers by tracking and analyzing their clickstreams, which is of great importance in e-commerce.

MATERIALS AND METHODS

Data preprocessing: An important task in any web mining application is the collection of target data set to which mining techniques can be applied. Data preprocessing is a pre-requisite phase and time-consuming process before the data can be mined to obtain useful and interesting patterns. As mentioned in (Suresh and Padmajavalli, 2006; Cooley *et al*., 1997; Srivatsava *et al*., 2000) data preprocessing involves data cleaning, user identification, session identification resulting in the creation of a user session file. A user session file is a collection of pageviews grouped by user sessions. Each user session can be considered in two ways: (i) a single transaction of many page references or (ii) a set of transactions each consisting of a single page reference. In this study, each session is considered as a single transaction consisting of a set of pageviews navigated by a user during a single visit to the site. Thus the session file consists of a sequence of user's request for pages $P = \{p_1, p_2, p_3,\ldots,p_n\}$and a set of m sessions, $S = \{s_1, s_2, s_3,\ldots,s_m\}$ where each $s_i \in S$ is a subset of P.

Example: 2 1 3 2 1 represents a session consisting of a sequence of page requests.

Conceptually, each page is associated with a weight representing its significance. The weights can be determined in various ways depending on the type of analysis. In most Web Usage Mining tasks, the weights may be based on a combination of factors such as the time that the user has spent on a page visited, number of visits to the page and size of the page.

Consider the web as a directed graph G, where a node p_i represents a web page visited and the edge e_i represents the successive linear path to p_i followed by the user. A session can be represented graphically as in Fig. 1.

In the context of web usage mining, a modified version of the Pagerank algorithm is used for assigning weights to the pages of a session based on their navigational behavior. Since it is a linear path (no parallel paths) each edge has a weight of 1. Hence the weight of a node p_i is the sum of the weights of the in degree of the node p_i:

$$W(p_i) = \sum w(e_i)$$

Hence, a session-pageview matrix is obtained. Each row represents a session and each column represents a frequency of occurrence of the pageview in the session. This is represented in Table 1 in which the first row represents the page id.

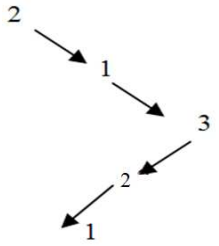

Fig. 1: Page traversal graph in a session

Table 1: Session-pageview

1	2	3	4	5	6
3	1	0	2	1	1
0	0	1	0	2	3
1	0	0	3	0	0
0	4	3	0	0	0
1	0	0	0	3	0
0	2	0	0	0	0
2	0	0	2	0	2

Table 2: Snapshot of the weights of the pages visited

1	2	3	4	5	6
0.166	0.000	0.000	0.000	0	0.166
0.545	0.090	0.000	0.000	0	0.000
0.375	0.000	0.000	0.250	0	0.000
0.400	0.000	0.000	0.000	0	0.400
0.0763	0.000	0.615	0.076	0	0.000

In our approach, the weight of the pageview is further determined by evaluating the importance of a page in terms of the ratio of the frequency of visits to the page with respect to the overall page visits in a session. A numerical weight is assigned to each pageview visited with the purpose of "measuring" its relative importance/interest within the session. If the page has not been visited, the weight of the page is assigned 0. The page visits repeated consecutively have been treated as a single visit to that page. The weights have been normalized to account for variances.

The above session file is represented by using the vector space model. Each session s_i is modeled as a vector over the n-dimensional space of pageviews. Each session s_i is represented as:

$$s_i = \{pf_1, pf_2, pf_3, \ldots, pf_n\}$$

where, each pf_j is the relative frequency of pageview j in session i. This type of weight normalization is referred to as transaction normalization which is beneficial since it captures the relative importance/interest of the pageview in a session.

Example 1: Suppose there are 17 pages and the user has visited page 1 twice (Table 2), page 2 twice and none of the other pages in session 1, then s_1 = 2/4, 1/4, 0, 0, 0,……,0.

Pattern discovery-model-based session clustering: Let S = $\{s_1, s_2, s_3, \ldots, s_m\}$be a set of m objects where each object is represented by a vector of pageviews. Our goal is to obtain sessions with "similar" interests. In contrast to the other clustering methods such as partitional clustering, hierarchical clustering wherein the similarity measure is distance-based, model-based clustering employs probability-based approach. The basis of the probability-based clustering approach is based on finite mixture model. Mixture models are able to capture more complex, dynamic user behavior.

A mixture is a set of k probability distributions, each of which governs the attribute values distribution of a cluster. Each individual distribution is referred to as a component distribution following a normal distribution. Each cluster is represented by a probability model. Model-Based clustering methods optimize the fit between the given data and a mathematical model.

A popular model-based clustering method is the Expectation-Maximization (EM) algorithm based on Bayesian probability theory which is an extension of the k-means algorithm. The Expectation Maximization algorithm is an efficient iterative method to determine the Maximum Likelihood (ML) estimate when missing or hidden data exists.

The EM algorithm is an iterative refinement algorithm that can be used for finding the mean and standard deviation parameter estimates. The expectation maximization algorithm assigns each object to a cluster according to a weight representing the probability of membership. Therefore, new means are computed based on weighted measures.

Each iteration of the EM algorithm (Han and Kamber, 2006) consists of two processes:

- E-step-Each object x_i is assigned to cluster C_k based on Bayesian probability. This is achieved using the conditional expectation. Assign:

$$P(x_i \in C_k) = p(C_k \mid x_i) = p(C_k)p(x_i \mid C_k) / p(x_i)$$

  The missing data are estimated given the observed data and current estimate of the model parameters. The probability of cluster membership of object x_i, for each of the clusters is calculated. These probabilities are the "expected" cluster membership for object x_i
- M-step-Assuming that the missing data are known, the likelihood function is maximized. The probability estimates from the E-step are used to re-estimate the model parameters:

$$m_k = (1 / n\sum_{i=1}^{n} x_i P(x_i \in C_k)) / \sum_i P(x_i \in C_j)$$

This step is the "maximization" of the likelihood of the distributions given the data.

It iterates until the parameters reach a stable, convergence point or until the Maximum Likelihood estimate reaches the maximum. The essence of the EM algorithm is that for every iteration, maximizing the conditional expectation leads to an increase of the log likelihood of the observed data for each iteration i. This determines the number of clusters.

RESULTS

Pattern analysis: The web log files of msnbc.com web site have been used for this research. This data set is publicly available through the UCI KDD Archive (2005) at the University of California. The web site includes the page visits of users who visited the "msnbc.com" web site on 28/9/99. The visits are

recorded at the level of URL category (for example sports, news and so on) and are recorded in time order. It includes visits to 17 categories (i.e., 17 distinct pageviews). The data is obtained from IIS logs for msnbc.com and news-related portions of msn.com. The client-side data is not available in the web log files. Each sequence in the dataset corresponds to a user's request for a page.

The 17 categories are:

Id	Category	Id	Category
1	Frontpage	10	Living
2	News	11	Business
3	Tech	12	Sports
4	Local	13	Summary
5	Opinion	14	BBS
6	On-air	15	Travel
7	Misc	16	Msn-news
8	Weather	17	Msn-sports
9	Health		

Example:

```
1 1
2
3 2 2 4 2 2 2
```

The above is a sequence of visits. Each record is a session. The first row indicates that the user has visited category 1 twice. The second row indicates that a user has visited category 2 once. The third row indicates that the user visited category 3 once, category 2 visited consecutively, then visited category 4 once and finally visited category 2 consecutively three times. About 10000 sessions have been selected randomly for this experiment. A portion of the dataset is as follows:

```
1 6 1
1 6 11
1 11 1 11 1 14 1 12 1 2 1
8 1
1 7 1
4 4
```

After suppressing the page visits repeated consecutively in a session (Table 3) using shell script in Linux, the sample dataset is as follows:

```
1 6 1
1 6 11
1 11 1 11 1 14 1 12 1 2 1
8 1
1 7 1
4
```

Table 3: Frequency of page visits in each session

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	2	1	0	1	0	0	0
1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Each page has been given a numerical weight for each session. This indicates the relative importance of each page in the session. If the page has not been visited, its weight is 0. This is represented in Table 4.

Weka tool has been used for the experimental evaluation. A model is estimated from the available samples in the dataset which are generally split into training and testing sets. The model is first designed using training samples and then it is evaluated based on the performance on the test samples. In the proposed approach, the dataset has been partitioned into 60% of training data and the remaining 40% as test data. The Expectation Maximization clustering algorithm has been applied. The experiment was performed within 10 iterations resulting in 9 clusters with a maximum likelihood estimate of 24.95867. Each cluster represents sessions of "similar" interest in the web pages or the usage profile. Hence an aggregate usage profile is determined using the formula:

$$Wt(p, up_c) = \sum_{s \in c} w_p^s / nc$$

Where:

w_p^s = The weight of the page in session s€c

nc = The number of sessions in cluster c

This is represented in Table 5.

Evaluation methods for session clustering: The Precision P(i,j), Recall R(i,j) and Purity evaluation measures of each cluster j for each web page i are calculated.

The Precision Measure is given by:

$$P(i, j) = w_{ij} / \sum_{i=1}^{k} w_i$$

where, w_{ij} represents the aggregate weight(user interest) of page i in cluster j. Since the weights have been normalized between [0,1], the weight of the cluster $\sum w_j$ is always equal to 1. Hence Table 5 also represents the Precision Measure of page i in cluster j.

Table 4: Session pageweight

F-page	News	Tech	Local	Opine	Misc	Weather	Health	Living	Businesss	Sports	Summery	Btin	Travel	Msn-N	Msn-S
0.667	0.000	0.000	0	0	0.33	0.00	0.0	0	0.000	0.00	0	0.00	0	0	0
0.333	0.000	0.000	0	0	0.33	0.00	0.0	0	0.333	0.00	0	0.00	0	0	0
0.545	0.091	0.000	0	0	0.00	0.00	0.0	0	0.182	0.09	0	0.09	0	0	0
0.5	0.000	0.000	0	0	0.00	0.00	0.5	0	0.000	0.00	0	0.00	0	0	0
0.667	0.000	0.000	0	0	0.00	0.33	0.0	0	0.000	0.00	0	0.00	0	0	0
0	0.000	0.000	1	0	0.00	0.00	0.0	0	0.000	0.00	0	0.00	0	0	0

Table 5: Aggregate usage profile

Cluster	Front page	News	Tech	Local	Opinion	On-air	Misc	Weather	Health	Living	Business	Sports	Summery	B'tin	Travel	Msn-N	Msn-S
C0	0.024	0.003	0.002	0.002	0.001	0.005	0.000	0.946	0.001	0.006	0.006	0.002	0.000	0.001	0.000	0.000	0.001
C1	0.146	0.075	0.010	0.008	0.084	0.028	0.003	0.001	0.002	0.238	0.221	0.009	0.003	0.002	0.112	0.001	0.058
C2	0.000	0.362	0.040	0.028	0.000	0.107	0.047	0.004	0.000	0.001	0.011	0.000	0.321	0.061	0.017	0.000	0.001
C3	0.004	0.000	0.000	0.000	0.001	0.003	0.001	0.007	0.000	0.005	0.000	0.472	0.004	0.502	0.001	0.000	0.000
C4	0.094	0.015	0.736	0.047	0.003	0.022	0.007	0.000	0.009	0.047	0.005	0.008	0.000	0.000	0.007	0.000	0.000
C5	0.061	0.008	0.044	0.065	0.035	0.021	0.037	0.035	0.563	0.000	0.018	0.054	0.035	0.009	0.013	0.000	0.000
C6	0.653	0.089	0.016	0.021	0.000	0.020	0.021	0.000	0.022	0.004	0.030	0.067	0.009	0.049	0.000	0.000	0.000
C7	0.028	0.012	0.011	0.000	0.000	0.884	0.000	0.037	0.018	0.002	0.000	0.007	0.000	0.002	0.000	0.000	0.000
C8	0.045	0.029	0.006	0.635	0.000	0.038	0.080	0.076	0.000	0.012	0.000	0.009	0.016	0.053	0.000	0.000	0.003

Table 6: Recall Measure of each cluster for each page

Cluster	Front page	News	Tech	Local	Opinion	On-air	Misc	Weather	Health	Living	Business	Sports	Summery	B'tin	Travel	Msn-N	Msn-S
C0	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.86	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.01
C1	0.14	0.13	0.01	0.01	0.67	0.02	0.01	0.00	0.00	0.75	0.76	0.02	0.01	0.00	0.74	1.00	0.93
C2	0.00	0.61	0.05	0.03	0.00	0.10	0.24	0.00	0.00	0.00	0.04	0.00	0.83	0.09	0.12	0.00	0.01
C3	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.02	0.00	0.75	0.01	0.74	0.00	0.00	0.00
C4	0.09	0.02	0.85	0.06	0.03	0.02	0.03	0.00	0.01	0.15	0.02	0.01	0.00	0.00	0.05	0.00	0.00
C5	0.06	0.01	0.05	0.08	0.28	0.02	0.19	0.03	0.92	0.00	0.06	0.09	0.09	0.01	0.09	0.00	0.00
C6	0.62	0.01	0.02	0.03	0.00	0.02	0.11	0.00	0.04	0.01	0.10	0.11	0.02	0.07	0.00	0.00	0.00
C7	0.03	0.02	0.01	0.00	0.00	0.78	0.00	0.03	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
C8	0.04	0.05	0.01	0.79	0.00	0.03	0.41	0.07	0.00	0.04	0.00	0.01	0.04	0.08	0.00	0.00	0.04

Table 7: Purity values of each cluster

Cluster	Purity
C0	0.95
C1	0.24
C2	0.36
C3	0.50
C4	0.74
C5	0.56
C6	0.65
C7	0.88
C8	0.64
Average purity	0.61

The modified Recall measure is given by:

$$R(i, j) = w_{ij} / \sum_{j=0}^{n} w_j$$

This is represented in Table 6. Purity is a simple and transparent evaluation measure. It represents the portion of the cluster corresponding to the largest aggregate weight of the page with respect to the cluster:

$$Purity(j) = \max(w_{ij}) / \sum_{i=1}^{k} w_i$$

This is shown in Table 7.

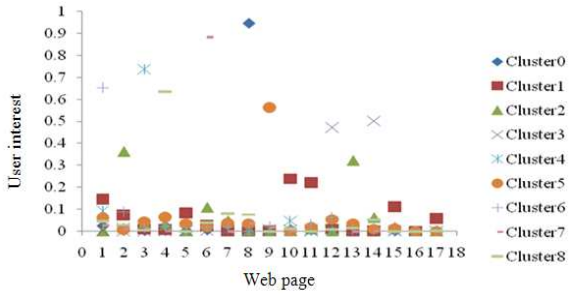

Fig. 2: Interest on the web pages in clusters

The average purity of the clustering is the ratio of sum of the purity values of all the clusters to the total number of clusters. It is found to be 61%. The larger the purity result, the better is the performance of the clustering result.

Interpretation of clusters: In Table 3, each cluster represents the cluster centroid. The centroid represents the mean values of the web pages contained in each cluster. The aggregate usage profile is represented in Fig. 2 which shows the user interest on the web pages. The centroids enable us to describe each cluster by assigning it a name. Among all the clusters, using the

Recall measure of each cluster for each web page shown in Table 7, we observe that the maximum recall of web page1 (FrontPage) is present in Cluster6. Similarly, in web page 2, it occurs in Cluster2. For web page 3, the maximum rating occurs in Cluster4. For web page 8, we find the maximum rating appears in Cluster0. From these observations, we infer that sessions in Cluster0 are interested in obtaining information about weather. Hence Cluster0 can be labeled as "Weather". This is represented in Fig. 3. However, sessions in Cluster1 indicate that users are randomly scanning the web pages. Hence Cluster1 may be labeled as "Random Surfers" as depicted in Fig. 4. Cluster2 is characterized by interest in web pages (News) and 13(Summary) as shown in Fig. 5. Hence cluster2 can be labeled as "News". Cluster3 is focused on web page 12 and 14. Hence this cluster can be labeled as "Sports" as shown in Fig. 6. Cluster4 can be labeled as "Technology" shown in Fig. 7. Cluster5 can be labeled as "Opinion" as represented in Fig. 8.

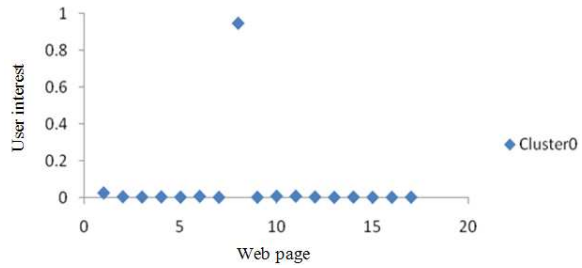

Fig. 3: Cluster0

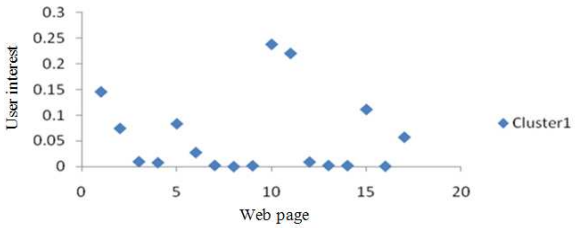

Fig. 4: Cluster1

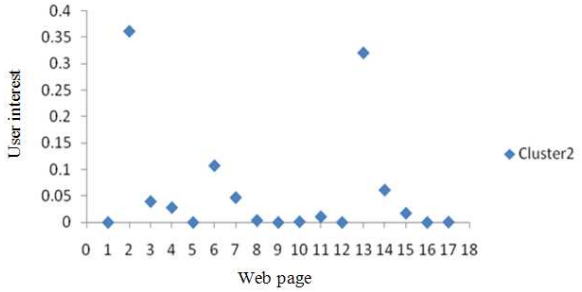

Fig. 5: Cluster2

Cluster6 is focused on "Frontpage" as seen in Fig. 9. Cluster7 can be labeled as "On-air" (Fig. 10). Cluster8 can be labeled as "Miscellaneous" (Fig. 11). A visual representation of the users' interest on the web pages is shown in Fig. 2-9. From this one can easily conclude that the user interests are either uni-focused (for instance Cluster0) or multi-focused (Cluster1).

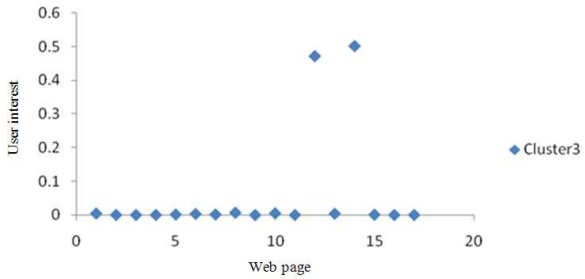

Fig. 6: Cluster3

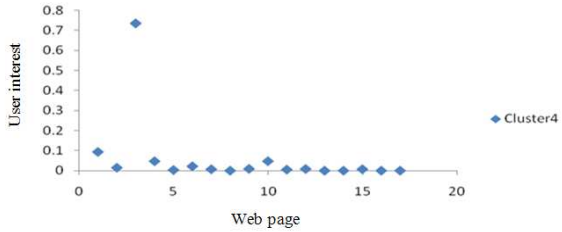

Fig. 7: Cluster4

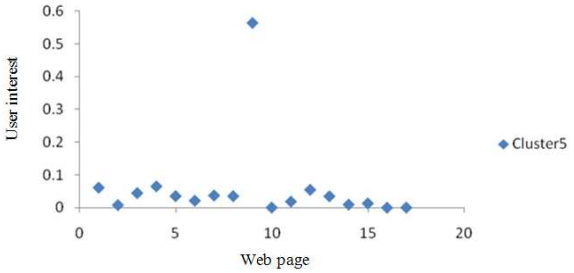

Fig. 8: Cluster5

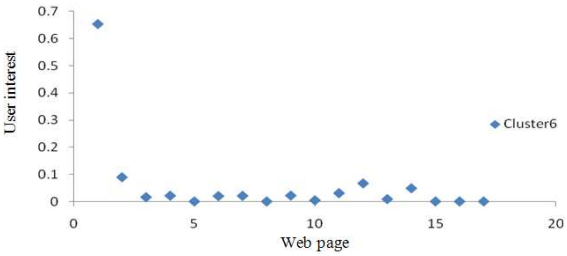

Fig. 9: Cluster6

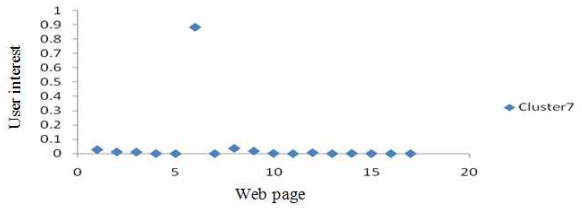

Fig. 10: Cluster7

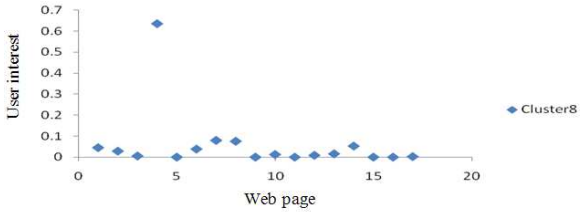

Fig. 11: Cluster8

DISCUSSION

Data preprocessing has been extensively done in this work with respect to user's interest in each session. It has been widely reported in the literature regarding the application of Expectation Maximization algorithm in the context of web usage mining domain and the use of statistical techniques to draw conclusions. This article deviates from the conventional statistical methods to derive the interpretations and employs the Aggregate Usage Profile and a modified Recall measure for analyzing the user interest in various clusters. Further cluster evaluation techniques like Precision and Purity have also been adopted.

CONCLUSION

In this study, the authors have suggested a probabilistic-based approach for grouping web usage transactions and generated user profiles. First, we mapped the frequency of page visits to the relative user interest within a session and obtained a weighted session-pageview matrix. Then the model-based Expectation Maximization clustering algorithm is employed to generate clusters or usage profiles. The aggregate usage profile has been used to analyze user interest in the web pages. Experiments are done on real world data set to discover the user interest in the web site. The significance of the results will be helpful for organizations for web site improvement based on their navigational interest and provide recommendations for page(s) not yet visited by the user. The future research will focus on using other methods for grouping web pages of user interests.

REFERENCES

Agarwal, R. and R. Srikant, 1994. Fast algorithms for mining association rules in large database. Proceeding of the 20th Conference on Very Large Data Bases, Sept. 12-15, Morgan Kaufmann Publishers Inc., San Francisco, CA., USA., pp: 487-499. DOI: 10.1234/12345678

Borges, J. and M. Levene, 2008. Mining users' web navigation patterns and predicting their next step. NATO Secur. Sci. Ser. D-Inform. Commun. Secur., 15: 45-55. http://direct.bl.uk/bld/PlaceOrder.do?UIN=229755922&ETOC=RN&from=searchengine

Chaofeng, L., 2009. Research on web session clustering. J. Software, 4: 460-468, DOI: 10.4304/jsw.4.5.460-468

Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence, Nov. 3-8, Newport Beach, CA., pp: 558-567. DOI: 10.1109/TAI.1997.632303

Etzioni, O., 1996. The world wide web: Quagmire or gold mine? Commun. ACM, 39: 65-68. DOI: 10.1145/240455.240473

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, San Francisco, CA., ISBN: 978-1-55860-901-3, pp: 770.

Hay, B., G. Wets and K. Vanhoof, 2004. Clustering navigation patterns on a website using a sequence alignment method. Knowl. Inform. Syst., 6: 150-163. DOI: 10.1007/BF02637153

Igor, V.C., D. Heckerman, C. Meek, P. Smyth and S. White, 2003. Model-based clustering and visualization of navigation patterns on a web site. Data Min., Knowl. Discov., 7: 399-424. DOI: 10.1023/A:1024992613384

Krishnapuram, R., A. Joshi and O. Nasraoui, 2001. Low-complexity fuzzy relational clustering algorithms for Web mining. IEEE Trans. Fuzzy Syst., 9: 595-607. DOI:10.1109/91.940971

Labroche, N., 2007. Learning web users profiles with relational clustering algorithms. Association for the Advancement of Artificial Intelligence, pp: 54-64. http://www.aaai.org/Papers/Workshops/2007/WS-07-08/WS07-08-007.pdf

Lee, C.H. and Y.H. Fu, 2008a. Two levels of prediction model for user's browsing behavior. Proceeding of the International Multi Conference of Engineers and Computer Scientists, Mar.19-21, National Science Council, Hong Kong, pp: 751-756. http://www.iaeng.org/publication/IMECS2008/IMECS2008_pp751-756.pdf

Lee, C.H. and Y.H. Fu, 2008b. Web usage mining based on clustering of browsing features. Proceeding of the 8th International Conference on Intelligent Systems Design and Applications, Nov. 26-28, IEEE Computer Society, Washington DC., USA., pp: 281-286. DOI: 10.1109/ISDA.2008.185

Mobasher, B., H. Dai, T. Luo, M. Nakagawa, Y. Sun and J. Wiltshire, 2000a. Discovery of Aggregate Usage Profiles for Web Personalization. Webkdd, Boston, USA, pp: 1-11.

Mobasher, B., R. Cooley and J. Srivatsava, 2000b. Automatic personalization based on Web usage mining. Commun. ACM, 43:142-151. DOI: 10.1145/345124.345169

Mustapha, N., M. Jalali and M. Jalali, 2009. Expectation maximization clustering algorithm for user modeling in web usage mining systems. Eur. J. Sci. Res., 32: 467-476. http://www.eurojournals.com/ejsr.htm

Pallis, G., L. Angelis and A. Vakali, 2005. Model-based cluster analysis for web users sessions. Lecture Notes Comput. Sci., 3488: 219-227. DOI: 10.1007/11425274_23

Shahabi, C., J. Abidi, A. M. Zarkesh and V. Shah, 1997. Knowledge discovery from users web-page navigation. Proceeding of the 7th IEEE International Workshop Research Issues in Data Engineering, Apr. 7-8, IEEE Computer Society, Birmingham, pp: 20-29. DOI: 10.1109/RIDE.1997.583692

Srivatsava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorat. Newsletter, 1: 12-23. DOI: 10.1145/846183.846188

Suresh, R.M. and R. Padmajavalli, 2006. An overview of data preprocessing in data and web usage mining. Proceeding of the IEEE International Conference on Digital Information Management, Dec. 2006, IEEE Computer Society, USA., pp: 193-198. DOI: 10.1109/ICDIM.2007.369352

UCI KKD Archive, 2005. UCI Knowledge Discovery in Databases Archive. http://kdd.ics.uci.edu/