Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 58 (2015) 558 – 564

Procedia Computer Science

Second International Symposium on Computer Vision and the Internet (VisionNet'15)

Clustering Of Web User Sessions To Maintain Occurrence Of Sequence In Navigation Pattern

Anupama D. S.[a], Sahana D. Gowda[b]

[a]PG Student, Dept. of CSE, BNMIT,Bengaluru-560070, India
[b] Professor, Dept. of CSE, BNMIT, Bengaluru - 560070, India

Abstract

Web log data available at server side helps in identifying the most appropriate pages based on the user request. Analysis of web log data poses challenges as it consists of abundant information of a web page. In this paper a novel technique has been proposed to pre-process the web log data to extract sequence of occurrence and navigation patterns helpful for prediction. Each URL in the web log data is parsed into tokens based on the web structure. Tokens are uniquely identified for the classification of URLs. The sequence of URLs navigated by a user for a period of 30 minutes is treated as a session. Session represents the navigation pattern of a user. Sessions from multiple users are clustered using hierarchical agglomerative clustering technique to analyze the occurrence of sequence in the navigation patterns. From each cluster, a session is identified as a representative as it holds most possible pages in the sequence, other sessions in the cluster are the subset of the representative session. Session representative navigation patterns are useful for predicting the most appropriate pages for the user request. The proposed model is tested on web log files of NASA and enggresources.

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)

Keywords: Hierarchical agglomerative clustering, session representative, occurrence of sequence

1. Introduction

Web log files are files that lists the actions of user that have been occurred when browsing website. These log files reside in the web server. Web log files contain information about User name, IP address, Timestamp, Access request, number of bytes transferred and User agent. Analysis of these log files gives navigation behaviour of the user. The data stored in the log files do not present an accurate picture of the user's accesses to the web site,

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)
doi:10.1016/j.procs.2015.08.073

hence pre-processing of web log data is one of the important phase of web mining for knowledge discovery. The phases in web mining are data collection, data pre-processing, pattern discovery and pattern analysis [1].

Collection of web log file from server by a process of authentication is known as data collection. Some web log files[2,3] are of free access for pattern analysis. Web log files are cleaned and pre-processed based on application domain[4]. During pre-processing relevant attributes are retained to reduce the size of web log file. Many works[5,6] have been reported in literature to pre-process the web log data to identify users sessions and navigation patterns which are useful for analysis for further prediction and ranking[7]. To discover useful patterns from the pre-processed data, data mining techniques are applied[8-14]. The existing techniques require prior knowledge for grouping sessions based on threshold. The most commonly used nearest neighbor is KNN in which centroids are randomly chosen to determine the clusters. The existing techniques do not cluster based on the nature of navigation.

To discover the navigation patterns based on the nature of occurrence, in this paper hierarchical agglomerative clustering technique has been adopted. Hierarchical agglomerative clustering technique clusters the patterns based on the sequence of occurrence of web pages. By this session representative navigation pattern is identified which contains a sequence of all possible page occurrence. Any new pattern with similar occurrences with the cluster is treated as subset of the session representative. A detailed pattern discovery has been explained in section 3.

2. Literature Review

Web log files of enggresources.com[18] and NASA[2] datasets are considered. Many works in literature target on pre-processing phase of web mining which poses number of challenges[1,5,6,7]. In the research carried out by Borges *et al.*, [5,] data cleaning was performed by removing the erroneous requests and image requests. Pirolli et al., [6] and Pitkow[7] discussed about difficulties in identifying users and sessions from web log files. Pitkow[7] discussed about criteria to define session. But in literature session time out is considered as 30 minutes[17].

Various techniques have been proposed in literature to cluster web user sessions. Chaitraa *et al.*, [8] proposed techniques to improve k-means algorithm by finding fixed centroids and then applying clustering algorithm, produces same clusters for every run. Om Prakash *et al.*, [9] discusses technique to predict next page based on clustering and Artificial Neural Network. Web user sessions are clustered using k-means algorithm. The prediction accuracy mainly depends on clustering quality. Dr. J.K.R. Sastry *et al.*,[10] proposed fuzzy clustering technique to cluster web user sessions where session may belong to more than one cluster, by doing this they are proving better prediction accuracy for common user profile. K-means clustering algorithm is used to cluster web user sessions. Chaofeng Li *et al.*,[11] proposed novel technique to cluster web user session based on increase of similarities. Initial centroids are found using ROCK and then Web Session Clustering based on Increase Similarity algorithm is applied to cluster sessions. G. Poornalatha *et al.*, [12] proposed improvements to K-means algorithm to cluster web uses sessions. This method takes care of variable length sessions. HuaXu *et al.*,[13] proposed user clustering based on vector matrix and K-Means algorithm. This approach is based on usage of web pages.

Wang *et al.*,[14] proposed an approach for clustering of web sessions based on sequence alignment method using dynamic programming. Hazarath Munaga *et al.*, [15] proposed trajectory clustering technique to predict user navigation. Based on trajectory dissimilarity between the transactions, clusters are formed. Dr.K.Duraiswamy *et al.*, [16] proposed matrix for calculation of similarity between sessions and then using agglomerative hierarchical clustering algorithm to cluster sessions.

It is evident from the survey that for session clustering, K-means, one of the most popular partition clustering algorithm[8] is used. But it requires prior knowledge about number of clusters and it is sensitive to initial centroids position selection. Many researchers concentrated on improving k-means clustering algorithm. The researches improves cluster quality by defining number of clusters based on application domain and by fixing initial centroids to extract usage pattern of web log data. But this improvement does not extract the occurrence of sequence of pages in the navigation pattern. To extract the sequence of occurrence of pages to define navigation patterns Hierarchical Agglomerative Clustering (HAC) algorithm is adopted which generates clusters with session representatives.

3. Proposed Methodology

Web log data consists of many irrelevant data. Three kinds of irrelevant data are image requests, erroneous requests and spider navigation requests. Using a logical data cleaning process irrelevant data are removed from the log files[5]. Users are identified through the IP address in the log entry. Based on the IP address, sessions are formulated for 30 minutes. Session consists of sequence of URLs navigated by the user. Each URL in the session is parsed into tokens to identify the pages navigated. By this navigation patterns are generated. Navigation pattern of every user is clustered to identify the session representative and its sub-sets.

Let $P_{1,} P_{2,}....P_n$ be the tokens of URLs of web pages parsed after pre-processing of web log file. If user visits pages $P_{1,} P_{2,} P_{4,} P_7$ in a sequence within 30 minutes of duration then session can be represented as,

$$S = \{ P_{1,} P_{2,} P_{4,} P_7\}$$

Navigation pattern of session is represented as,

$$Pat = \{P_1\ P_2\ P_4\ P_7\}$$

Such multiple sessions and their navigation patterns extracted from pre-processing of web log file are considered for clustering using hierarchical agglomerative clustering technique[19].

Hierarchical agglomerative clustering[19] is a bottom up approach. Bottom-up algorithms treat each sample as a singleton cluster at the outset and then successively merge (or *agglomerate*) pairs of clusters until all clusters have been merged into a single cluster that contains all samples. In this paper**,** hierarchical agglomerative clustering algorithm has been adopted with the inclusion of dynamic stopping criteria to merge clusters. In this method, initially each session and its navigation pattern is treated as single cluster. Pair of clusters are merged based on maximum value of subset occurrence, i.e., session could be a subset of another session. Subset is a set which contains the same pages and sequence of occurrence of page is also identical in the navigation pattern of session. The algorithm terminates when there is no subset sessions based on occurrence of sequence of pages in the navigation pattern. The algorithms of Hierarchical agglomerative clustering and distance measure (Similarity between Sessions(SBS)) used to cluster are explained below.

```
Algorithm: Hierarchical agglomerative clustering (HAC)
Input: A set of web sessions and its navigation patterns ws={$s_1$-$pat_1$,$s_2$-$pat_2$,$s_3$-$pat_3$....$s_n$-$pat_n$}
Output: Set of clusters c={c1,c2.....ck}
Method:
Repeat
    for each session $s_i$ from ws
        for each session $s_j$ from ws
            calculate $d_{(i,j)}$ =SBS($s_i$,$s_j$)
            if all entries in the $d_{(i,j)}$ is zero
            break;
    for i=1 to n
        for  j=1 to n
            merge($s_i$, $s_j$) where $d_{i,j}$=max value in the distance matrix
            replace $s_i$, $s_j$ entries in ws with $s_i$-$s_j$ and pattern should be longest pattern of two sessions
Until false
```

```
Function: Similarity between sessions(SBS)
Input: Two user sessions $s_i$-pat$_i$ and $s_j$-pat$_j$
Output: Similarity between $s_i$ and $s_j$
Method:
        if($s_i$'s pattern is proper subset of $s_j$'s pattern)
                return $p_i$.length>$p_j$.length ? return $p_j$.length : return $p_i$.length

        return 0;
```

To illustrate the function of SBS and working procedure of Hierarchical agglomerative clustering, Consider an example data set with 6 sessions and its navigation pattern.

Example:

S1: P_1 P_2 P_3 P_4 P_5
S2: P_1 P_2 P_3 P_5 P_6
S3: P_2 P_3 P_4 P_5
S4: P_3 P_4
S5: P_3 P_5 P_6
S6: P_3 P_2

	S1	S2	S3	S4	S5	S6
S1	-	0	4	2	0	0
S2	0	-	0	0	3	0
S3	3	0	-	2	0	0
S4	2	0	2	-	0	0
S5	0	3	0	0	-	0
S6	0	0	0	0	0	-

Table 1 Similarity matrix between sessions

Table 1 illustrates the distance between sessions of example dataset. The example clearly shows that S1 and S3 are having a similar navigation pattern which is of length 4. The similarity between sessions S1 and S6 is 0. Though session S1 and S6 share common pages, the order of occurrence of pages in S6 is different compared to S1.

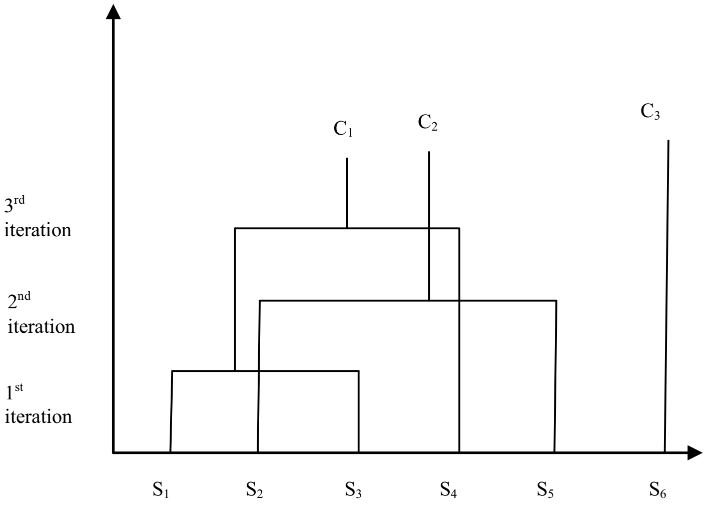

Figure 1: Dendrogram of merging sessions

The hierarchical agglomerative clustering technique uses SBS function to merge the sessions in iterations until there is no similarity between sessions. In the first iteration it merges sessions which are having highest similarity and continues merging until there are no similarity between sessions. Figure 1 illustrates dendrogram of merging of sessions. Three clusters are formed for the example dataset. The cluster representatives, c_1= {P_1 P_2 P_3 P_4 P_5}, c_2={P_1 P_2 P_3 P_5 P_6}, c_3={P_3 P_2} are unique navigation patterns of sessions formulated in cluster. The members of clusters are subset of the session representative navigation pattern.

4. Experimental results and analysis

The proposed model is tested on two data sets. The first dataset is the weblog file of www.enggresources.com. This web site focuses on engineering education and provides information related to engineering subjects, syllabus, courses, teaching guide lines, question banks, etc. Analysis of this web log file will be useful for engineering students, faculties. The log file consists of 2000 records of 19/July/2009. Each URL is parsed and 12 page categories are formulated, ex., admin, comedk, results, syllabus, etc.,

The second dataset is NASA log from NASA kennedy space center in Florida[2]. This weblog file consists of 10,00,000+ entries from 1/July/1995 to 31/July/1995. Each URL is parsed and 52 page categories are formulated., ex shuttle, shuttle/missions, elv, etc.,

Sl. No.	Dataset	No. of records	Relevant records	No. of sessions	No. of Clusters
1	ER	2000	1982	137	24
2	NASA	7000	4181	942	124
3	NASA	20000	6591	1548	265
4	NASA	40000	13347	3000	452
5	NASA	60000	17616	4000	543
6	NASA	75000	21951	5000	644

Table 2: Datasets used for experimentation

Table 2 illustrates the datasets used for experimental analysis, the number of sessions considered and the clusters formed for the identified sessions.

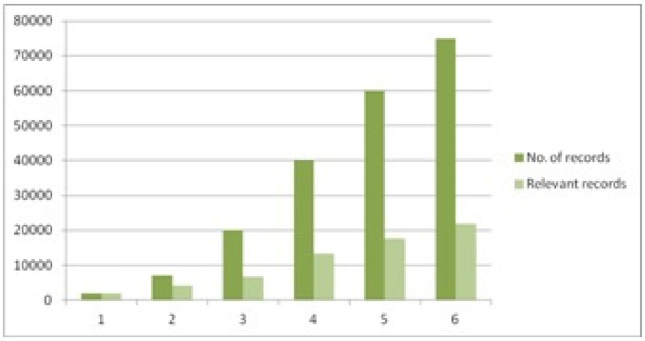

Figure 2: Number of relevant records

The data cleaning step of pre-processing phase retains only relevant data which are helpful for further knowledge discovery. Figure 2 illustrates number of records considered and number of relevant records useful for analysis. In figure 2, x-axis represents dataset used (SL No.) and y-axis represents number of relevant records after data pre-processing and total number of records considered.

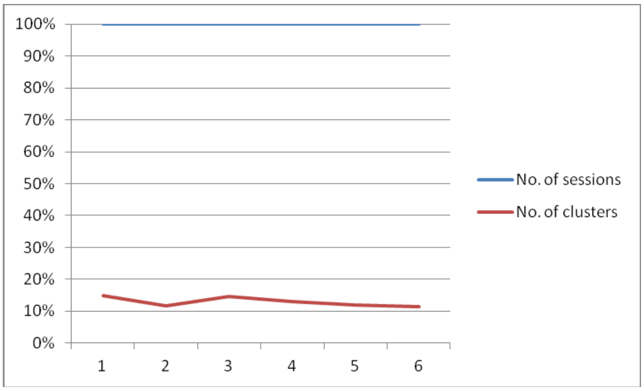

Figure 3 No. of sessions vs No. of clusters

Figure 3 illustrates number of sessions considered for testing and number of clusters formulated using hierarchical agglomerative clustering technique. X-axis represents dataset used and y-axis represents percentage of sessions considered and percentage of unique sessions based on the occurrence of sequence of pages in the navigation pattern. The sessions are clustered based on the occurrence of sequence of pages in the navigation pattern.

For the data sets considered for experimentation (Table 2), an average to all datasets, sessions with subset navigation patterns found are 89%. Unique navigation pattern in the sessions found are 11% (depicted in figure 3, representatives of clusters are session representatives), which are session representatives. The number of clusters formed are decreasing with number of sessions. Each cluster formed holds session representative and its sub set navigation patterns. These 11% of session's navigation patterns will be helpful for further pattern analysis, as each session representative consists of all possible subset navigation patterns.. The session representative navigation pattern will be given as input to prediction model, thus reducing the number of inputs.

5. Conclusion and Future Enhancements

Clustering sessions based on the nature of navigation patterns generates clusters which contains session representative and its sub set navigation patterns. By this all possible appropriate pages for user request are obtained which can be analyzed further for better prediction. As explained in the experimental results, the proposed model reduces inputs to prediction model by identifying session representatives and these session representatives maintains the occurrence of sequence in the navigation patterns and overcomes the drawbacks of partition based algorithms. And further it helps to pre-fetch or cache pages appropriately for user requests.

References

1. R. Kosala and H. Blockeel, "Web Mining Research: A Survey" *ACM SIGKDD Explorations*, 2000, 1-15
2. NASA dataset, "http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html"
3. MSNBC dataset, "http://kdd.ics.uci.edu/databases/msnbc/msnbc.html"
4. Ravindra Gupta, Prateek Gupta, "Application Specific Web Log Processig", *International Journal of Computer Techology & Applications*,Vol 3 (1),160-162
5. R. Cooley, B. Mobasher and J. Srivatsava, "Web mining: Information and pattern discovery on the World Wide Web", *9th IEEE International Conference on Tools with Artificial Intelligence.* CA, 1997, 558-567
6. J. Srivatsava, R. Cooley, M. Deshpande and P.N Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *ACM SIGKDD Explorat.* NewsLetter, 2000, 12-23
7. J. Pitkow, "In search of reliable usage data on the WWW", *Sixth International World Wide Web Conference*, Santa Clara,199, 451-463
8.V. Chitraa, Antony Selvadoss Thanamani, "An Enhanced Clustering Techniques for Web Usage Mining", *International Journal of Engineering Research and Technology,* ISSN:2278-0181, Vol 1, Issue 4, 2012
9.Om Prakash Mandal, Hiteshwar Kumar Azad, "Web Access Prediction Model using Clustering and Artificial Neural Network", *International Journal of Engineering Research and Technology,* ISSN:2278-0181, Vol 3, Issue 9, 2014
10.Dr. J.K.R. Sastry, K.Ruth Ramya, M. Devi Kavya Priya, "Random Indexing Based Web User Clustering for Faster Navigation", *IJSCIT,* Vol

4, 2013,541-545

11. Li, C.: "Algorithm of Web Session Clustering Based on Increase of Similarities", *Proceedings of International Conference on Information Management, Innovation Management and Industrial Engineering*, pp. 316–319. IEEE, Los Alamitos (2008)

12.G. Poornalatha and P. S. Raghavendra, " Web User Session Clustering Using Modified K-means Algorithm," *First International Conference on Advances in Computing and Communications (ACC – 2011),* CCIS(191),Springer-Verlag, pp.243-252, 2011

13. Xu, J.-H., Liu, H., "Web User Clustering Analysis based on KMeans Algorithm", *International conference on Information, Networking and Automation,*(ICINA), pp. V26–V29. IEEE, Los Alamitos 2010

14. W. Wang, O. R. Za¨ıane, "Clustering web sessions by sequence alignment "*University of Alberta Edmonton, Alberta, Canada.*

15 Hazarath Munaga, J. V. R. Murthy, N. B. Venkateswarlu, "A Hybrid Trajectory Clustering for Predicting User Navigation", *International Journal of Recent Trends in Engineering"*

16 Mayil, V. V. and Dr.K.Duraiswamy (2008). "Similarity Matrix Based Session Clustering by Sequence Alignment
Using Dynamic Programming." *Computer and Information Science*, Vol. 1, No. 3, August 2008

17 Ke Yiping, "A Survey on Preprocessing Techniques in Web Usage Mining", *The Hong Kong University of Science and Technology*, Dec-2003

18 http://www.enggresources.com

19 http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html