International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Dimension Reduction technique to discover the web usage clusters for Web Sessionization

R.Lokeshkumar[1],A.Bharathi[2], E.Maruthavani[3]

[12]Department of IT Bannari Amman Institute of Technology
Sathyamangalam, Erode,(India)
[3]Department of CSE Karpagam College of Engineering
Coimbatore, (India)

ABSTRACT

Clustering is one of the fundamental techniques to organize similar objects into proper groups based on features in the domain of data mining, machine learning and pattern recognition. In each cluster, objects are more similar to each other on the basis of particular features. Clustering has numerous applications in multiple domains such as information retrieval, data mining, machine learning, pattern recognition, mathematics, medical and bioinformatics. As a result of the unending extension of e-business, there is outstanding contention among relationship to pull in and hold customers. Examinations of the web server logs of these affiliations are essential for getting lots of information into web personalization lead, which can reinforce the arrangement of additionally engaging web structures. Web driven applications are growing step by step and the web has ended up one of the biggest information vaults. Similarity computation calculation among the information objects (web sessions) is mind boggling, however is a critical issue in unsupervised learning. This research is an attempt to overcome these challenges and problems. The objective of this research paper is to introduce a HAC based similarity measure to compute the similarity among the sessions. A HAC based approach is being applied to compute the statistically significant relationship between observed and expected frequencies of the number of pages visited and the time consumed by a user during a session. Also, Hierarchical agglomerative clustering (HAC) technique is proposed to extract useful knowledge from web log. This helps to improve the visualization of web logs and is equally important for website designers, developers and owners for the improvements of websites at each level. Experimental results with two different log files reveal that the proposed similarity measure with HAC algorithm has significantly improved the computation among data objects in web sessions.

Index Terms— Hybrid agglomerative clustering, Soft Computing, Web sessionization, Web Usage Mining, Web Log Mining,

I. INTRODUCTION

The World Wide Web as a large and dynamic information source is a fertile ground for data mining principles or Web Mining. Web mining is primarily aimed at deriving actionable knowledge from the Web through the application of various data mining techniques. Web Usage Mining is the discovery of user access patterns from Web server access logs [2]. Web Usage Mining analyses results of user interactions with a Web server, including Web logs and database transactions at a Web site. Web usage mining includes clustering to find

natural grouping of users or pages, associations to discover the URLs requested together and analysis of the sequential order in which URLs are accessed. The combination of innovation and the World Wide Web (WWW) has brought about bounty of digitized information and has opened new skylines for the Research group to investigate electronic information in various measurements. Therefore, the web has turned into a driving data hotspot for the worldwide group. With the progression of time since its commencement in 1990, the web is filling in as a mass travel course for the conveyance of administrations and assets to all parts of the world. The web is a system of systems of interrelated PCs, while sites and website pages give key data to its clients through the web. Sites are propelled on any web server over the web. There are two major issues, which are moving in parallel with the development of the web: (1) there is no understanding of a unified web server over the web and consequently there is no component to catch the client criticism and click history at a unified level, (2) a site can be composed and created with or without resulting standard improvement methodology [3]. This has opened a variety of issues over the web, for example, client conduct investigation, data recovery, prescribed framework, client relationship administration frameworks, profiling, forecast, and hacking.

Web log records are documents that rundowns the activities of users that have been happened when browsing site. These log records dwell in the web server. Web log documents contain data about User name, IP address, Timestamp, Access ask for, number of bytes exchanged and User operator. Examination of these log records gives route conduct of the client. The information put away in the log documents don't present an exact photo of the client's gets to th

The user’s click record is the key to investigate user trends and behavior on a specific website. The analysis of user click streams is useful in many ways such as website management website administration, fraud detection, web personalization, information retrieval systems and recommended systems. Due to decentralized web hosting, the user’s click record is also decentralized and has no centralized system to capture the user website traversing history. Consequently, we have to rely on web server log to study the user behavior and trends over a website. Web Usage Mining (WUM) consists of three main steps: preprocessing, knowledge extraction and results analysis [4]. The goal of the preprocessing step is to transform the raw web log data into a set of user profiles. Each such profile captures a sequence or a set of URLs representing a user session. Web usage data preprocessing exploit a variety of algorithms and heuristic techniques for various preprocessing tasks such as data cleaning, user identification, session identification etc.

There are three noteworthy sources for web log, for example, proxy web log, customer web log and web server log documents [7]. Every web log source is fragmented and has distinctive upsides and downsides. In the past, majority of the studies reveals that the web server log is utilized and considered as a legitimate source for the investigation of client click streams. In any case it is inadequate as the client may utilize the website pages from web cache too [8]. The program store support can break the client succession in web server log record. This issue can likewise be handled by utilizing the web site structure to finish the missing what's more, broken edges. The Web Usage mining (WUM) plays a key part in web site administration and web site organization. It is an expansion of information mining systems to remove the covered up information from web log and has various applications such as feature identification, design revelation, web personalization, recommender systems, frameworks and web user behavior analysis [9].

Consequently pre-handling of web log information is one of the essential period of web mining for learning revelation. The stages in web mining are information accumulation, information pre-preparing, design disclosure and example investigation. Gathering of web log document from server by a procedure of verification is known as information accumulation. Some web log records are of free access for example investigation. Web log records are cleaned and pre-prepared in view of use space. A mid pre-handling significant credits are held to lessen the span of web log record. Numerous works have been accounted for in writing to pre-handle the web log information to recognize user sessions and route designs which are valuable for examination for further forecast and positioning. To find helpful examples from the pre-prepared information, information mining methods are connected. The current strategies require earlier information for gathering sessions in light of limit. The most normally utilized closest neighbor is KNN as a part of which centroids are randomly define the clusters. The current methods don't group in light of the way of route. To find the route designs in light of the way of event, in this paper various leveled agglomerative grouping strategy has been embraced. Various leveled agglomerative bunching strategy groups the designs in light of the grouping of event of site pages. By this session agent route example is recognized which contains a succession of all conceivable page event. Any new example with comparative events with the group is dealt with as subset of the session agent. A point by point design revelation has been clarified in segment 3.

Hierarchical agglomerative clustering (HAC) plays an important part to group the web users with similar traversing patterns over a website. In HAC, each session is taken as a cluster itself and it is merged with the most similar clusters on the basis of the similarity measures of all the attributes of data objects of a session. Subsequently, the merged group is again combined with yet another group to form larger clusters on the basis of similarity among the session objects. The merging process is carried until the stopping criterion of the single largest group is obtained.

Web usage data are unlabeled so they do not contain any class information. The HAC algorithms can cluster similar user sessions in an efficient manner based on the frequencies at which URLs are accessed during user sessions. Web user session data usually contain inaccurate, inconsistent, and missing information. These weaknesses have a negative impact on the cluster discovery process. Thus, the clusters formed might not be reliable and trustworthy. However, the HAC techniques utilize the fuzzy membership concept in fuzzy sets, which are more robust against imperfections, so they are more suitable than traditional hard clustering techniques for pattern discovery in imperfect data.

Due to the non-deterministic browsing patterns of various web users, user session data do not have crisp boundaries and they often form overlapping clusters [10]. Because of the overlapping nature of web user session data, HAC clustering techniques can be applied very well to form overlapping clusters, where each user session object can belong to several clusters with different degrees of membership.

Moreover, HAC algorithms are simple, efficient, easy to implement, and they have been used widely for mining web usage data. The HAC algorithm has the added advantage that it is more robust against noise compared with other algorithms.

II. RELATED WORK

Web sessionization is an active research area to obtain unbiased and focused groups from web log for the identification of interesting patterns, which are previously unknown [11]. Whereas WUM is a complete process for mining hidden knowledge from web log files, and sessionization is a very important step as the rest of WUM process steps are solemnly depending on this step [12]. Moreover, clustering is a traditional data mining practice to knot the identical items based on the association similarity among the items. Another feature of clustering is that within the groups, inter object similarity is maximized and intra group objects similarity is minimized. Furthermore, for user click records, web sessionization clustering is an important process for the analysis of user behavior.

For clustering, similarity measure is significant and most of the web clustering literature revolves around the similarity measure [13]. In the subsequentparagraphs, we present the review of web sessionization and similarity measures used for sessionization clustering.

Web mining is facing different challenges such as robustness to noise, number of clusters, multi-resolution of the data, mining only good clusters, and efficiency. In their proposed research the hierarchical unsupervised niche clustering algorithm (H-UNC) with robust weights was applied for session clustering. For H-UNC, genetic algorithm (GA) was used to address the robustness issues [14]. The fitness function used for clustering is given in equation 1.

$$f_i = \frac{\sum_{j=1}^{N} w_{ij}}{\sigma_1^2} \qquad (1)$$

where Wij is robust scale dispersion measure. The fitness fun ij is the robust weight and $\sigma^2 1$ is the robust scale dispersion measure. The fitness function (equation 1) gives optimum results at the centroid of the cluster. The proposed H-UNC was 2-dimensional and used the Euclidean distance to find the similarity among the sessions. The Euclidean measure is widely criticized due to its nature and its application in web usage mining.

A scalable immune system clustering algorithm for user profiles mining in web log data under single pass was proposed to cluster the logs [15]. The proposed algorithm was inspired by the natural immune system to adopt dynamic changes. The web server was to act like a human body and click streams were marked as antigens. White blood cells (B-cells) detection and destroy system was used to detect the noisy click data in dynamic weighted B-cells (DWB). The weighted influence zone of each profile is calculated in equation 2.

$$\sigma_i^2 = \sum_{j=1}^{J} W_{ij} d_{ij}^2 \; / \; 2 \sum_{j=1}^{J} W_{ij} \qquad (2)$$

The euclidean distance, Cosine and Jaccard measures are not suitable measures for web session clustering due to the nature of user click stream data [16]. He proposed the time based and URL page similarity among the pages visited by different users. For any two web pages visited, the page viewing time was [0, 1] and for matching similarity, the similarity score is 20 and for mismatch and in between the gap, the similarity score is -10. To compute the similarity, dynamic programming was used. The only issue of match and mismatch among the sessions were considered while the similarity must be relative to sessions. Furthermore, hierarchical sessionization was not performed for focused visualization.

$$S_{time} = \frac{\min (t_{timeA}, t_{timeB})}{\max (t_{timeA}, t_{timeB})} \quad (3)$$

The web user session clustering by applying the agglomerative clustering algorithm is proposed to cluster the web users. Alignment score (Sa) and local similarity (Sb) are two major components to calculate the similarity between sessions. The applied dynamic programming on sessions and hierarchical clustering technique to pick the results. No comparative study and measure calculation justification was given. It is important to mention that no other preprocessing techniques were adapted for the complete preprocessing phase of WUM. The similarity was calculated by using the longest common subsequence (LCS) and applied the clustering algorithm to cluster the sessions.

$$S_a (S_1, S_2) = v / S(m) * M) \quad (4)$$

$$Sim (S_1, S_2) = Sa * S_b \quad (5)$$

The issue of time spent on a web page is discussed and used for web session clustering. It is very difficult to calculate the proper time utilization on a single page. A page consists of different web objects and each web object has a different worth to different users. For further details on web objects and web pages, some users may spend more time on that particular page while the other user may not [17]. Consequently, such type of approaches may work for a website consisting of a few pages, whereas for the larger websites this technique is not scalable.

$$S'' = (T_{LCS}^{\alpha} / T^{\alpha} * T^{\beta}_{LCS} / T^{\beta})^{1/2} \quad (6)$$

$$S_{\alpha\beta} = S' * S'' \quad (7)$$

The significance of similarity measure for web sessions were calculated with the similarity in two steps. In the first step, similarity among the web pages is calculated by tokenizing the pages' URL and by using the longest URL common string.

The string matching criteria stops when the URL of two completely mismatch [18]. For the matching web pages, similarity is marked as 1.0 and for the mismatch pair, it is 0.0. In the second step, the similarity among the web sessions is calculated and matching web pages in two sessions, the matching score is taken as 20 and for mismatch web pages it is -10. The higher the score between the sessions, the higher will be the similarity among the web sessions.

Today, the concept of dynamic web pages is common and their technique is silent. Moreover, the technique is not scalable for larger websites. Another limitation is that it is not necessary for the web page designer to design the website properly and follow the web pages naming conventions.

The clustering technique, whether it is supervised, semi-supervised, or unsupervised, is used to manage the efficiency and accuracy issues [19]. The author categorized the web usage data as heterogeneous because it is composed of different formats such as numerical and categorical. The session time, number of pages visited in a session and data downloaded in a session are numerical, while the pages visited are categorical. To find out the similarity among the web sessions in such a sparse nature of data is a tough task he used a two-step technique to compute the similarity among the sessions. A framework COWES was proposed by the author for the web user clustering based on evolutionary web sessions [20].The similarity among the users is calculated through the fractures and each web user is represented as a set of fractures. User similarity (US) is computed in the range [0,1] in the following equation

$$US(u_1, u_2) = \sum_{k=1}^{n} \delta_k FS_k(u_1, u_2) \Big/ \sum_{k=1}^{n} \delta_k \qquad (8)$$

where δ_k are shared fractures of two users. The clustering was performed by the standard agglomerative algorithm. The two major limitations were discussed for the proposed similarity such as common fractures and the denominator as total shared fractures.

The author applied the H-UNC algorithm to mine the evolving user profiles. The similarity score between the session and profile was calculated by cosine similarity, and the web session similarity was computed from URL to URL based on overlapping profiles P_i and P_j in the following equation.

$$S_u(I,j) = \begin{cases} 1 & \\ \min\left(1, |p_i \cap p_j| / \max(1, \min((|p_i|, |p_j|) - 1)\right) & \end{cases} \qquad (9)$$

It is evident from the survey that for session clustering, K-means, one of the most popular partition clustering algorithm 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.

III. FEATURE SELECTION AND SESSION WEIGHT ASSIGNMENT

The user session are mapped as vectors of URL references in the n-dimensional space. The unique URL from the set of preprocessed log file is taken for the experiment. The user sessions taken from the preprocessed log file is taken where each user session is represented as a set of weighted sessions. The weights assigned to the user sessions is represented as a binary and non-binary values which depends on the URL or the some feature of the URL sessions. Let $U = \{u1,u2,u3\ldots\ldots un\}$ be the unique URL and the user sessions are represented as $S = \{s1,s2,\ldots\ldots\ldots sm\}$.The weights of the user sessions are represented as wui

The feature weights are used to represent the user sessions instead of the fuzzy weights. The occurrence of the URL session is taken as a reference for the feature weights. The time a particular user spends on the web page or the number of bytes the user downloaded from the page is taken as the frequency of the occurrence. In the access logs taken from the user is of larger size, hence the URL of the user access is taken. The log dataset is a high dimensional data so distance based clustering is less efficient in clustering the user logs. In order to improve the clustering results the logs are filtered by removing the references to low support URLs which are not supported by the specified number of user sessions. A fuzzy based theoretic approach is used to improve the dimensionality reduction. We propose this approach to improve the clustering accuracy. We assign weights to all the URLs using the fuzzy set theoretic function. All the URLs with the session count less than α1 are assigned the weight 0.The support count higher than the session count is assigned the weight 1. The remaining weight is assigned between 0 to 1 which the having the count between α1 and α2.

$$W_u(X) = \begin{cases} 0, \text{ if } X <= \alpha_1 \\ 1, \text{ if } X >= \alpha_2 \\ \dfrac{X - \alpha_2}{\alpha_2 - \alpha_1} \\ 1, \text{ if } X >= \alpha_2 \\ \text{, otherwise} \end{cases} \quad (10)$$

Where

W_u *is weighed assigned to the* URL_u

α_1 *is the lower threshold on the session support count* α_2 *is the Upper threshold on the session support count x is the session support count of* URL_u

IV. ASSIGNING WEIGHTS TO THE USER SESSION : FUZZY APPROACH

The preprocessing of the session files were done to remove the noise and inconsistent data from the user clusters. The direct removal of the cluster data will result in the loss of significant amount of information when the dataset of the session file is large. To overcome this problem we used the Fuzzy set theoretic approach. Using this method the threshold is specified to remove only the unwanted sessions. The weights are assigned by using the Fuzzy membership function based on the URL accessed by the web users. The session weight using the linear fuzzy membership is carried using the equation as given above.

Algorithm: getLWPs (List SD, double MSLWP)

Input: A set of session-based web data SD; a user-specified minimum support MSLWP.

Output: A set of large web pages for each web user.

```
1. Define tmp_IP = SD1.IP;
2. Define i = 1;
3. Define out_LWP[][];
4. Define N_Sessioni = 0; // initialize the number of sessions for web user i
5. for each sequence data SDn in SD
6.  if (SDn.IP == tmp_IP)

7. N_Sessioni++;
8. for (int j = 1; j ¤ the number of web
   pages; j++)
9. if (SDn.URLs contain Pij)
   // Pij is the jth web page
   for web user i
10.        NPij++; // the visit time
   of web page j by web user
   i, add one
```

```
11. break;
12. end if
13. end
14. else
15. SDn.IP == tmp_IP;
16. i++;
17. N_Sessioni = 1;
18. end if
19. end
20. for each web user i
21. for each web page j
22. if (NPij {N_Sessioni>=MSLWP) //
      check if web page j for web user i is
      a Large web page
23. out_LWP[i][j].add(Pij);
24. end if
25. end
26. end
```

Sessionization issue is a vital stride in WUM process and it is considered as a substantial and solid answer for the accomplishment of WUM. In the event that we have temperamental sessionization, whatever remains of the WUM procedure may create ridicule comes about and at last make the framework blunder inclined what's more, defenseless. The framework may not look for the fancied position in the choice emotionally supportive network. The WUM handle includes various interrelated procedures what's more, these procedures are executed in various stages. A brief portrayal of these WUN steps is as beneath:

A. Web usage data and preprocessing

Web utilization information and preprocessing Web server log record is the essential natural information source for WUM strategy and the web get to document is a noteworthy wellspring of crude information. The diverse web server log documents has been talked .Web log is put away in plain content organization (ASCII) and that is a part of the working framework as opposed to a piece of web application. Get to log, operator log, blunder log, and referrer log are usually accessible web sign on web servers. Figure 1 demonstrates a nonspecific preview of the web log that conveys the qualities and significant data about the client crossing click history.

The log document records the client click stream while the client surfs the site, and because of the utilization of stateless convention HTTP, the log document records all articles (sound, video, pictures, robots) accessible on that solitary page along with the page URL. The greater part of the log sections are unessential for mining method. As the innovation has engaged us to catch a gigantic measure of web information, web log documents are a noteworthy wellspring of web information that store client click streams.

As per log documents contain 60 % immaterial information and that can't be utilized for information mining purposes. In this way we have utilized the college web log documents for our trials. Preprocessing step is essential and of web log record gets to be basic. For precise results, preprocessing is an essential stride in WebKDD. The purging was performed to have appropriate information for the WUM procedure. We expelled the sound, video, CSS, robots and crawlers passages because of the outline way of the site. The passages are unessential for the mining reason and must be dispensed with before applying the information mining procedures. These crude passages assume no part in mining and make comes about deride.

The passages, for example, picture records, CSS sheets, scripting, robots, crawlers, sound and video passages are recorded in a web log. Log documents likewise record the authoritative activities for example, overhaul, embed, or erases. Thusly, all these superfluous passages must be evacuated for nature of the WUM prepare.

We hold just valuable and mining required passages. The fruitful passages whose status code = "200" are kept while alternate passages are disposed of. The purifying stride helps us set up the web log for the following steps of WUM process. We applied various cleansing techniques to have a noise free web log file for further processes.

Sessionization is a basic issue and requires a legitimate research procedure to address it. We have embraced the chi-square based research strategy to address the sessionization issue. In the initial step, an broad writing survey was led concentrating on the sessionization issue that basically contains profiling and client conduct as its center issues. We too completely secured the commitments of the exploration group in such manner and that urged us to gadget the sessionization issue in a manner that all the partners might be recipients of this examination. In the following stride, we distinguished the sessionization issue observationally in light of the current restrictions found in the writing audit. Moreover, the proposed arrangement is critical to address the sessionization issue to beat the current disadvantages. In the following area, we executed the proposed chi-square based various leveled sessionization for the tenable and solid answer for the sessionization issue that entangles the pointed disadvantages. In the last area of our exploration philosophy, we approved the test comes about through standard, surely understood characterized measurements, for example, Accuracy and Recall. We additionally contrasted the outcomes and distributed results. In the following area, we are talking about the proposed arrangement in detail.

B. Web personalization and hierarchical clustering

Log sessionization is performed on the premise of IP address, nonetheless, clients have the choice to utilize diverse programs, diverse working frameworks and distinctive forms of HTTP. Clients likewise have an alternative to utilize sites from diverse topographical areas. These minor changes can be dealt with as hazard relief for the client examination furthermore, concentrating on the client patterns. In this proposed look into, we customized client's crosses, which are special what's more, unique in relation to the past snap history. This personalization recognizes the business rules for a particular site. For various leveled bunching of web log, we ascertained the quantity of site pages went by the client in a session and subsequent to performing preprocessing, we acquired 1987 sessions. In addition, we ascertained the chi-square values in view of the parameters of a number of website pages and a session time in every session. The chi-square estimation of every

session is figured with each other session and the most noteworthy chi-square esteem demonstrates the most grounded connection between's these two sessions. Assuming more than one sessions have the same higher esteem, then the in the first place event is viewed as a more suitable combine of related sessions. This is the principal level chain of importance. We moreover registered the normal of the most related sets for the estimation of next chain of importance level and for the tallness of related session in dendogram. We connected the accompanying proposed calculation (Figure 1) for chi based progressive sessionization of web log.

Day	No of Log Records (Log 1)	No of Log Records (Log 2)
1	17425	65536
2	16193	--
3	15214	--
4	11473	--
Total	60305	65536

Table 1: Raw web log entries

Hierarchical Levels	Data Objects (Sessions)	Did not participated
0	1985	0
1	1006	0
2	564	1
3	228	0
4	130	1
5	67	1
6	34	1
7	18	1
8	6	0
9	4	0
10	2	0
11	1	0

Table 2: Hierarchical levels of sessions

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

V. RESULTS AND DISCUSSION

Site log documents contain touchy information and site proprietors for the most part waver to uncover the site log records. Because of this prevention banks, online closeouts and web based shopping site proprietors don't impart their log documents to scientists. For the present study, distinctive site log documents of two unique colleges were chosen. Web log 1 contains a sum of 60302 client click steams in four days and web log 2 contains an aggregate of 65536 client navigates in one day

As web log records contain a gigantic measure of insignificant sections because of site structure, before playing out the trial we connected the information preprocessing strategy to set up the information for the genuine analysis. Amid preprocessing ventures around 40 % of passages were expelled as unessential. After the preprocessing stage, sessionization step was performed to make the client sessions. From log 1, we acquired 1738 extraordinary sessions and from log 2, 1987 sessions.This paper displays just the consequences of log 2.

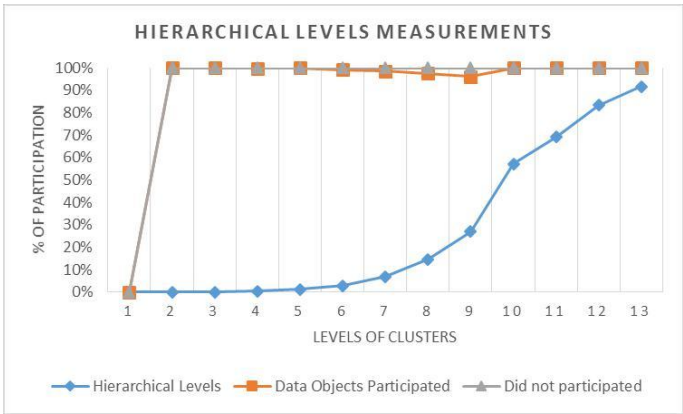

Figure 1: Hierarchy levels of web log sessionization

Table 2 and Figure 1 are illustrations of chi-square based sessionization of web log. In Table 2 we have also mentioned the number of sessions, which did not participate in session pairing based on chi. Each image in Figure 3 represents the hierarchical clustering combination of the session at each level. For hierarchical sessionization, we take the 1985 sessions as independent clusters themselves. We compute the measurement for each cluster with the other clusters and paired the clusters that have maximum chi-square values. We marked the computation as level 1 and an average linkage criterion was applied for hierarchy generation. The same step was repeated for the generation of 2^{nd} level hierarchy and so on. For this experiment we obtained 11 levels of hierarchy.

For the analysis of the proposed hierarchical clustering classifier, we used the precision and recall measures to evaluate the clustering results. We computed the true positive (TN), true negative (TN), false positive (FP) and false negative (FN) in each hierarchy level for the analysis of placements Figure 1: Hierarchy levels of web log sessionization of clusters in that particular level. The precision and recall results of 11 levels are shown in the Figures 2 and 3, respectively.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

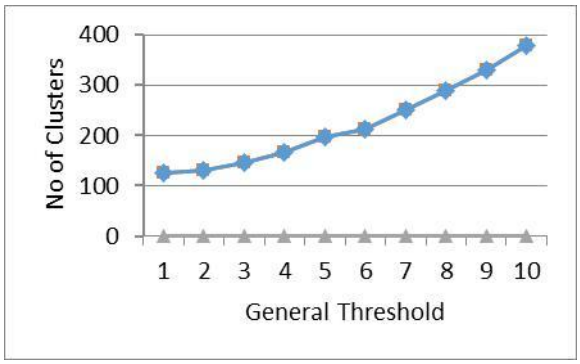

Figure 2: Relationship between number of clusters and general threshold

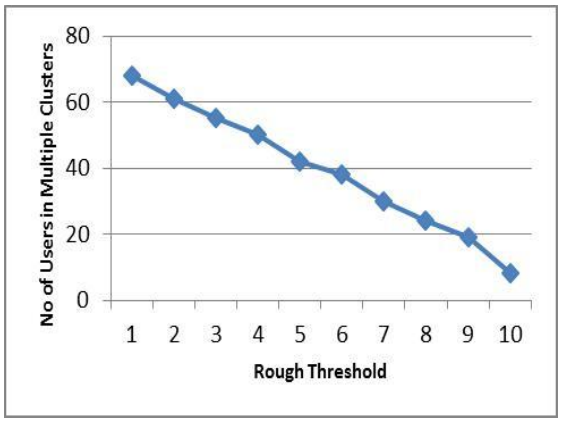

Figure3: Relationship between number of users in multiple clusters and rough threshold

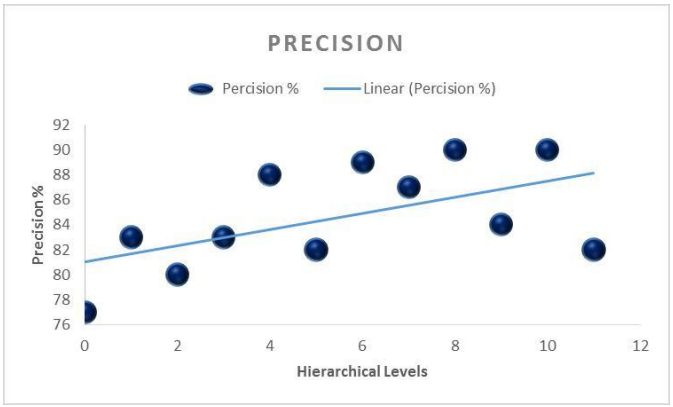

Figure 4: The Precision of clusters in each level

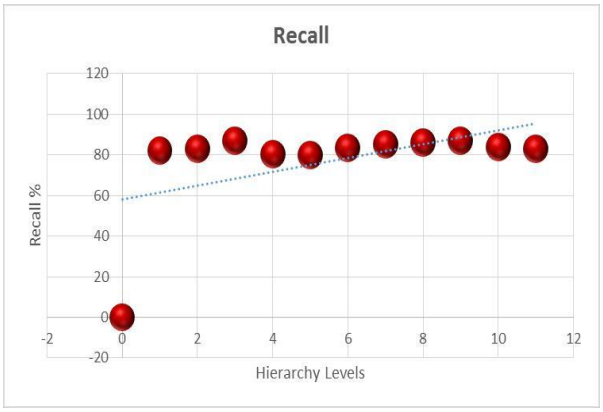

Figure 5: The Recall of clusters in each level

VI. CONCLUSION AND DISCUSSIONS

The proposed method classifier is simple and effective to improve the visualization of web log. The results are verified on two other published classifiers. It helps to analyze the web log for predefined objectives. Number of pages and time spent by a single user in a session are the two parameters on which the measurement values are calculated.

REFERENCES

[1] P. Kolari and A. Joshi, "Web mining: research and practice," Computing in Science and Engineering, vol. 6, no. 4, pp. 49–53, 2004.

[2] W. Tong and H. Pi-lian, "Web log mining by an improved aprioriall algorithm," in Intl proceeding of world academy of science, engineering, and technology, pp. 97–100, 2005.

[3] Hussain T. & Asghar S. "Web mining: approaches, applications and business intelligence", International Journal of Academic Research Part A 5: 211 – 217, 2013.

[4] B. Mobasher, "Data mining for web personalization", Lecture Notes in Computer Science, 4321:90, 2007.

[5] Wang Y.T. & Lee A.J.T., "Mining web navigation patterns with a path traversal graph", Expert Systems with Applications Vol.38 no.6, pp. 7112 – 7122, 2011.

[6] Vellingiri J., Kaliraj S., Satheeshkumar S. & Parthiban T., "A novel approach for user navigation pattern discovery and analysis for web usage mining", Journal of Computer Science vol.11 no.2,372 – 382, 2015.

[7] Chitraa V. & Davamani D.A.S., "A survey on preprocessing methods for web usage data", International Journal of Computer Science and Information Security Vol.7 no.3, pp. 78 – 83, 2010.

[8] Z. Ansari, M. Azeem, A.V. Babu, W. Ahmed, "A fuzzy approach for feature evaluation and dimensionality reduction to improve the quality of web usage mining results", Int. J. Adv. Sci. Eng. Inf. Technol. Vol.2 no.6 ,pp. 67–73, 2012.

[9] Z. Ansari, M.F. Azeem, A.V. Babu, W. Ahmed, "A fuzzy clustering based approach for mining usage profiles from web log data", Int. J. Comput. Inf. Sci. Secur. Vol.9 no.6, 70–79, 2011.

[10] A. Ketata, S. Mudur, N. Shiri, "Dependable performance analysis for fuzzy clustering of web usage data", IEEE Symposium on Computa-tional Intelligence and Data Mining, 2009, CIDM'09, pp.275–282, 2009.

[11] Park S., Suresh N.C. & Jeong B.K, "Sequence based clustering for web usage mining: a new experimental framework and ANN-enhanced K-means algorithm", Data and Knowledge Engineering vol.65 no.3, pp.512-543, 2008.

[12] Hasan T., Mudur S.P. & Shiri N, "A session generalization technique for improved web usage mining.", Proceedings of the 11th International Workshop on Web Information and Data Management, Hong Kong, China, 2 – 6 November, pp. 23 – 30, 2009.

[13] Kou G. & Lou C. "Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data", Annals of Operations Research vol.19 no.7, pp.123 – 134, 2012.

[14] Nasraoui O & Krishnapuram R., "One step evolutionary mining of context sensitive associations and web navigation patterns", SIAM International Conference on Data Mining, April 11-13, 2002. DOI: http://dx.doi.org/10.1137/1.9781611972726.31.

[15] Nasraoui O., Cardona C., Rojas C. & Gonzalez F, "Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm. Proceedings of WebKDD, Washington DC, USA, 2003.

[16] Li C, "Research on web session clustering", Journal of Software Vol.4 no.5, pp.460 – 468, 2009, DOI: http://dx.doi.org/10.4304/jsw.4.5.460-468.

[17] Duraiswamy K. & Mayil V.V, "Similarity matrix based session clustering by sequence alignment using dynamic programming", Computer and Information Science 1(3), 2008, DOI: http://dx.doi.org/10.5539/cis.v1n3p66.

[18] Wang W. & Zaiane O.R., "Clustering web sessions by sequence alignment", Proceedings of the 13th International Workshop on Database and Expert Systems Applications, 2 – 6 September, pp. 394 – 398, 2002.

[19] Alam S., Dobbie G., Koh Y.S. & Riddle P., "Clustering heterogeneous web usage data using hierarchical particle swarm optimization.", Proceedings of the IEEE Symposium on Swarm Intelligence (SIS), 16 – 19 April, Singapore, pp. 147 – 154, 2013.

[20] Chen L., Bhowmick S.S. & Nedjl W., "COWES: web user clustering based on evolutionary web sessions", Data and Knowledge Engineering 68(10): 867 − 885. DOI: http://dx.doi.org/10.1016/j.datak.2009.05.002,2009.