J.Natn.Sci.Foundation Sri Lanka 2016 44 (2): 211-222
DOI: http://dx.doi.org/10.4038/jnsfsr.v44i2.8002

RESEARCH ARTICLE

Chi-square based hierarchical agglomerative clustering for web sessionization

Tasawar Hussain[1]* and Sohail Asghar[2]
[1]*Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan.*
[2]*Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan.*

Revised: 05 October 2015; Accepted: 28 October 2015

Abstract: Clustering is one of the fundamental techniques to organise 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. Web centric applications are expanding day by day and the web has become one of the largest data repositories. During the last decade, information and knowledge retrieval from the web has become a challenging research area. Similarity computation among the data objects (web sessions) is complex, however is a significant problem in unsupervised learning. This research is an attempt to overcome these challenges and problems. The objective of this research paper is to introduce a chi-square based similarity measure to compute the similarity among the sessions. A chi-square 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. Moreover, a chi-square based hierarchical agglomerative clustering (Chi-HAC) technique is proposed to extract useful knowledge from web log. The Chi-HAC helps to improve the visualisation 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 Chi-HAC algorithm has significantly improved the computation among data objects in web sessions.

Keywords: Chi-square, hierarchical clustering, web sessionization, web usage mining.

INTRODUCTION

The fusion of technology and the World Wide Web (WWW) has resulted in profusion of digitised data and has opened new horizons for the research community to explore electronic data in different dimensions. The internet serves millions of people on a daily basis. According to Juliussen and Deegan (1986), the expected internet users in 2015 was 2.8 billion worldwide and millions of new pages are added to this mega repository every day. Consequently, the internet has become a leading information source for the global community. With the passage of time since its inception in 1990, the internet is working as a mass transit route for the delivery of services and resources to all parts of the world. The internet is a network of networks of interrelated computers, while websites and web pages provide key information to its users through the internet. Websites are launched on any web server over the internet. There are two big issues, which are moving in parallel with the growth of the internet: (1) there is no concept of a centralised web server over the internet and therefore there is no mechanism to capture the user feedback (Hussain *et al.,* 2010a) and click history at a centralised level, (2) a website can be designed and developed with or without ensuing standard development procedures. This has opened a varity of issues over the internet such as user behaviour analysis, information retrieval, recommended system, customer relationship management systems, profiling, prediction, and hacking.

*Corresponding author (tasawar70@yahoo.com)

The user's click record is the key to investigate user trends and behaviour on a specific website (Wang & Lee, 2011; Vellingiri *et al*., 2015). The analysis of user click streams is useful in many ways (Oliner *et al.,* 2012) such as website management (Vaarandi, 2003), website administration, fraud detection, web personalisation, information retrieval systems (Sote & Pande, 2015) and recommended systems (Hussain *et al*., 2010b). Due to decentralised web hosting, the user's click record is also decentralised and has no centralised system to capture the user website traversing history. Consequently, we have to rely on web server log to study the user behaviour and trends over a website.

There are three major sources for web log such as proxy web log, client web log and web server log files (Chitraa & Davamani, 2010). Each web log source is incomplete and has different pros and cons. In most of the studies, web server log is used and considered as an authentic source for the study of user click streams (Hussain *et al.,* 2010a; b; c). However it is incomplete as the user may use the web pages from cache as well. The browser cache support can break the user sequence in web server log file. This problem can also be tackled by using the website structure to complete the missing and broken edges. Web usage mining (WUM) plays a key role in website management and administration. It is an extension of data mining techniques to extract the hidden knowledge from web log and has numerous applications such as fraud detection, pattern discovery, personalisation, recommender systems and user behaviour analysis.

Hierarchical agglomerative clustering (HAC) plays a key role to group the users with similar traversing patterns over a website. In HAC, each session is a cluster itself and it is merged with the most similar clusters on the basis of some similarity measures of the attributes of data objects of a session. Subsequently, the merged group is again combined with another group to form bigger clusters on the basis of similarity among the objects (sessions). The merging process is carried on until the stopping criterion of the single largest group is obtained (Wei *et al*., 2008; Dimopoulos *et al*., 2010). HAC provides better control over log sessionization and research conducted by Johnson (1967), Lazzerini *et al*. (2003), Murray *et al*. (2006), Hussain *et al*. (2010a) and Hussain and Asghar (2013a) give detailed analysis about the use of HAC for WUM process.

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 (Park *et al*., 2008; Poornalatha & Raghavendra, 2011). 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 (Hasan *et al*., 2009; Kou & Lou, 2012). 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 maximised and intra group objects similarity is minimised. Furthermore, for user click records, web sessionization clustering is an important process for the analysis of user behaviour. For clustering, similarity measure is significant and most of the web clustering literature revolves around the similarity measure. In the subsequent paragraphs, we present the review of web sessionization and similarity measures used for sessionization clustering.

According to Nasraoui and Krishnapuram (2002), 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. The fitness function used for clustering is given in equation 1.

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

where W_{ij} is the robust weight and σ_1^2 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 criticised due to its nature and its application in web usage (Nasraoui *et al*., 2003; 2006; Li, 2009; Hussain *et al.*, 2010a).

Nasraoui *et al*. (2003) proposed a scalable immune system clustering algorithm for user profiles mining in web log data under single pass. 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.

$$\sigma_i^2 = \sum_{j=1}^{J} w_{ij} d_{ij}^2 \Big/ 2 \sum_{j=1}^{J} w_{ij} \quad ...(2)$$

Cosine similarity was used to calculate the sessions similarity. The proposed algorithm was not scalable and the said immune algorithm was improved by introducing Techno-Streams (Nasraoui *et al*., 2003) algorithm for robustness to noise.

According to Li (2009), the euclidean distance, Cosine and Jaccard measures are not suitable measures for web session clustering due to the nature of user click stream data. Li (2009) proposed the time based and URL page similarity among the pages visited by different users (equation 3).

$$S_{time=} \frac{\min(t_{time\ A}, t_{time\ B})}{\max(t_{time\ A}, t_{time\ B})} \quad ...(3)$$

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 visualisation.

Duraiswamy and Mayil (2008) performed the session clustering by applying the agglomerative hierarchical clustering algorithm. Alignment score (Sa) and local similarity (Sb) are two major components to calculate the similarity between sessions (equation 4). In equation 5, the similarity between sessions is calculated.

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

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

Duraiswamy and Mayil (2008) 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. Banerjee and Ghosh (2001) calculated the similarity by using the longest common subsequence (LCS) and applied the clustering algorithm to cluster the sessions. The authors calculated the similarity by using the following (equations 6, 7 and 8).

$$S' = \frac{1}{L}\sum_{i=1}^{L} Si = \frac{1}{L}\sum_{i=1}^{L} \min\left(\tau_{l^\alpha}^{\alpha}(i), \tau_{l^\beta}^{\beta}(i)\right) \Big/ \max\left(\tau_{l^\alpha}^{\alpha}(i), \tau_{l^\beta}^{\beta}(i)\right) \quad ...(6)$$

where $\tau_{l^\alpha}^{\alpha}(i)$ and $\tau_{l^\beta}^{\beta}(i)$ are time spent on a page $\gamma(i)$.

The importance of components and total similarity are calculated in equations 7 and 8, respectively.

$$S'' = \left(T_{LCS}^{\alpha}/T^{\alpha} * T_{LCS}^{\beta}/T^{\beta}\right)^{1/2} \quad ...(7)$$

$$S_{\alpha\beta} = S' * S'' \quad ...(8)$$

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 utilisation 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, the research work of Hussain *et al*. (2012) can be consulted. Some users may spend more time on that particular page while the other user may not. 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.

Wang and Zaïane (2002) highlighted the significance of similarity measure for web sessions and calculated 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. 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.

According to Alam *et al*. (2013) the clustering technique, whether it is supervised, semi-supervised, or unsupervised, is used to manage the efficiency and accuracy issues. Alam *et al.* (2013) categorised 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. Alam *et al.* (2013) used a two step technique to compute the similarity among the sessions.

In the first step, the user session is marked as XS_{ai} = XS (a_1, a_2, a_3, a_4, a_5). Where a_1 is the first attribute and the dissimilarity among the two sessions is calculated as follows:

$$d(XS, YS) = \left(\sum_{i}^{n} (XS_{ai} - YS_{ai})^2 \right)^{1/2} \quad ...(9)$$

where d (XS, YS) is the dissimilarity between the two sessions XS and YS.

In the second step, a hybrid of boolean and euclidean distance was used to compute the boolean distance among the user sessions.

$$b(XS, YS) = \left(\sum_{i}^{n} (XS_i \cap YS_i) \right) \quad ...(10)$$

The final distance among the various sessions is computed by the following equation

$$Dist(XS, YX) = d(XS, YS) + b(XS, YS) \quad ...(11)$$

Chen *et al*. (2009) proposed a framework COWES: web user clustering based on evolutionary web sessions. 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.

$$US(u_1, u_2) = \sum_{k=1}^{n} \delta_k FS_k(u_1, u_2) \Big/ \sum_{k=1}^{n} \delta_k \quad ...(12)$$

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.

Nasraoui *et al*. (2008) again applied the H-UNC (Nasraoui & Krishnapuram, 2002) 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.

$$S_u(I, j) = \begin{cases} 1 & \text{if } i = j \\ \min(1, \ |P_i \cap P_j| / \max(1, \min(|P_i|, |P_j|) - 1)) & \text{otherwise.} \end{cases} \quad ...(13)$$

The literature review has been summarised in Table 1. As clustering depends mainly on the selection of similarity measure and base classifier, the parameters selected are: similarity measure, base classifier, dataset and application area (Table 1). As user click streams contain the hidden knowledge about the users, organisations hesitate to open the access on web log data. During the review of literature, it has been observed that generally researchers face dataset issues and only relevant university data is available for experimental work. In Table 1, we have also discussed the application area, as one of the objectives was to review the literature for the identification of frauds. User profiling was the major area of research. Profiling is also helpful for the identification of frauds; however, we were unable to find the particular literature for fraud detection as there are various systems and tools, which are based on web usage mining for fraud detection. Furthermore, the focus was to review the literature for the identification of accurate and consistent user sessions. Sessionization was found in almost all the literature reviewed, but with complex or inappropriate measures and base algorithms.

The review of literature concludes that hierarchical clustering is a more appropriate and widely practiced technique as it provides in-depth visualisation and enhance the knowledge about click streams data of users in web usage mining. In most of the studies, evolutionary approaches have also been applied with the euclidean family of similarity measures. Therefore, the need arises for a strong measure to find the association and correlation among user sessions to improve the clustering results. Another common limitation about the clustering is that most of the work has been done to find out the clusters and after the cluster generation, knowledge discovery and pattern extraction portion is missing. Without complete cluster analysis the research work cannot

Table 1: Summary of web sessionization

Authors	Similarity measure	Base classifier	Data set	Application area
Nasraoui and Krishnapuram (2002)	$f_i = \frac{\sum_{j=1}^{N} w_{ij}}{\sigma_1^2}$	H-UNC	University website	User profiles
Nasraoui *et al.* (2003)	$\sigma_1^2 = \sum_{j=1}^{J} w_{ij} d_{ij}^2 / 2\sum_{j=1}^{J} w_{ij}$	Scalable immune learning	University website	User profiles
Li (2009)	$S_{time} = \frac{\min(t_{time\ A}, t_{time\ B})}{(\max(t_{time\ A}, t_{time\ B})}$	WSCBIS	University website	---
Duraiswamy and Mayil (2008)	$Sim((S_1,S_2) = Sa * Sb$	Agglomerative hierarchical clustering	---	Recommender system
Banerjee and Ghosh (2001)	$S_{\alpha\beta} = S' * S''$	Graph partitioning	Sulekha website	---
Wang and Zaïane (2002)	Tokenizing URLs	Sequence alignment (dynamic programming)	e-learning web log	e-learning
Alam *et al.* (2013)	$Dist(XS,YX) = d(XS,YS) + b(XS,YS)$	HPSO	University website	User prediction
Chen *et al.* (2009)	$US(u_1,u_2) = \sum_{k=1}^{n} \delta_k FS_k\ u_1,u_2) / \sum_{k=1}^{n} \delta_k$	COWES	Internet traffic Archive (website)	Web patronisation
Nasraoui *et al.* (2008)	$Su(i,j) = \{ \min^{1} (1, \lvert P_i \cap P_j \rvert / \max(1,\min(\lvert P_i$	UNC	University website	User Profiles

be effective in a domain. Hence, there is a need for a scalable hierarchical agglomerative algorithm that may help in-depth and enhanced knowledge visualisation of the web log file.

CHI-SQUARE BASED SIMILARITY MEASURE

Clustering is the most widely practiced technique of data mining and serves various application tools. There are a number of algorithms available for clustering. The selection of a proper clustering classifier depends upon a number of factors such as the dataset, data type, similarity measure or scoring function, number of parameters, environment, and the nature of applications, etc. In clustering, the objects are divided into different groups according to their relationship with each other based on properties and characteristics (similarity) (Yang & Wang, 2003). Each cluster possesses certain characteristics, such as objects within a cluster are more similar to each other as compared to the objects of other clusters. While in hierarchical agglomerative clustering (HAC), each object is considered as the cluster itself. These individual clusters are paired with each other based on the similarity between them. This process is repeated until we get the single cluster. These clusters form a tree like structure.

The user session is recorded as the minimum time between the login and logout of a user to a particular website. The session depends upon the time consumed by the user on the website and web pages traversed during that time. The web log file records all the user transactions along with the time stamp. Web log file contains a number of parameters, but the total time spent by a user and the web pages traversed are a key to analyse the user's traversing behaviour. The web log data is categorical in nature and chi-square computes the statistically related sessions based on the time consumed by a user on the website and the number of pages traversed by the user. The 2 × 2 contingency table of two sessions (Table 2) and chi-square (equation 14) are given below:

Table 2: The 2 × 2 chi-square contingency table

Sessions	Page visited	Time consumed	Row total
S_1	P_1	T1	P_1+T_1
S_2	P_2	T_2	P_2+T_2
Column total	P_1+P_2	T_1+T_2	$P_1+P_2+T_1+T_2$

$$\chi^2 = \frac{(P_1 * T_2 - P_2 * T_1)^2 (P_1 + P_2 + T_1 + T_2)}{(P_1 + P_2) * (P_1 + T_1) * (P_2 + T_2) * (T_1 + T_2)} \quad ...(14)$$

Similarly, the chi values are calculated for every session with the other sessions, and now two sessions can be joined by applying the hypothesis,

$H_0 = S_1$ *and* S_2 *are independent*
$H_1 = S_2$ *and* S_2 *are dependent*

The chi-square is computed between two sessions {(S_1, S_2), (S_1, S_3), (S_1, S_4), ..., (S_1, S_n)} in succession. The pair that holds the maximum chi value exhibits a maximum tendency of similar session. The marked pair is interrelated and dependent on each other and will not participate for the next iteration of finding similar sessions. The minimum value can also be taken into account for dependency test by taking them highly related. The selection of hypothesis is entirely based on the nature of data and test values, while the expert also has the right of hypothesis selection.

PROPOSED CHI-HAC (SESSIONIZATION)

Sessionization problem is an important step in WUM process and it is considered as a valid and reliable solution for the success of WUM. If we have unreliable sessionization, the rest of the WUM process may produce mock results and ultimately make the system error prone and vulnerable. The system may not seek the desired position in the decision support system. The WUM process comprises a number of interrelated processes and these processes are implemented in different phases. A brief description of these WUN steps is as below:

Web usage data and preprocessing

Web server log file is the primary unprocessed data source for WUM procedure and the web access file is a major source of raw data. The different web server log files has been discussed by Vellingiri *et al*. (2015). Web log is stored in plain text format (ASCII) and that is a part of the operating system rather than a part of web application. Access log, agent log, error log, and referrer log are commonly available web logs on web servers. Figure 1 shows a generic snapshot of the web log that delivers the attributes and relevant information about the user traversing click history.

The log file records the user click stream while the user surfs the website, and due to the use of stateless protocol HTTP, the log file records all objects (audio, video, images, robots) available on that single page along with the page URL. Most of the log entries are irrelevant for mining procedure. As the technology has empowered us to capture a huge amount of web data, web log files are a major source of web data that store user click streams. According to Alam *et al.* (2013) log files contain 60 % irrelevant data and that cannot be used for data mining purposes. Therefore we have used the university web log files for our experiments.

Preprocessing step is necessary and of web log file becomes imperative. For accurate results, preprocessing is a very important step in WebKDD. The cleansing was performed to have proper data for the WUM process. We removed the audio, video, CSS, robots and crawlers entries due to the design nature of the website. The entries are irrelevant for the mining purpose and has to be eliminated before applying the data mining techniques. These raw entries play no role in mining and make results mock. The entries such as image files, CSS sheets, scripting, robots, crawlers, audio and video entries are recorded in a web log. Log files also record the administrative actions such as update, insert, or deletes. Consequently, all these irrelevant entries have to be removed for quality of the WUM process.

We retain only useful and mining required entries. The successful entries whose status code = "200" are kept while the other entries are discarded. The cleansing step helps us prepare the web log for the next

```
127.0.0.1 - - [06/Dec/2008:14:09:32 +0500] "GET / HTTP/1.1" 200 1494
127.0.0.1 - - [06/Dec/2008:14:09:32 +0500] "GET /apache_pb.gif HTTP/1.1" 200 2326
127.0.0.1 - - [06/Dec/2008:14:09:32 +0500] "GET /favicon.ico HTTP/1.1" 404 283
127.0.0.1 - - [06/Dec/2008:14:13:00 +0500] "GET /phpinfo.php HTTP/1.1" 200 41959
127.0.0.1 - - [06/Dec/2008:14:13:01 +0500] "GET /phpinfo.php?=PHPE9568F34-D428-11d2-A769-
00AA001ACF42 HTTP/1.1" 200 2524
127.0.0.1 - - [06/Dec/2008:14:13:01 +0500] "GET /phpinfo.php?=PHPE9568F35-D428-11d2-A769-
00AA001ACF42 HTTP/1.1" 200 2146
127.0.0.1 - - [06/Dec/2008:14:41:30 +0500] "GET /phpinfo.php HTTP/1.1" 200 45265
```

Figure 1: Web log file

```
1.    Given WebServerLogFile R of n records.
2.    R = P <ip,date,url,agent,method,operation,status>.
3.    R=<r1,r2,r3,… ….rn> where n <> 0.
4.    Begin
5.        For i=1 to n
6.            Read R
7.            If r ( i ).url <> (images)
8.                AND ( r( i ).method=GET)
9.                AND ( r ( i ).status = 200)
10.               AND ( r ( i ). agent<>(crawler,Spide,Robot)) then
11.               Write r ( i ) into LogDatabase
12.           End if
13.       End for loop
14.   End
```

Figure 2: Web log cleansing algorithm

steps of WUM process. We applied various cleansing techniques to have a noise free web log file for further processes. One of the applied techniques is shown in Figure 2.

Sessionization is a critical issue and requires a proper research methodology to address it. We have adopted the chi-square based research methodology to address the sessionization problem. In the first step, an extensive literature review was conducted focusing on the sessionization problem that primarily comprises profiling and user behaviour as its core issues. We also thoroughly covered the contributions of the research community in this regard and that encouraged us to device the sessionization problem in such a way that all the stakeholders may be beneficiaries of this research. In the next step, we identified the sessionization problem empirically based on the existing limitations found in the literature review. Furthermore, the proposed solution is significant to address the sessionization problem to overcome the existing drawbacks. In the next section, we implemented the proposed chi-square based hierarchical sessionization for the credible and reliable solution to the sessionization problem that entraps the pointed drawbacks. In the last section of our research methodology, we validated the experimental results through standard, well known defined metrics such as Precision and Recall. We also compared the results with published results. In the next section, we are discussing the proposed solution in detail.

Web personilisation and hierarchical clustering

Log sessionization is performed on the basis of IP address, however, users have the option to use different browsers, different operating systems and different versions of HTTP. Users also have an option to use websites from different geographical locations. These minor changes can be treated as risk mitigation for the user analysis and studying the user trends. In this proposed research, we personalised user's traverses, which are unique and different from the previous click history. This personalisation helps to identify the business rules for a specific website. For hierarchical clustering of web log, we opted the research work of Hussain and Asghar (2013b). We calculated the number of web pages visited by the user in a session and after performing preprocessing, we obtained 1987 sessions. Moreover, we calculated the chi-square values based on the parameters of a number of web pages and a session time in each session. The chi-square value of each session is computed with every other session and the highest chi-square value shows the strongest correlation between these two sessions. If more than one sessions have the same higher value, then the first occurrence is considered a more appropriate pair of related sessions. This is the first level hierarchy. We also computed the average of the most related pairs for the calculation of next hierarchy level and for the height of related session in dendogram. We applied the following proposed algorithm (Figure 3) for chi based hierarchical sessionization of web log.

RESULTS AND EVALUATION

Web site log files contain sensitive data and website owners usually hesitate to expose the website log files (Mhamane & Lobo, 2012). Due to this hindrance banks, online auctions and online shopping website owners do not share their log files with researchers. For the present study, different website log files of two different universities were selected.

Web log 1 contains a total of 60302 user click steams in four days and web log 2 contains a total of 65536 user traverses in one day (Table 3).

As web log files contain a huge amount of irrelevant entries due to website structure, before performing the experiment we applied the data preprocessing technique to prepare the data for the actual experiment. During preprocessing steps around 40 % of entries were removed as irrelevant.

After the preprocessing phase, sessionization step was performed to create the user sessions. From log 1, we obtained 1738 unique sessions and from log 2, 1987 sessions. This paper presents only the results of log 2.

On the basis of the proposed algorithm (Figure 3), we obtained 11 levels of hierarchical clusters of web log.

```
Input sessionbase S of n sessions
Output Chi-Square based Hierarchical Clusters
S ={session_no, No_of_pages, session_time, R1, R2 Avg_value, level}
--Where R1 and R2 are related sessions
Begin
        level=0
    For i = level + 1 in total_sessions
            Max_Chi=0
            R1=i
        For j=i+1 in total_sessions
                R2=j
                Calculate Chi_Square(R1, R2 )
            if Chi_Square > Max_Chi then
                Max_Chi=Chi_Square
                R1=i
                R2=j
                Avg_value(R1, R2)
            End if
        End j loop
        updated the hierarchy Database with level
        level=level + 1
    End i loop
End
```

Figure 3: Proposed chi-square based hierarchical clustering algorithm

Table 3: Raw web log entries

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 4 and Figure 4 are illustrations of chi-square based sessionization of web log. In Table 4 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 1987

Table 4: Hierarchical levels of sessions

Hierarchical levels	Data objects (sessions)	Did not participated
0	1987	0
1	993	1
2	497	0
3	248	1
4	124	1
5	62	1
6	31	1
7	16	0
8	8	0
9	4	0
10	2	0
11	1	0

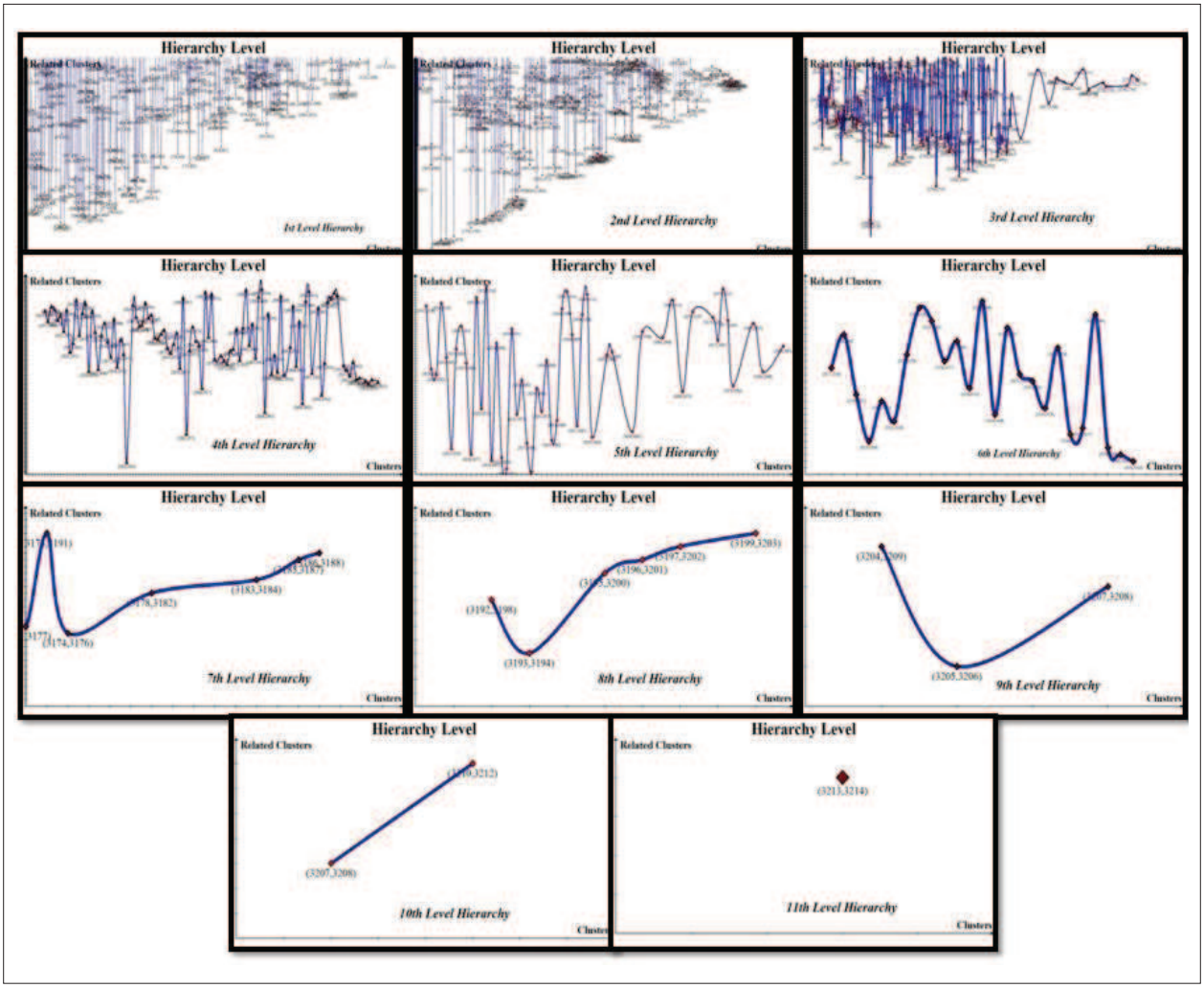

Figure 4: Hierarchy levels of web log sessionization

sessions as independent clusters themselves. We compute the chi-square of 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 2nd 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 of clusters in that particular level. The precision and recall results of 11 levels are shown in the Figures 5 and 6, respectively.

We also compared the proposed Chi-HAC with the research work of Murray *et al*. (2006) and Hussain *et al*. (2010b). We implemented the classifier of Murray (2006) with a few minor changes without affecting the original essence of the classifier. The dataset used for the experiment was the web log file of the university web site. We compared the three classifiers on the basis of precision and recall measure to find out the goodness of classifiers (Figures 7 and 8). The graphs indicate better results of the proposed Chi-HAC as comapred to published work.

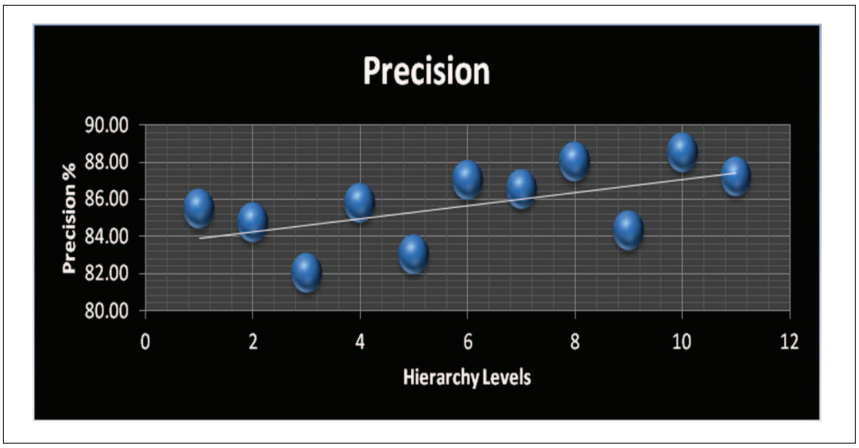

Figure 5: The precision of clusters in each level

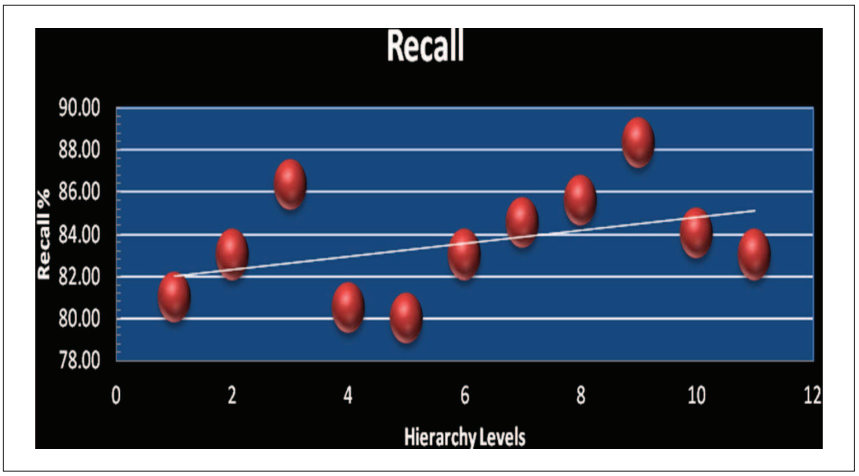

Figure 6: The recall of clusters in each level

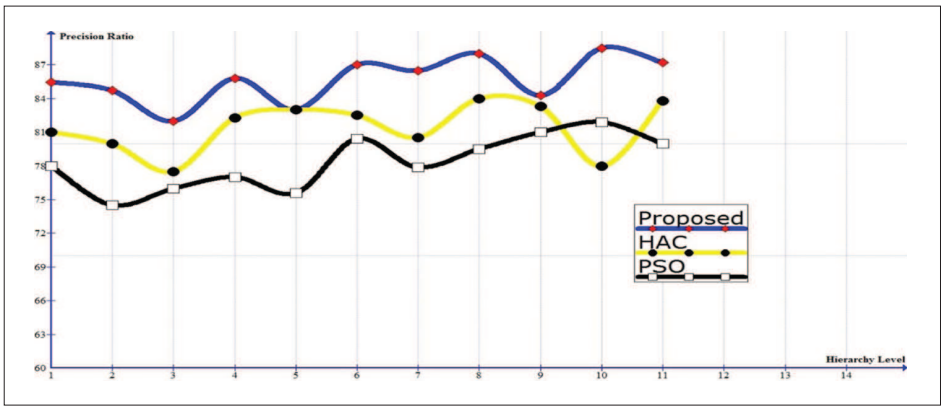

Figure 7: Comparison of proposed Chi-HAC with PSO and HAC at each hierarchy level (precision)

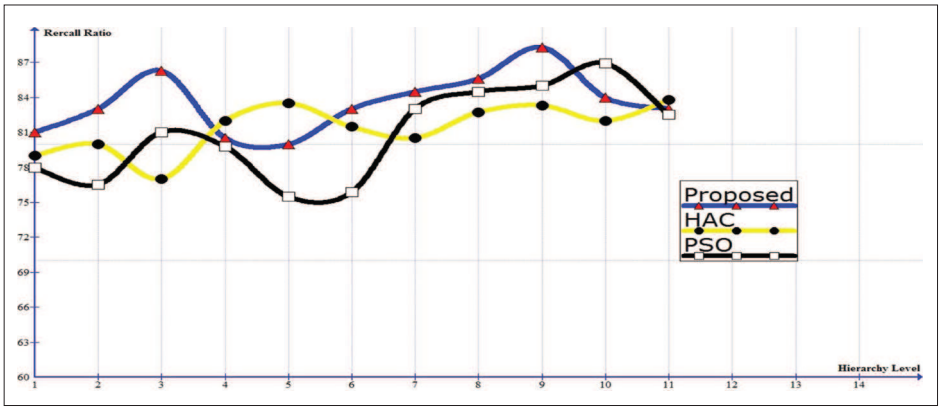

Figure 8: Comparison of proposed Chi-HAC with PSO and HAC at each hierarchy level (recall)

CONCLUSION

The proposed Chi-HAC classifier is simple and effective to improve the visualisation of web log. The results are verified on two other published classifiers. It helps to analyse 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 chi-square values are calculated.

REFERENCES

1. Alam S., Dobbie G., Koh Y.S. & Riddle P. (2013). 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. DOI: http://dx.doi.org/10.1109/sis.2013.6615172
2. Banerjee A. & Ghosh J. (2001). Clickstream clustering using weighted longest common subsequences. *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining,* volume 143, Chicago, USA, 5 – 7 April. Society for Industrial and Applied Mathematics, Philadelphia, USA, p. 144.
3. Chen L., Bhowmick S.S. & Nedjl W. (2009). 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
4. Chitraa V. & Davamani D.A.S. (2010). A survey on preprocessing methods for web usage data. *International Journal of Computer Science and Information Security* **7**(3): 78 − 83.
5. Dimopoulos C., Makris C., Panagis Y., Theodoridis E. & Tsakalidis A. (2010). A web page usage prediction scheme using sequence indexing and clustering techniques. *Data and Knowledge Engineering* **69**(4): 371 − 382. DOI: http://dx.doi.org/10.1016/j.datak.2009.04.010
6. Duraiswamy K. & Mayil V.V. (2008). Similarity matrix based session clustering by sequence alignment using dynamic programming. *Computer and Information Science* **1**(3): 66. DOI: http://dx.doi.org/10.5539/cis.v1n3p66
7. Hasan T., Mudur S.P. & Shiri N. (2009). 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. DOI: http://dx.doi.org/10.1145/1651587.1651595
8. Hussain T. & Asghar S. (2013a). Web mining: approaches, applications and business intelligence. *International Journal of Academic Research Part A* **5**: 211 − 217.
9. Hussain T. & Asghar S. (2013b). Evaluation of similarity measures for categorical data. *Nucleus* **50**(4): 387 − 394.
10. Hussain T., Asghar S. & Fong S. (2010a). A hierarchical cluster based preprocessing methodology for web usage mining. *Proceedings of the 6th International Conference on Advanced Information Management and Service (IMS)*, Seoul, Korea, 30 November – 02 December, pp. 472 − 477.
11. Hussain T., Asghar S. & Masood N. (2010b). Hierarchical sessionization at preprocessing level of WUM based on swarm intelligence. *Proceedings of the 6th International Conference on Emerging Technologies (ICET),* Islamabad, Pakistan, 18 – 19 October, pp. 21 − 26. DOI: http://dx.doi.org/10.1109/icet.2010.5638388
12. Hussain T., Asghar S. & Masood N. (2010c). Web usage mining: a survey on preprocessing of web log file. *Proceedings of the International Conference on Information and Emerging Technologies (ICIET),* Karachi, Pakistan, pp. 1 − 6. DOI: http://dx.doi.org/10.1109/iciet.2010.5625730
13. Hussain T., Qadir M.A. & Asghar S. (2012). Fuzzification of web objects: a semantic web mining approach. *International Journal of Computer Science* **9**(2): 61 − 67.

14. Johnson S.C. (1967). Hierarchical clustering schemes. *Psychometrika* **32**: 241 − 254.
DOI: http://dx.doi.org/10.1007/BF02289588
15. Juliussen F.E. & Deegan J. (1986). Computer Industry Almanac Inc., W. White Oak, Arlington Heights, IL, USA.
16. Kou G. & Lou C. (2012). Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data. *Annals of Operations Research* **197**(1): 123 − 134.
DOI: http://dx.doi.org/10.1007/s10479-010-0704-3
17. Lazzerini B., Marcelloni F. & Cococcioni M. (2003). A system based on hierarchical fuzzy clustering for web users profiling. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics,* volume 2, Washington DC, USA , 5 – 8 October, pp. 1995 − 2000.
DOI: http://dx.doi.org/10.1109/icsmc.2003.1244705
18. Li C. (2009). Research on web session clustering. *Journal of Software* **4**(5): 460 − 468.
DOI: http://dx.doi.org/10.4304/jsw.4.5.460-468
19. Mhamane S.S. & Lobo L. (2012). Use of hidden Markov Model as internet banking fraud detection. *International Journal of Computer Applications* **45**(21): 0975 – 8887.
20. Murray G.C., Lin J. & Chowdhury A. (2006). Identification of user sessions with hierarchical agglomerative clustering. *Proceedings of the American Society for Information Science and Technology* **43**(1): 1 − 9.
DOI: http://dx.doi.org/10.1002/meet.14504301312
21. Nasraoui O. & Krishnapuram R. (2002). One step evolutionary mining of context sensitive associations and web navigation patterns. *Proceedings of the 2002 SIAM International Conference on Data Mining*, Virginia, USA, 11 – 13 April.
DOI: http://dx.doi.org/10.1137/1.9781611972726.31
22. Nasraoui O., Cardona C., Rojas C. & Gonzalez F. (2003). Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm. *Proceedings of WebKDD*. Washington DC, USA.
23. Nasraoui O., Cardona C. & Rojas C. (2006). A framework for mining evolving trends in web data streams using dynamic learning and retrospective validation. *Computer Networks* **50**(10): 1488 − 1512.
DOI: http://dx.doi.org/10.1016/j.comnet.2005.10.021
24. Nasraoui O., Soliman M., Saka E., Badia A. & Germain R. (2008). A web usage mining framework for mining evolving user profiles in dynamic websites. *IEEE Transactions on Knowledge and Data Engineering* **20**(2): 202 − 215.
DOI: http://dx.doi.org/10.1109/TKDE.2007.190667
25. Ni X., Quan X., Lu Z., Wenyin L. & Hua B. (2011). Short text clustering by finding core terms. *Knowledge and Information Systems* **27**(3): 345 − 365.
DOI: http://dx.doi.org/10.1007/s10115-010-0299-7
26. Oliner A., Ganapathi A. & Xu W. (2012). Advances and challenges in log analysis. *Communications of the ACM* **55**(2): 55 − 61.
DOI: http://dx.doi.org/10.1145/2076450.2076466
27. Park S., Suresh N.C. & Jeong B.K. (2008). Sequence-based clustering for web usage mining: a new experimental framework and ANN-enhanced K-means algorithm. *Data and Knowledge Engineering* **65**(3): 512 − 543.
DOI: http://dx.doi.org/10.1016/j.datak.2008.01.002
28. Poornalatha G. & Raghavendra P. (2011). Alignment based similarity distance measure for better web sessions clustering. *Procedia Computer Science* **5**: 450 − 457.
DOI: http://dx.doi.org/10.1016/j.procs.2011.07.058
29. Sote A.M. & Pande S.R. (2015). Web page clustering using self-organizing map. *A Monthly Journal of Computer Science and Information Technology* **4**(1): 78 − 84.
30. Vaarandi R. (2003). A data clustering algorithm for mining patterns from event logs. *Proceedings of the 3rd IEEE Workshop on IP Operations and Management,* 1 – 3 October, pp. 119 − 126.
DOI: http://dx.doi.org/10.1109/ipom.2003.1251233
31. Vellingiri J., Kaliraj S., Satheeshkumar S. & Parthiban T. (2015). A novel approach for user navigation pattern discovery and analysis for web usage mining. *Journal of Computer Science* **11**(2): 372 − 382.
DOI: http://dx.doi.org/10.3844/jcssp.2015.372.382
32. Wang W. & Zaïane O.R. (2002). Clustering web sessions by sequence alignment. Proceedings of the 13th International Workshop on Database and Expert Systems Applications, 2 – 6 September, pp. 394 − 398.
DOI: http://dx.doi.org/10.1109/DEXA.2002.1045928
33. Wang Y.T. & Lee A.J.T. (2011). Mining web navigation patterns with a path traversal graph. *Expert Systems with Applications* **38**(6): 7112 − 7122.
DOI: http://dx.doi.org/10.1016/j.eswa.2010.12.058
34. Wei L., Yu-quan Z., Geng C. & Zhong Y. (2008). Clustering of web users based on competitive agglomeration. *Proceedings of the International Symposium on Computational Intelligence and Design*, volume 1, Wuhan, China, 17 – 18 October, pp. 515 − 519.
DOI: http://dx.doi.org/10.1109/iscid.2008.130
35. Yang J. & Wang W. (2003). CLUSEQ: efficient and effective sequence clustering. *Proceedings of the 19th International Conference on Data Engineering*, Bangalore, India, 5 – 8 March, pp. 101 − 112.
DOI: http://dx.doi.org/10.1109/icde.2003.1260785