International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 3
ISSN: 2088-5334

Augmented Session Similarity Based Framework for Measuring Web User Concern from Web Server Logs

Dilip Singh Sisodia

Department of Computer Science & Engineering, National Institute of Technology Raipur, India
E-mail: dssisodia.cs@nitrr.ac.in

***Abstract*— In this paper, an augmented sessions similarity based framework is proposed to measure the web user concern from web server logs. This framework utilized the best user session similarity between any sessions based on the relevance of accessed page in a particular session and syntactic structure of web URLs. The framework is tested using K-medoids clustering algorithms with independent and combined similarity measures. The merits of generated clusters are evaluated by measuring average intra-cluster and inter-cluster distances. The results demonstrated the superiority of combined augmented session dissimilarity metric over the independent dissimilarity measures using augmented session regarding cluster validity measures.**

***Keywords*— augmented web user sessions; user concerns; page relevance; syntactic structure; page stay time; the frequency of page; dissimilarity metric**

I. Introduction

With brisk enlargement and continued popularity of the World Wide Web, most of the institutes are presently putting their data on the web and offer online services [1]. So, many users now rely on the Web to seek information and gain knowledge by navigating websites. This user navigation results in an enormous information about their surfing behavior on the web hosting server in the form of weblogs [2]. A challenging aspect of web usage mining is to know in advance about user's habits, interest, and expectations [3] and deduction of knowledge from logs. This knowledge can be revealed by applying different web usages mining approaches on the weblog repositories which keep the record of past surfing behavior of users [4], [5]. The clustering is proved very effective in grouping users with similar browsing activities, navigational pattern, and access behavior [6], [7], [8]. However, clustering results are very much depending on similarity metric used for capturing user concerns and their accommodation. Normally, the higher precisions of similarity metric lead to the enhanced quality clusters of web user sessions and vice versa [9]. This paper discusses a combined dissimilarity based model to estimate a user's concern from augmented web user sessions. This dissimilarity metric is based on accessing page relevance and syntactic structure of page URLs. The page relevance is estimated using harmonic mean of the duration and frequency of the page while syntactic similarity is derived from the syntactic structure of URLs. The effective clustering of the web user session requires a precise definition of similarity metric between user sessions. In literature range of similarity measures are reported to recognize accurate similarity among users based on their access behavior. In [10] visitor behavior is consolidated by visited page content to built up likeness between various page groupings. In [11] common paths between two sessions isolated into the inward part and the external part of generalized web session and a new similarity measure is proposed from users' navigation patterns to identify the common paths. In [12] the element of uncertainty in user's navigation patterns was handled using belief function based similarity measure and Dempster-Shafer's theory of mass distribution for clustering. In [13] an enhanced similarity based user session clustering algorithm was presented to reduce the time and space complexity convention k-means [14] and ROCK algorithm [15]. The sequence alignment based similarity method and associated measures were used to cluster web user sessions [16]. In [9] comparative study of the effectiveness of most popular similarity measures including cosine, Jaccard, and Pearson coefficient was performed, and this work is further extended by [17] and [18]. In [19], [20] The user interest on a page was defined by implicit measures and further extended by [21] and proposed some implicit measures such as duration and frequency of the page for assessing the interest of a user for a particular accessed page. In [22] author used abstract similarity graph and comparative time used up on general greatest subsequence for clustering. In [23] the various leveled structure of universal resource locators (URLs) on the site was used to introduce the concept of sequence alignment and clustering. The relationship due to the concept hierarchy of web pages is quantified and incorporated it in similarity measure using a binary vector representation of web user sessions, assuming an equal degree of user concern are presented in [24]. The exact page viewing time of visiting pages and URL of the pages are used to find similarity between users in [13]. In [25] a website concept hierarchy based similarity scoring system was introduced and further integrated with other similarity measures like page stay time and page visiting sequence. In [26] the sessions intuitive augmented similarity concept is discussed, and the effectiveness of same was tested in [27] and [28].

The rest part of this paper is arranged under following sections: The materials and methods used in present work are discussed in Section II. In Section III experiments are performed with combined similarity approach and with individual measures and results are discussed in detail. Lastly, in Section IV, this study is concluded with some suggested future work.

II. Material and Method

The methodology adopted to carry out the present work is described in following subsections.

A. Pre-Processing of Web Server Logs

The explicit or implicit web access behavior of a user is recorded in weblogs. The weblogs maintain user information in the form of standard fields. The exact sequence and number of the field are varying from one log format to others. However essential information including address and the login name of the remote host, name of the user, request time, method, full path of the server, used protocol, status code, access data size, and agents used to access the information are present in every format [29]. All recorded information like entries made by automated software agents [30], default entries for embedded objects, unsuccessful requests and non-human user method are not useful. First, we remove these entries from log files. Second, user sessions are extracted from log entries. The user sessions are accumulated activity of a user on a web server. Whenever a fresh IP address is originated in the log file, a new session is generated and subsequently demand from the identical IP address is supplemented to the session based on some predefined elapse time. In this work 30 minutes elapsed time is used. Otherwise, a fresh session is initiated by closing the current session. Multiple sessions are possible for any web user because of the possibility of multiple visits and spending of an outlandish measure of time between back to back visits [30], [31].

B. Session Representation in Vector Space Model

For a particular website; we are assuming m usage sessions are identified $S_i=\{S_1,S_2,\ldots.S_m\}$ and n number of different URL's (pages) $P_i=\{P_1,P_2,\ldots.P_n\}$ are accessed in some time interval. Then each user session S_i may be represented by the following equation $S_i=\{S_i^1,S_i^2,\ldots S_i^n\}$, $\forall i=1,2,\ldots,m$. Where every S_k^i corresponds to a harmonic mean of the frequency of page P_k within the session S_i, and the duration of the page P_k in session S_i, as shown in Eq. (1) and Eq. (2).

$$S_k^i \leftarrow \begin{cases} \text{Page frequency} \\ \text{Page duration (in seconds)} \\ \text{Page size (in bytes)} \end{cases} \tag{1}$$

$$R[m,n]=\begin{pmatrix} S_1^1 & S_1^2 & \cdots & S_1^n \\ S_2^1 & S_2^2 & \cdots & S_2^n \\ \vdots & \vdots & \ddots & \vdots \\ S_m^1 & S_m^2 & \cdots & S_m^n \end{pmatrix} \tag{2}$$

C. Computation of Scale of Web User Concern for a Web Page

In [20] concept of implicit measure of user interest of a page was introduced and further extended by [21] and proposed Page stay time and page access frequency to measure the user concern for a page. The following metrics are used to compute the relevance of a page in any user session. Further, this relevance is used to measure the web user concern for a web page.

1) Duration of Page(DoP): The time used up by a user on any page is known as the duration of the page. Measured by the precise time difference between two consecutive requests for web pages in the session. A Higher value of page stay time implies more concern of the user to any page. However, sometimes the small size of a web page may lead to a swift transition to another page. Therefore, the time spent on the page is normalized by the page size. The time spent on the page is again normalized by the max page stay time in that session. The Eq. (3) is used to measure the DoP (P_i) in session (S_k)

$$(DoP)_{P_i}=\frac{\frac{\sum \textit{Time Spent on}(p_i)}{\textit{Size of }(P_i)}}{Max\left(\forall_{j\in S_k}\frac{\sum \textit{Time Spent on}(P_j)}{\textit{Size of }(P_j)}\right)} \tag{3}$$

Where $0\leq(\mathcal{D}\sigma\mathcal{P})_{\mathcal{P}_i}\leq 1$.

However, in the case of last access page, it is not feasible to compute the difference of requests time. Therefore, the average duration of the relevant session may be considered.

2) Frequency of Page (FoP): The count of visits of a page $\mathcal{P}_i$ in any session. The high value of this count indicates more concern of a user for any page. The frequency of a page is divided by the accumulated frequency in the session: The Eq. (4) is used to determined the FoP (P_i) in user session (S_k)

$$(FoP)_{P_i}=\frac{\sum \#\textit{ of visits to}(P_i)}{Max\left(\forall_{j\in S_k}\sum \#\textit{ of visits to}(P_j)\right)} \tag{4}$$

Where $0 \leq (\mathrm{FoP})_{\mathrm{P_i}} \leq 1$.

These two metrics are consolidated to measure the page relevance for a user. The harmonic mean of $(\mathrm{DoP})_{\mathrm{P_i}}$ and $(\mathrm{FoP})_{\mathrm{P_i}}$ is computed for estimating user concern because it will moderate the impact of large and small outliers [6]. After applying Eq. (1) and (2) on pre-processed log intermediate results are generated as shown in Table1

3) The Relevance of the Page (RoP): From Table1 the page relevance is computed by the harmonic mean of DoP and FoP. The Eq. (5) is used to calculate the relevance of a page (RoP) ($\mathcal{P}_i$) in user session ($\mathcal{S}_k$)

$$(\mathrm{RoP})_{\mathrm{P_i}}=\frac{2\times(\mathrm{DoP})_{\mathrm{P_i}}\times(\mathrm{FoP})_{\mathrm{P_i}}}{(\mathrm{DoP})_{\mathrm{P_i}}+(\mathrm{FoP})_{\mathrm{P_i}}} \tag{5}$$

Where $0 \leq (\mathrm{RoP})_{\mathrm{P_i}} \leq 1$.

The Eq. (5) demonstrates that high value of both DoP and FoP leads higher user concern for a particular page in any session.

4) Augmented Web User Sessions: first, we compute page relevance matrix ($RM_{m\times n}$) Eq. (6) of using equations (3) to (5). The relevance of each page in every session is calculated from this relevance matrix. The high value of relevance suggests more user concern for the page.

$$\mathrm{RM}_{m\times n}=\begin{pmatrix}(\mathrm{RoP})_{11} & (\mathrm{RoP})_{12} & \cdots & (\mathrm{RoP})_{1n}\\ (\mathrm{RoP})_{21} & (\mathrm{RoP})_{22} & \ldots & (\mathrm{RoP})_{12}\\ \vdots & \vdots & \ddots & \vdots\\ (\mathrm{RoP})_{m1} & (\mathrm{RoP})_{m2} & \cdots & (\mathrm{RoP})_{mn}\end{pmatrix} \tag{6}$$

The augmented user session is represented as $AS_a = \{(P_1, (RoP)_{P_{i1}}), (P_2, (RoP)_{P_2}) \ldots (P_n, (RoP)_{P_n})\}$, where P_i, $(RoP)_{P_i}$ and are the visited pages, and its relevance respectively.

5) Augmented Session Similarity Based on Page Relevance: The Eq. (7) shows the modified cosine similarity measure by incorporating relevance of a page. This is termed as page relevance based augmented session similarity measure and more realistic than binary cosine session similarity measure [9].

$$\mathrm{ASS}_{(\mathrm{AS_a},\mathrm{AS_b})}=\frac{\sum_{i=1}^{m}\mathrm{AS_a(RoP)_i}\times\mathrm{AS_b(RoP)_j}}{\sqrt{\sum_{i=1}^{m}\mathrm{AS_a(RoP)_i}^2}\sqrt{\sum_{i=1}^{m}\mathrm{AS_b(RoP)_j}^2}} \tag{7}$$

However, the key constraint of this measure will remain same as it abandoned the hierarchical grouping of web URL’s and placed in the same directory if web pages are related [32]

TABLE I
SESSION TABLE WITH PAGE FREQUENCY AND DURATION

Session ID	Host Name	Actual list of pages accessed during session	No. Of Pages accessed during session	Frequency of each page	Duration on each page	Size of each page
1	IP_1	P_{11}	N_1	f_{11}	d_{11}	s_{11}
		P_{12}		f_{12}	d_{12}	s_{12}
		.		.	.	.
		P_{1n}		f_{1n}	d_{1n}	s_{1n}
.		.	.	.	.	.
.		.		.	.	.
i	IP_i	P_{i1}	N_i	f_{i1}	d_{i1}	s_{i1}
		.		.	.	.
		P_{in}		f_{in}	d_{in}	s_{in}

D. URL based Syntactic Similarity between i^th^ and j^th^ Page URL’s

In [32] authors also consider the syntactic structure of URL’s and computed their similarity by their respective position in the web hierarchy by using Eq. (8). This measure will quantify the path overlapping between different visited pages.

$$\mathrm{USS}_{\left(\mathrm{US_a^{p_i}},\mathrm{US_b^{p_j}}\right)}=\mathrm{Min}\left(1,\frac{\left|\mathrm{LoP}\left(\mathrm{P_{(a,i)}}\right)\cap\mathrm{LoP}\left(\mathrm{P_{(b,j)}}\right)\right|}{\mathrm{Max}\left(1,\,\mathrm{Max}\left(\mathrm{LoP}\left(\mathrm{P_{(a,i)}}\right),\,\mathrm{LoP}\left(\mathrm{P_{(b,j)}}\right)\right)\text{-}1\right)}\right) \tag{8}$$

Where $\mathrm{LoP}\left(\mathrm{P_{(a,i)}}\right)$ is the length of URL (or a number of edges) of the path followed by the root node and a particular node of $\mathrm{P_i}$ in the user session $\mathrm{US_a}$. By incorporating this syntactic similarity of page URL’s, the similarity between two augmented web user sessions $\left(\mathrm{AS_a^{p_i}},\mathrm{AS_b^{p_j}}\right)$ is computed by Eq. (9).

$$\mathrm{AUSS}_{\left(\mathrm{AS_a^{p_i}},\mathrm{AS_b^{p_j}}\right)}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{AS_a(RoP)_i}\times\mathrm{AS_b(RoP)_j}\times\mathrm{USS}_{\left(\mathrm{US_a^{p_i}},\mathrm{US_b^{p_j}}\right)}}{\sum_{i=1}^{m}\mathrm{AS_a(RoP)_i}\times\sum_{j=1}^{m}\mathrm{AS_b(RoP)_j}} \tag{9}$$

Here, they assume a uniform URL based Syntactic Similarity of 1 for any node and its parent as well as all sibling nodes of any parent.

E. URL based Syntactic Similarity between i^th^ and j^th^ Page URL’s

Page relevance plays the major role in deciding the similarity of two web user sessions. If any web page pair has small $\mathrm{USS}_{\left(\mathrm{US_a^{p_i}},\mathrm{US_b^{p_j}}\right)}$ (URL based syntactic similarity) then $\mathrm{ASS}_{(\mathrm{AS_a},\mathrm{AS_b})}$ (Page relevance based augmented session similarity) will give good results and for a large value of $\mathrm{USS}_{\left(\mathrm{US_a^{p_i}},\mathrm{US_b^{p_j}}\right)}$, $\mathrm{AUSS}_{\left(\mathrm{AS_a^{p_i}},\mathrm{AS_b^{p_j}}\right)}$ (Augmented URL based syntactic similarity) Will be producing better results. Proposed combined augmented session similarity (CASS) measure utilizes the best characteristics of both the individual measures and considers the most idealistic accumulation to generate the better similarities between web user sessions using Eq. (10).

$$\mathrm{CASS}_{(AS_a,AS_b)}=\mathrm{Max}\left\{\mathrm{ASS}_{(AS_a,AS_b)},\mathrm{AUSS}_{\left(AS_a^{p_i},AS_b^{p_j}\right),}\right\} \quad (10)$$

As a requirement of relational clustering, this combined augmented session similarity is converted to the dissimilarity metric using Eq. (11).

$$D^2_{(AS_a,AS_b)}=\left(1-\mathrm{CASS}_{(AS_a,AS_b)}\right)^2 \quad (11)$$

Where $0 < D^2_{(AS_a,\, AS_b)} \leq 1$, for AS_a, AS_b=1, 2.....m.

If $D_{m\times n}$ is a dissimilarity relation denoted by $D^2_{(AS_a,AS_b)}$, Which satisfies the necessary conditions [18] of to be a Euclidean metric. (i) Non-Negativity (ii) Self Dissimilarity and (iii) Symmetry.

The above-described methodology for proposed framework is summarized in Algorithm 1. This algorithm is used to compute essential dissimilarity matrix for relational clustering of web user sessions.

F. K-medoids Clustering

To evaluate the performance of the proposed augmented user session (dis) similarity framework, we apply the most common implementation of K-medoids algorithm known as partition around medoids (PAM) on user session data with different session relational measures computed in this framework. The K-medoids is preferred for experimentation because; it selects actual user sessions as a cluster prototype to represent the cluster. At the same time in K-means[14], cluster prototypes (known as centroid) are the mean of sessions of that cluster. This fundamental difference makes the k-medoids algorithm more suitable for user session (dis)similarity metrics (relational data) [33].

Given a set of user sessions $S_i=\{S_1,S_2,\ldots.S_m\}$ for i=1,2,…m, where the vector of n-dimensions represents each session $S_i=\{s_i^1,s_i^2,\ldots s_i^n\}, \forall i=1,2,\ldots,m$. The objective of K-medoids clustering algorithm is to find k representative sessions known as medoids, to such an extent that the aggregate difference between different sessions to their nearest medoid is minimized. Let $C\leftarrow\{c_1,c_2,\ldots c_k\}$ be the set of medoids**.** The k-medoids objective function is shown as Eq.(12)

$$F_{\text{k-medoids}}=\sum_{i=1}^{m}\left(\sum_{c=1}^{k}\mu_{ci}\, d^2_{ci}(S_i,\delta_c)\right) \quad (12)$$

Where, S_i is ith user session. δ_c is the medoid of cluster C_c.

$$\mu_{ci}=\begin{cases}1, \text{ if } S_i\in C_c\\ 0 \text{ otherwise}\end{cases}$$

$d^2_{ci}(S_i,\delta_c)$ is the dissimilarity between session $\mathcal{S}_i$ and medoid of cluster $\mathcal{C}_c$ represented as δ_c. The membership functionμ_{ci} that minimizes $F_{\text{k-medoids}}$ can be derived from Eq. (13):

$$q=\mathrm{argmin}_{1\leq c\leq k}\, d^2_{ci}(S_i,\delta_c)\ ;\ \mu_{ci}=\begin{cases}1; & \text{if } c=q\\ 0; & \text{otherwise}\end{cases} \quad (13)$$

Once the membership matrix $\mathcal{U}=[\mu_{ci}]$ is fixed the new cluster medoid δ_c that minimize $F_{\text{k-medoids}}$ can be derived by the Eq.(14):

$$\delta_c=\mathrm{argmin}_{S_i\in C_c}\sum_{S_j\in C_c} d^2_{ij}\left(S_i,S_j\right) \quad (14)$$

Algorithm 1: Page relevance and *URL based syntactic (dis)similarity metric* of web user sessions.

```
Input:  {Log file:   $\mathcal{L}$ of $n$ records
          where $L\leftarrow\{r_1,r_2\ldots r_n\}$ , where $n\ggg 1$,
          $U_{n\times n}$ - URL based syntactic similarity matrix}
Output:  { $D_{m\times m}$ | (Dis) similarity matrix}

1) Pre-processing of web server access log
    • Removal of extraneous information:
      $L^c\leftarrow\{r_1,r_2\ldots r_n\}$ Log file after cleaning
    • Identification of web users $U_i=\{U_1,U_2,\ldots.U_n\}$ for i=1,2,…n.
    • Identification of web user sessions $S_i=\{S_1,S_2,\ldots.S_m\}$ for i=1,2,…m . Where m ≥ n
2) Vector representation of web user sessions $S_i=\{s_i^1,s_i^2,\ldots s_i^n\}, \forall i=1,2,\ldots,m$.
3) Computation of relevance of any web page in user sessions for each pair of $(S_k,P_i)$ in session $S_k$ and page $P_i$ where, i =1,2,…n. and k=1,2,…,m.
    • Compute the duration of a web page ($P_i$) in user session ($S_k$) using Eq. (3).
    • Compute the Frequency of web page ($P_i$) in user session ($S_k$) using eq.(4).
    • Compute the relevance of web page ($P_i$) in user session ($S_k$) using Eq. (5).
4) Computation of page relevance based augmented session similarity matrix by using Eq. (7).
5) Compute the page relevance and hierarchical URL similarity based augmented session similarity matrix by using Eq. (9).
6) Compute the combined augmented session similarity by using Eq. (10).
7) Compute the combined augmented web user session (dis)similarity matrix by using Eq. (11).
```

G. Validity of Generating Clusters

The unsupervised cluster assessment techniques are utilized to validate the worth of produced clusters. These techniques are based on compactness and separation which is measured by intra-cluster (Eq.(12)) and inter-cluster (Eq.(13) distances [34] [35], [36]. The low estimation of intra-cluster and high estimation of inter-cluster separation is desirable for worthy clusters [37].

$$D_{\text{intra}}=\frac{\sum_{S_{k\in C_i}}\sum_{S_{l\in C_{i,l\neq k}}} d^2_{kl}(AS_k,AS_l)}{|C_i|\,(|C_i|-1)} \quad (15)$$

$$D_{\text{inter}}=\frac{\sum_{S_{k\in C_i}}\sum_{S_{l\in C_{j},l\neq k,}} d^2_{kl}(AS_k,AS_l)}{|C_i|\,|C_j|} \quad (16)$$

Where, $d_{kl}^{2}(AS_k, AS_l)$ is the distance between two sessions in the cluster C_i and $|C_i|$ is the number of sessions in C_i.

III. Results and Discussion

The combined augmented session similarity framework is evaluated for its effectiveness and efficiency. The experiments are performed on the web server access log recorded from a heavy trafficked and global service provider portal. This portal provides information about visa, insurance, and other related services to international travelers to the USA. The proposed framework is implemented and tested using MATLAB [38]. Experiments were performed on Intel® Core™ i3 (CPU M370 2.40 GHz) with 4.00 GB RAM, and operating system Windows 7 (64-bit OS). In all experiments, k-medoid clustering was performed with different dissimilarity measures and computed both average intra-cluster and inter-cluster distances against a varying number of user sessions clusters. First, we applied pre-processing [29] on the used web log file and generated 468 sessions using 30 minute threshold time. We have filtered out 128 web robot sessions because autonomous software agents generate them and their access behavior is monotonous [31]. There are 216 very small sessions of length 1 or 2 are discarded from total identified session because they are not able to convey any useful information for clustering. After this process, we obtained 124 user sessions. These user sessions combined accessed more than 150 unique pages. We have reduced pages to 95 unique pages by removing very less frequent and very short duration pages visited by this session. First, we compute $D_{124\times124}$ using Eq. (7) and get augmented session dissimilarity $ASS_{(AS_a, AS_b)}$ Metric and run k-medoid clustering for 5, 10, 15 and 20 number of clusters. This algorithm assigns 124 sessions to a different number of clusters respectively. The summarized readings of experiments are recorded in Table 2. Then a 95×95 URL Similarity metric computed for all referred unique pages(URLs) by using Eq.(9) and by using this a URL based Syntactic dissimilarity $AUSS_{\left(AS_a^{p_i}, AS_b^{p_j}\right)}$ of 124×124 by Eq. (8) is computed. Again, the same procedure is adopted, and the results are reported in Table 2. Now we compute a combined augmented session dissimilarity metric $CASS_{(AS_a, AS_b)}$ by using Eq.(10) and (11) which takes the maximum value of dissimilarity from both dissimilarity matrices.

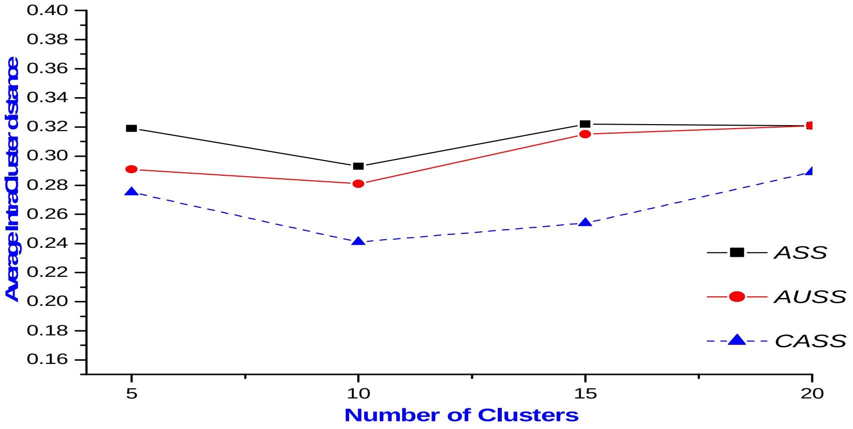

Fig. 1 Avg. Intra-Cluster distance Vs. No. of clusters for different dissimilarity measures

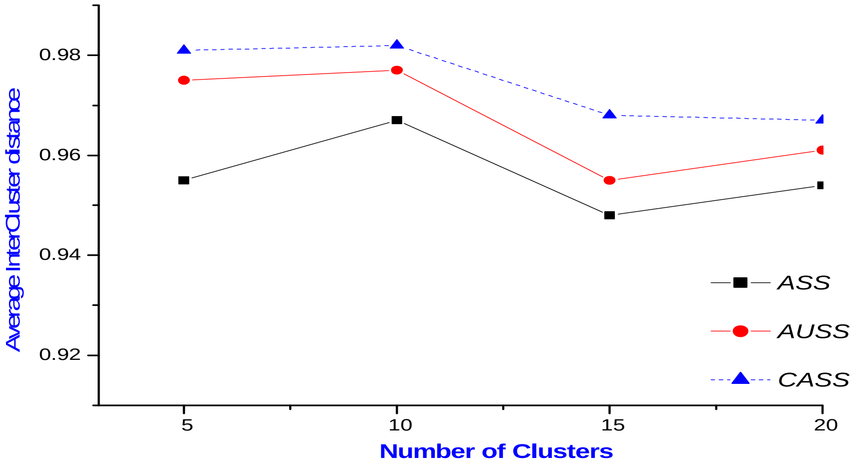

Fig. 2 Avg. Inter-Cluster distance Vs. No. of clusters for different dissimilarity measures

TABLE II
SUMMARY OF AVERAGE INTRA-CLUSTER AND INTER-CLUSTER DISTANCE FOR DIFFERENT (DIS) SIMILARITY MEASURES AND VARYING NO. OF CLUSTERS BY K-MEDOIDS CLUSTERING

K-Medoids Clustering with	No. Of Clusters	Avg. D_{intra}	Avg. D_{inter}
1. ASS$_{(AS_a,AS_b)}$	5	0.319	0.955
	10	0.293	0.967
	15	0.322	0.948
	20	0.321	0.954
2. AUSS$_{\left(AS_a^{p_i},AS_b^{p_j}\right)}$	5	0.291	0.975
	10	0.281	0.977
	15	0.315	0.955
	20	0.321	0.961
3. CASS$_{(AS_a,AS_b)}$	5	0.275	0.981
	10	0.241	0.982
	15	0.254	0.968
	20	0.289	0.967

IV. CONCLUSION

In this paper, the Page relevance based augmented session dissimilarity, URL based syntactic dissimilarity, and their combinations to take the advantages of both are evaluated. The dissimilarity metrics generated by these measures are passed as input to k-medoid clustering. The k-medoid generates the different number of clusters. The worthiness of produced clusters is evaluated by average intra-cluster and inter-cluster distance measures. The different value for input clusters is passed to k-medoid clustering to get optimized and stable results. The results suggest that combined session dissimilarity measure CASS$_{(AS_a,AS_b)}$ outperformed the individual ASS$_{(AS_a,AS_b)}$ and AUSS$_{\left(AS_a^{p_i},AS_b^{p_j}\right)}$ Dissimilarity measure on both intra-cluster and inter-cluster evaluation parameters. It is also observed that optimize and stable results are found with ten numbers of clusters. In this experiment we consider a small size log to avoid preprocessing overhead but in the same future experiment can be extended on the large log file. The same hypothesis can be tested with other clustering algorithms with a different set of cluster validation measures to generalize the results.

REFERENCES

[1] M. K. M. Nasution, "Social Network Mining (SNM): A Definition of Relation between The Resources and SNA," *International Journal of Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 975–981, 2016.

[2] F. M. Facca and P. L. Lanzi, "Mining interesting knowledge from weblogs: A survey," *Data and Knowledge Engineering*, vol. 53, no. 3, pp. 225–241, 2005.

[3] N. Saidin, D. Singh, Z. Akramin, M. Drus, and R. Hidayat, "Cultural Marker Identification for Web Application Design Targeted for Malaysian Multicultural Users," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 2016, pp. 959–965, 2016.

[4] F. Ramli, S. Azman, and M. Noah, "Building an Event Ontology for Historical Domain to Support Semantic Document Retrieval," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 1154–1160, 2016.

[5] B. Mobasher and R. Cooley, "Automatic Personalization Based on Web Usage Mining," *Communications of the ACM*, vol. 43, no. 8, pp. 142–151, 2000.

[6] A. Vakali, J. Pokorný, and T. Dalamagas, "An overview of web data clustering practices," in *Current Trends in Database WebKdd*, 2004, no. LNCS 3268, pp. 597–606.

[7] D. S. Sisodia, S. Verma, and O. Vyas, "A Discounted Fuzzy Relational Clustering of Web Users ' Using Intuitive Augmented Sessions Dissimilarity Metric," *IEEE Access*, vol. 4, no. 1, pp. 2883–2993, 2016.

[8] D. S. Sisodia, S. Verma, and O. P. Vyas, "A Conglomerate Relational Fuzzy Approach for Discovering Web User Session Clusters from Web Server Logs," *International Journal of Engineering and Technology*, vol. 8, no. 3, pp. 1433–1443, 2016.

[9] A. S. Joydeep, E. Strehl, J. Ghosh, R. Mooney, and A. Strehl, "Impact of Similarity Measures on Web-page Clustering," in *In Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 2000, pp. 58–64.

[10] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber, "A new similarity measure to understand visitor behavior on a website," *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, vol. E87–D, no. 1, pp. 1–8, 2004.

[11] Q. Y. Q. Yang, J. K. J. Kou, F. C. F. Chen, and M. L. M. Li, "A New Similarity Measure for Generalized Web Session Clustering," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2007, vol. 3.

[12] Y. Xie and V. V. Phoha, "Web user clustering from access log using belief function," in *Proceedings of the international conference on Knowledge Capture K-CAP 2001*, 2001, pp. 202–208.

[13] C. Li and Y. Lu, "Similarity measurement of web sessions by sequence alignment," in *Network and Parallel Computing Workshops*, 2007, pp. 716–720.

[14] A. K. Jain, "Data Clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[15] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," in *Proceedings of 15th International Conference on Data Engineering, IEEE*, 1999, pp. 512–521.

[16] B. Hay, G. Wets, and K. Vanhoof, "Segmentation of visiting patterns on websites using a sequence alignment method," *Journal of Retailing and Consumer Services*, vol. 10, no. 3, pp. 145–153, 2003.

[17] Y. Yaslan and Z. Cataltepe, "A comparison framework of similarity metrics used for web access log analysis," in *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2007, pp. 144–152.

[18] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008)*, 2008, pp. 49–56.

[19] P. K. Chan, "A non-invasive learning approach to building web user profiles," in *Proceedings of Workshop on Web Usage Analysis(KDD-99)*, 1999, pp. 7–12.

[20] J. Xiao and Y. Zhang, "Clustering of web users, using session-based similarity measures," in *International Conference on Computer Networks and Mobile Computing*, 2001, no. 2001, pp. 223–228.

[21] H. Liu and V. Kešelj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data & Knowledge Engineering*, vol. 61, no. 2, pp. 304–330, May 2007.

[22] A. Banerjee and J. Ghosh, "Clickstream clustering using weighted longest common subsequences," in *Proceedings of the Workshop on Web Mining SIAM Conference on Data Mining*, 2001, pp. 33–40.

[23] W. Wang and O. R. Zaiane, "Clustering Web sessions by sequence alignment," in *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, 2002, pp. 394–398.

[24] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites," *IEEE Transactions on Knowledge and data engineering*, vol. 20, no. 2, pp. 202–215, 2008.

[25] A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, “Incorporating concept hierarchies into usage mining based recommendations,” in *Proceedings of the 8th Knowledge discovery on the web international conference on Advances in web mining and web usage analysis*, 2006, pp. 110–126.

[26] D. S. Sisodia, S. Verma, and O. P. Vyas, “Augmented Intuitive Dissimilarity Metric for Clustering Of Web User Sessions,” *Journal of Information Science, DOI: 10.1177/0165551516648259*, pp. 1–12, 2016.

[27] D. S. Sisodia, S. Verma, and O. P. Vyas, “Performance Evaluation of an Augmented Session Dissimilarity Matrix of Web User Sessions Using Relational Fuzzy C-means Clustering,” *International Journal of Applied Engineering and Research*, vol. 11, no. 9, pp. 6497–6503, 2016.

[28] D. S. Sisodia, S. Verma, and O. P. Vyas, “Quantitative Evaluation of Web User Session Dissimilarity measures using medoids based Relational Fuzzy clustering,” *Indian Journal of Science and Technology*, vol. 9, no. 28, pp. 1–9, 2016.

[29] D. S. Sisodia and S. Verma, “Web usage pattern analysis through weblogs: A review,” in *IEEE 9th International Joint Conference on Computer Science and Software Engineering(JCSSE 2012)*, 2012, pp. 49–53.

[30] D. S. Sisodia, S. Verma, and O. P. Vyas, “A Comparative Analysis of Browsing Behavior of Human Visitors and Automatic Software Agents,” *American Journal of Systems and Software*, vol. 3, no. 2, pp. 31–35, 2015.

[31] D. S. Sisodia, S. Verma, and O. P. Vyas, “Agglomerative Approach for Identification and Elimination of Web Robots from Web Server Logs to Extract Knowledge about Actual Visitors,” *Journal of Data Analysis and Information Processing*, vol. 3, no. 2, pp. 1–10, 2015.

[32] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germin, “A web usage mining framework for mining evolving user profiles in dynamic websites,” *IEEE Transactions on Knowledge and data engineering*, vol. 20, no. 2, pp. 202–215, 2008.

[33] J.-P. Mei and L. Chen, “Fuzzy clustering with weighted medoids for relational data,” *Pattern Recognition*, vol. 43, no. 5, pp. 1964–1974, 2010.

[34] M. Sammour and Z. Othman, “An Agglomerative Hierarchical Clustering with Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya, Malaysia,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 1127–1133, 2016.

[35] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, no. 2–3, pp. 107–145, 2001.

[36] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “Cluster Validity Methods : Part I,” in *ACM SIGMOD Record*, 2002, vol. 31, no. 2, pp. 40–45.

[37] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty, “Model-based evaluation of clustering validation measures,” *Pattern Recognition*, vol. 40, no. 3, pp. 807–824, 2007.

[38] MATLAB (2012a), “Software package.” [Online]. Available: http://www.mathworks.com.