Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/89455, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Quantitative Evaluation of Web user Session Dissimilarity measures using Medoids based Relational Fuzzy clustering

Dilip Singh Sisodia[1]*, Shrish Verma[2] and Om Prakash Vyas[3]

[1]Department of Computer Science and Engineering, National Institute of Technology, Raipur - 492010, Chhattisgarh, India; dssisodia.cs@nitrr.ac.in
[2]Department of Electronics and Telecommunication, National Institute of Technology, Raipur - 492010, Chhattisgarh, India
[3]Department of Information Technology, Indian Institute of Information Technology, Allahabad - 211011, Uttar Pradesh, India

Abstract

Background/Objectives: Proficient relational clustering of web users' sessions not only depends on clustering algorithm's character but also profoundly influenced by the used dissimilarity measures. Therefore, determining the right dissimilarity measure to capture the actual access behaviour of the web user is imperative for the significant clustering.**Methods:** In this paper, the concept of an augmented session is used to derive different augmented session dissimilarity measures. The quantitative performance evaluation of different session dissimilarity measures are performed using a relational fuzzy c-medoid clustering approach. The intra-cluster and inter-cluster distance based cluster quality ratio is used for performance evaluation. **Findings:** The experimental results demonstrated that augmented web user session dissimilarity in general, and intuitive augmented session dissimilarity, in particular, performed better than the other dissimilarity measures. **Improvements:** It is argued that augmented session similarity measures are more realistic and represent session similarities based on the web user's habits, interest, and expectations as compared to simple binary session similarity measures.

Keywords: Augmented user Sessions, Cluster Evaluation, Dissimilarity Measures, Fuzzy Clustering, Page Relevance, Web User Sessions

1. Introduction

Web portals are real effectual means to interact with clients for any business entity. They are paramount not only in retaining the existing clients but also attracting potential customers in more effective and efficient way[1]. Client's web browsing behaviour is automatically recorded in web server logs. These web server logs can further analyse to extract useful knowledge, and this analysis is known as the web usage mining[2].

Various web usages mining techniques are used on web server log data and among them clustering is a very effective way to group users with common browsing activities, access pattern, and navigational behaviour[3]. The primary objective of web user sessions clustering is to group web sessions based on similarity and consists of maximising the intra-group similarity while minimising the inter-group similarity.However, the high dimensionality and sparseness in URLs accessing data, web user sessions are difficult to represent by features vectors. The relative distance measure is preferred over feature vector to store the pairwise relationship between web user's sessions[4].

Web user sessions were represented by a vector of the dimension of page URLs on any website. Different values are assigned to these dimensions (page URLs) for various

*Author for correspondence

user access behavior analysis. The extensively used technique is to assign binary values to these dimensions based on their user access (1) or not access (0) of a web page (URL) in a particular session[5].Suppose any website with n number of unique and valid URLs then i^{th} user session is represented as a binary vector in n - dimensional space of URLs by Eq. (1)

$$\mathcal{S}_h^i = \begin{cases} 1, if\ the\ user\ accessed\ the\ h^{th} URL\ in\ i^{th} session \\ 0 \qquad Otherwise \end{cases} \tag{1}$$

After, binary vectorization of web user sessions in URL space, authors defines three similarity measures between any two user sessions by incorporating URL accessed and their syntactic structure. Here, these similarity measures are abstracted as Binary Session Similarity (BSS), Binary URL Syntactic Similarity (BUSS) and Combined Binary Session Similarity (CBSS)

In this paper, the concept of an augmented session was used to derive different augmented session similarity measures. A relational fuzzy c-medoid clustering approach was used to evaluate the performance of different session dissimilarity measures.

The remaining of this paper is arranged as follows: In section 2 brief review of the various similarity/dissimilarity measures are discussed. Section 3, reintroduced the concept of relational fuzzy c-medoid clustering in the current context. In section 4, describes an essential pre-processing step. In section 5, vector representations of web user sessions in n-dimensional URL space is explained. In section 6, details of computation of page relevance and the concept of augmented web user sessions are discussed. In section 7, page relevance based augmented session similarity measure is introduced. In section 8, a URL based syntactic similarity between two augmented sessions is described. Section 9, discussed the idea of Intuitive augmented session similarity. In section 10, clusters quality assessment measures are discussed. Experimental setup and results are presented in section 11. Lastly, section 12 concluded this study with some proposed future works.

2. Related work

In literature, a variety of similarity measures are reported over the years for measuring web user's session similarity. Among them, simple binary session similarity measures are most popular and extensively used in previous studies[5-8]. In[9,10] authors proposed an alternative of binary weights as access duration and access frequency of the page based weights to page URLs, in particular, sessions. In[11,12] concept of implicit measure of user interest of a page was introduced and further extended by[13,14] and proposed Page stay time (duration) and page access frequency to measure the user concern for a web page. In[15]a used sequence alignment and associated measure based similarity method to cluster web user sessions. The particular page viewing time of visiting pages and URL of the pages presented in[16]. In[17] authors introduced a website concept hierarchy based similarity scoring system that further integrated with other similarity measures like browsing order and time spent on a page. In[18] a new similarity measure for generalized web session clustering is defined on the common paths of users' navigation patterns which divide the similarity of common paths between two sessions into the inner part and the outer part. In[19] introduced the concept of mass distribution in Dempster-Shafer's theory and the belief function similarity measure to provide the ability to clustering algorithm to capture the uncertainty in web user's navigation behaviour.

3. Fuzzy C-Medoids Clustering

In[7], fuzzy c-medoid clustering (FCMdd)was applied on web server log data using simple binary session similarity measure. In this section, the concept of fuzzy relational c-medoids[4] is reproduced in the present context. Given a set of augmented user sessions. $\mathcal{AS}_i = \{\mathcal{AS}_1, \mathcal{AS}_2, \ldots \mathcal{AS}_m\}\ for\ i = 1,2,\ldots m$ Where, each session is represented by vector of n-dimensions $\mathcal{S}_i = \{\mathcal{AS}_i^1, \mathcal{AS}_i^2, \ldots \mathcal{AS}_i^n\}, \forall i = 1,2,\ldots,m$. Let $d(\mathcal{AS}_i, \mathcal{AS}_j)$ represent the dissimilarity between session $\mathcal{AS}_i$ and session $\mathcal{AS}_j$. Let $\mathcal{V} \leftarrow \{v_1, v_2, \ldots v_c\}, v_i \in \mathcal{D}$ represent a subset of dissimilarity matrix $\mathcal{D}$ with cardinality c.

Algorithm 1: Pseudo code for augmented session dissimilarity based fuzzy c-medoids (FCMdd) algorithm[4].

Input: $\{D_{m\times m}|$Augmented session dissimilarity matrix, $\mathcal{C}$ Number of clusters,

t_{max} |maximum no of iterations }

Output: { $\mathcal{V} \leftarrow \{v_1, v_2, \ldots v_c\}$ | set of real session medoids,

μ_{ij} |Fuzzy membership matrix}

1: Fix the number of medoids $\mathcal{C} > 1$ and select first medoids randomly;

2: set $\mathcal{V} \leftarrow \{v_1\}$; $t \leftarrow 1$;

3: for $t \leftarrow 2, \ldots, c$

4: $q \leftarrow \arg\max_{1 \le i \le n; \mathcal{S}i \notin \mathcal{V}} \min_{1 \le h \le |\mathcal{V}|} d\left(v_j, \mathcal{A}\mathcal{S}i\right)$;

$\mathcal{V}_t \leftarrow \mathcal{S}_q$;

5: $\mathcal{V} \leftarrow \mathcal{V} \cup \{\mathcal{V}\}$;

6: $t \leftarrow t +$;

7: End for

8: set $t \leftarrow 0$;

9: chose aset of initial medoids: $\mathcal{V} \leftarrow \{v_1, v_2, \ldots v_c\}$ from $\mathcal{D}^c$

10: While $\{(t \le t_{max})$ or $\mathcal{V}_{old} \leftarrow \mathcal{V}\}$

Compute memberships μ_{ij} that minimizes $\mathcal{F}_{\mathbf{FCMdd}}$:

11: For j j ← 1, 2 ...*c do*

12: For $i \leftarrow 1, 2, \ldots n$

13: The membership function is calculate using Eq.(3)

14: End for

15: End for

16: Store the current medoid: $\mathcal{V}_{old} \leftarrow \mathcal{V}$;

Compute the new medoids v_i that minimize $\mathcal{F}_{\mathbf{FCMdd}}$:

17: For j ← 1, 2 ...*c do*

18: $q \leftarrow \arg\min_{1 \le h \le n; \mathcal{S}_i \ne v_i} \sum_{i=1}^{n} \mu_{ij}{}^f d\left(\mathcal{A}\mathcal{S}_j, \mathcal{A}\mathcal{S}_i\right)$; $v_j \leftarrow \mathcal{A}\mathcal{S}_q$;

19: End for

20: $t \leftarrow t + 1$;

21: End While

Let $\mathcal{D}^c$ represents the set of all c-subsets *V of D* of the objective function of relational fuzzy c-medoids algorithm seeks to c number of representative sessions (known as medoids), such that the total dissimilarity of other sessions to their closest medoid is minimized. The objective function of fuzzy c-medoids is defined as Eq. (2) and membership functions is given by (3)

$$\mathcal{F}_{FCMdd} = \sum_{j=1}^{c} \left(\sum_{i=1}^{m} \mu_{i=1}^{f} d\left(\mathcal{A}\mathcal{S}_i, v_j\right) \right) \quad (2)$$

$$\mu_{ij} = \frac{\left(d\left(\mathcal{S}_i, v_j\right)\right)^{-\frac{1}{(f-1)}}}{\sum_{j=1}^{c} \left(d\left(\mathcal{S}_i, v_j\right)\right)^{-\frac{1}{(f-1)}}} \quad (3)$$

Where, $d\left(\mathcal{A}\mathcal{S}_i, v_j\right)$ is the dissimilarity between augmented session $\mathcal{A}\mathcal{S}_i$ and medoid of cluster $\mathcal{A}\mathcal{S}_j$ and $f \in [1, \infty]$ is fuzzification coefficient.

The above-described clustering process is summarized in the form of pseudo code in Algorithm 1 and the steps involved in fuzzy relational c-medoids clustering to discover web user clusters from page relevance based relational matrix of augmented web user sessions.

4. Web Server Logs Pre-Processing

The web server access log keeps a record of all files accessed by users explicitly or implicitly. Each log entry consists of different fields such that remote host address, remote loginname, username, timestamp and time zone of the request, request method, path on the server, protocol version, service status code, size of the returned data, and referrer user agent,etc[20]. First, we remove those entries which are not germane to our purpose; Mostly these were implicit requests made by embedded objects within web pages, requests made by automated software agents[21,22]. Secondly, web user sessions are identified by adopting the methods presented in[23,24].

5. User Session in Vector Space Model

Suppose, for a given website; there are m usage sessions extracted from the web server $\log \mathcal{S}_i = \{\mathcal{S}_1, \mathcal{S}_2, \ldots . \mathcal{S}_m\}$, which Access n number of different URL's (pages) $\mathcal{P}_i = \{\mathcal{P}_1, \mathcal{P}_2, \ldots . \mathcal{P}_n\}$ on a websitein some time interval.

We represent each user session $\mathcal{S}_i$ by following equation $\mathcal{S}_i = \{\mathcal{S}_i^1, \mathcal{S}_i^2, \ldots \mathcal{S}_i^n\} \forall i = 1, 2, \ldots, m.$ where, each $\mathcal{S}_h^i$ represents a harmonic mean of the number of visits to the page $\mathcal{P}_h$ within the session $\mathcal{S}_i$, and the duration of the page (in seconds) $\mathcal{P}_h$ in session $\mathcal{S}_i$, which is represented by following matrices using Eq.(4) and (5).

$$\mathcal{S}_h^i \leftarrow \begin{cases} Number\ of\ visits\ to\ the\ page \\ Time\ spent\ on\ page (in\ seconds) \\ Size\ of\ the\ page\ (in\ bytes) \end{cases} \quad (4)$$

$$\mathcal{R}[m,n] = \begin{pmatrix} \mathcal{S}_1^1 & \mathcal{S}_1^2 & \cdots & \mathcal{S}_1^n \\ \mathcal{S}_2^1 & \mathcal{S}_2^2 & \cdots & \mathcal{S}_2^n \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{S}_m^1 & \mathcal{S}_m^2 & \cdots & \mathcal{S}_m^n \end{pmatrix} \quad (5)$$

6. Computation of Page Relevance in any Web user Session

Web users' interests for any page are computed by implicit measures[11,12]. These implicit measures are page stay time (duration) and page access frequency of a web page in web user session[13]. The following measures are computed to find the relevance of a web page in any user session to measure the web user concern for a web page.

6.1 Duration of Page (DoP)

Eq. (6) is used to compute the duration of a web page ($\mathcal{P}_i$) in user session ($\mathcal{S}_h$)

$$(\mathcal{DOP})_{\mathcal{P}i} = \frac{\frac{\sum \text{Time Spent on}(\mathrm{p}_i)}{\text{Size of}(\mathrm{P}_i)}}{\text{Max}\left(\forall_{j \in \mathrm{S_h}} \frac{\sum \text{Time Spent on}(\mathrm{P}_j)}{\text{Size of}(\mathrm{P}_j)}\right)} \quad (6)$$

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

6.2 Frequency of Page (FoP)

Eq. (7) is used to compute the Frequency of the web page ($\mathcal{P}_i$) in user session ($\mathcal{S}_h$)

$$(\mathcal{FOP})_{\mathcal{P}_i} = \frac{\sum \#\ of\ visits\ to(P_i)}{Max\left(\forall_{j \in S_h} \sum \#\ of\ visits\ to(P_j)\right)} \quad (7)$$

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

6.3 Relevance of the page (RoP)

The relevance of the page in any user session was computed by giving equal importance to the duration of page and frequency of page because this harmonic mean will moderate the impact of large and small outliers. Eq. (8) was used to measure the relevance of a web page ($\mathcal{P}_i$) in user session ($\mathcal{S}_h$)

$$(\mathcal{ROP})_{\mathcal{P}_i} = \frac{2 \times (\mathcal{DOP})_{\mathcal{P}_i} \times (\mathcal{FOP})_{\mathcal{P}_i}}{(\mathcal{DOP})_{\mathcal{P}_i} + (\mathcal{FOP})_{\mathcal{P}_i}} \quad (8)$$

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

6.4 Augmented Web user Sessions

Now, by applying equations (6) to (8) the page relevance matrix ($\mathcal{RM}_{m \times n}$) is computed This relevance matrix will define the relevance of each page in every session. If the page has high relevance means the user has more concern in this page. This relevance matrix is given by Eq. (9). By incorporating page relevance in web user session access behaviour matrix, simple web user sessions converted to augmented web user sessions.

$$\mathcal{RM}_{m \times n} = \begin{pmatrix} (\mathcal{ROP})_{11} & (\mathcal{ROP})_{12} & \cdots & (\mathcal{ROP})_{1n} \\ (\mathcal{ROP})_{21} & (\mathcal{ROP})_{22} & \cdots & (\mathcal{ROP})_{12} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathcal{ROP})_{m1} & (\mathcal{ROP})_{m2} & \cdots & (\mathcal{ROP})_{mn} \end{pmatrix} \quad (9)$$

The augmented web session is represented as

$$\mathcal{AS}_a = \{(\mathcal{P}1, (\mathcal{ROP})_{\mathcal{P}1}, (\mathcal{P}2, (\mathcal{ROP})_{\mathcal{P}2}) \ldots (\mathcal{P}n, (\mathcal{ROP})_{\mathcal{P}n})\}.$$

Where, $\mathcal{P}i$, and $(\mathcal{ROP})_{\mathcal{P}i}$ are the visiting page, and its relevance respectively.

7. Page Relevance based Augmented Session Similarity

Here relevance of pages accessed in user sessions is incorporated in simple cosine similarity measure Eq. (10). This augmented session similarity measure represented more

meaningful web user session similarity as compared to simple binary user session based cosine measure[5].

$$\mathcal{ASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)} = \frac{\sum_{i=1}^{n} \mathcal{AS}_a(\mathcal{RoP})_i \times \mathcal{AS}_b(\mathcal{RoP})_j}{\sqrt{\sum_{i=1}^{n} \mathcal{AS}_a(\mathcal{RoP})_i^{\,2}} \sqrt{\sum_{i=1}^{n} \mathcal{AS}_b(\mathcal{RoP})_j^{\,2}}} \quad (10)$$

This augmented session similarity measure is more realistic and represents session similarities based on the web user's habits, interest, and expectations as compared to simple binary cosine measure.

8. A URL Syntactic Similarity between ith and jth Page URL

In [5] authors defined an alternative URL based syntactic similarity measure to compute the syntactic similarity between any pair of URLs given by Eq. (11).

$$\mathcal{USS}_{\left(\mathcal{US}_a^{p_i},\mathcal{US}_b^{p_j}\right)} = \mathcal{M}in\left(1, \frac{\left|\mathcal{LoP}\left(\mathcal{P}_{(a,i)}\right) \cap \mathcal{LoP}\left(\mathcal{P}_{(b,j)}\right)\right|}{\mathcal{M}ax\left(1,\ \mathcal{M}ax\left(\mathcal{LoP}\left(\mathcal{P}_{(a,i)}\right),\ \mathcal{LoP}\left(\mathcal{P}_{(b,j)}\right)\right)-1\right)}\right) \quad (11)$$

Where $\mathcal{LoP}\left(\mathcal{P}_{(a,i)}\right)$ is length of URL (or number of edges) of path traversed from root node to respective node of $\mathcal{P}_i$ in user session $\mathcal{US}_a$. By applying this syntactic similarity of URL's, the similarity between two augmented web user sessions $\left(\mathcal{AS}_a^{p_i}, \mathcal{AS}_b^{p_j}\right)$ is computed by Eq. (12).

$$\mathcal{AUSS}_{\left(\mathcal{AS}_a^{p_i},\mathcal{AS}_b^{p_j}\right)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{AS}_a(\mathcal{RoP})_i \times \mathcal{AS}_b(\mathcal{RoP})_j \times \mathcal{USS}_{\left(\mathcal{US}_a^{p_i},\mathcal{US}_b^{p_j}\right)}}{\sum_{i=1}^{n} \mathcal{AS}_a(\mathcal{RoP})_i \times \sum_{j=1}^{n} \mathcal{AS}_b(\mathcal{RoP})_j} \quad (12)$$

9. Intuitive Augmented Session Similarity

Intuitive augmented session similarity utilizes the properties of two measures and considers the maximum optimistic aggregation of these measures to give remarkable similarities between web user sessions[25]. Intuitive augmented session similarity computed using Eq. (13).

$$\mathcal{IASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)} = Max\left\{\mathcal{ASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)}, \mathcal{AUSS}_{\left(\mathcal{AS}_a^{p_i},\mathcal{AS}_b^{p_j}\right)}\right\} \quad (13)$$

As a requirement of relational clustering, this augmented session similarity is converted to the dissimilarity/distance measure. This distance measure satisfies the necessary conditions of a metric[26]. The augmented session dissimilarity is computed using Eq. (14).

$$\mathcal{D}^2_{(\mathcal{AS}_a,\mathcal{AS}_b)} = (1 - \mathcal{ASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)})^2 \quad (14)$$

Where $0 < \mathcal{D}^2_{(\mathcal{AS}_a,\mathcal{AS}_b)} \leq 1$, for $\mathcal{AS}_a$, $\mathcal{AS}_b$ =1, 2.....m.

Table 1. Summary of Results

Parameters	Values
Number of initial sessions	2000
Number of valid sessions	1341
Size of FoP/DoP/RoP matrix	1341×589
Number of unique URLs in sessions	589
Size of URL syntactic similarity matrix	589×589
Size of IBSS/IASS/ ($\mathcal{D}_{m\times n}$)matrix	1341×1341

10. Clusters Quality Assessment

To assess the quality of produced fuzzy clusters[27],unsupervised evaluation method which is based on intra-cluster and inter-cluster distance measures is used[28,29]. Intra-cluster distance represented compactness of a cluster and computed as an average of the distance between all pair of sessions within the ith cluster. For good quality of clusters, the small value of the intra-cluster distance is expected. Inter-cluster distance is a measure of separation between clusters and computed as an average of the distance between sessions from ithcluster and sessions from jth cluster. The high value of inter-cluster distance represents good partition[30]. The intra-cluster and inter-cluster distance is computed using Eq. (15) and Eq. (16) respectively. The cluster quality index for measuring the goodness of partitions can be designed to consider both compactness and separation using Eq. (17)

$$\mathcal{D}_{intra} = \frac{\sum_{\mathcal{S}_{h\in Ci}} \sum_{\mathcal{S}_{\ell\in C_i,\ell\neq h}} d^2_{h\ell}\left(\mathcal{AS}_h, \mathcal{AS}_\ell\right)}{|c_i|\left(|c_i|-1\right)} \quad (15)$$

$$\mathcal{D}_{\text{int}ra} = \frac{\sum_{\mathcal{S}_h \in \mathcal{C}_i} \sum_{\mathcal{S}_{\ell \in \mathcal{C}_j, \ell \neq h,}} d_{h\ell}^2 \left(\mathcal{AS}_h, \mathcal{AS}_\ell \right)}{\left| c_i \right| \left| c_j \right|} \quad (16)$$

$$\text{Cluster Quality Ratio}\left(\text{c}\right) = \frac{\text{Compactness}}{\text{Separation}} \quad (17)$$

Where, $d_{h\ell}^2 \left(\mathcal{AS}_h, \mathcal{AS}_\ell \right)$ is the distance between two sessions in cluster c_i and $\left| c_i \right|$ is the number of sessions.

11. Experimental Results and Discussions

In this section, quantitative evaluation of different web user session dissimilarity measures are performed. All Experiments have been performed on publicly available NASA web server log data[31]. All similarity/dissimilarity measures used in this study and the relational fuzzy c-medoids algorithm are implemented in MATLAB (R2012a) package[32]. Experiments were performed on an HPZ420 workstation with an Intel(R) Xeon(R) CPU E51620 0 @ 3.60 GHz, and 4 GB RAM, running under the MS Windows-7 operating system(64-bit).After performing the essential preprocessing on NASA server log data only 2000 number of user sessions are considered to reduce the system processing overhead.The default root/and mini sessions of size 1 are filtered out from the total generated sessions as they did not contribute any significant information for user session clustering. Total useful, valid sessions are reduced to 1341, which access 589 unique URLS collectively. First, the different matrices (FoP, DoP, RoP,and USS) from the given web user sessions had been computed.Second, different similarity/dissimilarity measures were calculated by applying the notion of binary sessions(BSS,BUSS and CBSS)[5] and augmented sessions(ASS,AUSS and IASS).The summary of computed results are shown in Table 1.

To assume suitable values of some clusters for experimentation purpose, a Visual Assessment Tendency (VAT) tool[33] was used. The VAT plot suggests the number of clusters present in web user pair wise dissimilarity matrix without using any clustering algorithm. The number of dark blocks along the diagonal in image produced by VAT tool represents the number of potential clusters. However, this is not always possible if no compact group exists in the data then this is not feasible[34].

The VAT images of augmented sessions(ASS,AUSS, IASS and BSS,BUSS,CBSS) are shown in Figure 1.

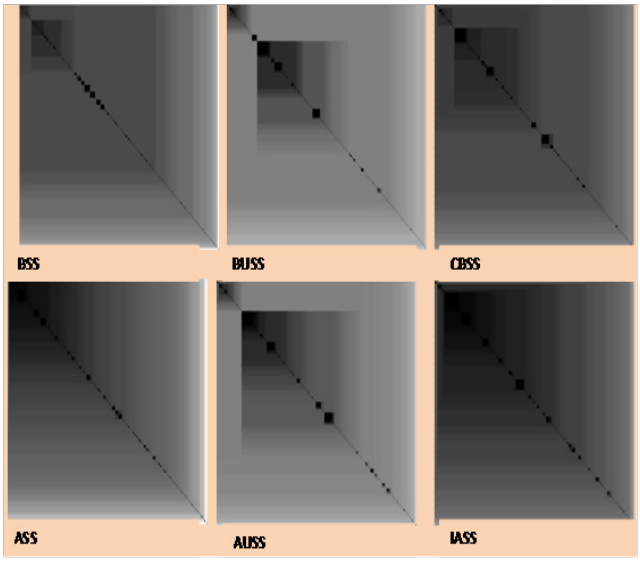

Figure 1. The VAT images of different augmented session dissimilarity matrices of 1341×1341.

Then, multiple runs of fuzzy relational c-medoids clustering algorithm was performed with six dissimilarity matrices of same size (1341×1341) generated for each dissimilarity measures. Default parameters were set during execution are fuzzifier coefficient ($f = 1.5 \text{ to } 2$), maximum number of iterations ($t_{max} = 100$) .For varying number of clusters (c=6, 8, 10), intra cluster and inter cluster distance was computed for each dissimilarity measure and to consider both compactness and separation cluster quality ratio (CQR) was computed. Table 2 summarized the performance of different dissimilarity measures. The low value of avg. Intra cluster distance represents good partition while a high value of inter cluster distance indicate the better cluster quality. It is very difficult to judge the cluster quality through any single measure. Therefore, cluster quality ratio was used to consider both and small value of this ratio will give better clusters. Figure 2(A) shows the average intra cluster distance obtained by RFCMdd with simple binary and augmented dissimilarity measures while Figure 2(B) represents average inter cluster distance for same. Figure 2(C) provides graphical representation of cluster quality ratio for different dissimilarity measures. It is evident from the Figure 2 that augmented dissimilarity measure gets optimum value of CQR. The intuitive augmented session similarity measure gives lowest value of CQR for cluster 6 and 8 while for cluster 10 it produce the nearby value of CQR.

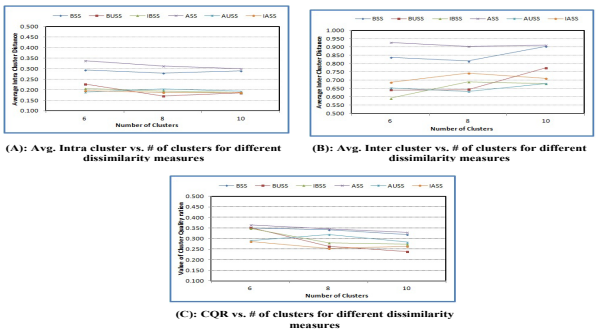

Figure 2. Performance comparison of different dissimilarity measures.

12. Conclusion and Future Work

This paper presented a quantitative performance evaluation of various user session dissimilarity measures using relational fuzzy c-medoid clustering.The concept of an augmented session was used to derive different augmentedsession dissimilarity measures (ASS, AUSS, and IASS). The performance of augmented dissimilarity measures was compared with popularly used simple binary session dissimilarity measures (BSS, BUSS, and CBSS). Multiple runs of relational fuzzy c-medoid clustering were performed with default parameters to generate a different number of clusters. The quality of generated clusters wasevaluated byaverage intra-cluster distance (compactness), average inter-cluster distance (separation) and cluster quality ratio. Experiments were performed with varying number clusters to generalize results. Experimental results demonstrate that intuitive augmented session dissimilarity measure (IASS) outperformed the other dissimilarity measures (both binary and augmented) on evaluation parameter of cluster quality ratio. In this study, only a small size log was considered to avoid pre-processing overhead, however, in the same future study may be extended to a large number of user sessions. The same hypothesis can be tested with other clustering algorithms with different web log data to generalize the results.

Table 2. Summary of average Intra and Intercluster distances for varying number of cluster and different dissimilarity measures

Number of Clusters	Dissimilarity Measures	Avg. Intra-Cluster Distance	Avg. Inter Cluster Distance	Cluster Quality Ratio
	BSS	0.293	0.838	0.350
	BUSS	0.226	0.642	0.353
C=6	CBSS	0.206	0.592	0.348
	ASS	0.339	0.928	0.365
	AUSS	0.190	0.654	0.291
	IASS	0.197	0.687	0.286
	BSS	0.278	0.816	0.341
	BUSS	0.170	0.644	0.264
C=8	CBSS	0.194	0.691	0.281
	ASS	0.313	0.903	0.346
	AUSS	0.203	0.634	0.320
	IASS	0.188	0.743	0.253
	BSS	0.289	0.904	0.320
	BUSS	0.185	0.774	0.239
C=10	CBSS	0.187	0.680	0.275
	ASS	0.300	0.911	0.329
	AUSS	0.193	0.681	0.284
	IASS	0.188	0.711	0.264

13. References

1. Lim M, Byun H, Kim J. A web usage mining for modeling buying behavior at a web store using network analysis. Indian Journal of Science and Technology. 2015; 8(25):1–7.
2. Guerbas A, Addam O, Zaarour O, Nagi M, Elhaj A, Ridley M, et al. Effective web log mining and online navigational pattern prediction. Knowledge-Based Systems. Elsevier B.V.; 2013; 49(12):50–62.
3. Mobasher B, Cooley R. Automatic Personalization Based on Web Usage Mining. Communications of the ACM. 2000; 43(8):142–51.
4. Krishnapuram R, Joshi A, Liyu Yi. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. IEEE International Fuzzy Systems Conference Proceedings (FUZZ-IEEE'99). IEEE; 1999. p. 1281–6.
5. Nasraoui O, Frigui H, Krishnapuram R, Joshi A. Extracting web user profiles using relational competitive fuzzy clustering. International Journal on Artificial Intelligence tools. 2000; 9(4):509–26.
6. Nasraoui O, Krishnapuram R, Anupam Joshi TK. Automatic web user profiling and personalization using robust fuzzy relational clustering. E-Commerce and Intelligent Methods. Physica-Verlag HD. 2002; 233–61.
7. Krishnapuram R, Joshi A, Nasraoui O, Yi L. Low-complexity fuzzy relational clustering algorithms for Web mining. IEEE Transactions on Fuzzy Systems. 2001; 9(4):595–607.
8. Nasraoui O, Cardona C. Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm. Proceedings of WebKDD. 2003; 71–81.
9. Yan TW, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. Computer Networks and ISDN Systems. 1996;28(7):1007–14.
10. Forsati R, Moayedikia A, Shamsfard M. An effective Web page recommender using binary data clustering. Information Retrieval Journal. Springer Netherlands. 2015; 18(3):167–214.
11. Chan PK. A non-invasive learning approach to building web user profiles. Proceedings of Workshop on Web Usage Analysis (KDD-99). 1999; 7–12.
12. Xiao J, Zhang Y. Clustering of web users using session-based similarity measures. In: International Conference on Computer Networks and Mobile Computing. 2001. p. 223–8.
13. Liu H, Keselj V. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data and Knowledge Engineering. 2007; 61(2):304–30.
14. Vakali A, Pokorny J, Dalamagas T. An overview of web data clustering practices. Current Trends in Database WebKdd. Springer Berlin Heidelberg. 2004; 597–606.
15. Hay B, Wets G, Vanhoof K. Clustering navigation patterns on a website using a Sequence Alignment Method. Intelligent Techniques for Web Personalization, IJCAI. 2001; 1–6.
16. Li C, Lu Y. Similarity measurement of web sessions by sequence alignment. In: IFIP International Conference on Network and Parallel Computing Workshops, NPC 2007. IEEE Computer Society. 2007. p. 716–20.
17. Bose A, Beemanapalli K, Srivastava J, Sahar S. Incorporating concept hierarchies into usage mining based recommendations. Proceedings of the 8th Knowledge Discovery on the Web International Conference on Advances in Web Mining and Web usage Analysis. 2006. p. 110–26.
18. Yang Q, Kou J, Chen F, Li M. A new similarity measure for generalized web session clustering. Proceedings - Fourth International Conference on Fuzzy Systems and Knowledge Discovery(FSKD 2007). 2007. p. 278–82.
19. Xie Y, Phoha V V. Web user clustering from access log using belief function. Proceedings of the International Conference on Knowledge Capture - K-CAP 2001. 2001. p. 202.
20. Sisodia D, Verma S. Web usage pattern analysis through web logs: A review. International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE. 2012. p. 49–53.
21. Sisodia DS, Verma S, Vyas OP. 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. 2015; 3(2):1–10.
22. Fernandez FMH, Ponnusamy R. Data preprocessing and cleansing in web log on ontology for enhanced decision making. Indian Journal of Science and Technology. 2016; 9(10):1–10.
23. Spiliopoulou M, Mobasher B, Berendt B, Nakagawa M. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. INFORMS Journal on Computing. 2003; 15(2):171–90.
24. Sisodia DS, Verma S, Vyas OP. A comparative analysis of browsing behavior of human visitors and automatic software agents. American Journal of Systems and Software. 2015; 3(2):31–5.
25. Sisodia DS, Verma S, Vyas OP. Augmented intuitive dissimilarity metric for clustering of web user sessions. Journal of Information Science. 2016; 1–12. Doi: 101177/0165551516648259.
26. Huang A. Similarity measures for text document clustering. Proceedings of the Sixth New Zealand. 2008 Apr; 49–56.

27. Revathy S, Parvaathavarthini B, Rajathi S. Futuristic validation method for rough fuzzy clustering. Indian Journal of Science and Technology. 2015; 8(2):120–7.
28. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. Journal of Intelligent Information Systems. 2001; 17(2-3):107–45.
29. Halkidi M, Batistakis Y, Vazirgiannis M. Cluster Validity Methods : Part I. ACM SIGMOD Record. 2002; 31(2):40–5.
30. Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, et al. Model-based evaluation of clustering validation measures. Pattern Recognition. 2007; 40(3):807–24.
31. NASA_Aug95. NASA Kennedy space centre's www server log data, Available at [Internet]. Available from: http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html
32. MATLAB(2012a). Software package [Internet]. Available from: http://www.mathworks.com.
33. Havens TC, Member S, Bezdek JC. An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm. IEEE Transactions on Knowledge and Data Engineering. 2012; 24(5):813–22.
34. Wang LA, Geng X, Bezdek J, Leckie C, Ramamohanarao K. iVAT and aVAT:Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning. Advances in