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

Dilip Singh Sisodia1*, Shrish Verma2 and Om Prakash Vyas3
1Department of Computer Science and Engineering, National Institute of Technology, Raipur - 492010, Chhattisgarh, India; dssisodia.cs@nitrr.ac.in
2Department of Electronics and Telecommunication, National Institute of Technology, Raipur - 492010, Chhattisgarh, India
3Department 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 way1. 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 mining2.

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 behaviour3. 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 sessions4.

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. Suppose any website with \( n \) number of unique and valid URLs then \( i^{th} \) user session is represented as a binary vector in \( n \times 1 \) dimensional space of URLs by Eq. (1)

\[
S_i = \begin{cases} 
1, & \text{if the user accessed the } h^{th} \text{ URL in } i^{th} \text{ session} \\
0, & \text{Otherwise} 
\end{cases} 
\]

(1)

After, binary vectorization of web user sessions in URL space, authors defined 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-medoids 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. In authors introduced a website concept of implicit measure of user interest of a page and proposed an alternative of binary

### 3. Fuzzy C-Medoids Clustering

In literature, 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 is reproduced in the present context. Given a set of augmented user sessions,

\[
\mathcal{A}_i = \{\mathcal{A}_{S_1}, \mathcal{A}_{S_2}, \ldots, \mathcal{A}_{S_m}\} \text{ for } i = 1, 2, \ldots, m
\]

Where, each session is represented by vector of n-dimensions

\[
\mathcal{S}_i = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n\}, \forall i = 1, 2, \ldots, m
\]

Let \( d(\mathcal{A}_i, \mathcal{A}_j) \) represent the dissimilarity between session \( \mathcal{A}_i \) and session \( \mathcal{A}_j \). Let \( \mathcal{V} = \{v_1, v_2, \ldots, v_c\}, v_i \in \mathcal{D} \) 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.

**Input:** \( \mathcal{D}_{\text{med}} \) Augmented session dissimilarity matrix, \( C \) Number of clusters,
\[ t_{\text{max}} \mid \text{maximum no of iterations} \]

**Output:** \[ \mathcal{V} \leftarrow \{v_1, v_2, \ldots, v_c\} \mid \text{set of real session medoids} \]

\[ \mu_{ij} \mid \text{Fuzzy membership matrix} \]

1. Fix the number of medoids \( C > 1 \) and select first medoids randomly;
2. set \( t \leftarrow 1 \);
3. for \( t \leftarrow 2, \ldots, c \)
4. \( q \leftarrow \text{argmax}_{1 \leq i \leq n, 1 \leq j \leq c} \min_{1 \leq s \leq p} d(A\mathcal{S}_j, v_j) \);\n5. \( \mathcal{V}_j \leftarrow \mathcal{S}_j \);
6. \( \mathcal{V} \leftarrow \mathcal{V} \cup \{\mathcal{V}_j\} \);
7. \( t \leftarrow t + 1 \);
8. end for
9. chose set of initial medoids: \( \mathcal{V} \leftarrow \{v_1, v_2, \ldots, v_c\} \)
from \( \mathcal{D}^c \)
10. While \( \{(t \leq t_{\text{max}}) \text{or } \mathcal{V}_{\text{old}} \leftarrow \mathcal{V}\} \)
11. Compute memberships \( \mu_{ij} \) that minimizes \( F_{\text{FCMdd}} : \)
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: \( \mathcal{V}_{\text{old}} \leftarrow \mathcal{V} \);
17. Compute the new medoids \( v_j \) that minimize \( F_{\text{FCMdd}} : \)
18. \( q \leftarrow \text{argmin}_{1 \leq i \leq n, 1 \leq j \leq c} \sum_{j=1}^{c} \mu_{ij} d(A\mathcal{S}_j, v_j) \);\n19. End for
20. \( t \leftarrow t + 1 \);
21. End While

Let \( \mathcal{D}^c \) represents the set of all \( c \)-subsets \( \mathcal{V} \) of \( \mathcal{D} \) 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)

\[ F_{\text{FCMdd}} = \sum_{j=1}^{c} \left( \sum_{i=1}^{m} \mu_{ij} d(A\mathcal{S}_i, v_j) \right) \]  \hspace{1cm} (2)

\[ \mu_{ij} = \frac{\left( d(A\mathcal{S}_i, v_j) \right)^{-1}}{\sum_{j=1}^{c} \left( d(A\mathcal{S}_i, v_j) \right)^{-1}} \]  \hspace{1cm} (3)

Where, \( d(A\mathcal{S}_i, v_j) \) is the dissimilarity between augmented session \( A\mathcal{S}_i \) and medoid of cluster \( A\mathcal{S}_j \) 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. 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. Secondly, web user sessions are identified by adopting the methods presented in.

### 5. User Session in Vector Space Model

Suppose, for a given website; there are \( m \) usage sessions extracted from the web server log \( \mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\} \), which Access \( n \) number of different URL’s (pages) \( \mathcal{P}_i = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\} \) on a website in some time interval.
We represent each user session \( S_i \) by following equation

\[
S_i = \{S_{i1}, S_{i2}, \ldots, S_{in}\} \quad \forall i = 1, 2, \ldots, m.
\]

where, each \( S_{ih} \) represents a harmonic mean of the number of visits to the page \( P_h \) within the session \( S_i \), and the duration of the page (in seconds) \( P_h \) in session \( S_i \), which is represented by following matrices using Eq. (4) and (5).

\[
(\text{DoP})_{ih} = \frac{\sum \text{Time Spent on}(P_i)}{\text{Size of}(P_i)} \quad \forall i = 1, 2, \ldots, m.
\]

Where \( 0 \leq (\text{DoP})_{ih} \leq 1. \)

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 \( (P_i) \) in user session \( (S_i) \)

\[
(\text{RoP})_{ih} = \frac{2 \times (\text{DoP})_{ih} \times (\text{FoP})_{ih}}{(\text{DoP})_{ih} + (\text{FoP})_{ih}}
\]

Where \( 0 \leq (\text{RoP})_{ih} \leq 1. \)

6.4 Augmented Web user Sessions

Now, by applying equations (6) to (8) the page relevance matrix \( (\text{R}_{\text{M}}) \) 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.

\[
\text{AS}_u = \{\{P_i_1, (\text{RoP})_{i_1}, (\text{RoP})_{i_2}, \ldots, (\text{RoP})_{i_n}\} \ldots \}
\]

Where, \( P_i \), and \( (\text{RoP})_{ih} \) 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. 

\[ \text{ASS}_{(AS,AS)} = \frac{\sum_{i=1}^{n} AS_i(\text{RoP})_i \times AS_j(\text{RoP})_j}{\sqrt{\sum_{i=1}^{n} AS_i(\text{RoP})_i^2 \times \sum_{i=1}^{n} AS_i(\text{RoP})_j^2}} \]  

(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 i\textsuperscript{th} and j\textsuperscript{th} Page URL

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

\[ uSS\{(a\textsuperscript{a}, a\textsuperscript{b})\} = \text{Min}\left\{ \frac{\text{LoP}(P_{a\textsuperscript{a}})}{\text{Max}\{\text{LoP}(P_{a\textsuperscript{a}}), \text{LoP}(P_{a\textsuperscript{b}})\}} \right\} \]  

(11)

Where \( \text{LoP}(P_{a\textsuperscript{a}}) \) is length of URL (or number of edges) of path traversed from root node to respective node of \( P_i \) in user session \( \text{US}_i \). By applying this syntactic similarity of URLs, the similarity between two augmented web user sessions \( (\text{AS}_a^p, \text{AS}_b^p) \) is computed by Eq. (12).

\[ aSS\{(a\textsuperscript{a}, a\textsuperscript{b})\} = \frac{\sum_{i=1}^{n} AS_i(\text{RoP})_i \times AS_j(\text{RoP})_j \times uSS\{(a\textsuperscript{a}, a\textsuperscript{b})\}}{\sum_{i=1}^{n} AS_i(\text{RoP})_i \times \sum_{i=1}^{n} AS_j(\text{RoP})_j} \]  

(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. Intuitive augmented session similarity computed using Eq. (13).

\[ IASS\{(a\textsuperscript{a}, a\textsuperscript{b})\} = \text{Max}\left\{ \text{ASS}_{(AS,AS)}, aSS\{(a\textsuperscript{a}, a\textsuperscript{b})\} \right\} \]  

(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. The augmented session dissimilarity is computed using Eq. (14).

\[ D^2_{(AS_a, AS_b)} = (1 - \text{ASS}_{(AS_a, AS_b)})^2 \]  

(14)

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

Table 1. Summary of Results

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of initial sessions</td>
<td>2000</td>
</tr>
<tr>
<td>Number of valid sessions</td>
<td>1341</td>
</tr>
<tr>
<td>Size of FoP/DoP/RoP matrix</td>
<td>1341×589</td>
</tr>
<tr>
<td>Number of unique URLs in sessions</td>
<td>589</td>
</tr>
<tr>
<td>Size of URL syntactic similarity matrix</td>
<td>589×589</td>
</tr>
<tr>
<td>Size of IBSS/IASS/ ( D_{\text{min}} ) matrix</td>
<td>1341×1341</td>
</tr>
</tbody>
</table>

10. Clusters Quality Assessment

To assess the quality of produced fuzzy clusters, unsupervised evaluation method which is based on intra-cluster and inter-cluster distance measures is used. Intra-cluster distance represented compactness of a cluster and computed as an average of the distance between all pair of sessions within the \( i \)\textsuperscript{th} 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 \( i \)\textsuperscript{th} cluster and sessions from \( j \)\textsuperscript{th} cluster. The high value of inter-cluster distance represents good partition. 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).

\[ D_{\text{intra}} = \frac{\sum_{c_i \in C} \sum_{c_j \in C, c_j \neq c_i} d^2(AS_i, AS_j)}{|c_i|(|c_i| - 1)} \]  

(15)
Quantitative Evaluation of Web user Session Dissimilarity measures using Medoids based Relational Fuzzy clustering

\[ D_{int ra} = \sum_{s_{int}} \sum_{s_{ra,j}} \frac{d^2_{med} (AS_{h}, AS_{j})}{|c_i|, |c_j|} \]  

(16)

Cluster Quality Ratio (c) = \frac{Compactness}{Separation} \]  

(17)

Where, \( d^2_{med} (AS_{h}, AS_{j}) \) is the distance between two sessions in cluster \( c_i \) and \( |c_j| \) 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\(^3\). All similarity/dissimilarity measures used in this study and the relational fuzzy c-medoids algorithm are implemented in MATLAB (R2012a) package\(^3\). 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](image_url)  

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 measure. Default parameters were set during execution are fuzzifier coefficient \( f = 1.5 \) 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 augmented session 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 was evaluated by average 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 Inter-cluster distances for varying number of cluster and different dissimilarity measures

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

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