International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 7 (2017) pp. 1142-1150
© Research India Publications. http://www.ripublication.com

A Subtractive Relational Fuzzy C-Medoids Clustering Approach To Cluster Web User Sessions from Web Server Logs

Dilip Singh Sisodia*
Department of Computer Science & Engineering,
National Institute of Technology Raipur, India
**Corresponding Author*

Shrish Verma
Department of Electronics & Telecommunications,
National Institute of Technology Raipur, India

Om Prakash Vyas
Department of Computer Science & Engineering,
International Institute of Information Technology Naya Raipur, India

Abstract

In this paper, a subtractive relational fuzzy c-medoids clustering approach is discussed to identify web user session clusters from weblogs, based on their browsing behavior. In this approach, the internal arrangement of data along with the density of pairwise dissimilarity values is favored over arbitrary starting estimations of medoids as done in the conventional relational fuzzy c-medoids algorithm. It is assumed that straightforward binary session dissimilarity measure proposed and utilized as a part of prior detailed work is not especially logical, and instinctive to speak to session dissimilarities instigated from web client's habits, interest, and expectations. The idea of augmented sessions is utilized to infer page relevance based web client's session dissimilarity matrix. The discussed approach is applied on an augmented session dissimilarity matrix obtained from an openly accessible NASA web server log data. The produced clusters are assessed by utilizing diverse fuzzy validity measures, and results are contrasted with conventional fuzzy c-medoids clustering. Experimental results demonstrated that quality of fuzzy clusters produced by utilizing proposed subtractive relational fuzzy c-medoids clustering is superior as compared with the conventional relational fuzzy c-medoids approach.

Keywords: dissimilarity, fuzzy clustering, fuzzy validity measures, relational c-medoids clustering, subtractive clustering, similarity, user sessions, web server logs.

INTRODUCTION

Web-Based applications are growing with phenomenal rate, and this growth is instigating the zeal of analyzing the web usage data to gain extra knowledge about the users' navigation patterns. The knowledge concerning the web users accesses behavior may be utilized to serve the needs of the user in a better way. The Web clients access pattern is recorded in the form of weblogs and logs are divided into user sessions for analysis. Any knowledge extraction technique used for knowing the user's access behavior is applied on Web client sessions. Web client sessions may contain incorrect, clashing and obscure data and medoid based relational clustering approaches demonstrate extremely helpful in the grouping of web session [1]. Because of high dimensionality and scantiness of URLs in weblogs, web sessions are hard to speak by component vectors. The relational distance measure is favored over component vector to store the pairwise connection between web sessions [2]. Though, the strategies based on relational grouping are extremely delicate to the underlying decision of group determination. In this way, the underlying estimation of group medoids might be a bottleneck to the grouping performance of these approaches [3]. In this paper, a subtractive grouping approach is used to estimate the initial value of cluster medoids. The discussed approach is applied on an augmented session dissimilarity matrix obtained from an openly accessible NASA web server log data [4].

The remaining part of this paper is arranged under following sections: The related work on web user clustering is discussed in section 2. In Section 3, background details of augmented session similarity are recapitulated. Section 4 presents the formulation of the idea of a proposed subtractive clustering based fuzzy relational c-medoids(SC-RFCMdd) approach. In section 5, fuzzy cluster validity measures are discussed in brief. In section 6, experiments are performed on openly available web log data, and results are discussed. Finally, this study is concluded with future work in section 7.

RELATED WORK

In this section, some significant contributions on web user sessions clustering reviewed in brief. In [5] K-means algorithm is used to cluster the user sessions. A clustering of uneven length user session based on hybrid sequence alignment measure is discussed in [6]. In [7] utilizing BIRCH algorithm author proposed a generalization based grouping strategy to develop an URL order and session clusters. In [3] a programmed revelation of client session clusters in a fuzzy and unverifiable condition of weblogs is done using competitive agglomeration for the relational data algorithm and further extended in [8]. In [9] a cube model is used for web user sessions representation and K-modes algorithm for clustering. For finding basic clusters in the weblogs and determining classifications demonstrating the inclinations of comparable clients authors proposed a fuzzy similarity measure the same in a relational fuzzy grouping algorithm [10]. For clustering of values in a dissimilarity matrix D, a relational fuzzy c-means (RFCM) clustering is discussed in [11]. In [1] a new fuzzy c- medoid (FCMdd) algorithm proposed which starts with arbitrary estimations of medoids and iteratively settled on some final values. The FCMdd and RFCM [11] algorithms are applied on weblogs utilizing straightforward binary dissimilarity measure as suggested in [3],[8]. The concept of augmented session and intuitive dissimilarity measure is introduced in [12], and performance of proposed measures are evaluated using RFCM [13] and FCMdd [14].

In this paper, a subtractive clustering based fuzzy relational c-medoids (SC-RFCMdd) approach is discussed. In the discussed approach internal arrangement of data along with the density of pairwise dissimilarity values are favored over arbitrary starting estimations of medoids. It is assumed that straightforward binary session dissimilarity measure proposed and utilized as a part of prior detailed work is not especially logical, and instinctive to speak to session dissimilarities instigated from web client's habits, interest, and expectations[3].

BACKGROUND DETAILS OF AUGMENTED SESSION SIMILARITY

For a better understanding of proposed approach, the concept of the augmented session similarity [12][13] is recapitulated in following subsections.

Web server log pre-processing

The weblogs keep a record of metadata of all documents accessed by clients unequivocally or certainly. Metadata is recorded in some fields and vary from one log format to other. Essentially, all format recorded information about host address and name, username, date and time, the method used for the request; protocol used for access, the status of service, data size and user agents, etc.[15]. However, all recorded entries are not useful always and increase the size of weblogs. Therefore, cleaning is performed on raw logs such as removal of entries made by embedded objects, web robots [16], failed requests, and requests made other than GET method, etc. After cleaning weblogs are divided into independent user sessions using session identification techniques discussed in [17],[18].

User sessions representation in vector space model

It is assumed that; there are m users sessions are identified from the web server log $\mathcal{S}_i = \{ \mathcal{S}_1, \mathcal{S}_2, \ldots . \mathcal{S}_m \}$. These m users' sessions accessing n number of different pages URL's $\mathcal{P}_i = \{ \mathcal{P}_1, \mathcal{P}_2, \ldots . \mathcal{P}_n \}$ on a given website in some time interval. Then each user session $\mathcal{S}_i$ are denoted 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, every $\mathcal{S}_k^i$ correspond to a harmonic mean of the frequency the page $\mathcal{P}_k$ within the session $\mathcal{S}_i$, and the time spent on the page (in seconds) $\mathcal{P}_k$ in session $\mathcal{S}_i$, and represented by following matrices Eq.(1) & (2).

$$\mathcal{S}_k^i \leftarrow \begin{cases} \text{Frequency of the page} \\ \text{Duration of the page(in seconds)} \\ \text{Page size(in bytes)} \end{cases} \tag{1}$$

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

Page relevance based augmented session similarity

Some implicit measures [19],[20] are used to determined the relevance of any page for particular user [21]. 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 [12].

Duration of Page($\mathcal{D}\sigma\mathcal{P}$)

Duration of page or Page stay time defined as the time spent on a page by a web user, and it is the difference between the exact time of the request of page Pi and the time of the request for the next web page in the session from the access log file. Eq. (3) is used to measure the duration of a web page ($\mathcal{P}_i$) in user session ($\mathcal{S}_k$).

$$(\mathcal{D}\sigma\mathcal{P})_{\mathcal{P}_i} = \frac{\frac{\sum \text{Time Spent on}(\mathcal{P}_i)}{\text{Size of }(\mathcal{P}_i)}}{\text{Max}\left(\forall_{j \in \mathcal{S}_k} \frac{\sum \text{Time Spent on}(\mathcal{P}_j)}{\text{Size of }(\mathcal{P}_j)} \right)} \tag{3}$$

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

Frequency of Page $(\mathcal{F}o\mathcal{P})$

Frequency is the number of times the web page $\mathcal{P}_i$has been visited in the session. It seems natural to assume that web pages with a higher frequency are of more concern to users. Eq. (4) is used to measure the frequency of a web page $(\mathcal{P}_i)$ in user session $(\mathcal{S}_k)$.

$$(\mathcal{F}o\mathcal{P})_{\mathcal{P}_i} = \frac{\sum \#\text{ of visits to}(\mathcal{P}_i)}{\text{Max}\left(\forall_{j \in \mathcal{S}_k} \sum \#\text{ of visits to}(\mathcal{P}_j)\right)} \tag{4}$$

Where $0 \leq (\mathcal{F}o\mathcal{P})_{\mathcal{P}_i} \leq 1$

The relevance of page $(\mathcal{R}o\mathcal{P})$

The relevance of page in any session is measured by giving equal importance to the duration of page and frequency of Page. We use Eq. (5) to measure the relevance of web page $(\mathcal{P}_i)$ in user session $(\mathcal{S}_k)$.

$$(\mathcal{R}o\mathcal{P})_{\mathcal{P}_\text{i}} = \frac{2 \times (\mathcal{D}o\mathcal{P})_{\mathcal{P}_i} \times (\mathcal{F}o\mathcal{P})_{\mathcal{P}_i}}{(\mathcal{D}o\mathcal{P})_{\mathcal{P}_i} + (\mathcal{F}o\mathcal{P})_{\mathcal{P}_i}} \tag{5}$$

Where $0 \leq (\mathcal{R}o\mathcal{P})_{\mathcal{P}_\text{i}} \leq 1$

Augmented web user sessions

Now, by applying equations (3) to (5) 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. (6). By incorporating page relevance in web user session access behaviour matrix, simple web user sessions converted to augmented web user sessions.

$$\mathcal{RM}_{m \times n} = \begin{pmatrix} (\mathcal{R}o\mathcal{P})_{11} & (\mathcal{R}o\mathcal{P})_{12} & \cdots & (\mathcal{R}o\mathcal{P})_{1n} \\ (\mathcal{R}o\mathcal{P})_{21} & (\mathcal{R}o\mathcal{P})_{22} & \dots & (\mathcal{R}o\mathcal{P})_{12} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathcal{R}o\mathcal{P})_{m1} & (\mathcal{R}o\mathcal{P})_{m2} & \cdots & (\mathcal{R}o\mathcal{P})_{mn} \end{pmatrix} \tag{6}$$

Now, the web sessions are converted into augmented web user sessions

$\mathcal{AS}_i = \{\mathcal{AS}_1, \mathcal{AS}_2, \dots \mathcal{AS}_m\}\, for\ i = 1,2, \dots m$. Where, every augmented web user session is represented by $\mathcal{AS}_a = \{(\mathcal{P}1, (\mathcal{R}o\mathcal{P})_{\mathcal{P}1}, (\mathcal{P}2, (\mathcal{R}o\mathcal{P})_{\mathcal{P}2}) \dots (\mathcal{P}n, (\mathcal{R}o\mathcal{P})_{\mathcal{P}n})\}$. Where, $\mathcal{P}i$ and $(\mathcal{R}o\mathcal{P})_{\mathcal{P}i}$ are the visiting page, and its relevance respectively.

Page relevance based augmented session similarity

Here relevance of pages accessed in user sessions is incorporated in simple cosine similarity measure Eq. (5). This augmented session similarity measure may represent more realistic and represents session similarities based on the web user's habits, interest, and expectations as compared to simple binary cosine measure [3].

$$\mathcal{ASS}_{(\mathcal{AS}_a, \mathcal{AS}_b)} = \frac{\sum_{i=1}^{m} \mathcal{AS}_a(\mathcal{R}o\mathcal{P})_i \times \mathcal{AS}_b(\mathcal{R}o\mathcal{P})_j}{\sqrt{\sum_{i=1}^{m} \mathcal{AS}_a(\mathcal{R}o\mathcal{P})_i^{\,2}} \sqrt{\sum_{i=1}^{m} \mathcal{AS}_b(\mathcal{R}o\mathcal{P})_j^{\,2}}} \tag{7}$$

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.

An URL Syntactic Similarity between i^th^ and j^th^ Page URL

In [3] authors also defined an alternative URL based syntactic similarity measure to compute the syntactic similarity between any pair of URLs given by Eq. (8).

$$\mathcal{USS}_{\left(\mathcal{US}_a^{\mathcal{P}_i}, \mathcal{US}_b^{\mathcal{P}_j}\right)} = \mathcal{M}in\left(1, \frac{\left|\mathcal{L}o\mathcal{P}(\mathcal{P}_{(a,i)}) \cap \mathcal{L}o\mathcal{P}\left(\mathcal{P}_{(b,j)}\right)\right|}{\mathcal{M}ax\left(1, \mathcal{M}ax\left(\mathcal{L}o\mathcal{P}(\mathcal{P}_{(a,i)}), \mathcal{L}o\mathcal{P}\left(\mathcal{P}_{(b,j)}\right)\right) - 1\right)}\right) \tag{8}$$

Where $\mathcal{L}o\mathcal{P}\left(\mathcal{P}_{(a,i)}\right)$ 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 $\left(\mathcal{AS}_a^{\mathcal{P}_i}, \mathcal{AS}_b^{\mathcal{P}_j}\right)$ is computed by Eq. (9).

$$\mathcal{AUSS}_{\left(\mathcal{AS}_a^{\mathcal{P}_i}, \mathcal{AS}_b^{\mathcal{P}_j}\right)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{AS}_a(\mathcal{R}o\mathcal{P})_i \times \mathcal{AS}_b(\mathcal{R}o\mathcal{P})_j \times \mathcal{USS}_{\left(\mathcal{US}_a^{\mathcal{P}_i}, \mathcal{US}_b^{\mathcal{P}_j}\right)}}{\sum_{i=1}^{n} \mathcal{AS}_a(\mathcal{R}o\mathcal{P})_i \times \sum_{j=1}^{n} \mathcal{AS}_b(\mathcal{R}o\mathcal{P})_j} \tag{9}$$

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. (12).

$$\mathcal{IASS}_{(\mathcal{AS}_a, \mathcal{AS}_b)} = Max\left\{\mathcal{ASS}_{(\mathcal{AS}_a, \mathcal{AS}_b)}, \mathcal{AUSS}_{\left(\mathcal{AS}_a^{\mathcal{P}_i}, \mathcal{AS}_b^{\mathcal{P}_j}\right)}\right\} \tag{10}$$

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 [22]. The augmented session dissimilarity is computed using Eq. (6).

$$\mathcal{D}^2_{(\mathcal{AS}_a, \mathcal{AS}_b)} = (1 - \mathcal{ASS}_{(\mathcal{AS}_a, \mathcal{AS}_b)})^2 \tag{11}$$

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

DESCRIPTION OF SUBTRACTIVE RELATIONAL FUZZY C-MEDOIDS(SC-RFCMDD) CLUSTERING

In this section, we are presenting the formulation of the idea of a proposed subtractive clustering based fuzzy relational c-medoids(SC-RFCMdd) algorithm in details.

The potential density based selection of cluster medoid[23].

To locate a delegate session as cluster medoids, we look for the centers of dense areas in the given pairwise relational matrix. A subtractive clustering (SC) method [24], is applied for initial assessment of a number of cluster medoids, The SC is a modified version of mountain clustering [25]. The subtractive grouping strategy accepts that every data point is a potential cluster center and, without earlier learning of the default number of centers, it ascertains the probability of a data point being characterized as a cluster center as indicated by the density of the encompassing data points [26]. We compute the value of potential density function value at every augmented session by applying Eq. (12).

$$\mathcal{P}_1(\mathcal{AS}_i) = \sum_{j=1}^{n} exp\left(-\frac{d_{ij}^2(\mathcal{AS}_i,\mathcal{AS}_j)}{r_a^2}\right), \forall i = 1,2,\dots,m. \quad (12)$$

If d_{ij}^2 is small then $\mathcal{AS}_j$ and $\mathcal{AS}_i$ would be tightly associated and had significant impact on the PDF value $\mathcal{P}_1(\mathcal{AS}_i)$; or else $\mathcal{AS}_j$ and $\mathcal{AS}_i$ remain loosely associated and had less impact on $\mathcal{P}_1(\mathcal{AS}_i)$. The parameter r_a^2 is a radius and characrized the region of intrest for the selected augmented session$\mathcal{AS}_i$. Sessions outside this span have little impact on the chose session's PDF value.

In the wake of registering the PDF value at each session; we picked the session with largest PDF value as the main illustrative cluster center v_1 by applying Eq. (10). On the off chance that there are numerous sessions with same largest PDF value then we can pick intuitively any of them.

$$\mathcal{P}_1(\mathcal{AS}_{k_1}) = \max_{i=1}\{\mathcal{P}_1(\mathcal{AS}_i)\}; \qquad v_1 \leftarrow \mathcal{P}_1(\mathcal{AS}_{k_1}) \quad (13)$$

To locate the second illustrative cluster center, we use Eq. (8) compute the discounted PDF value over the neighborhood defined by r_b^2. If d_{ij}^2 is smaller between $\mathcal{AS}_i$ and v_1 then effective potential of each sessions around v_1 will be reduced due to this subtraction.

$$\mathcal{P}_2(\mathcal{AS}_i) = \mathcal{P}_1(\mathcal{AS}_i) - \mathcal{P}_1(v_1)exp\left(-\frac{d_{ij}^2(\mathcal{AS}_i,v_1)}{r_b^2}\right), \forall i = 1,2,\dots,m. \quad (14)$$

Subsequently, for all sessions, PDF values are discounted over effectual region of control of r_b^2 , and highest discounted PDF value are considered as the next envoy cluster centre v_2 by using Eq.(12)

$$\mathcal{P}_2(\mathcal{AS}_{k_2}) = \max_{i=1}\{\mathcal{P}_2(\mathcal{AS}_i)\}; \qquad v_2 \leftarrow \mathcal{P}_2(\mathcal{S}_{k_2}) \quad (15)$$

In the same way, any t^th^ illustrative cluster center is chosen, and the potential density function value of every session on the effectual region of control throughout t^th^ iteration is calculated by applying Eq. (13)

$$\mathcal{P}_j(\mathcal{AS}_i) = \mathcal{P}_{j-1}(\mathcal{AS}_i) - \mathcal{P}_{j-1}(v_{j-1})exp\left(-\frac{d_{ij}^2(\mathcal{AS}_i,v_{j-1})}{r^2}\right), \forall i = 2,\dots,m. \quad (16)$$

A similar method will proceed until the c number of maximum illustrative cluster center v_j is choosen by applying Eq. (14)

$$\mathcal{P}_j\left(\mathcal{AS}_{k_j}\right) = \max_{i=2}\{\mathcal{P}_j(\mathcal{AS}_i)\}, \forall i = 2,\dots,m. \;\; v_j \leftarrow \mathcal{P}_j\left(\mathcal{AS}_{k_j}\right) \quad (17)$$

Fuzzy C-medoids clustering

Let's assume n number of user sessions $\mathcal{S}_i = \{\mathcal{S}_1, \mathcal{S}_2, \dots . \mathcal{S}_n\}\ for\ i = 1,2, \dots n$. Where, each session is characterized by a vector of m-dimensions $\mathcal{S}_i = \{\mathcal{S}_i^1, \mathcal{S}_i^2, \dots \mathcal{S}_i^m\}, \forall i = 1,2,\dots,n$. Suppose $d(\mathcal{S}_i, \mathcal{S}_j)$ stand for the dissimilarity among session $\mathcal{S}_i$ and $\mathcal{S}_j$ and $\mathcal{V} \leftarrow \{v_1, v_2, \dots v_c\}$, $v_i \in \mathcal{D}$ denotes a subset of dissimilarity matrix $\mathcal{D}$ with cardinality c. Where $\mathcal{D}^c$ represents the set of all c-subsets $\mathcal{V}\ of\ \mathcal{D}$. The fuzzy c-medoids clustering is used for selecting c representative medoids from the data optimizing objective fuction and membership functions so that the total disimiliarity within each cluster is minimized using Eq. (18) and Eq. (19)

$$\mathcal{F}_{FCMdd} = \sum_{j=1}^{c}\left(\sum_{i=1}^{n} \mu_{ij}^{f}\, d(\mathcal{S}_i, v_j)\right) \quad (18)$$

$$\mu_{ij} = \frac{\left(d(\mathcal{S}_i,v_j)\right)^{-\frac{1}{(f-1)}}}{\sum_{j=1}^{c}\left(d(\mathcal{S}_i,v_j)\right)^{-\frac{1}{(f-1)}}} \quad (19)$$

Where, $d(\mathcal{S}_i, v_j)$ is the dissimilarity between session $\mathcal{S}_i$ and medoid of cluster $\mathcal{S}_j$ and $f \in [1,\infty]$ is fuzzification coefficient [13], [14].

FUZZY CLUSTER VALIDITY MEASURES

To decide the appropriate number of clusters in given dissimilarity matrix, different fuzzy cluster validity measures have been proposed [27],[23].Some fuzzy validity measures used only fuzzy membership values while others considered both membership values as well as input relational data. The later validity measures including Xie and Beni (XB) index, Fukuyama and Sugeno (FS) index and separation index are considered more robust [28] and explained in following subsections:

Xie and Beni index (XB)

The Xie and Beni (XB) index [29] defined as Eq. (20), where, the numerator indicates the compactness of the fuzzy partition while the denominator indicates the strength of the separation between clusters.

$$XB = \frac{\sum_{j=1}^{c}\left(\sum_{i=1}^{m} \mu_{ij}^{f}\, d(\mathcal{S}_i , v_j)\right)}{m * \delta_{min}^{1}} \tag{20}$$

Where, δ_{min}^{2} is the squar of minimum Euclidean distance between the cluster centres given by Eq. (21).

$$\delta_{min}^{1} = \min_{l.k=1 \ldots\ldots c \wedge l \neq k} d(v_i, v_j) \tag{21}$$

The Fukuyama and Sugeno (FS) index[30]

FS index combines the properties of compactness and separation measurements. The FS index is defined by Eq. (22):

$$FS = \sum_{i=1}^{m} \sum_{j=1}^{c} \mu_{ij}^{f}\, d\left(\mathcal{S}_i , v_j\right) - \sum_{i=1}^{m} \sum_{j=1}^{c} \mu_{ij}^{f}\, d\left(v_j, \bar{v}\right) \tag{22}$$

Where the first term represents the geometrical compactness of the clusters, the second term indicates the separation between the clusters and v represents the mean of the cluster centroids Eq. (23).

$$\bar{v} = \sum_{j=1}^{c} \frac{v_j}{c} \tag{23}$$

Separation Index (SI)

The separation index used a minimum distance separation for partition validity [31],[32] and defined by Eq. (24).

$$SI = \frac{\sum_{j=1}^{c}(\sum_{i=1}^{m} \mu_{ij}^{f}\, d(\mathcal{S}_i , v_j) / \sum_{i=1}^{m} \mu_{ij})}{m * \delta_{min}^{2}} \tag{24}$$

Where, δ_{min}^{2} is the squar of minimum Euclidean distance between the cluster centres and mean of the cluster centroids given by Eq. (25)

$$\delta_{min}^{2} = \min_{l.k=1 \ldots\ldots c \wedge l \neq k} d(v_i, \bar{v}) \tag{25}$$

EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the clustering capability of the proposed approach is presented by applying it on publicly accessible NASA web server log data[4]. This data set (NASA_access_log_Aug95) contains one month's worth of all HTTP requests from the NASA Kennedy Space Centre's web server in Florida. The log was collected from 00:00:00 August 1, 1995, through 23:59:59 August 31, 1995. The uncompressed content of the dataset is 167.8 MB and contains 1,569,898 records with timestamps having a 1-second resolution.

The discussed approach is scripted in MATLAB [33].The MATLAB programs are run 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). The different number of web user sessions is randomly selected from a pre-processed NASA access log data set for experimentation.

The log entries due to embedded web objects files(image, icons, sound files, etc.) are removed from source log files for cleaning. The cleaned log file is further processed for session identification using thirty minutes threshold time. In this paper, merely 2000 sessions are used from preprocessed log file just to optimized the processing overhead of the system. It is supposed that very small session do not contribute any significant information for user session clustering. Therefore, After removing of the default root / and mini sessions of size, one from the total considered sessions valid sessions are reduced to 1341, which access 589 unique URLS collectively. First, we compute the different matrices such as frequency of page (FoP), duration of the page (DoP), the relevance of the page (RoP), and URL's syntactic similarity(USS) from the given web user sessions. Second, we calculated different similarity/dissimilarity by applying the notion of augmented sessions(ASS, AUSS, and IASS).The summary of computed results is presented in Table 1.

We used a visual assessment tendency technique [34], to visualize the number of underlying clusters in dissimilarity matrix. In visual assessment tendency method, the reordered dissimilarities between pairs of user sessions are represented using digital intensity images. The black spots on the main diagonal of generated images represent the number of clusters, and the spot size approximates the size of the cluster. In this way, the approximate number of clusters exist in relational data are known before applying clustering algorithm [35].The visual assessment tendency images of augmented sessions dissimilarity measures(ASS, AUSS, and IASS) are shown in Fig. 1.

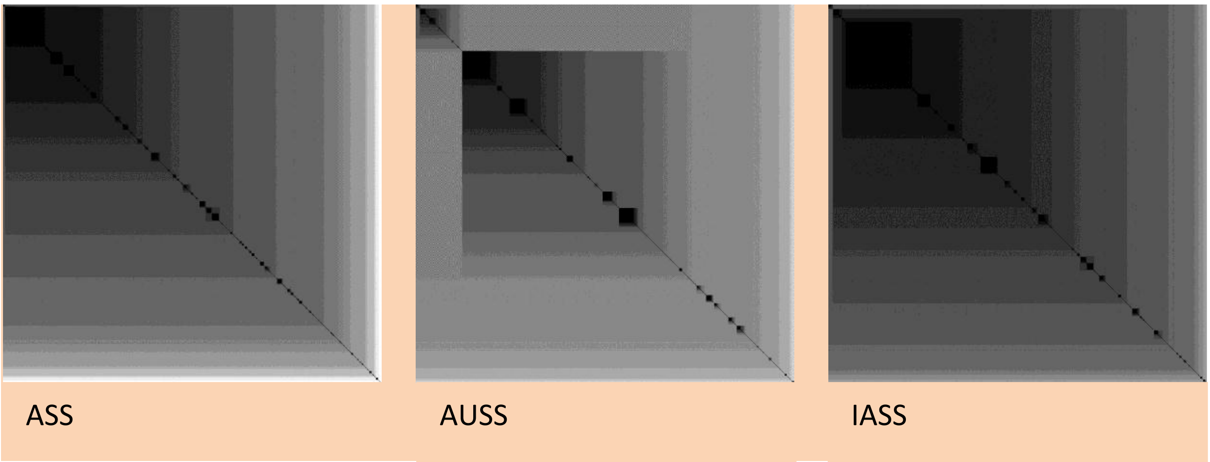

Figure 1: The VAT images for augmented sessions (ASS, AUSS and IASS) dissimilarity matrices of 1341×1341

Table 1: Summary of Results

Parameters	Values
# of preliminary sessions	2000
# of legitimate sessions	1341
Matrix size (FoP/DoP/RoP)	1341×589
# of unique page URLs	589
Size of USS	589×589
Size of IASS/ ($\mathcal{D}_{m\times m}$) matrix	1341×1341

To evaluate the performance of subtractive relational fuzzy c-medoids clustering algorithm, an intuitive augmented session dissimilarity matrix of size (1341×1341) is passed as input with default parameters. We perform multiple runs of RFCMdd and SC-RFCMdd algorithms with a changing number of clusters (from c=2 to 14), and fuzzifier coefficient ($\mathscr{f}$ = 1.5 to 2.5). Following default parameters are set during execution of RFCMdd, and SC-RFCMdd maximum number of iterations ($t_{max} = 100$), Neighbourhood radius (r_a^2=0.5 and r_b^2=1.5r_a^2).The value of fuzzy cluster validity measures (XB, FS and SI) are computed and results are shown in Table 2.

Table 2: RFCMdd Vs. SC-RFCMdd value of fuzzy cluster validity measures (XB, FS, and SI)

	XB-Index		FS-Index		S-Index	
Clusters	RFCMdd	SC-RFCMdd	RFCMdd	SC-RFCMdd	RFCMdd	SC-RFCMdd
	($\mathscr{f}$ =1.7)	($\mathscr{f}$ =1.5)	($\mathscr{f}$ =1.9)	($\mathscr{f}$ =1.9)	($\mathscr{f}$ =1.7)	($\mathscr{f}$ =1.7)
2	0.0387	0.0369	8918.78	9136.94	0.0000	0.0001
4	0.0738	0.0502	8423.70	6498.99	0.0001	0.0001
6	0.0339	0.0416	8020.36	6612.27	0.0001	0.0002
8	0.0934	0.0718	**6521.42**	**6099.93**	0.0001	0.0002
10	**0.0105**	**0.0134**	6569.92	7068.87	**0.0004**	**0.0003**
12	0.7448	0.0909	7366.01	7471.64	0.0002	0.0002
14	0.1404	0.5781	8226.26	7350.89	0.0002	0.0002

Figs. 2, 3 and 4 show the comparative values of fuzzy validity measures (XB, FS, and SI) for web user session clusters identified by RFCMdd and SC-RFCMdd with augmented session dissimilarity measures. It is evident from Figs. 2, 3 and 4 that quality of fuzzy clusters discovered using proposed SC-RFCMdd is better than that of those obtained with RFCMdd.

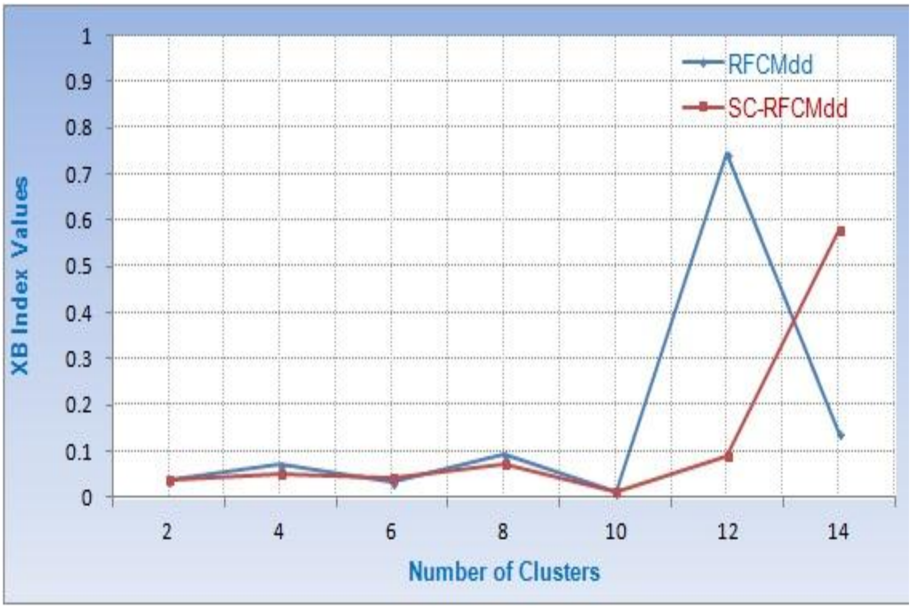

Figure 2 XB Index vs. No. of Clusters

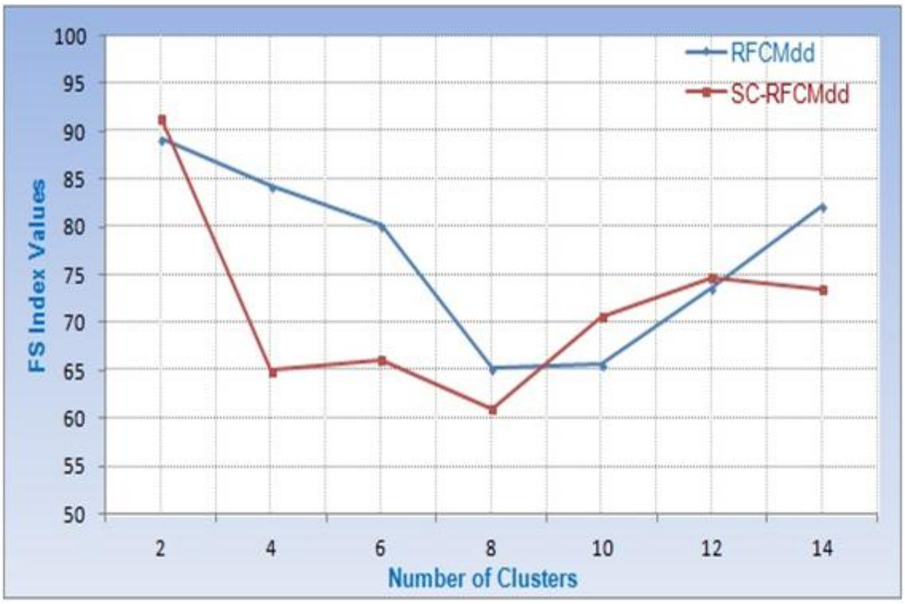

Figure 3: FS Index vs. No. of Clusters

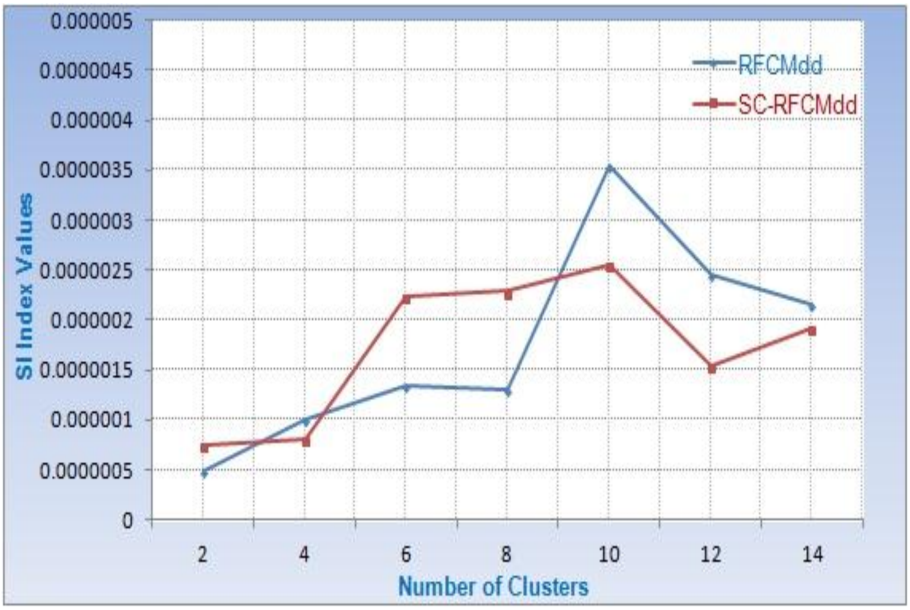

Figure 4: SI Index vs. No. of Clusters

CONCLUSION AND FUTURE WORK

A subtractive relational fuzzy c-medoids clustering/grouping approach is discussed to find the initial cluster medoids and consequently web user sessions clusters/groups. The connections between web client sessions are gotten from access importance of pages in any sessions. The relevance of a page is a measure of client's enthusiasm for any page URL and figured by applying harmonic mean of frequency and duration of pages in any session. The straightforward binary web client's sessions are changed into augmented sessions by consolidating relevance of page in getting to sessions. The augmented session comparability matrix is figured from page relevance matrix utilizing cosine closeness measure and changed over to dissimilarity matrix. A visual assessment tendency technique is utilized to imagine the conceivable clusters in client session dissimilarity matrix before applying any grouping algorithm. The picture produced visual assessment technique demonstrates the quantity of conceivable groups varying from eight to twelve. The trial comes about exhibited that nature of fuzzy clusters found by utilizing proposed subtractive relational fuzzy c-medoids grouping approach is superior to that of those acquired with relational fuzzy c-medoids clustering. In future, this approach can be more refined and presented in the algorithmic form. The effectiveness of this approach may be tested with some more data sets ranging from weblogs to other relational data sets.

REFERENCES

[1] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, “Low-complexity fuzzy relational clustering algorithms for Web Mining,” *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 4, pp. 595–607, 2001.

[2] R. Krishnapuram, A. Joshi, and Liyu Yi, “A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering,” in *IEEE International Fuzzy Systems. Conference Proceedings(FUZZ-IEEE’99)*, 1999, vol. 3, pp. 1281–1286.

[3] O. Nasraoui, F. Hichem, R. Krishnapuram, and A. Joshi, “Extracting web user profiles using relational competitive fuzzy clustering,” *International Journal on Artificial Intelligence Tools*, vol. 9, no. 4, pp. 509–526, 2000.

[4] NASA_SeverLog, “NASA Kennedy Space center’s www server log data, Available at,” 1995. [Online]. Available: http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.

[5] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, “From user access patterns to dynamic hypertext linking,” *Computer Networks and ISDN Systems*, vol. 28, no. 7, pp. 1007–1014, 1996.

[6] G. Poornalatha and S. R. Prakash, “Web sessions clustering using hybrid sequence alignment measure (HSAM),” *Social Network Analysis and Mining*, vol. 3, no. 2, pp. 257–268, 2013.

[7] Y. Fu, K. Sandhu, and M. Y. Shih, “A generalization-based approach to clustering of web usage sessions,” in *Web Usage Analysis and User Profiling*, 2000, pp. 21–38.

[8] T. K. Nasraoui, Olfa, Raghu Krishnapuram, Anupam Joshi, “Automatic web user profiling and

personalization using robust fuzzy relational clustering," in *In E-Commerce and Intelligent Methods*, 2002, pp. 233–261.

[9] Z. Huang, J. Ng, D. Cheung, M. Ng, and W. Ching, "A cube model for web access sessions and cluster analysis," *Proceedings of WEBKDD*, vol. 2001, no. 10203501, pp. 48–67, 2001.

[10] G. Castellano and M. A. Torsello, "Categorization of web users by fuzzy clustering," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5178, no. 2, pp. 222–229.

[11] J. C. B. R.J. Hathaway, J.W. Davenport, "Relational duals of the c-means clustering algorithms," *Pattern Recognition*, vol. 22, no. 2, pp. 205–212, 1989.

[12] 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.

[13] 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.

[14] 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.

[15] D. Sisodia and S. Verma, "Web Usage Pattern Analysis Through Web Logs: A Review," in *IEEE 9th International Joint Conference on Computer Science and Software Engineering(JCSSE 2012)*, 2012, pp. 49–53.

[16] 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.

[17] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis," *INFORMS Journal on Computing*, vol. 15, no. 2, pp. 171–190, 2003.

[18] 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.

[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, "Measuring similarity of interests for clustering Web-users," in *Proceedings 12th Australasian Database Conference. ADC 2001*, 2001, pp. 107–114.

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

[22] 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.

[23] 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.

[24] S. L. Chiu, "Fuzzy model identification based on cluster estimation.," *Journal of Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267–278, 1994.

[25] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 8, pp. 1279–1284, 1994.

[26] 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.

[27] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets and Systems*, vol. 158, no. 19, pp. 2095–2117, 2007.

[28] H. L. Shieh, "Robust validity index for a modified subtractive clustering algorithm," *Applied Soft Computing Journal*, vol. 22, pp. 47–59, 2014.

[29] Xuanli Lisa Xie and G. Beni, "A Validity Measure For Fuzzy Clustering," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[30] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in *Proceedings of the 5th Fuzzy Systems Symposium*, 1989, pp. 247–250.

[31] A. M. Bensaid, L. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh, "Validity Guided (re) clustering with applications to image segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 112–123, 1996.

[32] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering," *Pattern Recognition*, vol. 32, no. 7, pp. 1089–1097, 1999.

[33] MATLAB(2012a), "Software package." [Online]. Available: http://www.mathworks.com.

[34] T. C. Havens, S. Member, and J. C. Bezdek, "An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 813–822, 2012.

[35] L. A. Wang, X. Geng, J. Bezdek, C. Leckie, and K. Ramamohanarao, "iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning," in *Advances in Knowledge Discovery and Data Mining*, 2010, pp. 16–27.