IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2016, accepted August 30, 2016, date of publication October 7, 2016, date of current version November 8, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2611682

A Discounted Fuzzy Relational Clustering of Web Users' Using Intuitive Augmented Sessions Dissimilarity Metric

DILIP SINGH SISODIA[1], (Member IEEE), SHRISH VERMA[2], AND OM PRAKASH VYAS [3], (Member IEEE)

[1]Department of Computer Science and Engineering, National Institute of Technology Raipur, Raipur 492010, India
[2]Department of Electronics and Telecommunication Engineering, National Institute of Technology Raipur, Raipur 492010, India
[3]International Institute of Information Technology, Naya Raipur 492 002, India

Corresponding author: D. S. Sisodia (dssisodia.cs@nitrr.ac.in)

ABSTRACT Relational fuzzy clustering (RFC) algorithms prove very useful in Web user session clustering because Web user sessions may contain fuzzy, conflicting and imprecise information. Though RFC algorithms are very sensitive to cluster initialization and works only if the numbers of clusters are specified in advance. However, at all times, the prior initialization of a number of clusters is not feasible due to the dynamically evolving nature of user sessions. Therefore, estimating the number of clusters and initializing suitable cluster prototype are a significant performance bottleneck in this method. In this paper, the discounted fuzzy relational clustering (DFRC) algorithm is proposed to address the major constraint of RFC. The DFRC algorithm identifies Web user session clusters from Web server access logs, without initializing the number of clusters and prototypes of initial clusters. The DFRC algorithm works in two stages. In the first stage, DFRC automatically identifies the number of potential clusters based on the successively discounted potential density function value of each relational data and their respective centres. In the second stage, DFRC assigns fuzzy membership values to each data point and forms fuzzy clusters from the relational matrix. The DFRC algorithm is applied on an augmented session dissimilarity matrix obtained from a publicly accessed NASA Web server log data. The experimental results are evaluated using different fuzzy validity measures. The extensive experiments are performed to test the effect of various parameters, including accept/reject ratio and neighbourhood radius on the performance of DFRC algorithm. The results were also compared with fuzzy relational clustering algorithm using cluster quality measures. It is observed that the quality of generated clusters using DFRC is superior as compared with that of RFC.

INDEX TERMS Augmented sessions, cluster quality, fuzzy validity index, relational fuzzy clustering, similarity measures, subtractive clustering.

I. INTRODUCTION

Web portals are a means of interaction with clients for any business entity. They are highly significant not only in retaining the existing clients but also attracting potential customers in more effective and efficient way. Web portals can be used to enhance the clients' services as past browsing behaviours of clients automatically recorded in web server logs. These web server logs are analysed to extract hidden and potentially useful information. This process is known as the web usage mining [1].

Various web usages mining techniques are used on web server log data. Clustering is a very effective way of grouping web users or sessions, with common browsing activities, access pattern and navigational behaviours [2]. The primary objective of the web user sessions clustering is to group sessions based on similarity with the aim of maximising the intra-group and minimising the inter-group similarity. The user sessions with similar access patterns are clustered together and analysed by domain experts.

Clustering techniques are broadly classified into two major classes; One works with feature vectors data (object data clustering) and the other works with relational data (relational clustering). Though feature vector clustering is very popular and receives lots of attention from researchers, yet it is not

very much suitable for clustering of user sessions due to high dimensional and correlated feature space of web users' data [3].

Due to high dimensionality and sparseness in URLs accessing data, the generated user sessions are imprecise, inconsistent and indistinct. Relational fuzzy clustering algorithms are very useful in web user session clustering. But, RFC algorithms are very sensitive to cluster initialization and works only if the numbers of clusters are specified in advance. However, at all times the prior initialization of a number of clusters is not feasible due to dynamically evolving nature of user sessions. Therefore, estimating the number of clusters and initializing suitable cluster prototype is a significant performance bottleneck in this method. In this paper, the discounted fuzzy relational clustering (DFRC) algorithm is proposed to address the major constraint of relational fuzzy clustering. The DFRC algorithm identifies web user session clusters from web server access logs, without initializing the number of clusters and prototypes of initial clusters. The DFRC algorithm selects essential ideas from subtractive and relational fuzzy c-means clustering. For this, an augmented session dissimilarity based relational matrix is computed between all user sessions by calculating the various similarities/dissimilarity measures.

The rest of this paper is organised in following sections: Section 2 briefly reviews the existing relevant literature on user session clustering. In section 3, the methodology adopted for proposed approach is described in detail. Section 4 explains the need for underlying cluster formation through a hypothetical illustrative example. Section 5 describes the formulation of the idea of proposed discounted fuzzy relational clustering (DFRC) algorithm. Section 6 discusses different fuzzy validity measures. Section 7 discusses the measures for assessment of cluster quality. In section 8, experiments are set to demonstrate the performance of the DFRC clustering algorithm on NASA web server log data and results are discussed. Lastly, Section 9 concludes this study with a proposed future work.

II. RELATED WORK

In earlier reported research different clustering techniques have been extensively investigated to categorise web users/sessions based on their web access behaviours. In [4] Relational evidential c-means (RECM) was proposed to generates a credal partition. A credal partition is a new clustering structure based on belief functions and extends the existing concepts of hard, fuzzy and probabilistic partitions. In [5] a new credal c-means (CCM) clustering method was proposed to deal with the uncertain and imprecise data using credal partitions. In [6] credal classification method for incomplete pattern with adaptive imputation of missing values based on belief function theory. K-means algorithm was used to cluster the user sessions in [7]. In [8] authors proposed a Generalization-based clustering technique to construct a URL hierarchy and session clusters using BIRCH algorithm. A co-occurrence patterns of user transactions based method is used to compute overlapping groups of URL references [9]. Competitive Agglomeration for the relational data (CARD) algorithm is used for automatic discovery of user session groups in a fuzzy and uncertain environment of web log data in [10] and further extended in [11]. In [12] web user sessions were represented using cube model and clustered by K-modes algorithm. The self-organizing map-based visual analysis tool was used for clustering of web pages and to support the better understanding of characteristics and navigation behaviours of visiting pages [13]. In [14] a rough approximation based clustering is proposed to discover web page access patterns. In [15] authors proposed a fuzzy similarity measure and used the same in a relational fuzzy clustering algorithm to find underlying clusters in the Web usage data. The derived clusters model the preferences of similar users. In [16] authors proposed a new session clustering algorithm which takes advantage of ROCK algorithm to decide the initial points of each cluster and divides sessions into different groups. A new clustering approach based on logical path storing of web pages as similarity parameter and the conceptual relation between web pages is discussed in [17]. In [18] a matrix based fuzzy clustering approach is used to generate user clusters that can capture the web user's navigation behaviour depending on their interest. In [19] new similarity measure between two web pages and a fast optimal global sequence alignment algorithm were proposed to cluster the web user sessions in similar groups. In [20] and [21] Relational fuzzy c-means (RFCM) algorithm is used for gathering pairwise dissimilarity values in a dissimilarity matrix. Where RFCM is dual to the fuzzy c-means (FCM) [22], object data clustering algorithm with Euclidean distance matrix. The objective function of RFCM is based on computing representative clusters from the data so that the total dissimilarity between each group is minimised. However, RFCM works only when the numbers of potential clusters are specified in advance, that is not always feasible in user session clustering.

This paper proposes discounted fuzzy relational clustering (DFRC) algorithm for web user clustering. The DFRC uses key ideas from potential density based subtractive clustering (SC) [23] and relational fuzzy c-means (RFCM) [20], [21] algorithm.

III. PROPOSED METHODOLOGY

This section elaborates the methodology adopted in this paper. Table 1 briefly describes the notations used in this study.

A. WEB SERVER LOGS CLEANING

The record of all explicit and implicit requests made by users is stored in web server access logs. Where each log entry consists of different fields including remote host address, remote log name, 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. [24]. Weblog entries not germane to our purpose are removed using definition1. Mostly these are implicit requests

TABLE 1. Brief description of notations used in this study.

Notations used	Brief description
$\mathcal{AS}_i$	Set of augmented session
$\mathcal{AS}_i^n$	Vector representation of i^{th} augmented session in n-dimention
$\mathcal{A}_r$	Acceptance ratio
c	Number of clusters
d_{min}	Minimum distance between cluster prototype
$d_{\mathcal{R},ij}$	Relational Euclidean distance between j^{th} prototype and i^{th} session
$\mathcal{D}_{m\times m}$	a dissimilarity matrix
$\mathcal{F}_{RFCM}$	Objective function of RFCM
$\mathbb{f}_j$	Number of visits to j^{th} page
$\mathbb{f}$	Fuzzification coefficient
$\mathcal{L}$	Set of log records
$\mathcal{L}^c$	Set of cleaned log records
m	Number of user sessions
n	Number of URLs
$\mathcal{P}_j\left(\mathcal{AS}_{k_j}\right)$	j^{th} potential value of augmented session
$\mathcal{P}_i$	i^{th} web pages
r_i	i^{th} record of a log file
r_a^2	Neighbourhood radius
r_b^2	Neighbourhood radius
$\mathcal{R}_r$	Rejection ratio
$\mathcal{RM}_{m\times m}$	User session similarity matrix
$\mathcal{S}_i$	i^{th} user session
$\mathcal{S}_k^i$	Web user accessed the k^{th} *URL in* i^{th} *session*
$\mathcal{U}$	Fuzzy membership matrix
$v_{\mathcal{R},j}$	j^{th} cluster prototype
$\mathcal{V}_{\mathcal{R}}$	Set of cluster prototypes
μ_{ij}	Degree of membership between j^{th} cluster prototype and i^{th} session

made for embedded objects within web pages, requests made by automated software agents [25], unsuccessful requests of users, requests with access methods other than GET etc.

The web server log entries are grouped into user sessions, where session refers to the unit of interaction between a web user and a web server. The Web user sessions are identified using broadly accepted and practically implemented timeout based session identification method [26] as shown in Algorithm 1 which builds on the methods discussed in [27] and [28].

Definition 1 (Cleaning): Given a web server log file: $\mathcal{L}$ of n records where $\mathcal{L} \leftarrow \{r_1, r_2 \ldots r_n\}$,where $n \ggg 1$. Let $\mathcal{L}^c \leftarrow \{r_1, r_2 \ldots r_n\}$ and $\forall \ni r_i$ (r_i.url $\neq$(*.gif |*.jpeg |*.jpg |*.png |*.tif |*.bmp) and (r_i.method = "GET" |"POST") and (r_i.status) $\geq$ 200 && (r_i.status)$<$ 300) and (r_i.agent) $\neq$ (*. crawler.* |*.spider.* |*.bot.*) or (r_i.referrer) = "$-$")) then $\mathcal{L}^c \leftarrow \{r_1, r_2 \ldots r_n\}$ is a cleaned web log file.

B. VECTOR SPACE REPRESENTATION OF USER SESSIONS

Suppose, for a given website; there are m number of user sessions extracted from the web server logs $\mathcal{S}_i = \{\mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_m\}$, which Access n number of different URL's (pages) $\mathcal{P}_i = \{\mathcal{P}_1, \mathcal{P}_2, \ldots \mathcal{P}_n\}$ in a given website in a some specific time interval. The number of visits to the page or frequency of page and time spent on the page or duration of page these are the implicit measures and computed from weblog data to find interest of web users' for any page.

Algorithm 1 Pseudo Code for Web User Session Identification

Input: cleaned web log file: $\mathcal{L}^c$ *of* n *records where* $\mathcal{L}^c \leftarrow \{r_1, r_2 \ldots r_n\}$, *where* $n \ggg 1 \; \forall \ni r_i$ $<$ *ip, time, method, url, protocol, size, status, agent, referrer*$>$
τ_1:*User defined upper bound for page stay time*
τ_2:*User defined upper bound for session duration time*
Output: $\mathcal{S}_{set} \longleftarrow \{\mathcal{S}_1, \mathcal{S}_2 \ldots \mathcal{S}_n\}$

1: Initialization: $\mathcal{S}_0 \longleftarrow \{\varphi\}$; $\mathcal{S}_{set} \longleftarrow \{\varphi\}$
2: for i=1,2,...n, read (r_i) *from* $\mathcal{L}^c$
3: for $j = i + 1 mod\ n$.
4: if $((r_i.ip = r_j.ip) \wedge (r_i.agent = r_j.agent) \wedge (r_i.ref = r_j.ref) \wedge (r_j.time - r_i.time) \leq \tau_1)$;
5: then $\mathcal{S}_i \longleftarrow \{r_i \cup r_j\}$; Add record to the session;
6: else go to step 3; search new record for the session;
7: end of if
8: while $\left(\left(r_{last}.time - r_{first}.time\right) \leq \tau_2\right)$ do
9: $\mathcal{S}_{set} \longleftarrow \cup\{\mathcal{S}_i\}$; goto step 2; create new session
10: end of while
11: end of for
12: end of for
13: Update session set: $\mathcal{S}_{set} \longleftarrow \{\mathcal{S}_1, \mathcal{S}_2 \ldots \mathcal{S}_n\}$

The size of page is used to normalize the extreme values of these measures. Each user session($\mathcal{S}_i$) is represented by following equation $\mathcal{S}_i = \left\{\mathcal{S}_i^1, \mathcal{S}_i^2, \ldots \mathcal{S}_i^n\right\}$, $\forall i = 1, 2, \ldots, m$. where, each $\mathcal{S}_k^i$ represents a harmonic mean of the number of visits to the page $\mathcal{P}_k$ within the session $\mathcal{S}_i$, and the duration of the page (in seconds) $\mathcal{P}_k$ in session $\mathcal{S}_i$, and represented by Eq. (1) and Eq. (2)[26].

$$\mathcal{S}_k^i \leftarrow \begin{cases} \text{Number of visits to the page} \\ \text{Time spent on page(in seconds)} \\ \text{Size of the page (in bytes)} \end{cases} \tag{1}$$

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

C. PAGE RELEVANCE BASED INTUITIVE AUGMENTED SESSION SIMILARITY

The notion of intuitive augmented session similarity was discussed in [26] and evaluated in [29]. In this concept, two implicit measures, duration of the page (DoP) and frequency of page (FoP) [30], [31] are used to compute Web users' interest for any page in web user session [23]. The relevance of a page (RoP) in any session is measured by giving equal importance to the duration of the page (DoP) and frequency of Page (FoP). If the value of page relevance is high, it means the user has more concern for this page. Simple Web user sessions are converted to augmented Web user sessions by incorporating page relevance in the user sessions.

The relational dissimilarity matrices are computed between augmented sessions. This augmented session dissimilarity metric is more realistic and represented the session dissimilarities on the basis of web user's habits, interest, and expectations as compared to simple binary cosine measure. Intuitive augmented session similarity (IASS) utilises the properties of URL based syntactic similarity (USS) [3] and page relevance based augmented session similarity (ASS) measures to consider the maximum optimistic aggregation of these measures to give remarkable similarities between web user sessions as discussed in [26], [29], and [32].

Augmented Sessions	$\mathcal{AS}_1$	$\mathcal{AS}_2$	$\mathcal{AS}_3$	$\mathcal{AS}_4$	$\mathcal{AS}_5$	$\mathcal{AS}_6$	$\mathcal{AS}_7$	$\mathcal{AS}_8$	$\mathcal{AS}_9$	$\mathcal{AS}_{10}$
$\mathcal{AS}_1$	1.0	0.5	0.5	0.5	0.33	0.33	0.33	0.25	0.25	0.25
$\mathcal{AS}_2$	0.5	1.0	0.5	0.5	0.33	0.33	0.33	0.25	0.25	0.25
$\mathcal{AS}_3$	0.5	0.5	1.0	0.5	0.5	0.67	0.67	0.50	0.50	0.50
$\mathcal{AS}_4$	0.5	0.5	0.5	1.0	0.33	0.33	0.33	0.25	0.25	0.25
$\mathcal{AS}_5$	0.33	0.33	0.5	0.33	1.0	0.67	0.67	0.75	0.75	0.50
$\mathcal{AS}_6$	0.33	0.33	0.67	0.33	0.67	1.0	0.67	0.50	0.50	0.75
$\mathcal{AS}_7$	0.33	0.33	0.67	0.33	0.67	0.67	1.0	0.50	0.50	0.50
$\mathcal{AS}_8$	0.25	0.25	0.50	0.25	0.75	0.50	0.50	1.0	0.50	0.50
$\mathcal{AS}_9$	0.25	0.25	0.50	0.25	0.75	0.50	0.50	0.50	1.0	0.50
$\mathcal{AS}_{10}$	0.25	0.25	0.50	0.25	0.50	0.75	0.50	0.50	0.50	1.0

FIGURE 1. Example of page relevance based augmented session (dis)similarity matrix.

IV. A HYPOTHETICAL ILLUSTRATIVE EXAMPLE

The mere description of the abstract algorithm may be indistinct to understand the intuitive nature of extraction of clusters through the proposed algorithm. At first, it is explained with the help of an illustrative example. In this example, the expected output of DFRC is shown in Figure 1. This hypothetical relational table shows the mutual relationship between 10 augmented web user sessions by their accessing page relevance. The relationship between sessions $\mathcal{AS}_i$ *and* $\mathcal{AS}_j$ may be a similarity matrix ($\mathcal{RM}_{mm}$) or a dissimilarity matrix ($\mathcal{D}_{m\times m}$).

A careful observation of Figure 1 finds that there are only two homogeneous dense regions in the given relation. In these dense regions, sessions $\mathcal{AS}_3$ and $\mathcal{AS}_8$ are apparently located at the centre of a set of homogeneous sessions and surrounded by other sessions including $\mathcal{AS}_1$, $\mathcal{AS}_2$, $\mathcal{AS}_4$, $\mathcal{AS}_5$ and $\mathcal{AS}_6$, $\mathcal{AS}_7$, $\mathcal{AS}_9$, $\mathcal{AS}_{10}$ respectively. A prospective clustering algorithm is expected to select $\mathcal{AS}_3$ and $\mathcal{AS}_8$ sessions as representative cluster centres, and $\{\mathcal{AS}_1, \mathcal{AS}_2, \boldsymbol{\mathcal{AS}_3}, \mathcal{AS}_4, \mathcal{AS}_5\}$ and $\{\mathcal{AS}_6, \mathcal{AS}_7, \boldsymbol{\mathcal{AS}_8}, \mathcal{AS}_9, \mathcal{AS}_{10}\}$ as the cluster members.

V. DESCRIPTION OF PROPOSED CLUSTERING ALGORITHM

This section discusses the idea of discounted fuzzy relational clustering (DFRC) algorithm. Suppose a given set of augmented user sessions is $\mathcal{AS}_i = \{\mathcal{AS}_1, \mathcal{AS}_2, \ldots \mathcal{AS}_m\}$ *for* $i = 1, 2, \ldots m$. Where, each session is represented by vector of n-dimensions $\mathcal{S} = \left\{\mathcal{AS}_i^1, \mathcal{AS}_i^2, \ldots \mathcal{AS}_i^n\right\}, \forall i = 1, 2, \ldots, m.$

A. THE ESTIMATION OF NUMBER OF REPRESENTATIVE CLUSTER CENTRES

Clustering algorithms search for the centres of dense regions in the given relation to finding a typical session. A subtractive clustering method [23], [33] is used for estimation of the number of cluster centres. The subtractive clustering assumes each data point as a potential cluster centre; it calculates the possibility of each data point which could be a cluster centre according to the densities of the surrounding data points. The Eq. (3) is used for computation of potential density function (PDF) at every augmented session.

$$\mathcal{P}_1\left(\mathcal{AS}_i\right) = \sum_{j=1}^{n} exp\left(-\frac{d_{\mathcal{R},ij}\left(\mathcal{AS}_i, \mathcal{AS}_j\right)}{r_a^2}\right), \quad \forall i = 1, 2, \ldots, m. \tag{3}$$

Where, $d_{\mathcal{R},ij}$ is the distance between $\mathcal{AS}_j$ and $\mathcal{AS}_i$. If $d_{\mathcal{R},ij}$ has less value than $\mathcal{AS}_j$ and $\mathcal{AS}_i$ will be more related and will have major influence on the potential density value $\mathcal{P}_1\left(\mathcal{AS}_i\right)$; otherwise $\mathcal{AS}_j$ and $\mathcal{AS}_i$ will be less related and will have no significant influence on $\mathcal{P}_1\left(\mathcal{AS}_i\right)$. The parameter r_a^2 is a radius and it defines the neighbourhood region of the selected augmented session $\mathcal{AS}_i$. The sessions outside this radius have little influence on the potential density value of selected session.

After computing the PDF values at every session, the session highest PDF value is selected as the first representative cluster centre $v_{\mathcal{R},1}$ by using Eq. (4). If there exists multiple sessions with the same PDF value, then any one of them may be randomly chosen.

$$\mathcal{P}_1\left(\mathcal{AS}_{k_1}\right) = \max_{i=1}\left\{\mathcal{P}_1\left(\mathcal{AS}_i\right)\right\}; \quad v_{\mathcal{R},1} \leftarrow \left(\mathcal{AS}_{k_1}\right) \tag{4}$$

The Eq. (5) is used to find the second representative cluster centre and compute the discounted PDF values in the neighbourhood region defined by r_b^2. If $d_{\mathcal{R},ij}$ the Euclidean distance between $\mathcal{AS}_i$ and $v_{\mathcal{R},1}$ is small the effective potential of each sessions around $v_{\mathcal{R},1}$will be reduced due to this subtraction.

$$\mathcal{P}_2\left(\mathcal{AS}_i\right) = \mathcal{P}_1\left(\mathcal{AS}_i\right) - \mathcal{P}_1\left(v_{\mathcal{R},1}\right) \times exp\left(-\frac{d_{\mathcal{R}\vec{y}}\left(\mathcal{AS}_i, v_{\mathcal{R},1}\right)}{r_b^2}\right), \quad \forall i = 1, 2, \ldots, m. \tag{5}$$

After discounting PDF values for all sessions in the effective zone of the influence of r_b^2 , the highest discounted PDF value is selected as the second representative cluster centre $v_{\mathcal{R},2}$ by using Eq.(6)

$$\mathcal{P}_2\left(\mathcal{AS}_{k_2}\right) = \max_{i=1}\left\{\mathcal{P}_2\left(\mathcal{AS}_i\right)\right\}; \quad v_{\mathcal{R},2} \leftarrow \mathcal{P}_2\left(\mathcal{S}_{k_2}\right) \tag{6}$$

Similarly, to select any t^{th} representative cluster centre, the PDF value of each user session over an effective zone of influence during t^{th} iteration is computed

using Eq. (7).

$$\mathcal{P}_j\left(\mathcal{AS}_i\right) = \mathcal{P}_{j\text{-}1}\left(\mathcal{AS}_i\right) - \mathcal{P}_{j\text{-}1}\left(v_{\mathcal{R},(j\text{-}1)}\right) \times exp\left(-\frac{d_{\mathcal{R},ij}\left(\mathcal{AS}_i, v_{\mathcal{R},(j\text{-}1)}\right)}{r^2}\right), \quad \forall i=2,\ldots,m. \tag{7}$$

The same procedure will continue until the ratio of highest potential (during t^{th} iteration), and maximum potential (during 1^{st} iteration) is greater than to the acceptance ratio $\mathcal{A}_r$. The t^{th} representative cluster centre $v_{\mathcal{R},j}$ is selected by using Eq. (8)

$$\mathcal{P}_j\left(\mathcal{AS}_{k_j}\right) = \max_{i=2}\left\{\mathcal{P}_j\left(\mathcal{AS}_i\right)\right\}, \quad \forall i = 2,\ldots,m.\ v_{\mathcal{R},j} \leftarrow \mathcal{P}_j\left(\mathcal{AS}_{k_j}\right) \tag{8}$$

If the ratio is less than the reject ratio $\mathcal{R}_r$ then it will reject the session as representative cluster centre. If this value falls between $\mathcal{A}_r$ and $\mathcal{R}_r$ we check that session is how far from the existing representative cluster centre.

B. THE DISCOUNTED FUZZY RELATIONAL CLUSTERING

In this section, the idea of discounted fuzzy relational clustering (DFRC) using augmented session dissimilarity metric is discussed. In the present context, the essentials prerequisites are used from the relational fuzzy c-means clustering (RFCM) [21]. Let $d_{\mathcal{R},ji}$ is the relational distance between cluster prototype and augmented session $\mathcal{AS}_i$. Let $\mathcal{V}_{\mathcal{R}} \leftarrow \{v_{\mathcal{R},1}, v_{\mathcal{R},2},\ldots v_{\mathcal{R},c}\}$ represent a set of relational cluster centres in dissimilarity the matrix. The objective function of relational fuzzy c-means algorithm seeks to find number of representative sessions as relational cluster centres (known as centroid), so that the total distance of other sessions to their closest centroid is minimized. The objective function of relational fuzzy c-means (RFCM) [34] is defined as Eq. (9) and membership functions is given by (10) where, $f \in [1,\infty]$ is fuzzification coefficient.

$$\mathcal{F}_{\mathcal{R}FCM} = \sum_{j=1}^{c} \frac{\sum_{i=1}^{n}\sum_{k=1}^{n}\mu_{ij}^{f}\mu_{kj}^{f}d_{\mathcal{R},ji}}{2\sum_{i=1}^{n}\mu_{ij}^{f}} \tag{9}$$

$$\mu_{ij} = \frac{\left(d_{\mathcal{R},ji}\right)^{-\frac{1}{(f\text{-}1)}}}{\sum_{j=1}^{c}\left(d_{\mathcal{R},ji}\right)^{-\frac{1}{(f\text{-}1)}}} \tag{10}$$

The Euclidean distance $d_{\mathcal{R},ji}$ is the relational distance between cluster prototype and augmented session$\mathcal{AS}_i$. This distance is calculated on the basis of memberships in fuzzy membership matrix $\mathcal{U}$ and dissimilarities in dissimilarity matrix $\mathcal{D}$. The Euclidean distance is computed using Eq. (11) and the relational cluster centres are updated by using Eq. (12).

$$d_{\mathcal{R},ji} = \left(\mathcal{D}v_{\mathcal{R},j}^{t\text{-}1}\right)_i - \frac{1}{2}\left(v_{\mathcal{R},j}^{t\text{-}1}\right)^{\mathcal{T}}\mathcal{D}v_{\mathcal{R},j}^{t\text{-}1} \quad \text{for } 1 \le j \le c \text{ and } 1 \le i \le m \tag{11}$$

$$v_{\mathcal{R},j}^{t} = \frac{\left(\mu_{j1}{}^{f},\ \mu_{j}{}_{2}{}^{f},\ldots\ \mu_{jm}{}^{f}\right)}{\sum_{i}^{m}\mu_{ij}{}^{f}} \text{ for } 1 \le j \le c \tag{12}$$

Algorithm 2 Pseudo Code for Discounted Relational Fuzzy Clustering (DFRC) Algorithm

Input: 1. $\{\mathcal{D}_{m\times m}|$Dissimilarity matrix$\}$
2. Neighborhood parameters: $r_b > r_a > 0$;
3. accept ratio: $\mathcal{A}_r$; reject ratio: $\mathcal{R}_r$;

Output: $\{\ \mathcal{V}_{\mathcal{R}} \leftarrow \{v_{\mathcal{R},1}, v_{\mathcal{R},2},\ldots v_{\mathcal{R},c}\}$ |set of relational cluster centres, $\mathcal{U}$|Fuzzy membership matrix $\}$

t ← 1;
for $i = 1, 2, \ldots, m$
calculate potential density function (PDF) value of each session using Eq.(3)
end for
select the session with Maxpdf $\mathcal{P}_1\left(\mathcal{S}_{k_1}\right) = max_{i=1}\{\mathcal{P}_1\left(\mathcal{S}_i\right)\}$
set it as first cluster centre $v_{\mathcal{R},1} \leftarrow \mathcal{P}_1\left(\mathcal{S}_{k_1}\right)$
compute the discounted PDF of each session using Eq.(5)
if $\frac{\mathcal{P}_j\left(\mathcal{S}_{k_j}\right)}{\mathcal{P}_1\left(\mathcal{S}_{k_1}\right)} > \mathcal{A}_r$ then add $\mathcal{S}_{k_j}$ as the new cluster centre;
t ← t + 1 & set $v_{\mathcal{R},j} \leftarrow \mathcal{P}_j\left(\mathcal{S}_{k_j}\right)$ go to step 5
else if $\frac{\mathcal{P}_j\left(\mathcal{S}_{k_j}\right)}{\mathcal{P}_1\left(\mathcal{S}_{k_1}\right)} < \mathcal{R}_r$ then discard $\mathcal{S}_{k_j}$ and terminate
else let $d_{min} = min_{j=1}^{c-1} d_{jc}^2\left(v_{\mathcal{R},j}, v_{\mathcal{R},c}\right)$
if $\frac{d_{min}}{r_a} + \frac{\mathcal{P}_j\left(\mathcal{S}_{k_j}\right)}{\mathcal{P}_1\left(\mathcal{S}_{k_1}\right)} > 1$ then add $\mathcal{S}_{k_j}$ as the new cluster centre
t ←t + 1 & set $v_{\mathcal{R},j} \leftarrow \mathcal{P}_j\left(\mathcal{S}_{k_j}\right)$, go to step 3
else discard $\mathcal{S}_{k_j}$ and 0 ← $\mathcal{P}_j\left(\mathcal{S}_i\right)$ and select $\mathcal{P}_{next}\left(\mathcal{S}_i\right)$, go to step 3
end if
end if
Estimated cluster centres: $\mathcal{V}_{\mathcal{R}} \leftarrow \{v_{\mathcal{R},1}, v_{\mathcal{R},2},\ldots v_{\mathcal{R},c}\}$
Calculate Euclidean distance ($d_{\mathcal{R},ji}$) between augmented sessions $\mathcal{AS}_i$ and centroid of clusters using Eq.(11)
For $i \leftarrow 1,2,\ldots m$ do
If $d_{\mathcal{R},ji} \neq 0 \quad \forall j$
calculate membership function matrix ($\mathcal{U}$) using Eq.(10)
else
Set $\mu_{ij} > 0\ for\ d_{\mathcal{R},ji} = 0,\ \ \mu_{ij} \in [0,1]$ and $\sum_{j=1}^{c}\mu_{ij} = 1$
End If
End For
Apply defuzzification by assigning sessions to nearest neighbour clusters

The pseudo code for the above-described procedure is given as Algorithm 2 to summarise the concept of discounted fuzzy relational clustering (DFRC) algorithm. The working flow of DFRC algorithm is shown as a block diagram in Figure 2.

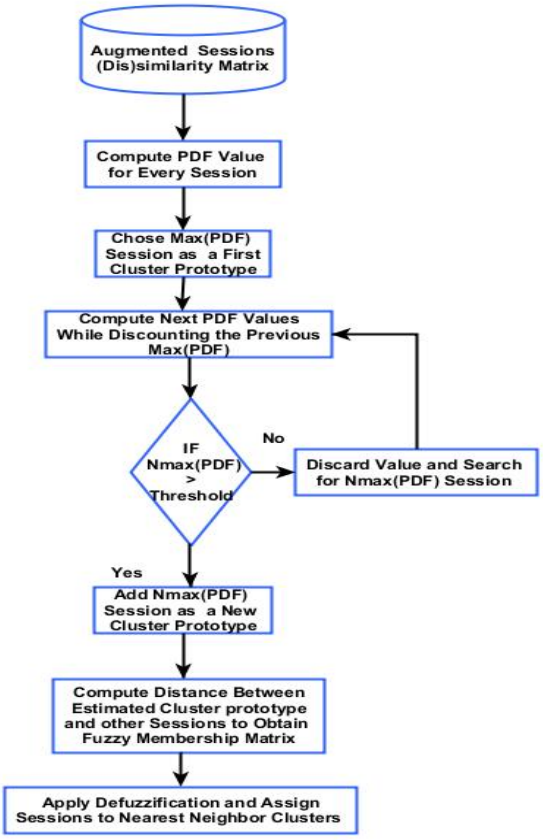

FIGURE 2. Working of DFRC algorithm using Augmented Session Dissimilarity Metric.

VI. FUZZY CLUSTER VALIDITY MEASURES

Different fuzzy cluster validity measures [35] are used to decide the appropriate number of clusters in given dissimilarity matrix based on the compactness and separation. Compactness measures the solidity of data in each cluster. Separation is used to characterise coupling between clusters. For a data clustering algorithm, the high value of compactness indicates a good partition. In contrast, low coupling suggests a weak relationship between clusters that indicates good separation. The validity index for measuring the goodness of partitions may be derived by using both compactness and separation [36] as shown in Eq.(13) and (14) where $2 \leq c \leq c_{\max}$ is the number of clusters

$$\text{Validity index } (c) = \frac{\text{Compactness}}{\text{Separation}} \tag{13}$$

$$\text{Validity index } (c) = (\text{Compactness } - \text{Separation}) \tag{14}$$

The most commonly used fuzzy clustering validity measures based on fuzzy membership matrix as well as relational data are described as follows:

A. Xie AND Beni INDEX (XB)

The Xie and Beni (XB) index [37] (Eq. 15), where, the numerator indicates the compactness of the fuzzy partition while the denominator indicates the strength of the separation between the clusters.

$$\mathcal{XB} = \frac{\sum_{j=1}^{c}\left(\sum_{i=1}^{m}\mu_{ij}^{f} d\left(\mathcal{AS}_i, v_j\right)\right)}{m^{*}\,\delta_{min}^{1}} \tag{15}$$

Where, δ_{min}^{2} (Eq. 16) is the square of minimum Euclidean distance between the cluster centres.

$$\delta_{min}^{1} = \min_{\ell.k=1\ldots\ldots c \bigwedge \ell \neq k} d\left(v_i, v_j\right) \tag{16}$$

The low value of XB suggests compact and well-separated clusters. So, the best fuzzy partition can be obtained by minimising XB on $c = 2,3,\ldots\ c_{\max}$.

B. THE Fukuyama AND Sugeno (FS) INDEX

Fukuyama and Sugeno (FS) index [38] (Eq.17) combines the properties of compactness and separation measures.

$$\mathcal{FS} = \sum_{i=1}^{m}\sum_{j=1}^{c}\mu_{ij}^{f} d\left(\mathcal{AS}_i, v_j\right) - \sum_{i=1}^{m}\sum_{j=1}^{c}\mu_{ij}^{f} d\left(v_j, \bar{v}\right) \tag{17}$$

Where the first term represents the geometrical compactness of the clusters, the second term indicates the separation between the clusters and $\bar{v}$ represents the mean of the cluster centroids and it is defined by Eq. (18).

$$\bar{v} = \sum_{j=1}^{c}\frac{v_j}{c} \tag{18}$$

The low value of FS indicates the fuzzy partition with well-separated and compact clusters..

C. SEPARATION INDEX (SI)

The separation index (Eq.19) uses a minimum distance for separation of fuzzy partition [39], [40].

$$\mathcal{SI} = \frac{\sum_{j=1}^{c}\left(\sum_{i=1}^{m}\mu_{ij}^{f} d\left(\mathcal{AS}_i, v_j\right)\right)/\sum_{i=1}^{m}\mu_{ij}}{m^{*}\delta_{min}^{2}} \tag{19}$$

Where, δ_{min}^{2} (Eq. 20) is the square of minimum Euclidean distance between the cluster centres and mean of the cluster centroids.

$$\delta_{min}^{2} = \min_{\ell.k=1\ldots\ldots c \bigwedge \ell \neq k} d(v_i, \bar{v}) \tag{20}$$

The high value of SI indicates the well-separated and compact fuzzy partitions.

VII. GENERATED CLUSTER QUALITY MEASURES

An unsupervised evaluation method based on intra-cluster and inter-cluster distance measures is used [41], [42] to assess the quality of formed fuzzy clusters. Intra-cluster distance represents compactness of a cluster and is computed as an average of the distance between all pair of sessions within the i^{th} cluster.For good quality of clusters, the low value of the intra-cluster distance is expected. Inter-cluster distance is a measure of separation between clusters and is computed as an average of the distance between sessions from i^{th} cluster

to sessions from j^{th} cluster. The high value of inter-cluster distance represents good partition [43]. The intra-cluster and inter-cluster distance is computed using Eq. (21) and Eq. (22) respectively. The cluster quality ratio for measuring the goodness of partitions can be designed to consider the both average intra cluster and average inter cluster distance using Eq. (23)

$$\mathcal{D}_{intra} = \frac{\sum_{\mathcal{S}_{k\in c_i}} \sum_{\mathcal{S}_{\ell\in c_{i,\ell\neq k}}} d^2_{k\ell}\left(\mathcal{AS}_k, \mathcal{AS}_\ell\right)}{|c_i|\ (|c_i|\text{-}1)} \tag{21}$$

$$\mathcal{D}_{inter} = \frac{\sum_{\mathcal{S}_{k\in c_i}} \sum_{\mathcal{S}_{\ell\in c_{j,\ell\neq k,}}} d^2_{k\ell}\left(\mathcal{AS}_k, \mathcal{AS}_\ell\right)}{|c_i|\ |c_j|} \tag{22}$$

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

$$\text{Cluster Quality Ratio (c)} = \frac{\text{Avg. Intra Cluster Distance}}{\text{Avg. Inter Cluster Distance}} \tag{23}$$

VIII. EXPERIMENTAL RESULTS AND DISCUSSIONS

The extensive experiments were performed to evaluate the clustering performance of DFRC algorithm by applying it on augmented session dissimilarity metric derived from publicly accessible NASA web server log data [26]. 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 numbers of records with timestamp having a one-second resolution. The DFRC, other algorithms and fuzzy validity measures are implemented using MATLAB (R2012a) package [44]. The experiments are 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).

A. PRE-PROCESSING OF WEBLOG DATA

The irrelevant entries including image, icons, and sound files, etc. were removed from the original log file. Then the sessions were identified by setting 30 minutes time threshold, as it was widely used to identify the user sessions in most of the weblog dataset. This study considered only 1000 number of the sessions from pre-processed log data to reduce the system processing overhead. The default root (/) and mini sessions of size one are filtered out from the total generated sessions as they did not contribute any significant information for efficient user session clustering. Total useful, valid sessions were reduced to 665, 1341, and 2048 respectively which access 419,589, and 731 unique URLs collectively. First, the different matrices (FoP, DoP, RoP, and USS) were computed from the web user sessions. Consequently, different augmented session's dissimilarity matrices were calculated by using the notion of augmented sessions as discussed in [27]. The summary of calculated results is shown in Table 2.

TABLE 2. Summary of derived matrices.

Parameters	DS-1	DS-2	DS-3
Number of initial sessions	1000	2000	3000
Number of valid sessions	665	1341	2048
Size of FoP/DoP/RoP matrix	665×419	1341×589	2048×731
Number of unique URLs in sessions	419	589	731
Size of URL syntactic similarity matrix	419×419	589×589	731×731
Size of ASS/AUSS/IASS/ ($\mathcal{D}_{m\times m}$) matrices	665×665	1341×1341	2048×2048

B. VISUAL ASSESSMENT OF POTENTIAL CLUSTERS HIDDEN IN THE DISSIMILARITY MATRICES

A visual assessment tendency (VAT) tool is used [45], [46] to visualise the number of potential dormant clusters in the relational dissimilarity matrix. In VAT plot the reordered dissimilarities between pairs of sessions are represented using digital intensity images, where darker pixels indicated smaller dissimilarities between sessions. These dark blocks appear only when a compact group exists in the relational data, and the size of each block represents the approximate size of the cluster [47], [48]. The VAT plot suggests the approximate number of clusters present in the data before applying any clustering algorithm.

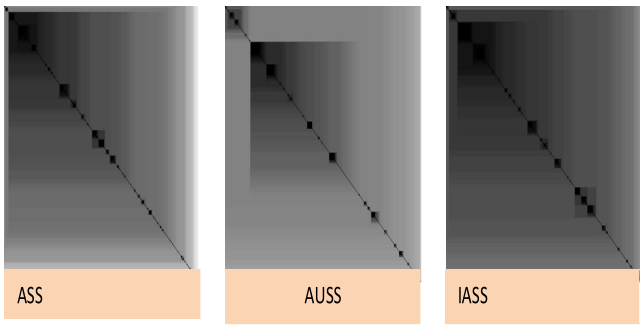

FIGURE 3. VAT images for various size of augmented sessionąęs dissimilarity matrices of size 665 × 665.

In this experiment 6 to 14, significant clusters are indicated by the VAT tool for different size of (such as 665×665, 1341×1341, and 2048 × 2048) augmented session dissimilarity matrices (ASS, AUSS and IASS) of the given data sets. The Figure 3 shows VAT images for various augmented session's

TABLE 3. Empirically identified number of clusters using RFCM with fuzzy cluster validity measures.

Number of Cluster	XB Index Values	FS Index Values	S Index Values
2	0.13759	11.24396	0.019732
4	0.14186	6.542782	0.049348
6	0.1584	5.67719	0.065477
8	**0.08497**	**3.668357**	**0.182718**
10	13.3192	6.085762	0.182222
12	1070.85	6.199901	0.101721
14	46396	5.011466	0.100454

dissimilarity matrices of size 665 × 665. Similarly, VAT images are produced for other matrices of size 1341 × 1341, and 2048 × 2048.

C. USER SESSION CLUSTERING WITH RFCM

An intuitive augmented session dissimilarity matrix of size (665 × 665) was given as input in relational fuzzy c-means clustering (RFCM) algorithm. Multiple runs of RFCM were performed with a varied number of clusters (C=2 to 14 with an interval of 2), and fuzzifier coefficient (f=1.5 to 2.5 with an interval of 0.2). The default parameters including maximum number of iterations (t_{max}=100), Step size (ϵ=0.0001) were set during execution of RFCM.

The empirical values of different fuzzy cluster validity measures were obtained for a number of clusters generated by RFCM as shown in Table 3. In each iteration, the RFCM algorithm the various values for three validity measures XB, FS, and SI were produced. The primary objective of this experiment was to obtain the number of clusters with a minimum value of XB and FS and the maximum value of SI index. So, from the table, the highlighted values of validity indices empirically shows the eight numbers of clusters in the augmented dissimilarity matrix for the fuzzifier coefficient values ranging from 1.5 to 1.9. Similarly, for other matrices of size 1341×1341 and 2048×2048, respectively ten and twelve numbers of clusters are decided empirically using RFCM algorithm.

D. USER SESSION CLUSTERING WITH DFRC

The same intuitive augmented session dissimilarity ($\mathcal{D}_{m\times m}$) matrix as used in RFCM along with the default parameters, are passed as input arguments in the DFRC algorithm. The summary of used default parameters and there values are shown in Table 4. For given augmented session dissimilarity matrix of size 665 × 665, DFRC algorithm generates eight numbers of clusters. Eight numbers of clusters are suggested by VAT tool as well as they were empirically identified by RFCM using fuzzy validity measures.

TABLE 4. Summary of used default parameter values.

Parameters	Symbols	Choose Values	Range
Neighbourhood radius	r_a^2	0.3	$0.2 \le r_a^2 \le 0.5$
Neighbourhood radius	r_b^2	1.5	$r_b^2 \ge r_a^2$
Accept Ratio	$\mathcal{A}_r$	0.7	$0.3 \le \mathcal{A}_r \le 0.9$
Reject Ratio	$\mathcal{R}_r$	0.15	$0.05 \le \mathcal{R}_r \le 0.4$
Fuzzyfier	$\oint$	1.5	$\oint \in [1,\infty]$

TABLE 5. Clustering results of DFRC for dissimilarity matrix of size ($\mathcal{D}_{m\times m}$ = 665 × 665).

Cluster Number	Prototype Session of the cluster	Cardinality of Cluster	Potential Density Value	Number of Unique URLs
1	256	320	243.43	322
2	156	85	234.94	117
3	143	42	224.02	79
4	395	7	198.67	35
5	171	46	197.91	72
6	200	4	192.97	66
7	609	159	179.47	98
8	49	2	164.01	25

The DFRC algorithm produces output as potential density value for an index of prototype session of the cluster in descending order. The detailed output generated by DFRC algorithm is shown in Table 5. The table consists of some clusters, prototype session for each cluster, the number of sessions in each cluster, the total number of unique URLs in each cluster and the value of potential density function for each cluster in descending order. Similarly, for other matrices of size 1341 × 1341 and 2048 × 2048, respectively ten and twelve numbers of clusters are produced by DFRC algorithm.

a: EFFECT OF ACCEPT RATIO AND REJECT RATIO ON DFRC

In DFRC, accept ratio ($\mathcal{A}_r$) and reject ratio ($\mathcal{R}_r$) are used to control the degree of generated clusters. Petite values of $\mathcal{A}_r$and $\mathcal{R}_r$ lead to produce large number of clusters including some cluster with insignificant PDF values. Huge values of $\mathcal{A}_r$and $\mathcal{R}_r$ may result in very less number of clusters excluding some significant clusters.

Experiments were conducted to evaluate the effect of accept ratio ($\mathcal{A}_r$) and reject ratio ($\mathcal{R}_r$) on performance of DFRC algorithm. Different discrete values of $\mathcal{A}_r$ (0.4 to 0.9),

TABLE 6. Performance of DFRC with different values of accept ratio and reject ratio.

Accept Ratio	Reject Ratio	Number of Clusters	XB Index Values	FS Index Values	S Index Values
0.4	0.4	11	0.33537	8.26582	0.045927
0.4	0.35	11	0.33537	8.265827	0.045927
0.5	0.3	10	0.34084	8.265827	0.045927
0.6	0.25	9	0.35521	8.023081	0.051197
0.7	**0.2**	**8**	**0.28197**	**7.839719**	**0.079761**
0.8	**0.15**	**8**	**0.28197**	**7.886633**	**0.074428**
0.9	0.1	4	0.3154	9.056722	0.028099

$\mathcal{R}_r$ (0.1 to 0.4) and their combinations were used to generate number of clusters. Multiple runs of the DFRC algorithm on dissimilarity matrix of size 665 × 665 is performed to get the best combination of $\mathcal{A}_r$and $\mathcal{R}_r$ for optimal number of clusters. The value of fuzzy cluster validity indices (XB, FS and SI) are recorded for each generated cluster. It is observed from results that minimum value of XB, FS and maximum value of SI is repeated for cluster 8, with accept ratio 0.7,0.8 and reject ratio is 0.15, 0.2 as shown in Table 6.

b: EFFECT OF ACCEPT RATIO AND REJECT RATIO ON DFRC

The output of DFRC algorithm is highly dependent on selection of neighbourhood radius parameters r_a^2 and r_b^2 values. However the DFRC algorithm discounted the influence zone of already identified cluster prototypes and seeks to find all sessions with significant PDF values. Therefore, the value of $r_b^2 \geq r_a^2$ and selected as shown in Table 4.

An evaluation of the influence of the neighbourhood radius (Radii) on the performance of the DFRC algorithm is done with fuzzy validity measures. The neighbourhood radius has a value between 0 and 1 and specifies the size of the cluster in each of the data dimensions. Table 7 shows the summary of experimental results performed with different values of neighbourhood radius on dissimilarity matrix of size 665 × 665. It was observed that the values of radii 0.4 and 0.5 might be the best choice for the optimum cluster selection.

E. COMPARATIVE PERFORMANCE OF FUZZY RELATIONAL CLUSTERING ALGORITHMS

A relative performance of aforementioned fuzzy relational clustering algorithms is discussed in this section. Extensive experiments were performed to evaluate the quality of clusters generated by RFCM and DFRC algorithms using augmented session dissimilarity matrix of size ($\mathcal{D}_{m\times m}$ = 665 × 665). The average intra-cluster, inter-cluster distance and the cluster quality ratio is used to measure the quality of generated clusters. The values default parameters as shown

TABLE 7. Performance of DFRC with varying neighbourhood radius (radii).

Neighbourhood Radius (radii)	Number of Cluster	XB Index Values	FS Index Values	S Index Values
0.1	7	0.477581	10.56814	0.036
0.2	5	0.355785	11.15827	0.027
0.3	5	0.137875	11.16697	0.039
0.4	8	0.203199	**10.8574**	**0.085**
0.5	8	**0.097493**	10.75966	0.084
0.6	8	0.103376	10.9634	0.064
0.7	10	0.120251	11.00783	0.043
0.8	9	0.115967	11.06791	0.05
0.9	10	0.122373	11.36719	0.027

TABLE 8. Summary of used parameter values.

Parameters	Symbols	Choose Values	Remarks
Step size	ε	0.0001	Only for RFCM
Maximum number of iterations	t_{max}	1000	Only for RFCM
Fuzzifier	$\oint$	1.5 to 2	For both
Neighbourhood radius-1	r_a^2	0.4	Only for DFRC
Neighbourhood radius-2	r_b^2	1.5 r_a^2	Only for DFRC
Accept ratio	$\mathcal{A}_r$	0.7	Only for DFRC
Reject ratio	$\mathcal{R}_r$	0.2	Only for DFRC

TABLE 9. Summary of comparative performance of RFCM and DFRC.

Measures / Algorithms	Avg. Intra cluster distance	Avg. Inter cluster distance	Cluster quality ratio
RFCM	0.177563588	0.779901033	0.227675
Proposed DFRC	0.179212938	0.931432394	**0.192406**

in Table 8 were used during execution of fuzzy relational clustering algorithms.

The summary of results obtained from experiments is given in Table 9. The table shows the average Intra-cluster and Inter-cluster distances of both RFCM and DFRC algorithm and their average cluster quality ratio. It is evident from

the results that DFRC algorithm improved performance over RFCM regarding cluster quality ratio.

IX. CONCLUSION AND FUTURE WORK

This paper described a discounted relational fuzzy clustering (DFRC) algorithm and used it for clustering of web user sessions. The relationships between web user sessions were derived from access relevance of pages in any sessions. The importance of a page or users' interest was computed by applying harmonic mean of access frequency of pages and duration of pages in any session. The simple binary web user's sessions were transformed into augmented sessions by incorporating relevance of page in accessing sessions. The augmented session dissimilarity matrix was computed from page relevance matrix using cosine similarity measure and converted to dissimilarity matrix. The intuitive augmented session dissimilarity matrix derived from a publicly accessible NASA web server log data were used to perform the experiments with visual assessment tool (VAT) and relational fuzzy c-means (RFCM) algorithm. The experimental results suggest eight numbers of clusters in the dissimilarity matrix. The generated clusters were evaluated using different fuzzy validity measures. The DFRC algorithm was applied on same intuitive augmented session dissimilarity matrix which identifies the eight numbers of clusters and their prototypes respectively without assuming any initial values. The effect of various parameters including accept/reject ratio and neighbourhood radius were evaluated on the performance of DFRC algorithm. The clusters generated by DFRC were also compared with existing fuzzy relational clustering algorithm using cluster quality measures. Experimental results suggest that quality of clusters generated using DFRC is better than that of those obtained from existing fuzzy relational clustering algorithms. The effectiveness and generalisation capability of DFRC algorithm is proposed to be evaluated with different benchmark relational data sets.

Acknowledgment

O. P. Vyas is on leave from the Indian Institute of Information Technology Allahabad-211012, UP, India.

REFERENCES

[1] A. Guerbas *et al.*, "Effective Web log mining and online navigational pattern prediction," *Knowl.-Based Syst.*, vol. 49, no. 12, pp. 50–62, Sep. 2013.

[2] B. Mobasher, J. Srivastava, and R. Cooley, "Automatic personalization based on Web usage mining," *Commun. ACM*, vol. 43, no. 8, pp. 142–151, Aug. 2000.

[3] R. Krishnapuram, A. Joshi, and L. Yi, "A fuzzy relative of the k-medoids algorithm with application to Web document and snippet clustering," in *Proc. IEEE Int. Fuzzy Syst. Conf. (FUZZ-IEEE)*, vol. 3. Aug. 1999, pp. 1281–1286.

[4] M. Marie-Hélène and T. Denæux, "RECM: Relational Evidential c-means algorithm," *Pattern Recognit. Lett.*, vol. 30, no. 11, pp. 1015–1026, 2009.

[5] Z.-G. Liu, Q. Pan, J. Dezert, and G. Mercier, "Credal c-means clustering method based on belief functions," *Knowl.-Based Syst.*, vol. 74, pp. 119–132, Jan. 2015.

[6] Z.-G. Liu, Q. Pan, J. Dezert, and A. Martin, "Adaptive imputation of missing values for incomplete pattern classification," *Pattern Recognit.*, vol. 52, pp. 85–95, Apr. 2016.

[7] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From user access patterns to dynamic hypertext linking," *Comput. Netw. ISDN Syst.*, vol. 28, nos. 7–11, pp. 1007–1014, May 1996.

[8] Y. Fu, K. Sandhu, and M.-Y. Shih, "A generalization-based approach to clustering of Web usage sessions," in *Proc. Web Usage Analysis and User Profiling*, 1999, pp. 21–38.

[9] B. Mobasher, R. Cooley, and J. Srivastava, "Creating adaptive Web sites through usage-based clustering of URLs," in *Proc. Workshop Knowl. Data Eng. Exchange (KDEX)*, Nov. 1999, pp. 19–25.

[10] O. Nasraoui, F. Hichem, R. Krishnapuram, and A. Joshi, "Extracting Web user profiles using relational competitive fuzzy clustering," *Int. J. Artif. Intell. Tools*, vol. 9, no. 4, pp. 509–526, Dec. 2000.

[11] O. Nasraoui, R. Krishnapuram, A. Joshi, and T. Kamdar, "Automatic Web user profiling and personalization using robust fuzzy relational clustering," in *E-Commerce Intelligent Methods*. Germany: Springer-verlag, 2002, pp. 233–261.

[12] J. Z. Huang, M. Ng, W.-K. Ching, J. Ng, and D. Cheung, "A cube model and cluster analysis for Web access sessions," in *Proc. WEBKDD*, 2001, pp. 48–67.

[13] K. A. Smith and A. Ng, "Web page clustering using a self-organizing map of user navigation patterns," *Decision Support Syst.*, vol. 35, no. 2, pp. 245–256, May 2003.

[14] S. K. De and P. R. Krishna, "Clustering Web transactions using rough approximation," *Fuzzy Sets Syst.*, vol. 148, no. 1, pp. 131–138, Nov. 2004.

[15] G. Castellano and M. A. Torsello, "Categorization of Web users by fuzzy clustering," in *Knowledge-Based Intelligent Information and Engineering Systems*(Lecture Notes in Computer Science). Germany: Springer Berlin Heidelberg, vol. 5178. 2008, pp. 222–229.

[16] L. Chaofeng, "Research on Web session clustering," *J. Softw.*, vol. 4, no. 5, pp. 460–468, Jul. 2009.

[17] H. Mamosian, A. M. Rahmani, and M. A. Dezfouli, "A new clustering approach based on page's path similarity for navigation patterns mining," *Int. J. Comput. Sci. Inf. Secur. (IJCSIS)*, vol. 7, no. 2, pp. 009–014, 2010.

[18] G. Sudhamathy and C. J. Venkateswaran, "Matrix based fuzzy clustering for categorization of Web users and Web pages," *Int. J. Comput. Appl. (IJCA)*, vol. 43, no. 14, pp. 43–47, 2012.

[19] A. Chakraborty and S. Bandyopadhyay, "Clustering of Web sessions by FOGSAA," in *Proc. Recent Adv. Intell. Comput. Syst. (RAICS)*, Dec. 2013, pp. 282–287.

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

[21] M. A. Khalilia, J. Bezdek, M. Popescu, and J. M. Keller, "Improvements to the relational fuzzy c-means clustering algorithm," *Pattern Recognit.*, vol. 47, no. 12, pp. 3920–3930, Dec. 2014.

[22] J. C. Bezdek, W. Full, and R. Ehrlich, "FCM: The fuzzy c-means clustering algorithm," *Comput. & Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.

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

[24] D. S. Sisodia and S. Verma, "Web usage pattern analysis through Web logs: A review," in *Proc. IEEE 9th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, May/Jun. 2012, pp. 49–53.

[25] D. S. Sisodia, S. Verma, and O. P. Vyas, "A comparative analysis of browsing behavior of human visitors and automatic software agents," *Amer. J. Syst. Softw.*, vol. 3, no. 2, pp. 31–35, 2015.

[26] D. S. Sisodia, S. Verma, and O. P. Vyas, "Augmented intuitive dissimilarity metric for clustering of Web user sessions," *J. Inf. Sci.*, pp. 1–12, May 2016, doi: 10.1177/0165551516648259.

[27] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in Web-usage analysis," *INFORMS J. Comput.*, vol. 15, no. 2, pp. 171–190, 2003.

[28] 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," *J. Data Anal. Inf. Process.*, vol. 3, no. 1, pp. 1–10, 2016.

[29] 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," *Int. J. Appl. Eng. Res.*, vol. 11, no. 9, pp. 6497–6503, 2016.

[30] P. K. Chan, "A non-invasive learning approach to building Web user profiles," in *Proc. Workshop Web Usage Anal. (KDD)*, 1999, pp. 7–12.

[31] J. Xiao, Y. Zhang, X. Jia, and T. Li, "Measuring similarity of interests for clustering Web-users," in *Proc. 12th Australasian Database Conf. (ADC)*, 2001, pp. 107–114.
[32] D. S. Sisodia, S. Verma, and O. P. Vyas, "Quantitative evaluation of Web user session dissimilarity measures using medoids based relational fuzzy clustering," *Indian J. Sci. Technol.*, vol. 9, no. 28, pp. 1–9, Jul. 2016.
[33] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *J. Intell. Fuzzy Syst.*, vol. 2, no. 3, pp. 267–278, 1994.
[34] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Trans. Syst., Man and*, vol. 24, no. 8, pp. 1279–1284, Aug. 1994.
[35] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets Syst.*, vol. 158, no. 19, pp. 2095–2117, Oct. 2007.
[36] H.-L. Shieh, "Robust validity index for a modified subtractive clustering algorithm," *Appl. Soft Comput. J.*, vol. 22, pp. 47–59, Sep. 2014.
[37] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.
[38] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in *Proc. 5th Fuzzy Syst. Symp.*, 1989, pp. 247–250.
[39] A. M. Bensaid *et al.*, "Validity-guided (re)clustering with applications to image segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 4, no. 2, pp. 112–123, May 1996.
[40] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering," *Pattern Recognit.*, vol. 32, no. 7, pp. 1089–1097, 1999.
[41] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, Dec. 2001.
[42] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster Validity Methods?: Part I," *ACM SIGMOD Rec.*, vol. 31, no. 2, pp. 40–45, Sep. 2002.
[43] M. Brun *et al.*, "Model-based evaluation of clustering validation measures," *Pattern Recognit.*, vol. 40, no. 3, pp. 807–824, Mar. 2007.
[44] J.-P. Mei and L. Chen, "LinkFCM: Relation integrated fuzzy c-means," *Pattern Recognit.*, vol. 46, no. 1, pp. 272–283, Jan. 2013.
[45] J. C. Bezdek and R. J. Hathaway, "VAT: a tool for visual assessment of (cluster) tendency," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 3, 2002, pp. 2225–2230.
[46] T. C. Havens, and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 813–822, May 2012.
[47] L. A. Wang, X. Geng, J. Bezdek, C. Leckie, and R. Kotagiri, "Enhanced visual analysis for cluster tendency assessment and data partitioning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1401–1414, Oct. 2010.
[48] L. Wang, U. T. V. Nguyen, J. C. Bezdek, C. A. Leckie, and R. Kotagiri, "iVAT and aVAT: Enhanced visual analysis for cluster tendency assessment," in *Proc. Adv. Knowl. Discovery Data Mining*, 2010, pp. 16–27.

DILIP SINGH SISODIA (M'12) received the B.E. and M.Tech. degrees respectively in computer science and engineering and information technology with a focus on artificial intelligence from the Rajiv Gandhi Technological University, Bhopal, India, and the Ph.D. degree in computer science and engineering from the National Institute of Technology Raipur, India. He is currently an Assistant Professor with the Department of Computer Science Engineering, NIT Raipur. He has over 13 years of experience of various reputed institutes in the field of academics and research. He has published over 15 referred articles and served as a reviewer for several international journals, and conferences. His current research interests include web usage mining, machine learning, and computational intelligence. He is actively associated with various professional societies including ACM, CSI, IETE, and IE (India).

SHRISH VERMA received the master's degree in electronics and electrical communication engineering from the IIT Kharagpur, India, with a focus on computer engineering, and the Ph.D. degree in engineering from Pt. Ravi Shankar Shukla University, Raipur, India. He is currently a Professor with the Department of Electronics and Telecommunication, NIT Raipur. He has authored over 50 research papers in various journals and conferences in the field of computer and communication networks, distributed processing, data mining and analysis, text analytics and software engineering. He has served as a Reviewer of several journals. His current research interests include digital system design, data mining and its applications, and software fault prediction.

OM PRAKASH VYAS (M'14) received the M.Tech. degree in computer science from the IIT Kharagpur, India, and the Ph.D. degree in computer networks from the IIT and the Technical University of Kaiserslautern, Kaiserslautern, Germany. He is currently a Professor with the International Institute of Information Technology, Naya Raipur. He is on leave from IIIT Allahabad. He is also a Visiting Professor with the University of Paderborn, Paderborn, Germany. He has authored over 100 research papers and three books. He completed an Indo-German Project under the Department of Science and Technology, Bundesministerium für Bildung und Forschung and an Indo-French Project with Inria-France. His current research interests include data analytics, software engineering, and smart city technologies. He was a recipient of the Deutscher Akademischer Austauschdienst German Academic Exchange Service Fellowship (Technical University of Kaiserslautern) and the Association for Overseas Technical Scholarship Fellowship (Center of the International Cooperation for Computerization, Japan).