A Conglomerate Relational Fuzzy Approach for Discovering Web User Session Clusters from Web Server Logs

Dilip Singh Sisodia[#1], Shrish Verma[$2], Om Prakash Vyas[*3]
[#]Department of Computer Science & Engineering,
[$]Department of Electronics & Telecommunications,
[*]Departments of Information Technology,
[1,2]National Institute of Technology Raipur, Raipur, India
[3]Indian Institute of Information Technology Allahabad, Allahabad, India
[1]corresponding author email: dssisodia.cs@nitrr.ac.in

Abstract-Clustering of web user sessions is extremely significant to comprehend their surfing activities on the internet. Users with similar browsing behaviour are grouped together, and further analysis of discovered user groups by domain experts may generate usable and actionable knowledge. In this paper, a conglomerative clustering approach is presented to identify web user session clusters from web server access logs, based on their browsing behaviour. Presented algorithm captures essential ideas from subtractive and relational fuzzy c-mean clustering algorithm.This algorithm works in two phases, in the first phase, it automatically identifies the number of potential clusters based on the successively subtractive potential density function value of each relational data and their respective centres (centroid). In the second phase, it assigns fuzzy membership values to from fuzzy clusters from a relational matrix. The presented algorithm is applied on an augmented session dissimilarity matrix obtained from an openly accessible NASA web server log data.

Keywords: relational clustering, subtractive, similarity, user sessions, user profile.

I. INTRODUCTION

Clustering techniques are broadly classified into two major classes, one which works with feature vector based clustering(object data clustering) and other works with relational data(relational clustering). Even though feature vector clustering is very popular and received lots of attention from researchers, it is not very much suitable for clustering of user sessions due to high dimensional and correlated feature space of web-related data [1].

In this paper, a conglomerate relational data clustering approach is used to cluster user sessions based on their browsing behaviour. For this, an augmented session dissimilarity based relational matrix is computed between all user sessions by defining and calculating the various similarities/dissimilarity measures.

The remaining of this paper is arranged as follows: Section 2 briefly reviews the existing literature on user session clustering. After that in section 3, the methodology adopted for proposed approach is described in details. Section 4 presents the formulation of the idea of proposed conglomerate fuzzy clustering algorithm in details. In section 5, experiments are set to demonstrate the performance of the given clustering approach on NASA web server log data and results are discussed. Finally, this study is concluded with future work in section 6.

II. RELATED WORK

Clustering techniques have been extensively investigated in the web usage mining to categorise web users/sessions according to their web access behaviour with varying precision and accuracy in reported results. 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 [2] and further extended in [3]. [4] Propose a fuzzy similarity measure and used the same in a relational fuzzy clustering algorithm to find underlying clusters in the web usage data and derive categories modelling the preferences of similar users. In [5] 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 [6], [7] Relational fuzzy c-means (RFCM) algorithm is presented for gathering pairwise dissimilarity values in a dissimilarity matrix D. Where RFCM is dual to the fuzzy c-means (FCM) [8] object data algorithm when D is a Euclidean matrix. The objective function of RFCM was based on computing c representative clusters from the data so that the total dissimilarity between each group is minimised. But RFCM works when we specify the number of potential clusters in advance, and this is not always feasible in user session clustering.

In this paper, we formulated a conglomerative web user clustering approach which captures key idea from potential density based subtractive clustering (SC)[23] and relational fuzzy c-means RFCM[6],[7] algorithm and subsequently known as conglomerate fuzzy relational c-mean (DFRCM) clustering algorithm. In Table 1 we are presenting the summary of different notations used in the successive development of this study.

III. PROPOSED METHODOLOGY

In this section, a detailed description of adopted procedure is presented. The flowchart in Figure 1 gives the brief outline of the proposed methodology.

A. Web server logs pre-processing

The web server access log keeps a record of all files accessed by users explicitly or implicitly. Each log entry consists of different fields like 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.[9].First we remove those entries which are not germane to our purpose. Mostly these are implicit requests made by embedded objects within web pages [10], requests made by automated software agents [12], unsuccessful requests of users, requests with access methods other than GET etc. In Second step web user sessions are identified by adopting the methods presented in [11], [12].

B. User Session in Vector Space Model

Suppose that, for a given website, there are m usage sessions extracted from the web server log $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots \mathcal{S}_m\}\ for\ i = 1,2, \dots m$, Accessing n number of different URL's (pages) $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \dots \mathcal{P}_n\}\ for\ j = 1,2, \dots n$ of a website in some time interval. We can represent each user session $\mathcal{S}_i\ (for\ i = 1,2, \dots m)$ in n - dimensional vector over the space of web pages. $\mathcal{S} = \{(\mathcal{P}_1, t_1, \mathbb{f}_1, s_1), (\mathcal{P}_2, t_2, \mathbb{f}_2, s_2), \dots (\mathcal{P}_n, t_n, \mathbb{f}_n, s_n)\}\ for\ j = 1,2, \dots n.$ Where $\mathcal{P}_j, t_j, \mathbb{f}_j\ and\ s_j$ are the accessed page, time spent on page (in seconds), # of visits to page and size of page (in bytes) respectively.

C. Page relevance based augmented session similarity

Web users' interests for any page are computed by implicit measures [13], [14]. These implicit measures are page stay time (duration) and page access frequency of a web page in web user session [15]. 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.

1) Duration of Page($\mathcal{D}o\mathcal{P}$):

Duration of page or Page stay time is defined as the time spent on a page by 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 webpage in the session from the access log file. Following Eq. (1) is used to measure the duration of a web page ($\mathcal{P}_i$) in user session ($\mathcal{S}_k$)

$$(\mathcal{D}o\mathcal{P})_{\mathcal{P}_i} = \frac{\frac{\sum \text{Time Spent on}(\mathcal{P}_i)}{\text{Size of}\ (\mathcal{P}_i)}}{\text{Max}\left(\forall_{j \in \mathcal{S}_k} \frac{\sum \text{Time Spent on}(\mathcal{P}_j)}{\text{Size of}\ (\mathcal{P}_j)}\right)} \quad (1)$$

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

2) 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

$$(\mathcal{F}o\mathcal{P})_{\mathcal{P}_i} = \frac{\sum \#\ \text{of visits to}(\mathcal{P}_i)}{\text{Max}\left(\forall_{j \in \mathcal{S}_k} \sum \#\ \text{of visits to}(\mathcal{P}_j)\right)} \quad (2)$$

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

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

Relevance of page in any session is measured by giving equal importance to duration of page and frequency of Page. We use Eq. (3) to measure the relevance of web page ($\mathcal{P}_i$) in user session ($\mathcal{S}_k$)

$$(\mathcal{R}o\mathcal{P})_{\mathcal{P}_i} = \frac{2 \times (\mathcal{D}o\mathcal{P})_{\mathcal{P}_i} \times (\mathcal{F}o\mathcal{P})_{\mathcal{P}_i}}{(\mathcal{D}o\mathcal{P})_{\mathcal{P}_i} + (\mathcal{F}o\mathcal{P})_{\mathcal{P}_i}} \quad (3)$$

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

Table I. Brief Description of Notations used in this article

Notations used	Brief description
$\mathcal{AS}_i$	Set of augmented session
$\mathcal{AS}_i^n$	Vector representation of i^{th} session in n-dimention
$\mathcal{ASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)}$	Augmented session similarity between sessions $\mathcal{AS}_a$ and $\mathcal{AS}_b$
$\mathcal{A}_r$	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
$\mathcal{D}^2_{(\mathcal{AS}_a,\mathcal{AS}_b)}$	An augmented dissimilarity matrix
$(\mathcal{D}o\mathcal{P})_{\mathcal{P}_i}$	Duration of page $\mathcal{P}_i$
$\mathcal{F}_{\mathcal{RFCM}}$	Objective function of RFCM
$(\mathcal{F}o\mathcal{P})_{\mathcal{P}_i}$	Frequency of page $\mathcal{P}_i$
f_j	Number of visits to j^{th} page
f	Fuzzification coefficient
$\mathcal{L}$	Set of 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
$(\mathcal{R}o\mathcal{P})_{\mathcal{P}_i}$	Relevance of each accessing page
r_i	i^{th} record of a log file
r_a^2	Neighbourhood radius
r_b^2	Neighbourhood radius
$\mathcal{R}_r$	Rejection ratio
$\mathcal{RM}_{m\times n}$	Page Relevance Matrix
s_j	size of j^{th}page (in bytes)
t_j	time spent on j^{th} page (in seconds)
$\mathcal{S}_i$	i^{th} user session
$\mathcal{U}$	Fuzzy membership matrix
$v_{\mathcal{R},j}$	j^{th} cluster prototype
$\mathcal{V}_{\mathcal{R}}$	Set of cluster prototypes
$\mathcal{W}_{ij}$	Weight of j^{th} URL in i^{th} cluster
μ_{ij}	Degree of membership between j^{th} cluster prototype and i^{th} session

D. Augmented web user sessions

Now, by applying equations (1) to (3) 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. (4). By incorporating page relevance in web user session access behaviour matrix, simple web user sessions converted to augmented web user sessions.

$$\mathcal{RM}_{m\times n}=\begin{pmatrix}(\mathcal{R}o\mathcal{P})_{11} & (\mathcal{R}o\mathcal{P})_{12} & \cdots & (\mathcal{R}o\mathcal{P})_{1n}\\ (\mathcal{R}o\mathcal{P})_{21} & (\mathcal{R}o\mathcal{P})_{22} & \ldots & (\mathcal{R}o\mathcal{P})_{12}\\ \vdots & \vdots & \ddots & \vdots\\ (\mathcal{R}o\mathcal{P})_{m1} & (\mathcal{R}o\mathcal{P})_{m2} & \cdots & (\mathcal{R}o\mathcal{P})_{mn}\end{pmatrix} \quad (4)$$

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

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

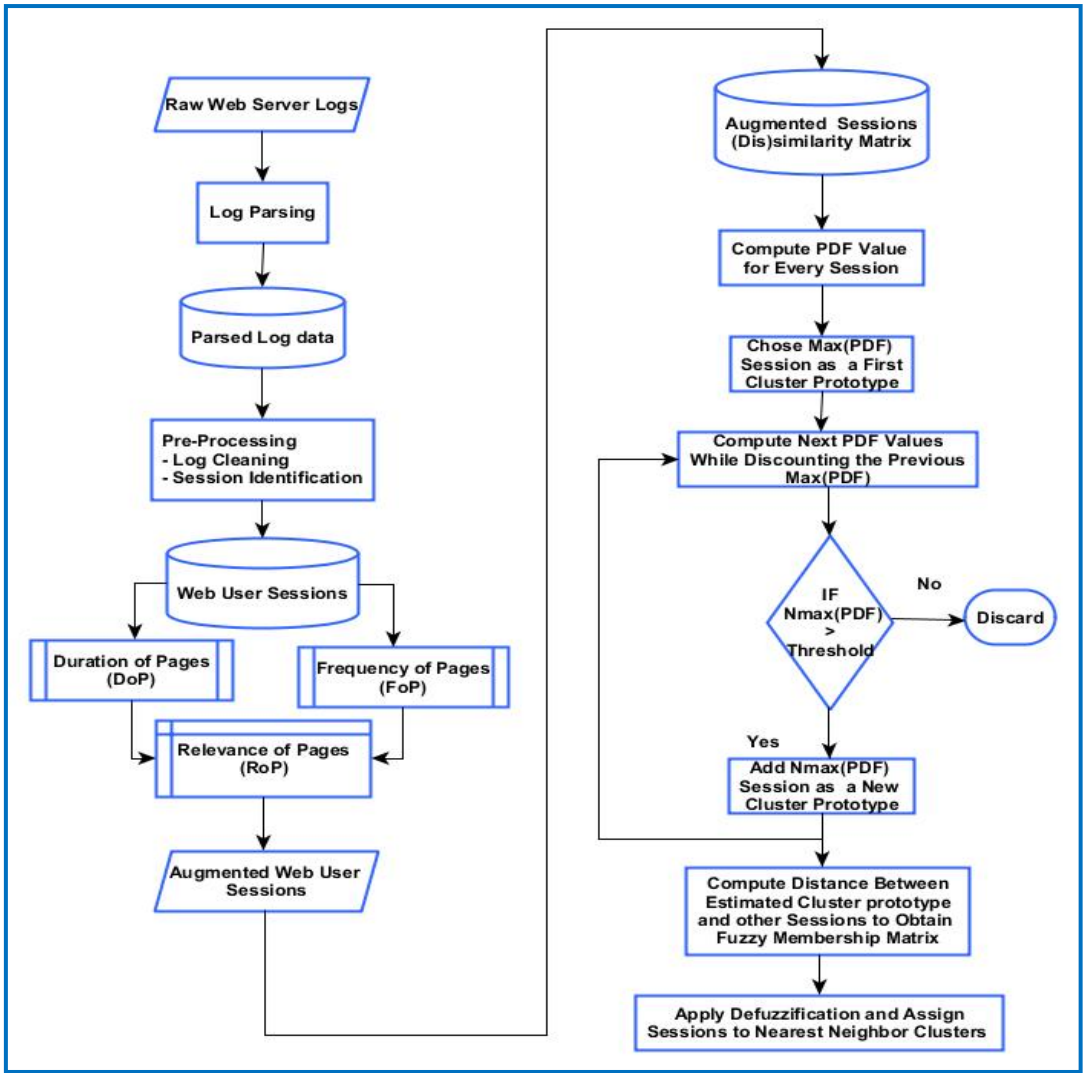

Fig 1: Proposed methodology for web user session clustering

E. 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 [2].

$$\mathcal{ASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)} = \frac{\sum_{i=1}^{n} \mathcal{AS}_a(\mathcal{RoP})_i \times \mathcal{AS}_b(\mathcal{RoP})_j}{\sqrt{\sum_{i=1}^{n} \mathcal{AS}_a(\mathcal{RoP})_i{}^2}\sqrt{\sum_{i=1}^{n} \mathcal{AS}_b(\mathcal{RoP})_j{}^2}} \quad (5)$$

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

$$\mathcal{D}^2_{(\mathcal{AS}_a,\mathcal{AS}_b)} = (1 - \mathcal{ASS}_{(\mathcal{AS}_a,\mathcal{AS}_b)})^2 \quad (6)$$

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

The pseudo code for above-described procedure is presented a Algorithm1.Which summarises the steps involved in the computation of page relevance based dissimilarity matrix of web user sessions

Algorithm 1: pseudo code for Page relevance based augmented session (dis)similarity matrix.

Input: $\{web\ server\ log\ file:\ \mathcal{L}\ of\ n\ records\ where\ \mathcal{L} \leftarrow \{r_1, r_2 \ldots r_n\}\ ,where\ n \ggg 1, \forall \ni r_i < ip, time, method, url, protocol, size, status, agent, referrer >\}$

Output: $\{ \mathcal{D}_{m\times m}|$ augmented session dissimilarity $\}$

1: Pre-processing of web server access log
   a. Removing of extraneous information- $\mathcal{L}^c \leftarrow \{r_1, r_2 \ldots r_n\}$ is a cleaned web log file
   b. Identification of web users $\mathcal{U}_i = \{ \mathcal{U}_1, \mathcal{U}_2, \ldots . \mathcal{U}_n \}\ for\ i = 1,2, \ldots n.$
   c. Identification of web user sessions $\mathcal{S} = \{ \mathcal{S}_1, \mathcal{S}_2, \ldots . \mathcal{S}_m \}\ for\ i = 1,2, \ldots m.$
   Where, $m \geq n$

2: Vector representation of web user sessions:
$\mathcal{S} = \{(\mathcal{P}_1, t_1, f_1, s_1), (\mathcal{P}_2, t_2, f_2, s_2), \ldots (\mathcal{P}_n, t_n, f_n, s_n)\}\ for\ j = 1,2, \ldots n.$
Where, $\mathcal{P}_j, t_j, f_j\ and\ s_j$ are the accessed page, time spent on page (in seconds), # of visits to page and the size of the page (in bytes) respectively.

3: Computation of relevance of any web page in user sessions for each pair of $(\mathcal{S}_k, \mathcal{P}_i)$ session $\mathcal{S}_k$and page $\mathcal{P}_i\ where\ i = 1,2, \ldots n\ and\ k = 1,2, \ldots m.$
   a. Compute the duration of a web page $(\mathcal{P}_i)$ in user session $(\mathcal{S}_k)$ using Eq. (1).
   b. Compute the Frequency of web page $(\mathcal{P}_i)$ in user session $(\mathcal{S}_k)$ using Eq. (2).
   c. Compute the relevance of web page $(\mathcal{P}_i)$ in user session $(\mathcal{S}_k)$ using Eq. (3).

4: Computation of page relevance based augmented session similarity matrix by using Eq. (5)

5: Calculation of page relevance based augmented session dissimilarity matrix $\mathcal{D}_{m\times m} = 1 - \mathcal{RM}_{m\times \mathrm{m}}$ using Eq.(6).

IV. DESCRIPTION OF PROPOSED CLUSTERING ALGORITHM

In this section, we are presenting the formulation of the idea of proposed conglomerate fuzzy relational c-mean (DFRCM) clustering algorithm in details.

F. The estimation of number of representative cluster centres

To find a representative cluster prototype , clustering algorithms search for the centres of dense regions in the given relation. We applied subtractive clustering method [16] is which more useful extension of the mountain clustering [17] for estimation of some cluster centres. The subtractive clustering method assumes that each data point is a potential cluster centre and, without prior knowledge of the default number of centres, it calculates the likelihood of a data point being defined as a cluster centre according to the density of the surrounding data points. We compute the value of potential density function at every augmented session using Eq. (7)

$$\mathcal{P}_1(\mathcal{AS}_i) = \sum_{j=1}^{n} exp\left(-\frac{d_{\mathcal{R},ij}(\mathcal{AS}_i, \mathcal{AS}_j)}{r_a^2}\right), \forall i = 1,2, \ldots, m. \quad (7)$$

If $d_{\mathcal{R},ij}$ is small then $\mathcal{AS}_j$ and $\mathcal{AS}_i$ will be more related and had major influence on the potential density value $\mathcal{P}_1(\mathcal{AS}_i)$; otherwise $\mathcal{AS}_j$ and $\mathcal{AS}_i$ will be less related and had no significant influence on $\mathcal{P}_1(\mathcal{AS}_i)$. The parameter r_a^2 is a radius and defines the neighbourhood of the selected augmented session $\mathcal{AS}_i$. The sessions outside this radius have little influence on the selected session's potential density value.

After computing the PDF at every session, we chose the highest PDF value session as the first representative cluster centre $v_{\mathcal{R},1}$ by using Eq. (8). If there are multiple sessions with highest PDF value then we choose subjectively any one of them.

$$\mathcal{P}_1\left(\mathcal{AS}_{k_1}\right) = \max_{i=1}\{\mathcal{P}_1(\mathcal{AS}_i)\}; \qquad v_{\mathcal{R},1} \leftarrow \left(\mathcal{AS}_{k_1}\right) \quad (8)$$

To find the second representative cluster centre, we use Eq. (9) compute the subtractive PDF value over the neighbourhood defined by r_b^2. If $d_{\mathcal{R},ij}$ is smaller between $\mathcal{AS}_i$ and $v_{\mathcal{R},1}$ then effective potential of each sessions around $v_{\mathcal{R},1}$will be reduced due to this subtraction.

$$\mathcal{P}_2(\mathcal{AS}_i) = \mathcal{P}_1(\mathcal{AS}_i) - \mathcal{P}_1\left(v_{\mathcal{R},1}\right) exp\left(-\frac{d_{\mathcal{R},ij}(\mathcal{AS}_i, v_{\mathcal{R},1})}{r_b^2}\right), \forall i = 1,2, \ldots, m. \quad (9)$$

After discounting PDF values for all sessions over effective zone of influence of r_b^2 , we select the highest subtractive PDF value as the second representative cluster centre $v_{\mathcal{R},2}$ by using Eq.(10)

$$\mathcal{P}_2\left(\mathcal{AS}_{k_2}\right) = \max_{i=1}\{\mathcal{P}_2(\mathcal{AS}_i)\}; \qquad v_{\mathcal{R},2} \leftarrow \mathcal{P}_2\left(\mathcal{S}_{k_2}\right) \quad (10)$$

Similarly, to select any j[th] representative cluster centre, the PDF value of each user session over effective zone of influence during j[th] iteration is computed using Eq. (11)

$$\mathcal{P}_j(\mathcal{AS}_i) = \mathcal{P}_{j-1}(\mathcal{AS}_i) - \mathcal{P}_{j-1}\left(v_{\mathcal{R},(j-1)}\right) exp\left(-\frac{d_{\mathcal{R},ij}(\mathcal{AS}_i, v_{\mathcal{R},(j-1)})}{r^2}\right), \forall i = 2, \ldots, m. \quad (11)$$

The same procedure will continued 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. (12)

$$\mathcal{P}_j\left(\mathcal{AS}_{k_j}\right) = \max_{i=2}\{\mathcal{P}_j(\mathcal{AS}_i)\}, \forall i = 2, \dots, m. \;\; v_{\mathcal{R},j} \leftarrow \mathcal{P}_j\left(\mathcal{AS}_{k_j}\right) \tag{12}$$

If this value is less than the reject ratio $\mathcal{R}_r$ then we 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 far from the existing representative cluster centre.

G. Relational Fuzzy C-mean clustering

Given a set of augmented user sessions $\mathcal{AS}_i = \{\mathcal{AS}_1, \mathcal{AS}_2, \dots . \mathcal{AS}_m\}$ $for\ i = 1,2, \dots m$. Where, each session is represented by vector of n-dimensions $\mathcal{S} = \{\mathcal{AS}_i^1, \mathcal{AS}_i^2, \dots \mathcal{AS}_i^n\}$, $\forall i = 1,2, \dots, m$. Let $d_{\mathcal{R},ji}$ is the relational distance between cluster prototype and augmented session $\mathcal{AS}_i$. Let $\mathcal{V}_\mathcal{R} \leftarrow \{v_{\mathcal{R},1}, v_{\mathcal{R},2}, \dots v_{\mathcal{R},c}\}$ represent a set of relational cluster centres in $\mathcal{D}$. The objective function of relational fuzzy c-means algorithm seek to c representative sessions as relational cluster centres (known as centroid), such that the total distance of other sessions to their closest centroid is minimized. The objective function of relational fuzzy c-means (RFCM) [19] is defined as Eq. (13).

$$\mathcal{F}_{\mathcal{R}FCM} = \sum_{j=1}^{c} \frac{\sum_{i=1}^{n}\sum_{k=1}^{n} \mu_{ij}^{f}\mu_{kj}^{f} d_{\mathcal{R},ji}}{2\sum_{i=1}^{n}\mu_{ij}^{f}} \tag{13}$$

And membership functions is given by (14)

$$\mu_{ij} = \frac{\left(d_{\mathcal{R},ji}\right)^{-\frac{1}{(f-1)}}}{\sum_{j=1}^{c}\left(d_{\mathcal{R},ji}\right)^{-\frac{1}{(f-1)}}} \tag{14}$$

Where, $f \in [1, \infty]$ is a fuzzification coefficient and The Euclidean distance $d_{\mathcal{R},ji}$ is the relational distance between cluster prototype and augmented session $\mathcal{AS}_i$ and this distance is calculated based on memberships in $\mathcal{U}$ and dissimilarities in $\mathcal{D}$ using Eq.(15)

$$d_{\mathcal{R},ji} = \left(\mathcal{D}v_{\mathcal{R},j}^{t-1}\right)_i - \frac{1}{2}\left(v_{\mathcal{R},j}^{t-1}\right)^T \mathcal{D}v_{\mathcal{R},j}^{t-1} \quad \text{for } 1 \le j \le c \text{ and } 1 \le i \le m \tag{15}$$

The relational cluster centres are updated by using Eq.(16)

$$v_{\mathcal{R},j}^{t} = \frac{(\mu_{j1}{}^{f}, \mu_{j2}{}^{f}, \dots \mu_{jm}{}^{f})}{\sum_{i}^{m}\mu_{ij}{}^{f}} \text{ for } 1 \le j \le c \tag{16}$$

The pseudo code for above-described procedure is presented as algorithm 2. Which summarises the steps involved in conglomerate relational fuzzy clustering on estimating representative cluster centres from page relevance based relational matrix of augmented web user sessions.

Algorithm 2: pseudo code for conglomerate relational fuzzy clustering for augmented web user sessions

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

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

1: t ← 1;
2: for $i = 1,2, \dots, m$
3: calculate potential density function (PDF)of each session using Eq.(7)
4: end for
5: select the session with Maxpdf $\mathcal{P}_1\left(\mathcal{S}_{k_1}\right) = max_{i=1}\{\mathcal{P}_1(\mathcal{S}_i)\}$
6: set it as first cluster centre $v_{\mathcal{R},1} \leftarrow \mathcal{P}_1\left(\mathcal{S}_{k_1}\right)$
7: compute the subtractive PDF of each session using Eq.(11)
8: 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
9: 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
10: else let $d_{min} = min_{j=1}^{c-1} d_{jc}^2\left(v_{\mathcal{R},j}, v_{\mathcal{R},c}\right)$
11: 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
12: t ←t +1 & set $v_{\mathcal{R},j} \leftarrow \mathcal{P}_j\left(\mathcal{S}_{k_j}\right)$, go to step 3
13: else discard $\mathcal{S}_{k_j}$ and $0 \leftarrow \mathcal{P}_j(\mathcal{S}_i)$ and select $\mathcal{P}_{next}(\mathcal{S}_i)$, go to step 3

14: end if
15: end if
16: Estimated cluster centres: $\mathcal{V}_{\mathcal{R}} \leftarrow \{v_{\mathcal{R},1}, v_{\mathcal{R},2}, \ldots v_{\mathcal{R},c}\}$
17: Calculate Euclidean distance ($d_{\mathcal{R},ji}$)between augmented sessions $\mathcal{AS}_i$ and centroid of clusters using Eq.(15)
18: For $i \leftarrow 1,2, \ldots m$ do
19: If $d_{\mathcal{R},ji} \neq 0 \quad \forall j$
20: calculate membership function matrix ($\mathcal{U}$) using Eq.(14)
21: else
22: Set $\mu_{ij} > 0 \; for \; d_{\mathcal{R},ji} = 0, \; \mu_{ij} \in [0,1]$and $\sum_{j=1}^{c} \mu_{ij} = 1$
23: End If
24: End For
25: Apply defuzzification by assigning sessions to nearest neighbour clusters

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the clustering capability of the proposed approach is presented by applying it to open access NASA web server log data[20]. 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 proposed algorithms are implemented using MATLAB (R2012a) package [21].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) with a randomly selected different number of web user sessions from a pre-processed NASA access log data set.

After cleaning of irrelevant entries(image, icons, sound files, etc.) from the original log file, it has been reduced to 525981 entries results are shown in the figure. However, we did not find any entries for automated software (web robots, spiders, crawlers, indexers, etc.) probably due to old log files. In this log data, a total number of 75060 unique hosts'/IP are requesting 4030 individual pages.

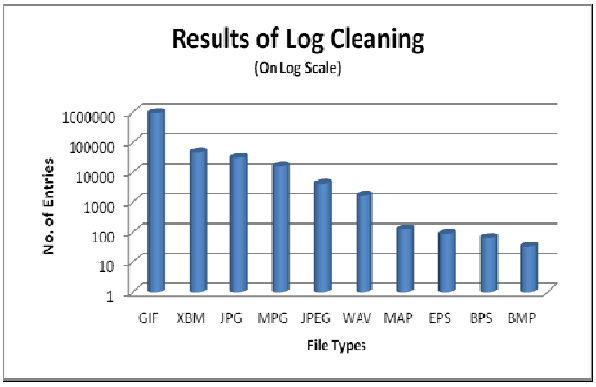

Fig 2 Web server log cleaning results

The sessions are identified by setting 30 minutes threshold time, as it is widely acceptable to capture the user notion in most of the weblog dataset, and a total of 139,086 sessions are identified. To alleviate processing overhead, randomly selected user sessions of size 1000, 2000 and 3000 are considered from an enormous number of sessions for further processing. The default root / and mini sessions of size 1 are filtered out from the total generated sessions as they did not contribute any significant information for user session clustering. Table 2 presents a summary of cleaned and pre-processed sample logs.

Table II. Summary of pre-processed sessions and matrices

Number of initial sessions	Number of valid sessions	Number of unique URLs in sessions	Size of FoP/DoP /RoP matrix	Size of ASS/ ($\mathcal{D}_{m\times m}$) matrix
1000	665	419	665×419	665×665
2000	1341	589	1341×589	1341×1341
3000	2048	731	2048×731	2048×2048

From pre-processed sample logs frequency of page (FoP), duration of the page (DoP) and consequently relevance of page (RoP) matrix is computed. The page relevance matrix is harmonic mean of the frequency of page(FoP) and duration of the page(DoP) and gives equal importance to both matrices.The Augmented session (dis)similarity(ASS) is derived from this matrix and converted to dissimilarity matrix($\mathcal{D}_{m\times m}$) using algorithm 1. The size of FoP, DoP, RoP and ASS matrix for sample data sets is given in Table 2.

To visualise, potential clusters buried in relational dissimilarity matrix, a visual assessment of tendency (VAT) tool is used [22]. This VAT tool maps the values of dissimilarity matrix into the densities of a reordered VAT image along the diagonal. The number of dark blocks in diagonal which possess visual clarity represents the number of potential clusters. However, this is not always possible. These dark blocks appear only when a compact group exists in the data; however, this is not always possible [23]. The VAT images of augmented session (dis)similarity (ASS) are shown in Figure 3.

The augmented session dissimilarity ($\mathcal{D}_{m\times m}$) for different size of user session data sets and other parameters (as given in Table 3), are used as input arguments in the proposed conglomerate fuzzy c-mean clustering algorithm (Algorithm 2). The selection of right neighbourhood radius parameters r_a^2 and r_b^2 affect the output of clustering algorithm.Generally, the value of $r_b^2 \geq r_a^2$ because algorithm seeks to find all sessions with significant potential density values, while discounting the influence zone of already detected potential cluster prototypes.

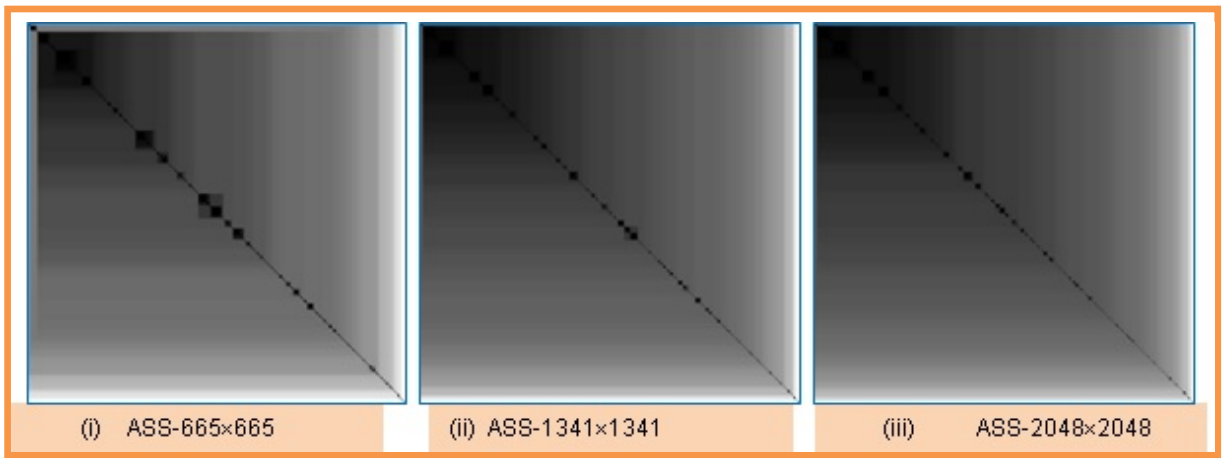

Fig 3: VAT images of augmented session pair wise dissimilarity matrix of different sizes

The value of accept ratio and reject ratio governs the number of generated clusters, if $\mathcal{A}_r$and $\mathcal{R}_r$ is too low then it will generate more clusters and some of them may be of insignificant potential density values.if the value of $\mathcal{A}_r$and $\mathcal{R}_r$ is very high then only few clusters will be generated and some significant clusters may be left undetected.We conducted multiple experiments on selected data sets with different values and combinations of these parameters and empirically selected the best values for these data set.

Table III. Summary of used default parameter values

Parameters	Symbols	Choose Values	Range
Neighbourhood radius	r_a^2	0.3	$0.2\leq r_a^2 \leq 0.5$
Neighbourhood radius	r_b^2	1.5 r_a^2	$r_b^2 \geq r_a^2$
Acceptance ratio	$\mathcal{A}_r$	0.7	$0.3\leq \mathcal{A}_r \leq 0.9$
Rejection ratio	$\mathcal{R}_r$	0.15	$0.05\leq \mathcal{R}_r \leq 0.4$
Fuzzyfier	$\oint$	1.5	$\oint \in [1,\infty]$

This algorithm produces output as Potential density value for an index of prototype session of the cluster in descending order. For our experimental log sessions data the number of generated clusters and time taken to produce these clusters are summarised in Table IV.

Table IV. Time required for generate clusters from different size dissimilarity matrix.

Number of initial sessions	Size of ASS/ ($\mathcal{D}_{m\times m}$) matrix	Number of generated clusters	Time required creating clusters (In Seconds)
1000	665×665	8	5.71
2000	1341×1341	10	65.41
3000	2048×2048	12	103.63

After finding the number of clusters and cluster centres (prototype sessions), the degree of membership of each session in every group is determined. By applying defuzzification, sessions with the highest level of accession is assigned crisply to nearest neighbour clusters to decide cardinality of clusters. All designated sessions are merged in a cluster, to represent the aggregate view of uniquely accessing URLs in that cluster. The detailed results are summarised in Table 5,6, and 7.

Table V. Clustering results for dissimilarity matrix of size 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

Table VI. Clustering results for dissimilarity matrix of size 1341×1341.

Cluster Number	Prototype Session of the cluster	Cardinality of Cluster	Potential density Value	Number of Unique URLs
1	1104	396	506.28	400
2	256	223	503.91	264
3	982	19	495.54	58
4	156	270	488.39	182
5	1068	18	477.03	63
6	1105	41	431.48	95
7	1034	19	418.44	70
8	200	8	387.21	66
9	609	337	362.37	210
10	94	10	349.35	43

Now, the clusters are representing the aggregate view of all belonging members (web user sessions) and their accessing page URLs. We computed the weight of URLs in any cluster by using Eq. (17) as given in [2].

$$\mathcal{W}_{ij} = \frac{\#\ of\ \text{occurrences of the } j^{\text{th}} \text{ URL in cluster } i}{cardinality\ of\ cluster\ i} \tag{17}$$

Table VII. Clustering results for dissimilarity matrix of size 2048×2048.

Cluster Number	Prototype Session of the cluster	Cardinality of Cluster	Potential density Value	Number of Unique URLs
1	1104	566	737.43	477
2	256	344	735.03	377
3	982	79	734.06	148
4	1378	58	727.05	138
5	1452	23	714.28	82
6	156	400	688.73	264
7	1034	13	634.10	30
8	1057	17	625.06	111
9	1353	51	590.10	82
10	609	480	560.69	257
11	94	10	531.90	45
12	1992	7	469.61	20

The most important pages in any clusters are represented by URL weight value($\mathcal{W}_{ij}$), and the prototype URL of the cluster has the highest weight value. The cluster profile is represented by group of URLs who have $\mathcal{W}_{ij}$ greater than some predefined threshold weight values and are more similar to the prototype URL of cluster. Due to space constraint only the summary of Prototype URL's and their respective relevance value $\mathcal{W}_{ij}$ for each cluster of 1000 size log data are presented in Table 8.

Table VIII. Summary of generated cluster profiles for a matrix of size 665×665.

Cluster No.	Prototype URL	URL Relevance $\mathcal{R}_{ij}$	Remarks
1	/shuttle/missions/*	0.31	URLs accessed for different shuttle missions
2	/history/Apollo/*	0.60	URLs accessed for historical information
3	/cgi-bin/imagemap/astrohome/*	0.37	Dynamic web pages
4	/facilities/*	0.85	URLs giving information related with facilities
5	/shuttle/resources/*	0.56	URLs related to shuttle resources
6	/facts/*	1.00	pages giving facts
7	/ksc.html	0.53	Main page of Kennedy space centre
8	/persons/astronauts/*	1.00	URLs related with persons and astronauts

VI. CONCLUSION AND FUTURE WORK

In this paper, a conglomerate relational fuzzy clustering approach is presented for clustering of web user sessions. This method is based on potential density based sub-clustering and relational fuzzy clustering techniques. The relationships between web user sessions are derived from access relevance of pages in any sessions. The relevance of a page is a measure of user's interest for any page URL and calculated by applying harmonic mean of access frequency of pages and duration of pages in any session. The simple binary web user's sessions are transformed into augmented sessions by incorporating relevance of page in accessing sessions.The augmented session similarity matrix is computed from page relevance matrix using cosine similarity measure and converted to dissimilarity matrix. To visualise the number of potential clusters from an of web user relational dissimilarity matrix, a (VAT) tool is used. The VAT image shows the number of possible clusters ranging from 8 to 12. This dissimilarity matrix along with the values of different parameters are passed to the algorithm and get the output as some clusters, their potential density values and index of prototype session of clusters. The number of clusters generated by proposed approach are 8, 10 and 12 respectively for the different size of matrices (665×665, 1341×1341 and 2048×2048). The aggregate user cluster profile for 665×665 augmented dissimilarity matrix with 8 number of cluster is discovered and presented in results.

REFERENCES

[1] T 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] Nasraoui, Olfa, Hichem Frigui, Raghu Krishnapuram, and Anupam Joshi. "Extracting web user profiles using relational competitive fuzzy clustering." International Journal on Artificial Intelligence Tools 9, no. 04, 2000, 509-526.

[3] O. Nasraoui, R. Krishnapuram, A. Joshi, and T. Kamdar, "Automatic Web User Profiling and Personalization using Robust Fuzzy Relational Clustering," in E-Commerce and Intelligent Methods, Springer-Verlag, 2002.

[4] Castellano, Giovanna, and Maria Alessandra Torsello. "Categorization of web users by fuzzy clustering." Knowledge-Based Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2008.

[5] Sudhamathy, G., and C. Jothi Venkateswaran. "Matrix based Fuzzy Clustering for Categorization of Web Users and Web Pages." International Journal of Computer Applications 43.14 (2012): 43-47.

[6] R.J. Hathaway, J.W. Davenport, J.C. Bezdek, Relational duals of the c-means clustering algorithms, Pattern Recognit. 22 (2) (1989) 205–212 (Jan.).

[7] Khalilia MA, Bezdek J, Popescu M, Keller JM. Improvements to the relational fuzzy c-means clustering algorithm. Pattern Recognition. 2014 Dec 31; 47(12):3920-30.

[8] Bezdek, James C., Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm." Computers & Geosciences 10.2 (1984): 191-203.

[9] Sisodia, Dilip Singh, and Shrish Verma. "Web usage pattern analysis through web logs: A review." In Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on, pp. 49-53. IEEE, 2012.

[10] Sisodia, Dilip Singh, Shrish Verma, and Om Prakash Vyas. "A Comparative Analysis of Browsing Behavior of Human Visitors and Automatic Software Agents." American Journal of Systems and Software 3.2 (2015): 31-35.

[11] 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.

[12] 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.

[13] P. K. Chan. A non-invasive learning approach to building web user profiles. In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, 1999, pp. 7–12.

[14] Xiao, Jitian, Yanchun Zhang, Xiaohua Jia, and Tianzhu Li. "Measuring similarity of interests for clustering web-users." In Proceedings of the 12th Australasian database conference, IEEE Computer Society, 2001, pp. 107-114.

[15] 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, 61(2): 2007, pp. 304–330.

[16] Chiu, Stephen L. "Fuzzy model identification based on cluster estimation." Journal of Intelligent and Fuzzy Systems 2(3), 1994, pp. 267-278.

[17] Yager, Ronald R., and Dimitar P. Filev. "Approximate clustering via the mountain method." IEEE Transactions on Systems, Man and Cybernetics, 24(8), 1994, pp. 1279-1284.

[18] Huang A. Similarity measures for text document clustering. In Proceedings of the sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), pp. 49-56, 2008.

[19] Mei JP, Chen L. LinkFCM: Relation integrated fuzzy c-means, Pattern Recognition. 2013 Jan 31; 46(1):272-83.

[20] NASA Kennedy Space centre's www server log data (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html).

[21] MATLAB (R2012a) Software. http://www.mathworks.com.

[22] 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.

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

AUTHORS PROFILE

*Dilip Singh Sisodia is with the Department of Computer Science and Engineering as an Assistant Professor at National Institute of Technology, Raipur. He has completed his M.Tech. From RGTU, Bhopal, India with specialization in Artificial Intelligence. His current research interests include web usage mining, Machine learning, and computational intelligence. He has more than twelve years of experience in various academic institutes. He has published over 10 referred articles and served as reviewer of various international journals and conferences. He is a Member of IEEE, ACM and CSI.

Shrish Verma is Professor in the department of Electronics & Telecommunication, National Institute of Technology, Raipur. He has completed his Post graduation in Computer Engineering from Indian Institute of Technology, Kharagpur. He has completed his PhD in Engineering from Pt. Ravi Shankar Shukla University Raipur. His area of interest is Image processing, Web mining, Software fault prediction models and Software bug classification. He has published over 20 referred articles and served as reviewer of several journals.

Om Prakash Vyas is Professor at Indian Institute of Information Technology Allahabad. Prof. O. P. Vyas has pursued M.Tech. in Computer Science from IIT Kharagpur and Ph.D. in Computer Networks from IIT Kharagpur in joint collaboration with Technical University of Kaiserslautern (Germany). Before joining IIIT-Allahabad, he worked at Pt. R. S. University Raipur, where he successfully founded the School of Computer Science and was instrumental in introducing Computer Education in area of Chhattisgarh region. Prof. O.P. Vyas has been active researcher and had published more than 80 research papers, three books and completed one Indo-German Project under DST-BMBF. He has received DAAD Fellow (Technical University of Kaiserslautern-Germany) and AOTS Fellow (CICC Japan).