

# Fuzzy Clustering based Web Users Analysis through Weighted Session PageView Matrix in Web Usage Mining

<sup>1</sup>Serin . J, <sup>2</sup>R. Lawrance

<sup>1</sup>Research Scholar and Associate Professor, <sup>2</sup>Director,  
<sup>1</sup>Research & Development center, Bharathiar University, Coimbatore,  
Women's Christian College, Chennai

<sup>2</sup>Department of Computer Applications,  
Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India  
[serin.j@gmail.com](mailto:serin.j@gmail.com), [lawrancer@yahoo.com](mailto:lawrancer@yahoo.com)

**Abstract:** The rapid growth of Internet users are increasing ubiquitous with the necessity of identify the similar users behavioral pattern. Web Usage Mining is used to discover the users behavioral pattern in order to identify the potential customers for business organizations and also helpful to quantify the success of marketing campaign through the click stream data or log file records. In this paper, fuzzy clustering technique is used to analyze the similar users browsing pattern. The proposed model constructs the Weighted Session Page View matrix during the preprocessing stage in order to improve the accuracy of clustering results. Evaluation of the proposed method is done on UCI repository MSNBC datasets. The experimental results exhibit that the proposed method produces significantly higher performance than the traditional method.

**Keywords:** Behavioral Pattern, Fuzzy Clustering, Web Usage Mining, Weighted Session Page View Matrix.

## 1. INTRODUCTION

The number of people using the Internet has surged over the past year, with more than one million people coming online for the first time. The time spent on the internet has also gone up over the past 12 months and the online buyers' rate has also increased rapidly [1]. But with more people shopping online, more online businesses emerge creating a fierce and competitive marketing place. This makes it harder for online businesses to stand out from competitors and the need for new creative ways to market the product rises. To satisfy the needs of the users, the recent advancement in business analytics is Behavioral analytics which reveal new insights to the customers on E-commerce platforms. Web Mining is used to extract hidden knowledge from capture the click stream data. Web Usage Mining identifies the user behavioral pattern recorded in Logfiles from the server. The patterns or aggregate user profiles are discovered in the preprocessed data by applying data mining techniques and provide input to the recommendation engine which recommends pages based on their intelligence acquired in the user profiles. Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large set to produce a concise representation of a system's behavior. Normally the user's interest will be in more than one page. Fuzzy clustering is more appropriate to choose the overlapping of clusters. In Fuzzy clustering, dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree. For example, certain data point that lies close to the

center of the cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster. It allows the objects to belong to several clusters simultaneously with different degrees of membership. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather assigned membership degrees between 0 and 1 indicating their partial membership.

This paper is structured in the following way: Section 2 discusses about the related works carried out in the field of Web Usage Mining by using Fuzzy Clustering. Section 3 introduces the methodology describing about the preprocessing algorithm to get the better accuracy of the clustering results. Section 4 presents the experimental results which implements Fuzzy Clustering. In Section 5, it analysis the similar uses through Fuzzy Web Inference System. Finally, section 6 summarizes conclusion with the proposed work and future research work direction in Web Usage Mining.

## 2. LITERATURE SURVEY

Suresh et.al [2] proposed an improved Fuzzy Clustering to initialize the cluster centers by using information entropy and also introduced weighting parameters to find the location of cluster centers and noise problems. Fuzzy algorithm is applied in MSNBC web navigation dataset in order to find the Web clusters Web Usage Mining. Castellano et. al [3] presented an approach to cluster the website users into different groups and produces common user profiles. These profiles used to provide

recommendations based on the users navigation behavior. Fuzzy Clustering technique is applied to cluster the session categories after identifying the session identification. Each session cluster represents the similar browsing patterns and finds similar interest. Their proposed approach can successfully identify user profiles from web usage data.

Ajith Abraham [4] introduced a novel approach called intelligent miner (i-miner) which discovers web data clusters by using Fuzzy Clustering algorithm. The resultant cluster be analysed by using Takagi-sugeno fuzzy inference system inorder to identify the trend of website visitors. The chromosome of the i-miner framework has three layers, the first layer represents the optimal number of clusters and initial cluster centers. Second layer is responsible for the total number of rules. Final year is accountable for the selection of optimal learning parameters.

Tomas Chovanak et al.[5] proposed a novel method called HyPBMine to identify the behavioural pattern of the users while browsing. The method which identifies behavioral of all browsers called global patterns and these patterns dynamically identified group of similar users called group patterns by using Clustering and it recommended the next user actions within the selected sessions. The proposed method is applied data from e-learning and News domains. The performance of their work reaches higher precision.

Pawan et. al [6] specified the need of analyze the data from web log files. In this paper, fuzzy c-means clustering algorithm is applied to educational sites to

group the web visitors. This algorithm proved the ability to distinguish different characteristics of the users with similar interest efficiently.

In this literature study, it evidently defined that fuzzy clustering is suitable to find the user behavioral pattern in web usage mining. The user's interest might be in overlapping clusters. Clusters of similar user's pattern can help the website owners to provide web personalization, to recommend web pages, to enhance the marketing strategy and there by improve their web structure.

### 3. METHODOLOGY

In this section, the proposed methodology explains about the working of fuzzy clustering as in Fig 3.1 through preprocessed data with weighted session page view matrix to improve the accuracy of clustering in web usage mining. The information available in the web is heterogeneous and unstructured. Therefore, the preprocessing phase is a prerequisite for discovering patterns. The primary goal of preprocessing is to transform the click stream data into set of aggregate user profiles.

#### 3.1 Log Files

A log file is a file which records the browsing information of the user. Log files are also important to keeping track of applications that have little to human interaction, such as server applications. Log files analysis is generally performed by some kind of computer program that makes the log file information more concise and readable format.

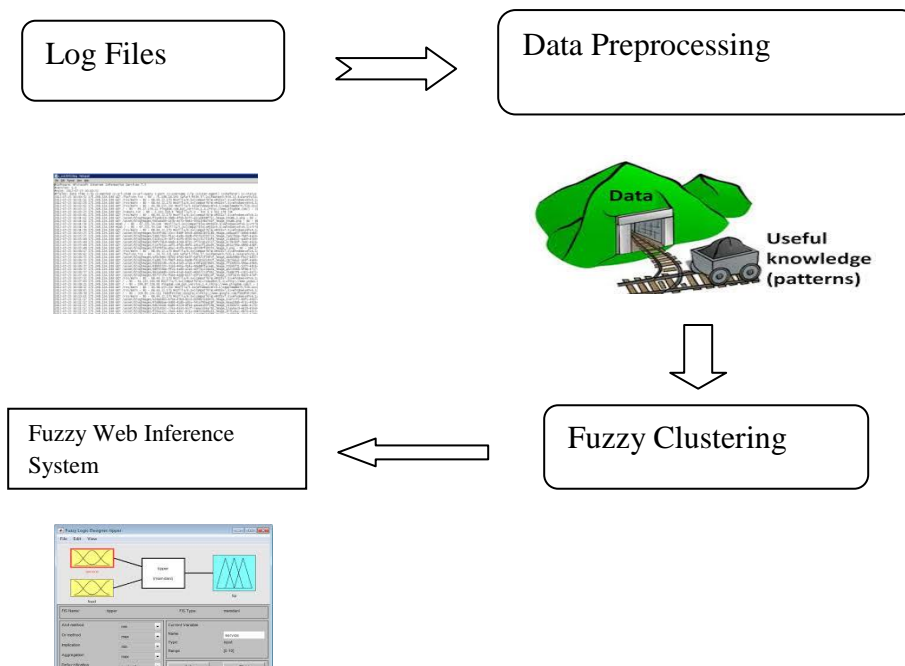


Fig 3.1 : Methodology for Fuzzy Clustering in Web Usage Mining

### 3.2 Data Preprocessing

Data preprocessing is the crucial step in any of the data mining techniques. Each webpage in the website can be mapped into numerals. Clustering can't be applied since it contains many empty cells. User-Session matrix is constructed based on the number of pages in the website. Each row represents the session of the user and the user session matrix is constructed by assigning weights to each page request. If the page is not visited, then assign zero. Each row represents a session and the

pages specify the request given by the user. Consider a session and the weights assigned for each page. The weight can be assigned by using the formula

$$W_i = \frac{m}{\sum_{i=1}^n np}$$

Where m is the frequency of the each page and np be the total number of page requests in a session and then construct the user session matrix to perform normalization of the data. The algorithm for Weighted Session Page View Matrix can be constructed as follows:

Algorithm: Weighted Session Page View Matrix

Input : User Session Page Request

Output: Weighted Matrix

Step:1 Create Column for each page in the website

Step:2 For each Session

Begin

Calculate the frequency

Calculate the sum of weights for each session

Calculate weight  $W_i$

$$W_i = \frac{m}{\sum_{i=1}^n np}$$

End

The user session matrix can be presented as

$$US_i = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{m1} & W_{m2} & \dots & W_{mn} \end{bmatrix}$$

The session can be given as : 3 4 4 1 1 4 2 1 3

The session represents the user's sequence of page visit and the Table 3.1 presents the weight assigned to each page request in the given session.

Table 3.1 : Weight assigned for session

Pages	1	2	3	4	5
Frequency	3	1	2	2	0
Weight	0.33	0.11	0.22	0.33	0

### 3.3 Fuzzy Clustering

One of the unsupervised learning, fuzzy clustering technique was developed by Dunn and improved by Bezdek[7]. Fuzzy partitions and prototypes for any set of numerical data can be generated by this

algorithm. The objective of this algorithm is to find c fuzzy groups of the collection vectors  $X_i$ ,  $i=1,2,\dots,n$  and find the cluster center in each group. The fuzzy partition equation of the cost function of dissimilarity matrix is reduced. Each cluster is represented by the vector V. The objective function of FCM is given by:

$$\sum_{i=1}^n \mu_{ik}^m d_{ik}^2$$

U represents the membership function and  $\mu_{ik}$  is the elements of U. V is the cluster center of the vector,  $d_{ik}$  represents the distance between the data points and the cluster center. m is the fuzziness parameter. The pseudo code of the FCM algorithm is as follows:

Step: 1 Input the cluster center and randomly selects the initial cluster center, V,  $m=2$ .

Step :2 Calculate cluster center for each step

Step:3 Update the membership matrix as Q(k), Q(k+1)

Step: 4 Check the new membership matrix with previous value.

$$Q(U, V) =$$

If  $(Q(k) - Q(k+1)) < \epsilon$  then Stop, Otherwise goto Step 3

#### 4. EXPERIMENTAL DESIGN

The weblog files of the website (msnbc.com) have been used for this experimental purpose [8]. Each page in the website is represented as an integer value for the smooth processing of data. The msn data is available in the website UCI KDD Archive at the University of California. The log file contains the page visit by the user on September 28<sup>th</sup> 1999.

Table 4.1 : Weighted Session Page View Matrix

Front page	1
News	2
Tech	3
Local	4
Opinion	5
On-air	6
Misc	7
Weather	8
Health	9
Living	10
Business	11
Sports	12
Summary	13
Bbs	14
Travel	15
msn-news	16
msn-sports	17



3	2	2	4	2	2	2	3	3	
6	7	7	7	6	6	8	8	8	8
6	9	4	4	4	10	3	10	5	10
1	1	11	1	1	1				
8	8	8	8	8					

The index table of the MSN dataset is represented in in table format. In Fig 4.1, the first table denotes the mapping of each page and each page in the website is represented as numeric and each row in the second table represents the sequence of pages visited by the user. For example, third row represents that the user visit tech page and then sports page twice and so on. Fuzzy Clustering can't be applied with this dataset because the dataset needs normalization.

Front Page	News	Tech	Local	Opinion	On-air	Misc	Weather	Health	Living	Business	Sports	Summary	BBs	Travel	msn-news	msn-sports
0	0.55	0.33	1.99	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.3	0.3	0.4	0	0	0	0	0	0	0	0	0
0	0	0.1	0.3	0.1	0.1	0	0	0.1	0.3	0	0	0	0	0	0	0
0.83	0	0	0	0	0	0	0	0	0	0.16	0	0	0	0	0	0

Fig 4.1 : Dataset of MSNBC.com

User session matrix is constructed by consider each page as a separate column. The weight of each column can be calculated by the frequency of the page visit divided by

the total number of pages in each session. The weighted Session page view matrix is represented in the table 4.1.

The step by step procedure for fuzzy clustering is explained clearly with sample data points.

Data Points

1	2
2	4
3	2
6	7
6	9

Initial Cluster Center

1	1	1	0	0
0	0	0	1	1

The sample data points for each row represent the sequence of visit done by the user. For example, the first user visited *frontpage* and then *News*. Second user

1	1	1	0	0
0	0	0	1	1

visited *News* & then *local* page and so on. The initialization of initial cluster center is taken as follows.

This cluster center represents first three points belong to Cluster 1 and the 4<sup>th</sup> & 5<sup>th</sup> datapoint belong to Cluster 2.

### Step : 1

Cluster points to be calculated by using the formula as:

$$V_{ij} = \sum_{k=1}^n \mu_{ik}^2 \cdot x_{kj} / \sum_{k=1}^n \mu_{ik}^2$$

$$V_{ij} = \mu_1^2 \cdot x_{1j} + \mu_2^2 x_{2j} + \mu_3^2 x_{3j} + \mu_4^2 x_{4j} + \mu_5^2 x_{5j} / (\mu_1^2 + \mu_2^2 + \mu_3^2 + \mu_4^2 + \mu_5^2)$$

$$V_{11} = 1^2 \cdot (1) + 1^2 \cdot (2) + 1^2 \cdot (3) + 0 \cdot (6) + 0 \cdot (6) / 1^2 + 1^2 + 1^2 + 0 + 0 = 2$$

$$V_{12} = 1^2 \cdot (2) + 1^2 \cdot (4) + 1^2 \cdot (2) + 0^2 \cdot (7) + 0^2 \cdot (9) / 1^2 + 1^2 + 1^2 + 0^2 + 0^2 = 2.66$$

$$V_{21} = 0^2 \cdot (1) + 0^2 \cdot (2) + 0^2 \cdot (3) + 1^2 \cdot (6) + 1^2 \cdot (6) / 0^2 + 0^2 + 0^2 + 1^2 + 1^2 = 6$$

$$V_{22} = 0^2 \cdot (2) + 0^2 \cdot (4) + 0^2 \cdot (2) + 1^2 \cdot (7) + 1^2 \cdot (9) / 0^2 + 0^2 + 0^2 + 1^2 + 1^2 = 8$$

$$C_1 = (2, 2.66) \text{ and } C_2 = (6, 8)$$

The distance can be calculated by using the formula

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

#### Centroid 1 :

$$(1, 2) (2, 2.66) D_{11} = \sqrt{(2-1)^2 + (2.66-2)^2} = 1.20$$

$$(2, 4) (2, 2.66) D_{12} = \sqrt{(2-2)^2 + (2.66-4)^2} = 1.16$$

$$(3, 2) (2, 2.66) D_{13} = \sqrt{(2-3)^2 + (2.66-2)^2} = 1.20$$

$$(6, 7) (2, 2.66) D_{14} = \sqrt{(2-6)^2 + (2.66-7)^2} = 4.51$$

$$(6, 9) (2, 2.66) D_{15} = \sqrt{(2-6)^2 + (2.66-9)^2} = 7.49$$

#### Centroid 2:

$$(1, 2) (6, 8) = \sqrt{(6-1)^2 + (8-2)^2} = 7.81$$

$$(2, 4) (6, 8) = \sqrt{(6-2)^2 + (8-4)^2} = 5.65$$

$$(3, 2) (6, 8) = \sqrt{(6-3)^2 + (8-2)^2} = 6.7$$

$$(6, 7) (6, 8) = \sqrt{(6-6)^2 + (8-7)^2} = 1$$

$$(6, 9) (6, 8) = \sqrt{(6-6)^2 + (8-9)^2} = 1$$

$$\begin{aligned}\mu_{11} &= (d_{11}/d_{11})^{1/(2-1)} / (1/d_{11})^{1/2-1} + (1/d_{21})^{1/2-1} \\ &= (1/1.20) / (1/1.20) + (1/7.81) = 0.86 \\ \mu_{12} &= (1/d_{12}) / (1/d_{12}) + (1/d_{22}) \\ &= (1/1.16) / (1/1.6) + (1/5.65) = 0.833 \\ \mu_{13} &= (1/d_{13}) / (1/d_{13}) + (1/d_{23}) \\ &= (1/1.20) / (1/1.20) + (1/6.7) = 0.85 \\ \mu_{14} &= (1/d_{14}) / (1/d_{14}) + (1/d_{24}) \\ &= (1/4.51) / (1/4.51) + (1/1) = 0.18 \\ \mu_{15} &= (1/d_{14}) / (1/d_{14}) + (1/d_{24}) \\ &= (1/7.49) / (1/7.49) + (1/1) = 0.097\end{aligned}$$

$$\begin{aligned}\mu_{21} &= (1/d_{21}) / (1/d_{21}) + (1/d_{22}) \\ &= (1/7.81) / (1/7.81) + (1/1.20) = 0.12 \\ \mu_{22} &= (1/d_{22}) / (1/d_{12}) + (1/d_{22}) \\ &= (1/5.65) / (1/1.16) + (1/1.565) = 0.16 \\ \mu_{23} &= (1/d_{23}) / (1/d_{13}) + (1/d_{23}) \\ &= (1/6.7) / (1/1.20) + (1/1.67) = 0.15 \\ \mu_{24} &= (1/d_{24}) / (1/d_{14}) + (1/d_{24}) \\ &= (1/1) / (1/4.51) + (1/1) = 0.81 \\ \mu_{25} &= (1/d_{25}) / (1/d_{15}) + (1/d_{25}) \\ &= (1/1) / (1/7.49) + (1/1) = 0.88\end{aligned}$$

$$Q_{21} = \begin{bmatrix} 0.86 & 0.83 & 0.85 & 0.18 & 0.097 \\ 0.12 & 0.16 & 0.15 & 0.81 & 0.88 \end{bmatrix}$$

Repeat this process until we get the convergence.

## 5. RESULTS AND DISCUSSION

The dataset is taken from the UCI repository in the website msnbc.com on September 28, 1999. It consists of 17 categories. Fuzzy clustering is applied

only after preprocessing the dataset with weighted session pageview matrix and the output of the clustered membership matrix grouped as five clusters and it is shown in Fig 5.2.

```
0.64304159 0.04194270 0.27196380 0.015190636 0.02786127
0.56033455 0.11570984 0.21190907 0.037842438 0.07420411
0.22208283 0.46310719 0.11535956 0.054953081 0.14449734
0.00000000 0.00000000 1.00000000 0.000000000 0.00000000
0.00000000 1.00000000 0.00000000 0.000000000 0.00000000
0.10509466 0.03099497 0.82732529 0.013449552 0.02313553
0.00000000 1.00000000 0.00000000 0.000000000 0.00000000
```

Fig 5.1: Five Cluster resultant membership matrix

In this membership matrix, each row represented as each user visit. Each column value in that row represented as the probability of being in that cluster. The user group

can be identified by taken a sample of 100 records and find the clustered group. The Fuzzy web inference rule can be generated by analyzing the cluster.

**Fuzzy Web Inference**  
 If (2,3,4,9) THEN UG1  
 If (6,7,8) THEN UG2  
 If (1,12) THEN UG3  
 If (13,14) THEN UG4  
 If(1,4,7,8) THEN UG5

Fuzzy inference rule can be mapped to the given dataset is represented in Table 5.1

Table 5.1: Fuzzy inference rule results

User Group1	News	Tech	Local	Health
User Group 2	On-air	Misc	Weather	
User Group 3	Frontpage	Sports		
User Group 4	Summary	BBs		
User Group 5	Weather	Frontpage	Misc	

### Validating Clusters

The Cluster will be validated through some internal measures called Silhouette and Dunn Index. The silhouette width measures the degree of confidence in the clustering assignment of particular observation and the dunn index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-

cluster distance. The results are validated before and after the preprocessed dataset. It is clearly evident that, both the internal measures such as Silhouette and Dunn Index are high if it used the preprocessed data. In Silhouette width, the well clustered observations having value near to one and the preprocessed algorithm used in this methodology which is very close to one ie.0.88.

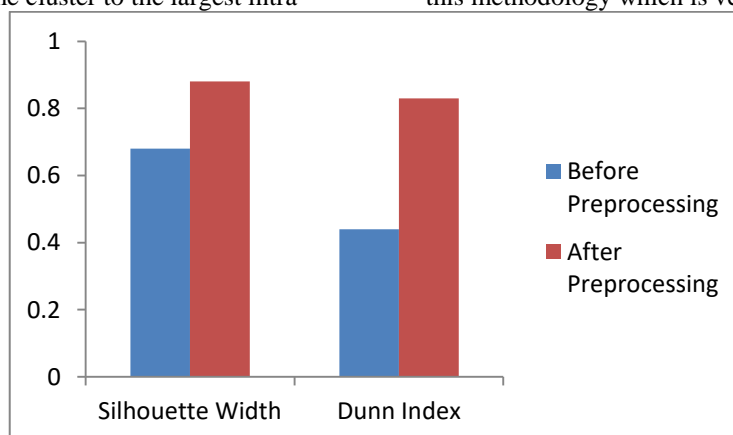


Fig 5.1 : Comparison of cluster measures

## 6. CONCLUSION

In this paper, the web browsers of similar behavior pattern be identified with the proposed methodology by constructing the Weighted Session Page view matrix. This approach improves the accuracy of the clustering to identify the similar users behavioral pattern. The experimental result proved that the clustering result improved. The results be validated through the internal measures of clustering techniques. In future, the various fuzzy clustering techniques will be used in order to improve the noise detection.

## REFERENCES:

- [1] <https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates>
- [2] James Bezdek, Robert Ehrich, Willam Full, "The fuzzy c-means clustering algorithm", Computers and Geosciences, Vol 10, Pg: 191 – 203
- [3] K. Suresh, R. Madana Mohana, A. Rama MohanReddy, A. Subrmayam, "Improved FCM

algorithm for clustering on Web Usage Mining", IEEE,2001

- [4] G. Castelano, Fanelli Mand Torsello, "Mining usage profiles from access data using fuzzy clustering", proceedings of the 6<sup>th</sup> WSEAS International conference on Simulation, Modelling and Optimization, Lisbon, Portugal, September 22-24, 2006.

- [5] Ajith Abraham, "Business Intelligence from Web Usage Mining", Journal of Information & Knowledge Management", Volume 2, No. 4(2003) 375 – 390

- [6] Tomas Chovanak, Ondrej Kassak, MichalKompan, Maria Bielikova, "Fast streaming Behavioural Pattern Mining", New Generation Computing, Ohmsha, Ltd. And Springer Japan KK, part of Springer Nature 2018

- [7] Pawan Lingras, Rui Yan and Chad West, "Fuzzy C-means Clustering of Web Users for Educational Sites", Springer – Verlag Berlin Heidelberg 2003, LNAI 2671, pp. 557-562, 2003.

- [8] James C.Bezdek, Rober Ehrlich, William Full, "FCM: The fuzzy c-means clustering algorithm",

Elsevier, Computers & Geosciences Vol. 10, No2-3,  
pp.191- 203,1984.  
[9][https://archive.ics.uci.edu/ml/datasets/msnbc.com+an  
onymous+web+data](https://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data)