International Journal of Engineering & Technology, 7 (4.17) (2018) 10-13

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



An Effective Model for Analysis Web User's Behavior from Web Log Data

Jayanti Mehra[1]*, Dr R.S. Thakur[2]

[1]*Maulana Azad National Institute of Technology, Bhopal*
[2]*Maulana Azad National Institute of Technology, Bhopal*
**Corresponding author Email: mehra.jayanti109@gmail.com*

Abstract

The regular explosion of e-Commerce, there is strong competition amongst companies and other sectors to be a focus for the customers. Web server analysis is very difficult to find out the web user behavior for any organization. It is useful for future web site improvement and design. In this paper, has to propose a PDFCM-based approach for assigning weights to client sessions for considering the high dimensionality of client session information. To analyzing the client session cluster, we use fuzzy c-means (FCM) algorithm. A major challenge for these methods is selection of suitable cluster center, so it has to propose PDFCM-based algorithm to solve this problem. Clustering is also used to approximation the number of clusters this clustering is similarly used to calculate the number of clusters. Our outcomes demonstrate that the quality of the clusters framed utilizing the proposed algorithm is much superior. So, outcome shows our proposed methodology is much batter then the other algorithms.

Keywords: *Fuzzy cluster validation, Probability Density Function, t location fit, User session clustering.*

1. Introduction

The World Wide Web is a vast source of information and important resource for data mining applications [1]. Web usage mining refers the automated use and analysis of us-age patterns supported on data gathered from user interactions with web resources on one or multiple web sites. [2][3] Clustering algorithms widely use to determine user session clusters that signify similar URL access patterns [5][6]

FCM method is their compassion to the preliminary as-sortment of the cluster centers one of the most important problems related with the FCM method [12] [30]. Hence, approximating proper values for the preliminary cluster centers is a major challenge connected with these methods [4]. in this examination, it proposes the use of a PDFCM-based approach to find a proper set of initial cluster centers.

PDFCM-based clustering is also used to approximation suited values for the number of clusters parameter c. A novel user session clustering framework is proposed, which incorporate the fuzzy dynamic cluster center initialization scheme with the FCM algorithm. The proposed PDFCM-based fuzzy clustering framework for discovering web user session clusters, which aims to develop the power of the clusters selection and give a point by point representation of the PDFCM-based fuzzy user session clustering system and its principal statistical model. It has represented to the PDFCM-based

clustering algorithm for approximation a correct and incentive for the number of clusters. In Section 2. This clarify the validity indexes as a part of quantitative examinations of the nature of the found fuzzy clusters.

2. Literature Review

S K Dwivedi et al. [32] have carried out different types of data preprocessing techniques to convert raw data into suitable format. Z Ansari et al. [7] They proposed MDF-based FCM (MDFCM) and FCMed (MDFCMed) algorithms and compare different validity index with FCM and FCMed algorithm. B. Maheswari and P. Sumathi [8] have investigated about the achievement of preprocessing and clustering of web log. Zhou Jiadi and Geng Hai [2] have described user identification and recommended a different user identification algorithm, from one perspective, utilize IP address and user access time to distinguish diverse users in the logs, in particular, heuristic rule-based user identification algorithm. A. Gupta, A. Khandekar [9] Investigated two algorithm named fuzzy c means clustering and adaptive fuzzy clustering and also described advantages of those algorithms [19]. A kumar et al. [3] Proposed a model for predict user behavior from the web log data and also discussed different techniques used in data mining and described how to apply these techniques in web log analysis. N. Anand and S. Hilal [4] proposed a way to deal with user access pattern to design from web log data. In the primary stage, the server raw log data is preprocessed. In second stage examination is performed to identify access pattern of users. V. Anitha and P. Isakki Devi [11] they predict the user behavior users through the web server log files. Users using website pages, a continuous access ways and incessant access pages, links are store in web server log files. A Weblog alongside the independence of the client captures their browsing behavior on a site and talking about with respect to the behavior from investigation of various algorithms and distinctive techniques. G. Neelima and S. Rodda [5] incorporated procedure of three phases and actualizing these three stages. Contingent on the recurrence of users visiting by each page min-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ing is performed. By finding the session of the client and investigated the user behavior when time spent on a specific page. Dimitrios Koutsoukos et al. [12] Described session identification algorithm and fuzzy c means clustering using web log data. Li Chaofeng [13] has described web session cluster, data preprocessing technique and identified the session cluster. Next, they talked about the procedure of Web session clustering and also gives a few rules in each period of data preprocessing so as to plan and actualize them naturally. They reduce the log file estimate as well as increment the nature of the accessible data. H Gulat et al [23] Investigated Indian crime records demonstrate that the proposed strategy normally does better than the current procedures in clustering of such multivariate time series data. A Zahid el. [25] In this paper, different data preprocessing techniques have described here also described feature subset selection through session vector and session weight assignment.

3. Proposed PDFCM- based Fuzzy C-Means Clustering Algorithm

It is another algorithm named probability density based fuzzy c-means clustering algorithm (PDFCM) pro-posed keep in mind the two weaknesses of the FCM algorithm initially, it requires the number of clusters c and the initial membership matrix to be indicated as from the earlier, and also, it is exceedingly delicate to the determination of the two parameters. These two weaknesses make FCM difficult to decide the best number of clusters and the result of FCM is unstable. The determination of appropriate initial cluster is a noteworthy challenge for these techniques, so this utilize PDFCM algorithm to describe the number of clusters. The outcome clearly shows the proposed algorithm is vastly improved as far as different index measures compared and the FCM. Determining the web user session clusters in web usage data aimed to develop the quality of the clusters extracted to get improved performance and quality of PDFCM- based user session clustering in conditions of the index function indices by the assignment of fuzzy weights to user sessions and URL items. The complete framework of proposed PDFM is shown in Fig.1.

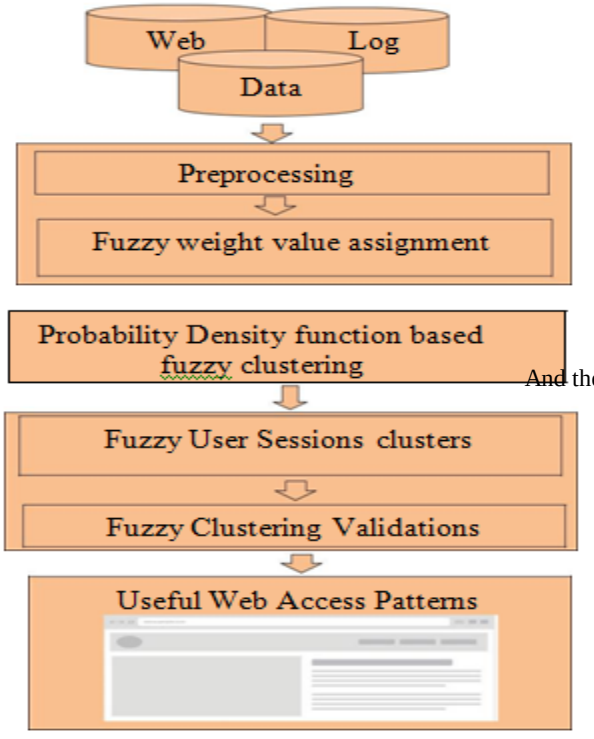

Figure 1: Framework of Probability Density function based fuzzy session clustering

4. Probability Density Function (PDF):

PDF. Let $S = \{s1, s2, \ldots, sm\}$ be the set of m user sessions in an n-dimensional Euclidean space. In order to identify the first initial cluster center, each user session si is treated as a potential candidate and the first PDF value for user session si , which is denoted as $P1(si)$, is computed using:

$$P1(si) = \sum_{k=1}^{m} \exp(-E^2(si,sk)/R^2 \quad \text{for all } i=1\ldots\ldots m \qquad (1)$$

Neighborhood radius R was set to $\sqrt{n}$ (here n is number of URLs)

E 2(si, sk) Euclidean distance between session si and center sk

Where R is a positive fixed that characterizes the neighbor-hood of the client session si . The PDF estimation of the client session si is a estimate of the density of all the client sessions in the neighborhood of si . Client sessions outside the radial distance have little effect on its PDF value. The client session with the uppermost PDF value is picked as the main cluster focus v1 as takes after.

$$i1 \leftarrow \operatorname{argmax}_{i=1}^{m} \{P1(si)\}; \; v1 \leftarrow {}^{s}i1 \qquad (2)$$

For selecting second cluster center reduce the PDF Value. PDF around the first cluster center $v1$, as follows.

$$P2(si \leftarrow P1(si) - P1(v1) \exp\left(-\frac{E2(si,v1)}{R2}\right) \quad \text{for all } i=1\ldots\ldots m \qquad (3)$$

Second cluster center is selected which is the highest PDF value as follows

$$i2 \leftarrow \operatorname{argmax}_{i=1}^{M} \{P1(si)\}; \; v2 \leftarrow {}^{s}i2 \qquad (4)$$

Also, to pick the j th cluster center, the PDF value for each user session is revised as follows:

$$Pj(si \leftarrow Pj\text{-}1(si) - Pj\text{-}1(vj\text{-}1) \exp\left(-\frac{E2(si,vj-1)}{R2}\right) \quad \text{for all } i=1\ldots\ldots m \qquad (5)$$

And the jth cluster center vj is selected is following

$$ij \leftarrow \operatorname{argmax}_{i=1}^{m} \{Pj(si)\}; \; vj \leftarrow sij \qquad (6)$$

Algorithm for PDFCM

Input: neighborhood radius R $(\sqrt{n})$,c, maximum iterations η(100), error threshold (0.01),and set of user sessions $S = \{s1, \cdots, sm\}$

Output: Set of c cluster centers $V = \{v1, \cdots, vc\}$ and partition matrix P

Initialize the set of cluster centers $V(0)$

for $i \leftarrow 1$, m do

Calculate the Probability Density values *P1(si)* using (1)
end for
Calculate the first cluster center *v1(0)* using (2)
for *j* ← 2, *c* do
for *i* ← 1, *m* do
Calculate the revised probability values *Pj (si)* us-ing (5)
end for
Calculate the *j* th cluster center *vj (0)* using (6)
end for
t ← 1
repeat
Calculate the partition matrix *P(t)* entries:
for *i* ← 1, *m* do
for *j* ← 1, *c* do
Calculate μij *(t)*
end for
end for
Calculate the set of new cluster centers *V (t)*:
for *j* ← 1, *c* do
Compute *vj (t)*
end for
Calculate the objective function JFCM (t)
t ← t + 1
until |JFCM (t) − JFCM (t − 1)| < t = η

5. Experimental Results

The performance evaluations of proposed approaches are done using standard database taken from NASA-HTTP [30]
The sample of web log data set is shown in Table 1. The proposed approaches are developed in Windows7 environment using MATLAB (version 8.3) The Evaluation Index such as Partition Coefficient, Classification Entropy Index, Partition Index, Separation Index, Xie and Beni Index, Dunn Index and Alternative Dunn Index are used to evaluate the performance of proposed approaches.
The probability density based fuzzy c means algorithm is implemented in MATLAB for session clustering in web log data. The Fig.2 shows clustering results. T location fit values shows for set of records in Table 1 and different parameters like mean, variance, std. err. etc. values shown in Table 2.

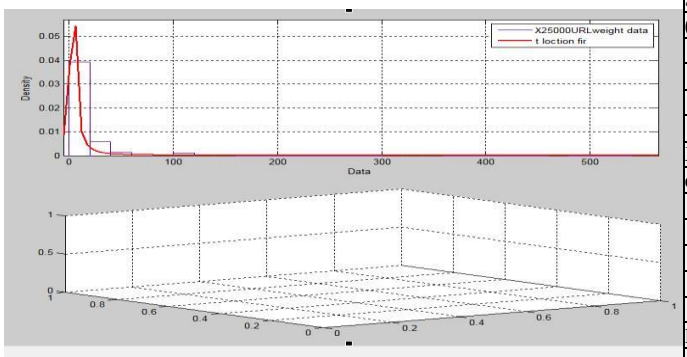

Figure 2: Graph of PDFCM-based Clustering

Table 1: T location fit value Vs. Number of records calculated by PDFCM

X(records)	t location fit f(X)
0	0.0354593003793722
56	0.000491318882463387
112	0.000148823218694935
168	7.50199755153678e-05
224	4.63137459682924e-05
280	3.19082367371482e-05
336	2.35525675162622e-05
392	1.82283066662835e-05
448	1.46037992493204e-05
504	1.20122514037579e-05
560	1.00876561236190e-05

Table 2: Different parameter values for PDFCM

Domain:	-Inf < y < Inf	
Mean:	NaN	
Variance:	Inf	
Parameter	Estimate	Std. Err.
Mu	3.48962	0.30896
Sigma	2.1834	0.354393
Nu	0.646433	0.0744372
mu	sigma	nu
mu	0.0954563 0.0721756 0.0121692	
sigma 0.0721756	0.125594 0.0165231	
nu	0.0121692 0.0165231 0.0055409	

Different values used by PDFCM based clustering
Log likelihood: -865.891
Estimated covariance of parameter estimates:
Validity measures are very prevalent for the validation of hard or crisp clusters; however, they are not appropriate validation of overlapping clusters with fuzzy cluster partitions [28]. Table 3 shows values of measures for PDFCM.

Table 3: Cluster Validity measures were computed for PDFCM

Evaluation index	Clusters	
		PDFCM Value
Partition Coeffi-	10	0.839116
cient (PC)	20	0.877896
	30	0.838533
	40	0.85019
	50	NaN
Classification En-	10	0.330371
tropy index(CE)	20	0.267376
	30	0.252924
	40	0.259667
	50	NaN
SC(partition index)	10	0.086428
	20	0.019806
	30	1.75E-06
	40	3.49E-06
	50	NaN
Separation index	10	0.002709
(S)	20	5.74E-04
	30	5.15E-08
	40	8.49E-08
	50	Nan
Xie and Beni in-	10	2.200598
dex(XB)	20	1.815612
	30	Inf
	40	Inf
	50	Inf
	10	0.25
	20	0.301511
Dunn Index(DI)	30	NaN
	40	NaN
		NaN
	50	NaN
Alternative Dunn	10	3.43E-04
Index(ADI)	20	0.00E+00
	30	NaN
	40	NaN
	50	NaN

6. Conclusion

Web log mining is one of the current areas of research in Data mining. Web Usage Mining turns into an important aspect in the present time on the basis that the quantity of information is every time expanding. To improve minimization of clustering execution

file this utilizes fuzzy dynamic approach for it. PDFCM algorithm is one of the algorithms that short out two short coming of FCM algorithm first selection of initial cluster center and second estimate the number of clusters. Its construction the validity index better to without weight task. The task of fuzzy weights to client sessions and URL items enhanced the execution and quality of PDFCM-based client session clustering regarding the different indexes. This demonstrates fuzzy weight task reduced the unfriendly impacts of insignificant client sessions and URLs, which given about the arrangement of enhanced quality groups.

References:

[1] TT Aye, "Web log cleaning for mining of web us-age patterns."In: Proc. of 3rd International Conference on Computer Research and Development (ICCRD) IEEE, vol. 2, pp. 490-494, 2011.

[2] Z Jiadi and G Hai, "Research on User Identification Algorithm based on Rewriting URL.", International Journal of Security and Its Applications ,Vol.10(3),pp. 215-222,2016.

[3] A Kumar, V Ahirwar, RK Singh, "A Study on Pre-diction of User Behavior Based on Web Server
Log Files in Web Usage Mining', International Journal Of Engineering And Computer Science, Vol. 6 (2), pp. 20233-20236,2015.

[4] N Anand and S Hilal, "Identifying the User Access Pattern in Web Log Data", International Journal of Computer Science and Information Technologies, Vol.3 (2), pp. 3536-3539, 2012.

[5] Neelima and S Rodda, "Predicting user behavior through sessions using the web log mining." In: Proc. of International Conference on Advances in Human Machine Interaction (HMI), 2016 pp. 1-5. IEEE, 2016.

[6] T Vandana and M. K. Rao, "Linking Emotional In-telligence to Knowledge Sharing Behaviour: Or-ganizational Justice and Work Engagement as Me- diators", Global Business Review, Vol.18 (6), pp.1580-1596, 2017.

[7] Z Ansari, SA Sattar, A.V. Babu, and M.F. Azeem, "Mountain density-based fuzzy approach for discovering web usage clusters from web log data, Fuzzy Sets and Systems", Vol.279(1), pp.40-63,2015.

[8] B. Maheswari and P. Sumathi, "A New Clustering and Preprocessing for web log mining." In: Proc. of World Congress on Computing and Communication Technologies (WCCCT) IEEE, pp. 25-29, 2014.

[9] A Gupta and A Khandekar, "Development of Web Log Mining Based On Improved Fuzzy C-Means Clustering Algorithm," International Journal of Science, Engineering and Technology Research (IJSETR), Vol.5 (3), March 2016.

[10] N Pushpalatha and SSS Reddy, "Towards an extensible web usage mining framework for actionable knowledge", In: Proc of International Conference on Inventive Communication and Computational Technologies (ICICCT) IEEE, pp. 35-40, 2017.

[11] V Anitha and P. Isakki, "A survey on predicting user behavior based on web server log files in a web usage mining", In: Proc. of International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE) IEEE, pp. 1-4, 2016.

[12] D Koutsoukos, G Alexandridis, G Siolas, and A Stafylopatis, "A new approach to session identifi-cation by applying fuzzy c-means clustering on web logs", In: Proc. of Symposium Series on In Computational Intelligence (SSCI) IEEE, pp. 1-8, 2016.

[13] L Chaofeng, "Research on web session clustering", Journal of Software, Vol. 4(5), pp. 460-468, 2009.

[14] A Zahid, AV Babuy, W Ahmed, and M F Azeemz, "A fuzzy set theoretic approach to discover user sessions from web navigational data", Recent Ad-vances in Intelligent Computational Systems (RA-ICS) IEEE, pp. 879-884, 2011.

[15] B Chandra, M Gupta, and MP Gupta, "A multivariate time series clustering approach for crime trends prediction", In: Proc of International Con-ference on Systems, Man and Cybernetics, SMC IEEE., pp. 892-896, 2008.

[16] MS Bhuvaneswari and K Muneeswaran, "Extracting usage patterns from web server log", In: Proc. of 2nd International Conference on Green High Performance Computing (ICGHPC) IEEE, pp.1-7, 2016.

[17] P Sampath, and M. Prabhavathy, "Web page access prediction using fuzzy clustering by local approximation memberships (flame) algorithm",ARPN Journal of Engineering and Applied Sciences", Vol. (10) 7, ISSN 1819-6608, 2015.

[18] NP Gopalan, and J. Akilandeswari, "A distributed, fault-tolerant multi-agent web mining system for scalable web search",In: Proc. of WSEAS 5th In-ternational conference on Applied Informatics and Communications, pp. 15-7, 2005.

[19] K Venkateswarlu, K Muni, A. Kandasamy, and K. Chandrasekaran. "An Energy-Efficient Hybrid Clustering Mechanism for Wireless Sensor Net-work," Unmanned Systems, Vol.3 (2), pp.109-125, 2015.

[20] SK Jauha and M Pant, "Recent trends in supply chain management: A soft computing approach", In: Proc. of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012) Springer, pp. 465-478, 2013.

[21] C. Saravanan ,S Mishra , VK Dwivedi, and K. K. Pathak, "Discovering flood recession pattern inhydrological time series data mining during the post monsoon period", International Journal of Computer Applications, Vol. 90(8), 2014.

[22] M Srivastava, R Garg, and P. K. Mishra, "Analysis of data extraction and data cleaning in Web usage mining", In: Proc .of International Conference on Advanced Research in Computer Science Engi-neering & Technology (ICARCSET 2015) ACM, pp. 13-13, 2015.

[23] H Gulat, and P. K. Singh, "Clustering techniques in data mining: A comparison", In: Proc. of 2nd International Conference on Computing for Sus-tainable Global Development (INDIACom) IEEE, pp.410-415, 2015.

[24] GS Chandel, K Patidar and MS Mali, "A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm", In: Proc of International Journal of Computer Science and Network Security (IJCSNS), Vol.16(1),2016.

[25] Ansari, Zahid, MF Azeem, AV Babu and W Ah-med, "A fuzzy approach for feature evaluation and dimensionality reduction to improve the quality of web usage mining results", International Journal on Advanced Science Engineering Information Technology, Vol.2 (6), pp. 67-73, 2012.

[26] D. P. Shrivastava, "Mining Frequent term set for Web Document Data using Genetic Algorithm", International Journal of Pharmacy and Technology, Vol. 8(2), pp. 4038-4054, 2016.

[27] L Wang and F Xiuju, "Rule extraction using a novel gradient-based method and data dimension-ality reduction", In:Proc. of International Joint Conference on Neural Networks, 2002. IJCNN'02 IEEE, Vol. 2, pp. 1275-1280, 2002.

[28] K Das, G Ivan and MJN Arani, "Relations between distance–based and degree–based topological indices", Applied Mathematics and Computation, Vol. 270 (1), pp. 142-14, 2015.

[29] Hu, Yating, Chuncheng Zuo, Yang Yang and Fuheng Qu, "A cluster validity index for fuzzy c-means clustering", In System Science, In:Proc. of International Conference on Engineering Design and Manufacturing Informatization (ICSEM) IEEE ,vol. 2, pp. 263-266 , 2011.

[30] Weblog dataset downloaded from http://ita.ee.lbl .gov/html/contrib/NASA-HTTP.html (Retrieved on July 2015).

[31] J Han and Kamber ,"Data Mining: Concepts and Techniques", 3rd edition, 2011.

[32] S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process", In: Proc. of International Conference on Green Computing and Internet of Things (ICGCIoT) IEEE, pp. 506-510, 2015.