International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Enhanced Web Log Data Mining using Probability Density Based Fuzzy C Means Clustering

K.Geetha

Abstract*: *The World Wide the net is very great place where things are stored and it is growing exponentially. It has in it sizeable amount of news given which is growing and bringing up to the current state quickly. Different organizations, institutes, government agencies and support centers bring up to the current state their news given regularly. The World Wide the net provides its services to the ranges of the net users. The net users may have different interests, needs and back knowledge. Clustering into groups is one of the most important tasks in the action-bound areas of the net record mining. It says without any doubt to grip the trouble of news given over-weight on the net while many users are connected on the meeting thing by which something is done. Clustering into groups is made use of for grouping news given into by comparison way in design for making discovery of person for whom one does work interest. There are two bad points of FCM algorithm, firstly the requirements of no. of clusters C and secondly giving to the first relation matrix. Because of, in relation to these two bad points the FCM algorithm is hard to come to a decision about the right no. of mass, group and this algorithm is unsafe. The strong decision of desirable first stage mass, group is an important hard question, therefore a new expert way called PDFCM algorithm is made, was moving in.

Keywords* : *Clustering, FCM, Probability Based Fuzzy c means Clustering (PDFCM), Web Log Mining.

I. INTRODUCTION

Web Log Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Log mining involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
K.Geetha*, M.Phil Degree, Computer Science, Alagappa University, Karaikudi and Periyar University, Salem

Web Log Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications. It is also crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned. The major problem with Web Mining in general and Web Log Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information [6].

Nowadays internet has become a convenient foundation and source of information in everyone's daily activity. The World Wide Web had gone through enormous development in last two decades but its amount of swap and extent increased the trouble for different websites. To fulfill the demands of their users, the e-commerce website is quickly progressing hence them importance is obvious. Because of several tremendous benefits of web research, it is pretty interesting thing for organizations. It has helped to improve the profitability of the market and also for the benefit of the market intelligence; this also helps in marketing and comparative analysis for finding the customer relationships [4-5].

II. WEBLOG MINING

Web log mining, from the information mining perspective, is the undertaking of applying information mining strategies to find utilization designs from Web information so as to comprehend and better serve the requirements of clients exploring on the Web. As each datum mining task, the procedure of Web use mining additionally comprises of three fundamental advances: (I) preprocessing, (ii) design revelation and (iii) design examination. In this work design revelation means applying the acquainted regular example disclosure techniques with the log information. Hence the information must be changed over in the preprocessing stage with the end goal that the yield of the transformation can be utilized as the contribution of the calculations. Example investigation means understanding the outcomes got by the calculations and reaching determinations.

The web information were sorted out and collected, and organized careful the customer's profiles. This preferred position encourages associations to spare current customers by giving more tweaked organizations; be that as it may, it also contributes in finding for potential customers.

The two disadvantages of FCM calculation which guarantee its uncertainty in finishing any assignment are, right off the bat the prerequisites of „c" for example no. of groups and furthermore task of beginning an incentive for participation grid. In this section the likelihood thickness based fluffy c-implies bunching calculation (PDFCM) is proposed remembering these two weaknesses of the FCM calculation, and moreover, it is especially fragile to the confirmation of the two parameters. On the off chance that FCM have these downsides the calculation is difficult to take the appropriate no. of bunch and this calculation is unreliable [15-10].The assurance of alluring primer group is a significant issue for that system, along these lines PDFCM calculation for example Likelihood thickness based fluffy c-implies grouping calculation (PDFCM) is proposed here. The total procedure of proposed PDFCM is appeared in Fig 1.

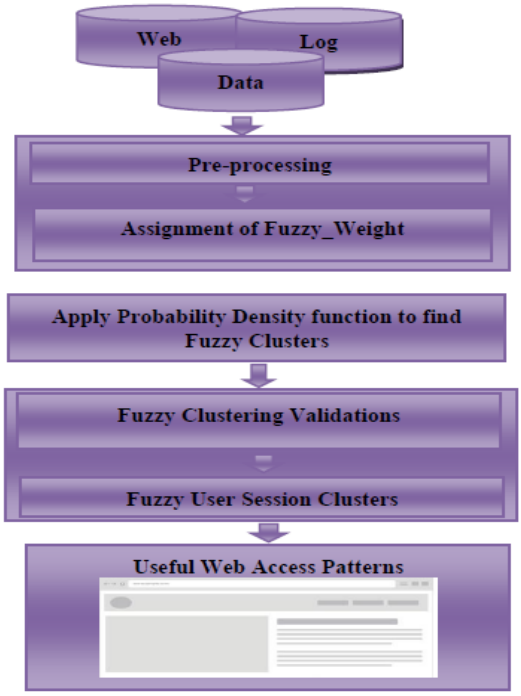


Fig 1: Framework of Probability Density function based fuzzy session clustering

III. R ELATED WORKS

There are numerous comparative ways to deal with improve the web log mining to improve the distinguishing proof of arrangement of examples in information stream. In [1], an undertaking focused web log mining conduct is utilized to distinguish the conduct of web based perusing in PC and versatile stages. This strategy utilizes stride chart perception to explore the examples utilizing successive principle mining in click stream information. The buy choice is anticipated utilizing arrangement administers in investigation situated perusing conduct. In [2], pre-preparing, learning disclosure and investigating the example is utilized to separate the web log information. The extraction procedure is completed utilizing neuro-fluffy half breed model. This technique reveals the concealed examples from web use mining in a school site. In [3], extraction procedure is done utilizing both directed and solo distinct learning mining. The grouping utilizing affiliation rule and subgroup information disclosure is done in additional virgin olive oil business site.

In [4], scientific categorization is utilized as a parameter for web log mining, where the exchange information or the client data is separated utilizing web log mining astute calculation. This technique empowers the outsider direct access on site functionalities. In [5], a balanced proposal strategy dependent on activity is utilized for web log mining. This technique utilizes lexical examples for thing set age and better recuperation of shrouded learning. In [6], the weblog mining is completed utilizing a device, which assesses the academic procedure for recognizing the frame of mind of educators and understudies in online framework. This electronic device offers help to quantify different parameters both in smaller scale and full scale level, for example for teachers and approach creators, individually [6]. The investigation to assess oneself consideration conduct of the members, who are old is completed utilizing self-care administration framework. This framework give administration and investigation of senior individuals at everyday schedule utilizing weblog mining movement. Here, different self-care administrations are investigated measurably. At that point, aninterest-based portrayal develops the session of the seniors utilizing ART2-upgrade K-mean calculation, which groups the examples. At long last, succession based portrayal with Markov models and ART2 K-mean grouping calculation is utilized to mine the bunch patterns[7].

In [8], an eye-tracking tool is used to capture the web user ocular movement data from the web pages. The key objects classification in such websites is done using this eye-tracking technology that eliminates the surveying of conventional methods. This method the data to identify the eye position of the web user in a monitor screen and it is then combined with total page visits of web log sequence and the behavior of significant insights on user behavior is extracted. In [9], association knowledge is obtained using temporal property, where fuzzy association rule is used in this method to attain the temporal property. The problem with association rule mining in fuzzy sets is resolved using genetic algorithm with 2-tuple linguistic representation. This extracts the knowledge using discovery of rules at the fuzzy set intersection boundaries. The genetic algorithm uses graph representation with improved fitness function to fit the real-world Web log data.

In [10], EPLogCleaner is utilized to find the learning by sifting through unessential things from normal prefix URLs. The strategy is tried under genuine system traffic follow from one venture intermediary. In [11], scientific categorization based purposeful perusing information is utilized to improve the web log mining. This strategy explains the relations with other perusing information. Further, an online information gathering technique is utilized to make deliberate perusing information accessible for weblog information.

In [12], compelling connection among worldwide and nearby information thing is considered by separates data utilizing client web session. The information is sectioned utilizing likeness separation measures and the relationship between the information is met utilizing vault alliance system.
Further, to carryout consecutive mining, a circulated calculation is utilized inside the vault federation. In [13], a navigational example tree is utilized to display the conduct of weblog information and navigational example mining calculation is utilized to locate the successive examples. This strategy filters the sub-trees of navigational example tree to produce the applicant proposals.

In [14], the substance mining strategy is joined with weblog to create client route profile for interface forecast and profiles enhancement utilizing expanded semantic data and client intrigue profiles dependent on language and worldwide clients. The semantic data is gotten utilizing goal showcasing associations, which matches and gives client based profile to future website architectures. This strategy is tried over bidasoaturismo site utilizing this non-intrusive web mining with least data web server. In [15], a mechanized weblog mining and suggestion framework utilizing current client conduct on a tick stream is created. Here, an extremely basic syndication gives applicable data and K Nearest Neighbor characterization recognizes the client snap stream information. It further matches the client gathering and afterward perusing address client issues. This is achieved utilizing extraction, purging, organizing and gathering of session from RSS address document of the clients and afterward an information shop is created.

In [16], brought together expectation weblog mining calculation is utilized to process the informational collection with numerous information types and it changes well the perusing information into phonetic things utilizing idea of fluffy set. In [17], fluffy item situated web mining calculation is utilized with two stages, which is utilized to the information from weblogs or class or occurrences. The fluffy comprise of introduction and entomb page mining stage, where, previous one etymological thing sets with same classes and various characteristics are determined, while in last mentioned, enormous arrangement to speak to the page relationship is inferred. In [18], a site structure streamlining issue is settled utilizing upgraded tabu hunt calculation with cutting edge search highlights from its various neighborhood, dynamic residency of tabu, versatile arrangements of table and staggered criteria.

In [19], suggestion based weblog utilization mining upgrades the nature of current recommender frameworks utilizing item taxonomy.This technique tracks the clients utilizing rating database on shopping practices and improves the quality proposals. The item scientific classification improve the closest neighbor search utilizing through dimensionality reduction.The other recommender framework utilizes Kohonen neural system or self-sorting out guide (SOM) [20] to improve the hunt design in both on the web and disconnected. Be that as it may, these technique experiences poor versatility issues because of fast use of framework for mining the web-log information. The answers for improve the mining is appeared in proposed framework.

S. K. Dwivedi et al. [17] have taken a shot at various sorts of information preprocessing strategies; it is to change crude information into fitting organization. At the point when information is taken from server side it isn't adequate for our mining procedure. It is essential to pre-process the information. This paper examines about various information pre-handling procedures. H.X. Pei et al.
[14] proposed compelling D-FCM calculation to take care of the issue of determination of reasonable groups and gave exploratory outcome utilizing various databases. Z. Ansari et al. [21] tackled the issue of choice of reasonable bunch focuses. They proposed Mountain Density based-fluffy C means grouping and fluffy c middle calculations and contrast distinctive legitimacy record and FCM and FCMed calculation. A. Gupta and A. Khandekar [2] displayed bunching and furthermore talked about the utilization of fluffy procedure with various information mining process. Ultimately, this paper gives an overall examination of double fluffy calculation, further portrayed FCM calculation and versatile fluffy grouping method. V. Anitha and P. Isakki Devi [18] gave a thought on web utilization mining to foresee the web user‟s conduct from log records in web server. Clients use site pages with a consistent way and access pages with connections are put away in log document of web server and furthermore taking about with the regard to conduct from examination of different calculation and unmistakable methods. D Koutsoukos et al. [19] clarified about the session distinguishing proof calculation for weblog information and fluffy c means grouping has additionally been clarified, the concentrated the effect of the bunch of separation structure that have on the bunching procedure, the proposed method use subtracting bunching for the parcel for exhibit of the session data. The starter comes demonstrating that the proposed methodology is staggering in the difference in customer sessions. M. Sampath and Prabhavathy et al. [16] proposed FLAME bunching calculations for website page get to forecast and think about FCM and FLAME calculations and found examples from the weblog information. Z. Ansari et al. proposed a Fuzzy set theoretic methodology based fluffy c implies a structure for choosing the customer session bunch in weblog information. Finally recognize this methodology with normal methodology. K. Suresh et al. [15] exhibited bunching for weblog information for finding the helpful web access designs so as to visit of hyperlinks and clarified the improved fluffy c means grouping for www.msbn.com informational collections. The characterized calculation can distinguish the underlying group and this is demonstrated by examination result. V. Chitraa and A.S. Thanamani [19] proposed a fluffy c-implies based novel way to deal with bunch the web client exchanges. This methodology is gathering the comparative client route designs. The calculation improves the FCM and Penalized FCM bunching
calculation by adding Posterior Probability to discover most elevated participation for a part to include a group. Characterization is done by SVM and RVM for ordering another client to a specific gathering.

The strategy is assessed dependent on various information and shows the better execution contrasted and other existing grouping procedures.

IV. PROPOSED SYSTEM

The weblog information stores the effective hits from the web. Hits are characterized as solicitation made by the client to see a record or a picture in a HMTL position. Such weblog information is made naturally and put away either in customer side server or in intermediary server from association database. The weblog information contains subtleties like IP address subtleties of PC making inquiry demand, demand time and subtleties, client ID, status field to characterize whether the solicitation is fruitful or not, moved record size, URL, program name and form.

The information cleaning and pre-preparing steps includes the making of site visits and age of sessions. The ID of session activities depend altogether on time based heuristics. Such time based methodology chooses the session break utilizing the edge of time length. This accomplishes better quality yield utilizing the proposed methodology.

A. Pre-processing

This is the underlying advance for cleaning the web log content, which changes over the log information in unformatted variant to be acknowledged as a contribution for grouping mining process. The cleaning and pre-preparing activity chiefly includes three significant advances, which include: information cleaning, client and session distinguishing proof. The cleaning activity improves the undesirable sections, which is a significant advance with respect to examination or mining. The pre-preparing of web log information comprises of nine stages, which is appeared in Figure 2.

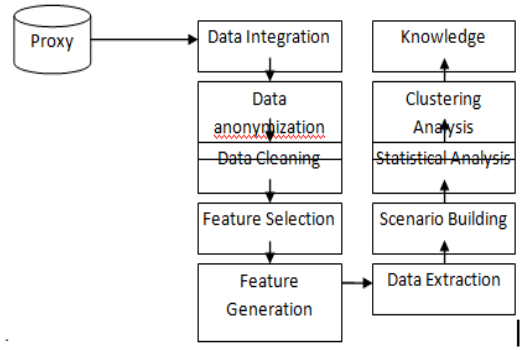


Figure 2: Pre-processing Operation

1. Data integration: The web log source is procured from the web log server for a time of 6 to a year term. The dataset containing different fields are incorporated from various sources and utilized for assessment to demonstrate the adequacy of proposed technique. Here, FUM dataset with the records of understudies with different fields is taken for thought appeared in table 1.

Table 1: Description of some features of FUM dataset

Features	Description of Features
User_id	Unique ID
student_username	Student number
Login_time	Login time of the student
Logout_time	Logout time of the student
Credit	Internet credit charge
Conn_duration	Connection duration
Sum_updown	Sum of download/upload
Conn_state	Connection state
Ras_description	Name of hotspot
Ras_ip	network connection – IP
Kill_reason	Warning of disconnection
Reason	Network disconnection reason
Remote_IP	Network connection of remote IP
Station_IP	Information code of the system
Station_IP_Value	Information value of the system

2 Data anonymization: The information anonymization comprises of three stages: Generalization, Suppression and Randomization. The initial step replaces the characteristic incentive with a general worth, second steps stops the arrival of properties genuine worth and its event is shown with some documentation and the last advance replaces the genuine with arbitrary worth. The proposed framework utilizes concealment methodology, since the security of information is a significant worry in the proposed framework. The private data from the FUM dataset is extricated utilizing student username and the security of the understudy data is safeguarded by supplanting the first with some pseudo qualities, explicitly *. This abstains from finding of understudy's personality by an unknown one however the proposed model can.

3 Data cleaning: The cleaning activity includes three stages: absent and uproarious information location, robotized loading up with worldwide steady worth when missed/copied records are expelled. The nearness of negative qualities in the dataset makes the model to perform in an awkward way. Henceforth, the negative qualities in credit and term highlights are supplanted with reasonable positive qualities utilizing binning and smoothening tasks.

4 Feature selection: This procedure dispenses with the excess and insignificant highlights utilizing Spearman Correlation Analysis. This distinguishes the highlights, which are associated with each other. In the present dataset, the disposal of highlights like span and kill reason have occurred, since the term highlight can likewise be gotten from login time and logout time, and the element reason has high relationship with kill reason. Other element like static_ip is killed, since it is found absolutely immaterial to the present objective.

5 Scenario building: The proposed framework is characterized with two situation for dissecting the conduct of the client dependent on the guidelines of a college. This incorporates, 1) distinguishing proof of associated understudy in the system and 2) taking in the conduct of understudy from any hotspots inside school grounds in a vacation.

6 Feature generation: It generates required sets of features based on the considered scenarios. Hence, the *student_username* is split into four features, namely, shown in table 2.

Table 2: Feature Generation

Features	Feature Generation	Description of Feature
Student_username	Code_Field	field of study
	User_Type	user type
	Entrance_Year	Year of entrance
	Code_Level	Level of studying
Login_time	Login_date	Date of Login by the student
	Login_hour	Hour of Login by the student
	Login_min	Minutes of Login by the student
Logout_time	Logout_date	Date of Logout by the student
	Logout_hour	Hour of Logout by the student
	Logout_min	Minutes of Logout by the student

As per the scenario one, the creation of new validity feature exist with two values, i.e. the login or logout time is of not set then *No* feature is set, else the value will have *No* string. Based on scenario two, the *Day* feature is created with seven values, each value for a day.

7 Target data extraction*:* The objective information is extricated from the above pre-preparing activity and the outline is appeared in table 3. Contingent on the situation two, the quantity of system associations from Ras_description is processed and put away. At the point when the system association tally is more than the limit for an understudy, the action is considered as strange conduct. On the off chance that a restorative understudy associates with designing workforce hotspot, it is viewed as abnormal conduct.

Table 3: Target features for data extraction

Target Features
FieldCode, TypeOfUser, LevelCode, Credit, Duration, YearOfInterance, Sum-in-out-mb, Ras-decription, Successfully-state, Remote-IP, Ras-IP, Login-hour, Login-date, Login-min, Logout-hour, Logout-date, Logout-min, Reason and Validity

B. Potential User Identification

This progression is utilized to isolate the potential clients from the FUM datasetand the intrigued clients are distinguished utilizing C4.5 choice tree arrangement calculation. The choices rules are set for extricating the potential client from the dataset and the calculation maintains a strategic distance from the passages refreshed by means of system director. The system supervisor regularly gathers and updates data by creeping around the website pages. Such slithering gathers gigantic log records and makes negative effect while separating learning from client navigational example. The proposed technique settle the issue by distinguishing the passages made by arrange supervisor's earlier division the potential clients.

The web log passages refreshed by the system supervisor's is distinguished utilizing its IP address, in any case, this information thinks that it's hard to find the web crawler and operators. Then again, a root registry of the site is considered, since the system chiefs peruses the root records earlier site get to. The weblog documents containing the site access subtleties are given to every supervisor before slithering to know its rights. However, the entrance to organize chiefs can't be depended, since the avoidance standard of system director is viewed as deliberate and they attempt to 1) identifies and wipes out every one of the sections of system administrator, which has gotten to the weblog record. 2) Detects and dispenses with all the system supervisor access inside 12 PM. 3) Eliminates the system chief sections during Head mode and 4) processes the speed of perusing and dispenses with the system administrator's speed not exactly the edge esteem T and furthermore when the all out visited pages surpassing the limit esteem.

The speed of perusing is assessed dependent on all out number of pages perused and all out session time. To deal with the absolute number of passages by the system directors, a lot of choice standards are applied. This gatherings the client into potential and non-potential ones. With the substantial weblog traits, the characterization calculation orders the clients dependent on the preparation information. The choice of traits is done inside 30 seconds and session time to elude the complete pages is 30 minutes. Further, the choice principle to distinguish the potential client is set under 30 minutes and the complete pages access is set under 5. The entrance code Post is utilized for ordering the clients and it diminishes the weblog record size, which improves the grouping forecast and exactness.

C. Clustering Process

$S = \{s1, s2, \ldots, sm\} \subseteq m$ user sessions

In sequence to recognize the primary cluster center, from all user session *si* is deal with as a probable candidate and the starting PDF value for user session *si*, it is indicating like$P1(si)$, is calculated using Eq.(4.1).

$$P1(s_i) = \sum_{k=1}^{m} \exp(-E^2(s_i, s_k)/R^2 \text{ for all } i=1..m \quad \ldots\ldots\ldots(4.1)$$

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

$E_2(si,sk)$ =Euclidean distance(si,sk)

Where R is a +fixed that characterizes a neighbour-hood for client session si.

The PDF estimation of the client session *si* is an estimate density of every client sessions in the neighborhood of *si*. Client sessions external the circular distance have small effect on its PDF value. The client session with the topmost PDF value is select as the main cluster center point v1 as takes after.

$$i_1 \leftarrow \underset{i=1}{\overset{M}{\text{argmax}}}\{P1(s_i)\};\ v_1 \leftarrow {}^{s}i_1 \quad \ldots\ldots\ldots 4.2$$

the second PDF value calculating using Eq. (4.3).

$$P_2(s_i \leftarrow P_1(s_i) - P_1(v_1)\exp(-\frac{E2(si,v1)}{R2}) \text{for all } i=1..m \quad \ldots\ldots\ldots(4.3)$$

Continue work on PDF value for all user sessions, and next cluster center is selected with highest PDF value using Eq.(4.4).

$$i_2 \leftarrow \operatorname{argmax}_{i=1}^{m}\{P1(s_i)\};\ v_1 \leftarrow s_{i_2} \quad \ldots..4.4$$

Also, to select the *j*th cluster center, the PDF value calculates using Eq. (4.5).

$$P_j(s_i \leftarrow P_{j-1}(s_i) - P_{j-1}(v_{j-1})\exp\left(-\frac{E2(si,vj-1)}{R2}\right) \text{ for all } i=1..m \ldots\ldots\ldots(4.5)$$

And the jth cluster centervj is selected is using Eq. (4.6).

$$i_j \leftarrow \operatorname{argmax}_{i=1}^{m}\{P_j(s_i)\}; v_j \leftarrow s_{ij} \quad \ldots\ldots\ldots..4.6$$

Algorithm for PDFCM

Input: *neighborhood radius R(√n), c, maximum iterations η (100), error threshold (0.01), and set of user's sessions S= {$s_1, \cdots, s_m$}*

Output: *Set of c cluster centers V= {$v_1, \cdots, v_c$}and partition matrix P*

define the fixed of cluster centers V(0)

for *i←1, m* ***do***

Calculate the Probability Density values $P_i(s_i)$ using (4.1)

end for

Calculate the Ist cluster center $v_1(0)$ equation (4.2)

for *j←2, c* ***do***

for *i←1, m* ***do***

Calculate the revised probability values $P_j(s_i)$ using (4.5)

end for

Calculate the jth cluster center$v_j(0)$ using (4.6)

end for

t ←1

repeat

Calculate the partition matrix P(m) entries:

for *i←1, m* ***do***

for *j←1, c* ***do***

Calculate μij(m)

Stop for

Stop for

After that new cluster center calculate through V(m):

for *j←1, c* ***do***

Compute $v_j(m)$

end for

Calculate the objective function $J_{FCM}(m)$

m ← m +1

until *$J_{FCM}(m) - J_{FCM}(m-1) <$ t =η*

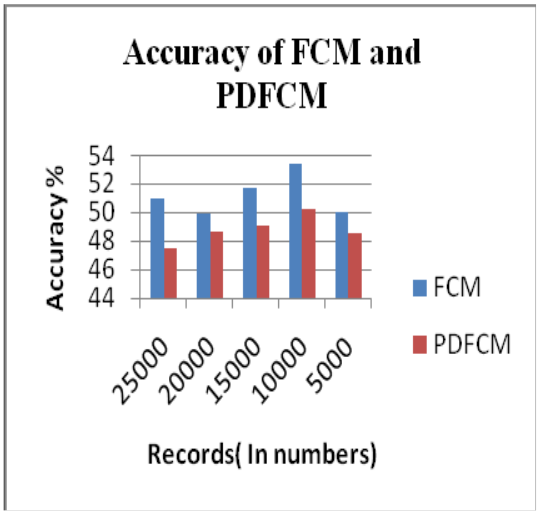
V. EXPERIMENTAL RESULTS

The likelihood thickness based fluffy c implies calculation is actualized in MATLAB for session bunching in weblog information. The test work is done on processor-intel(R), RAM-4GB, framework type-64-piece working framework in windows 8.1 condition.

A. Analysis in Terms of Accuracy

The examination of the conventional FCM calculation with proposed PDFCM shows that the PDFCM calculation perform well over the FCM, which expresses that the alteration in the FCM has improved the exhibition of the ordinary FCM calculation. It is chosen to look at the precision of the two calculations over the weblog informational indexes. The correlation is demonstrated diagrammatically.

It clearly shows that accuracy of PDFCM is better than FCM for different number of datasets. The accuracy of both FCM and PDFCM on various sizes of weblog datasets is shown in Fig. 2.

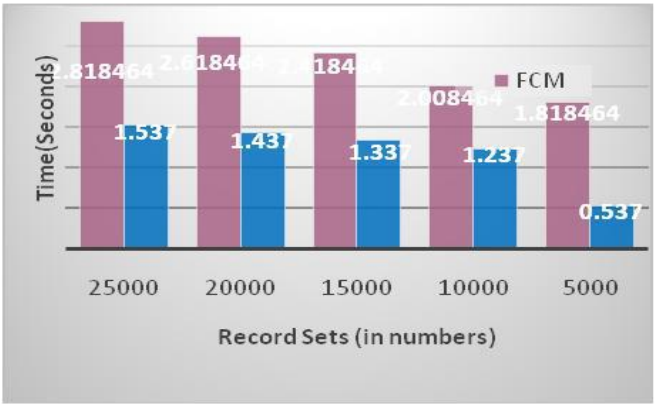


B. Analysis in Term of Execution Time

The execution time required by both FCM and PDFCM calculations are looked at. Which is appeared in Table 1 the outcomes are demonstrated diagrammatically in fig.3 likewise from the outline, result shows that computational time of PDFCM is lesser than contrasted with the FCM for each dataset.

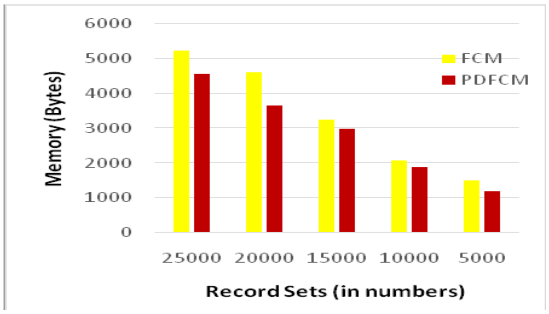
Table 1 Running time (In Seconds) of FCM & PDFCM

Record Sets	Running Time (in Seconds)	
	FCM	PDFCM
25000	2.818464	1.537
20000	2.618464	1.437
15000	2.418464	1.337
10000	2.008464	1.237
5000	1.818464	0.537



C. Analysis in Terms of Memory Requirement

Finally, the memory requirement in terms of bytes is compared between FCM and PDFCM in Table 2 The results are shown diagrammatically as in the Fig.4 From this diagram, it is clear that PDFCM requires less memory than the FCM for all size data set records.



VI. CONCLUSION

In this book division, a gave greater value to way in for making ready of number of persons in a society and mass, group number for FCM algorithm is presented. And general random given undertaking of first parameter to the FCM algorithm is changed in this way in. based on based on experience results it is clear that PDFCM way in provides better act of having no error, get changed to other form running time and take less iteration to complete the experiments PDFCM algorithm is used to discover the net user meetings mass, group. It is better than the FCM because in it before part of right mass, group inside is done which was FCM the bad point of algorithm In PDFCM no of iteration gets changed to other form from 46-49 to 3. After this it is making out by using list of words in a book purpose, use. That is why PDFCM algorithm works better than FCM by looking at different measurement.

REFERENCES

1. A. Gupta and A. Khandekar, "Development of Weblog Mining Based on Improved Fuzzy C-Means Clustering Algorithm", International Journal of Science, Engineering and Technology Research, Vol.5 (3), pp.688-693, March 2016.
2. A. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms", In Proc. of 3rd International Conference on Computational Intelligence & Communication Technology, IEEE, pp. 1-6, 2017.
3. A. Zahid, A. V. Babuy, W Ahmed and M F Azeemz, "A fuzzy set theoretic approach to discover user sessions from web navigational data", In Proc. of Recent Advances in Intelligent Computational Systems, IEEE, pp. 879-884, 2011.
4. B. Chandra, M. Gupta, and M.P. Gupta, "A multivariate time series clustering approach for crime trends prediction", In Proc of International Conference on Systems, Man and Cybernetics, IEEE, pp. 892-896, 2008.
5. B. Maheswari and P. Sumathi, "A New Clustering and Preprocessing for weblog mining" In Proc. of World Congress on Computing and Communication Technologies, IEEE, pp. 25-29, 2014.
6. B. S. Shedthi, Shetty and M. Siddappa, "Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data", In Proc. of International Conference on Inventive Communication and Computational Technologies, IEEE, pp. 105-108,2017.
7. C. Baviskar and S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", In Proc. of International Conference on In Computation of Power, Energy Information and Communication, IEEE, pp. 139-147, 2016.
8. C. T. Baviskar and S. S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", In Proc. of International Conference on Computation of Power, Energy Information and Communication (ICCPEIC)IEEE, pp. 139-147, 2016.
9. C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan and C. Li, "Advances in research of Fuzzy c-means clustering algorithm", In Proc. of International Conference on Network Computing and Information Security, IEEE, vol. 2, pp. 28-31, 2011.
10. Chen, Y.L. and Huang, C.K., "Discovering fuzzy time-interval sequential patterns in sequence databases", IEEE Transactions on Systems, Man and Cybernetics, Vol. 35(5), pp. 959-972, 2005.
11. D. Koutsoukos, G. Alexandridis, G. Siolas, and A. Stafylopatis, "A new approach to session identification by applying fuzzy c-means clustering on weblogs", In Proc. of Symposium Series on Computational Intelligence, IEEE, pp. 1-8, 2016.
12. G. S. Chandel, K. Patidar and M. S. 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, Vol.16(1). pp.135-140, 2016.
13. H. Gulat, and P. K. Singh, "Clustering techniques in data mining: A comparison", In Proc. of 2nd International Conference on Computing for Sustainable Global Development, IEEE, pp.410-4
14. 15, 2015. 14. H. X. Pei, Z. R. Zheng, C. Wang, C. Li, and Y. H. Shao, "D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation", Procedia Computer Science, Vol.122(1), pp. 407-414, 2017.
15. K. Suresh, R. M. Mohana, A. Rama Mohan Reddy, and A. Subramanyam, "Improved FCM algorithm for clustering on web usage mining." In Proc. of International Conference on Computer and Management, pp. 1-4. 2011.
16. P. Sampath and M. Prabhavathy, "Web Page Access Prediction Using Fuzzy Clustering by Local Approximation Memberships (Flame) Algorithm", Vol.10 (7), pp.3217-3220, 2006.
17. 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, IEEE, pp. 506-510, 2015.
18. 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, IEEE, pp. 1-4, 2016.
19. V. Chitraa, and A. S. Thanamani, "Weblog Data Analysis by Enhanced Fuzzy C Means Clustering", International Journal on Computational Sciences & Applications, Vol.4 (2), pp. 81-95, 2014.
20. Y. Hu, Chuncheng Y. Y. Zuo, and F. 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.

AUTHORS PROFILE

Mrs. Geetha Krishnagandhi is an Assistant Professor in Computer Applications since July2019 at Karpagam Academy of Higher Education. She worked as an Assistant Professor inComputer Applications Department and also has been acting as an Assistant Controller of Examinations at GTN Arts College (Autonomous), Dindigul from 2007 to June,2019. She received M.Sc. and M.Phil., degree in Computer Science from the Alagappa University, Karaikudi and Periyar University, Salem. She has been approved as a project guide for Distance Education, Alagappa University and guide.