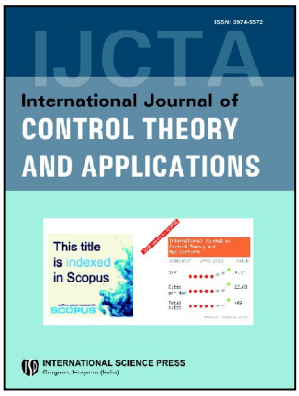

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 11 • 2017

An Optimal Web Usage Mining Technique for Personalizing the Web Directory with The Help of Possibilistic Fuzzy C-means and Relevance Feed Back Algorithm

[1]Sumanth V, and [2]N. Guruprasad

[1]Asstt Prof., Department of Computer Science & Engineering, Bangalore, India
[2]Professor, Department of Computer Science & Engineering, Bangalore, India
E-mail: [1]sumanthphd2016@gmail.com, [2]guruprasadn@newhorizonindia.edu

Abstract: The rapid growth of the Web in terms of Web sites and their users during the last decade has put lots of pressure for Web site owners in reducing the latency of Web pages. The main objective of our proposed web usage mining technique is to recognize interesting behavioral patterns. Initially the web usage data are collected from the access log files after that we will perform preprocessing the data with aid of IP address and these web directories to be initialized by analyzing the association of the web pages after that we will identify the behavioral patterns. These patterns will be identified with the aid of Possibilistic fuzzy c-means algorithm. Consequently the gain and loss of the web users will be determinate based on these only the web directories will be modified with the aid of relevant feedback technique.

Keywords: Web page, Behavioral patterns, Log files, Preprocessing, Web directory, Possibilistic fuzzy c-means, Relevant feedback.

1. INTRODUCTION

As there are very fast development and wide range of application of the internet, World Wide Web has become an interesting medium for pool, exchange, sharing of information and efficient channel for collaborative work. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining aims to extract and mine useful knowledge from the Web [1-3]. The web is a universal information platform space which can be accessed by companies, universities, businessman etc. Generally, web hold there are numerous sources of information like internal sources and external sources. Internal sources are those which include personal information of any organization and external sources are those which include information of clients, vendors, suppliers, intranet and extranet [4-5].

Web Mining is defined as automatic crawling and extraction of relevant information from the artifacts, activities, and hidden patterns found in WWW. Web Mining is used for tracking customers' online behavior, most importantly, cookies tracking and hyperlinks correlations. Unlike search engines, which send agents to

crawl the web searching for keywords, Web Mining agents are far more intelligent [6-9]. Web Mining work by sending intelligent agents to certain targets, like competitors' sites. These agents collect information from the host web server and collect as much information from analyzing the web page itself. Mainly they look for the hyperlinks, cookies, and the traffic patterns. Using this collected knowledge enterprise can establish better customer relationships, offers and target potential buyers with exclusive deals. The WWW is very dynamic, and web crawling is repetitive process where contentious iteration will achieve effective results. Web Mining is used for business, stochastic, and for criminal and juridical purposes mainly in network forensics [10-14].

Web mining is used to capture relevant information, creating new knowledge out of relevant data, personalization of the information and learning about Consumers or individual users and several others. Web mining can be divided into three categories depending on the type of data as :(i) Web usage mining, (ii) Web content mining and (iii) Web structure mining. Web usage mining is also known as Web log mining. It is the process of extracting interesting patterns in Web access logs. It analyzes navigational activities of web users. This usage data provides the paths leading to accessed Web pages [15-17]. This information is often gathered automatically into access logs via the Web server. Web content mining is the process of extracting knowledge from the content of documents or their description, available on the World Wide Web .Web structure mining is the process of inferring knowledge from the World Wide Web .Which is further divided into two types based on the structure information. 1. Hyperlinks: Hyperlink connects a Web page either in the same Webpage or on different Web pages. 2. Document Structure: Web page can also be arranged in a tree-structured format, based on HTML and XML tags within the page [18-19].

The Web Mining can be decomposed into the following subtasks, namely : 1.Resource finding, 2. Information selection and pre-processing,3.Generalization and 4.Analysis. Resource finding is the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on the web. 2. Information selection and pre-processing involves the automatic selection and pre-processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in the training corpus. 3. Generalization automatically discovers general patterns at individual web sites as well as across multiple sites. 4. Analysis is termed as the validation and interpretation of the mined patterns. It plays an important role in pattern mining [20].

2. OBJECTIVES

1. To reduce the loss of the end users
    - In a web directory, if the user does not find out the content that he looking for then it will be a loss to him. Increasing of loss will decrease the usage of web directory.
    - We are creating the web directory to extract the information what user is looking for. So increase of loss while searching in the web directory will arise the question that what is the purpose of creating the web directory.
    - So in our proposed technique the loss is considered as an important parameter and by using the relevance feedback technique web directory is modified to reduce the loss of the end users.
2. To find out the interesting behavioral patterns
    - To build a useful web directory, interesting behavioral patterns need to be found out.
    - So by using the Possibilistic Fuzzy C-Means technique, interesting behavioral patterns will be found out.

3. LITERATURE REVIEW

Peng Yeng *et al.* [21] have proposed a formulation for the website structure optimization (WSO) problem based on a comprehensive survey of existing works and practice considerations. An Enhanced Tabu Search (ETS) algorithm was projected with advanced search features of multiple neighborhoods, adaptive tabu lists, dynamic tabu tenure, and multi-level aspiration criteria. The experimental result on 24 real-world problem instances have shown that their proposed ETS algorithm was obtained a better value of web usage estimation than a genetic algorithm method. Moreover, ETS was computationally efficient due to the strategy that handles problem constraints on-the-fly when constructing the solution.

Stephen Matthews *et al.* [22] have proposed the solution for losing temporal fuzzy association rules on real-world Web log data. GA and the 2-tuple linguistic representation was improved by transforming the dataset to a graph, which ensures a valid item-sets were discovered, and modifying the fitness function. The GA-based approach was recommended because it discovered extra rules that a traditional algorithm does not. The decision to use this complementary approach relied on understanding what temporal changes may be present in the application domain (e.g., seasonal and/or scheduled events). It was important to note that lowering minimum support/confidence would overcome the problem of losing rules with traditional approaches, however, the number of rules increases, which was undesirable in association rule mining.

Olatz Arbelaitz *et al.* [23] have proposed web usage and con-tent mining techniques, with the three main objectives: generating user navigation profiles to be used for link prediction; enriching the profiles with semantic information to diversify them, which provides the DMO with a tool to introduce links that will match the users taste; and moreover, obtaining global and language-dependent user interest profiles, which provides the DMO staff with important information for future web designs, and allows them to design future marketing campaigns for specific targets. The system performed successfully, obtaining profiles which fit in more than 60% of cases with the real user navigation sequences and in more than 90% of cases with the user interests.

Dimitrios Pierrakos *et al.* [24] have presented a knowledge discovery framework for the construction of Community Web Directories. In that context, the Web directory was viewed as a thematic hierarchy and personalization was realized by constructing user community models on the basis of usage data. In contrast to most of the work on Web usage mining, the usage data that were analyzed here correspond to user navigation throughout the Web, rather than a particular Web site, exhibiting as a result a high degree of thematic diversity. For modeling the user communities, they introduced a methodology that combines the users' browsing behavior with thematic information from the Web directories. The resulting community models take the form of Community Web Directories. Their proposed personalization methodology was evaluated both on a specialized artificial and a general-purpose Web directory, indicating its potential value to the Web user. The experiments assessed the effectiveness of the different machine learning techniques on the task.

Leonidas Akritidis *et al.* [25] have presented a rank aggregation method, which takes into consideration additional information regarding the query terms, the collected results and the data correlated to each of these results (title, textual snippet, URL, individual ranking and others). They have implemented and tested Quad Rank in a real-world Meta search engine, Quad Search, a system developed as a test bed for algorithms related to the wide problem of Meta searching. They have exhaustively tested Quad Rank for both effectiveness and efficiency in the real-world search environment of Quad Search and also, using a task from the recentTREC-2009 conference. The results they presented in their experiments has revealed that, in most cases Quad Rank outperformed all component engines, another meta search engine (Dogpile) and two successful rank aggregation methods, Borda Count and the Outranking Approach.

4. MOTIVATION OF THE RESEARCH

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Many techniques have been

proposed in the field of web usage mining to reduce the time of the users while looking any data in the web. Dimitrios *et al.* [24] have proposed a technique for personalizing the web directory to make the access of the internet in an easy way. Although it produced effective performance, it has some drawbacks. Also, they have used PLSA (Probabilistic Latent Semantic Analysis) technique for identifying latent factors in data at the time of retrieving the community web directory. But the PLSA model suffers from the limitation of being unable to discover enough semantic and discriminative information for action recognition. Also, they have used FCM while discovering the community web directory, but the FCM has the following drawbacks. Many of the fuzzy degrees are zero leads to a very sparse vector while using the FCM with the PLSA which implies the uncertainty. The FCM algorithm assigns memberships to the input which is inversely related to the relative distance of the input to the cluster centers in the FCM model. Suppose, If the input is equidistant from two prototypes, the membership of the input in each cluster will be the same, regardless of the absolute value of the distance of input from the two centroids as well as from the other points in the data. The problem this creates is that noise points, far but equidistant from the central structure of the two clusters, can nonetheless be given equal membership in both, when it seems far more natural that such points be given very low or even no membership in either cluster. Those above mentioned problems have motivated me to do the research work in this area.

5. PROPOSED METHODOLOGY

The main aim of this research is to provide a better web usage mining technique by solving the drawbacks that currently exist in the literary works. Thus, I have intended to propose an efficient Web Usage Mining technique with the aid of PFCM (Possibilistic Fuzzy C-Means) and Relevance Feed Back for web personalization. The principle objective is to recognize interesting behavioral examples in the gathered use information and build group Web indexes focused around those examples. Initially, the web usage data will be collected from the access log files of ISP cache proxy servers. After collecting the usage data, the preprocessing approach will be carried out to clean the unwanted data from the collected data in order to reduce the time. Subsequently, user sessions will be identified from the preprocessed data with the help of the IP address and a predefined threshold. Next, the web directory will be initialized by analyzing the association of the web pages. After determining the mapping and the associations between Web pages, user sessions, and Web directory categories, interesting behavioral patterns will be identified with the help of the Possibilistic Fuzzy C-Means. Subsequently, the gain and the loss of the web users will be determined and based on the gain (the users find what they are looking for) and the loss (the users do not find what they are looking for) the web directory will be modified with the help of the Relevance Feedback technique which in turn reduce the loss of the end users. The proposed web usage mining technique will be implemented in the working platform of JAVA and the performance will be analyzed.

5.1. Preprocessing

Initially, the web usage data will be collected from the access log files of ISP cache proxy servers. After collecting the usage data, the preprocessing approach will be carried out to clean the unwanted data from the collected data in order to reduce the time. The main aim of preprocessing an input data is the data which is obtained from the logs may be incomplete noisy and inconsistent. The raw proxy server log files are unsuitable for access pattern analysis. The proxy server log requires effective preprocessing to remove irrelevant data from the proxy server log file for analysis. It is important to remove all the requests from the web proxy log file that are explicitly requested by the user. When the user requests any page using the browser, there are a number of log entries created in the log file as a page contains other web objects like image and java script files. Subsequently, user sessions will be identified from the preprocessed data with the help of the IP address and a predefined threshold. Next, the web directory will be initialized by analyzing the association of the web pages.

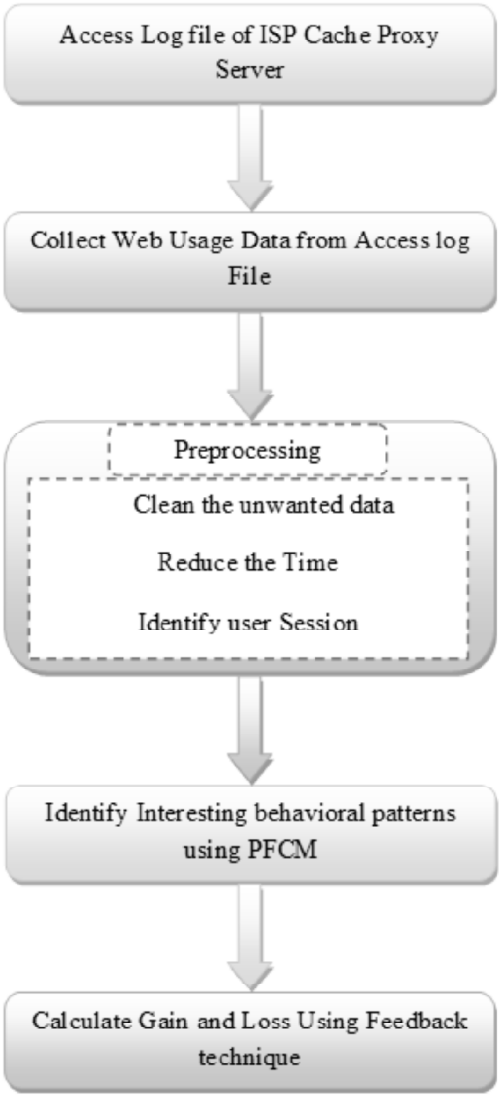

Figure 1: Proposed PFCM Based Web usage mining for personalizing a web directory

5.2. Session Identification

In Session Identification each unique IP address is identified as a different user. Session identifications process construct tuple of the form {IPi, Pagesj} i.e., set of pages visited by the particular user from the particular machine is identified. The set of URLs which is forming a session should satisfy the requirement and the time elapsed between two consecutive requests is smaller than a given Dt.

5.3. PFCM based clustering module

The input dataset coordinated to the intrusion detection framework en-compose large gigantic measure of data which makes the handling to a great degree mind boggling, furious and time consuming. Executing this sweeping number of data can furthermore provoke having poor results about by the increment of mistakes. From this time forward, it will have checked effect on the adequacy of the framework and in the end inciting reduced quality intrusion detection framework. To insignificant this issue, the dataset is preprocessed clustering technique is used before characterization. In pre-processing stage, we layout typical esteemed properties of KDD glass 99

dataset to numeric-esteemed qualities. Then the pre-processed input data is connected to a clustering method called PFCM. The application of Possibilistic Fuzzy C-means clustering methods will improve the clustering process and accuracy of the intrusion detection system. The vital motive of the probabilistic fuzzy c-means (PFCM) cluster module is devoted to the segmentation of a specified set of data into clusters. By means of the clustering module, the training set (*TR*) is grouped into various subsets. As the dimension and intricacy of each and every training subset are scaled down, the effectiveness and efficacy of subsequent RNN module can be incredibly incremented. The Probabilistic Fuzzy c-means constitutes a data clustering technique where each data point is a part of the cluster to a level indicated by the membership grade. In the clustering module, it is dependent on the reduction of the objective function which is illustrated in the following equation (1).

$$J_{PFCM}(U,T,V;X)=\sum_{k=1}^{n}\sum_{i=1}^{c}(aU_{ik}^{m}+bT_{ik}^{\eta})\times\left\|x_k-v_i\right\|_A^2+\sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-T_{ik})^{\eta} \tag{1}$$

Where;

$U \rightarrow$ Membership matrix

$T \rightarrow$ Possibilistic matrix

$V \rightarrow$ Resultant cluster center

$X \rightarrow$ Set of all data point

The constants a and b represent the comparative significance of fuzzy membership and typicality values in the objective function. The gradual procedure of PFCM is effectively elucidated below.

Step 1: Calculation of distance matrix

At the outset, the number of cluster (C) is furnished by the user, which is identical in respect of every segment. When the number of cluster is determined, the evaluation of the distance between the centroids and data point for each segment is carried out. In the document, Euclidian distance function as illustrated in equation (2) shown below is elegantly employed to evaluate the distance between centroids and data points with the ultimate motive of evaluating the distance matrix. Further, the distance matrix is determined for each and every cluster.

$$D(x_k,v_i)=\sqrt{\sum_{i,k=1}^{i=n,k=n}(x_i-v_i)^2} \tag{2}$$

$$D_{ki}=\begin{bmatrix} d_{11} & d_{12} & d_{1i} & d_{1C-1} & d_{1C} \\ d_{21} & d_{22} & d_{2i} & d_{2\,C-1} & d_{2C} \\ d_{k1} & d_{k2} & d_{kk} & d_{k\,C-1} & d_{k\,C} \\ d_{n-11} & d_{n-12} & d_{n-1\,i} & d_{n-1\,C-1} & d_{n-1\,C} \\ d_{n1} & d_{n2} & d_{ni} & d_{n\,C-1} & d_{n\,C} \end{bmatrix} \tag{3}$$

Here, the values characterize the distance of k^{th} data point in relation to the i^{th} centroid. Similarly, the distance function is effectively utilized to estimate the distance between every data point x_k with every centroid v_i value. At last, the distance matrix is created which is indicated in the equation (3).

Step 2: Calculation of typicality matrix

After the estimation of the distance matrix, the typicality matrix is evaluated. The typicality matrix, in turn, is obtained from PCM. As illustrated in equation (4) shown below, the probability value of each data point in

relation to every centroid is completed. Subsequently, typicality matrix is created as demonstrated in equation (5).

$$T_{ik} = \frac{1}{1+\left[\frac{D^2(x_k, v_i)}{\eta_i}\right]^{1/(m-1)}} \tag{4}$$

$$T_{ki} = \begin{bmatrix} t_{11} & t_{12} & t_{1i} & t_{1C-1} & t_{1C} \\ t_{21} & t_{22} & t_{2i} & t_{2\,C-1} & t_{2C} \\ t_{k1} & t_{k2} & t_{ki} & t_{k\,C-1} & t_{k\,C} \\ t_{n-11} & t_{n-12} & t_{n-1\,i} & t_{n-1\,C-1} & t_{n-1\,C} \\ t_{n1} & t_{n2} & t_{ni} & t_{n\,C-1} & t_{n\,C} \end{bmatrix} \tag{5}$$

The equation (5) effectively depicts the typicality matrix. Here, the value of T_{ik} characterizes the probability of i^{th} data point chance to move towards the k^{th} centroid. Similarly, the distance function is efficiently utilized to estimate the distance between every data point with every centroid value.

Step 3: Calculation of membership matrix

The evaluation of the membership matrix U_{ik} is performed by means of assessing the membership value of data point which is gathered from the FCM [3]. As illustrated in the help of the following equation (6), the membership value of each data point in relation to each centroid is completed. Subsequently, the membership matrix is created as illustrated in the ensuing equation (7).

$$U_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|}\right)^{\frac{2}{m-1}}} \tag{6}$$

$$U_{ik} = \begin{bmatrix} u_{11} & u_{12} & u_{1i} & u_{1C-1} & u_{1C} \\ u_{21} & u_{22} & u_{2i} & u_{2\,C-1} & u_{2C} \\ u_{k1} & u_{k2} & u_{k\,i} & u_{k\,C-1} & u_{k\,C} \\ u_{n-11} & u_{n-12} & u_{n-1\,i} & u_{n-1\,C-1} & u_{n-1\,C} \\ u_{n1} & u_{n2} & u_{ni} & u_{n\,C-1} & u_{n\,C} \end{bmatrix} \tag{7}$$

Step 4: Updation of centroid

After the generation of the clusters, the modernization of the centroids is performed in accordance with equation (8) shown below.

$$v_i = \frac{\sum_{k=i}^{n}\left(U_{ik}^m + T_{ik}^\eta\right)x_k}{\sum_{k=1}^{n}\left(U_{ik}^m + T_{ik}^\eta\right)}, \quad 1 \le i \le c \tag{8}$$

Subsequent to the modernization of the centroids in respect to each and every cluster, the task of evaluating the distance with the lately modernized centroids is started and continued till the evaluation of the modernization of the centroids. The relative procedure is performed again and again till the modernized centroids of each and every cluster becomes identical similar in successive iterations.

5.4. Possible Outcome

By using the proposed web usage mining technique, interesting behavioral patterns will be obtained with the help of Possibilistic Fuzzy C-Means and the user's loss will be reduced using Relevance Feed Back technique. The performance of the proposed technique will be assessed in terms of user's gain and loss. Also the performance of the Possibilistic Fuzzy C-Means utilized in the proposed technique will be analyzed in terms of interesting behavioral patterns.

5.5. Relevance Feedback

The idea of relevance feedback (RF) is to involve the user in the retrieval process to improve the final result set. In particular, the user gives feedback on the relevance of documents in an initial set of results.

The basic procedure is:

1. The user issues a (short, simple) query
2. The system returns an initial set of retrieval results
3. The user marks some returned documents as relevant or non-relevant.
4. The system computes a better representation of the information need based on the user feedback.
5. The system displays a revised set of retrieval results

Relevance feedback can go through one or more iterations of this sort. The process exploits the idea that it may be difficult to formulate a good query when you don't know the collection well, but it is easy to judge particular documents and so it makes sense to engage in iterative query refinement of this sort. In such a scenario, relevance feedback can also be effective in tracking a user's evolving information need: seeing some documents may lead users to refine their understanding of the information they are seeking.

6. RESULT AND DISCUSSION

An experimental result is used to evaluate the effectiveness of the proposed systems and to justify theoretical and practical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology.

Table 1
Proposed study obtained time for whole study based on the threshold

Threshold	*Time*
3	165487
4	154785
5	151236
6	146321

Table I explain the total no of Time taken for our study based upon a threshold value. We have mentioned some threshold values based on that the time to be calculated. The threshold 3 has taken 165487ms to complete

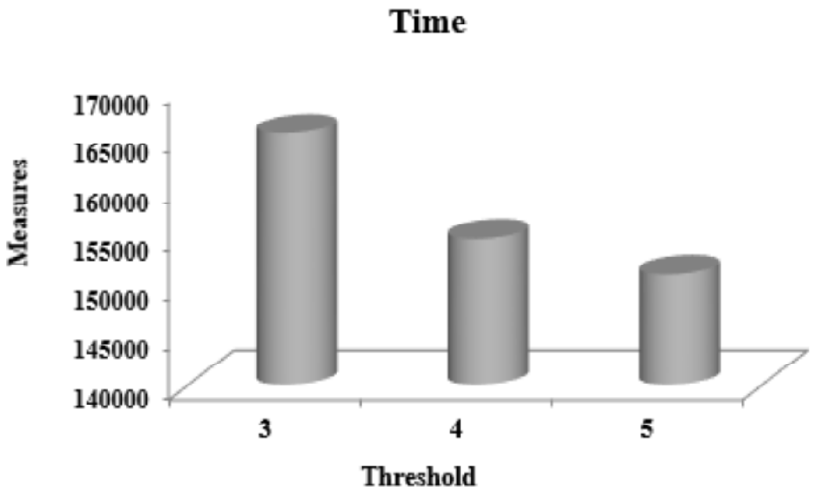

Figure 2: Graph for measure an overall time based on threshold value

a process and in threshold 4 has taken 154785 ms to complete a process and threshold 5 has taken a 151236ms to complete a process and in threshold 6 has taken 146321ms to complete a process. Our research has taken the less amount of time, which is depicted in the table and the graphical representation is highlighted in fig 2.

Table 2
Illustrated the memory space obtained for our proposed based on iteration

Threshold	*Time*
3	1248578
4	1124545
5	1115477
6	1011446

Table II explains the total no of memory space occupied for our proposed study based on the threshold values 3, 4, 5, and 6. In threshold value 3 will occupied a 1248578 memory space. In threshold 4 have occupied a 1124545 memory space and the threshold 5 have occupied a 1011446 memory space.

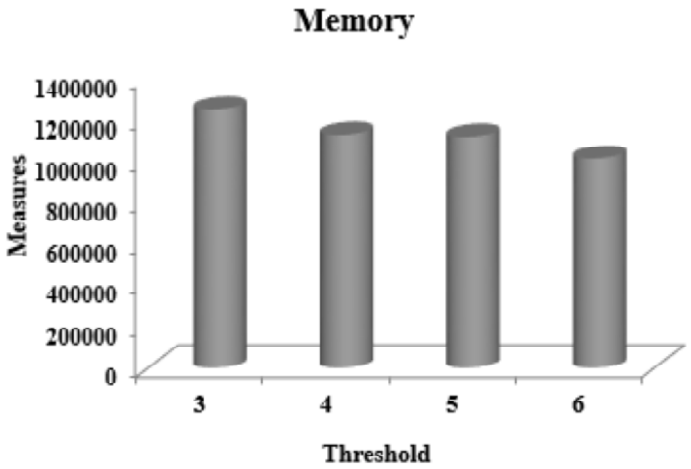

Figure 3: Graph for memory space taken for our whole process

Table II explains the total memory space taken for a study based upon a threshold value. We will give a memory space occupied for each and every process is given here initially the threshold value for our process is 3, 4, 5, and 6; based on these they obtained memory space as 1248578, 1124545, 1115477, and 1011446. In our research, we have occupied a low level memory space which has shown in the table and the graphical diagram is demonstrated in fig. 3.

Table 3
Illustrated the loss and gain of data's based on the threshold

Threshold	*Loss data's*	*Gain data's*
3	21	49
4	23	47
5	25	45
6	30	40

Table II explains a loss and gain of an input data's measured based on the threshold. In threshold 3 the loss data is 21 and the gain data is 49. Then in the threshlold 4, the loss data is 23 and the gain data is 47. In the threshold 5, the losss data is 25 and the gain data is 45. Finally, in the threshold 6 the loss data is 30 and the gain data is 40. These loss and gain data is graphically represented in the fig. 4.

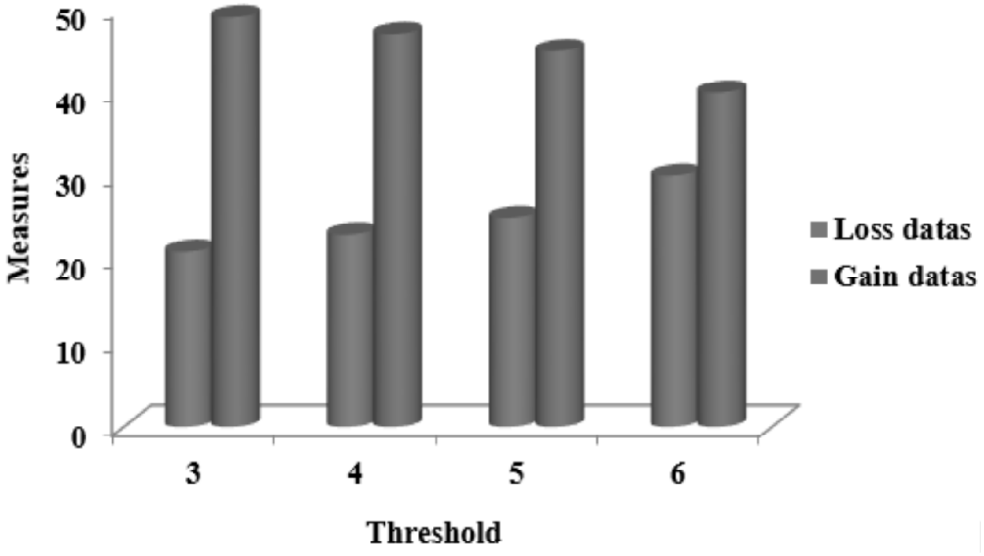

Figure 4: Graph for proposed input data loss and gain measures taken based on the threshold

Table 4
Propose study cluster accuracy measures

No of Clusters	*Cluster Accuracy*
2	71.26
3	72.35
4	73.45
5	75.61

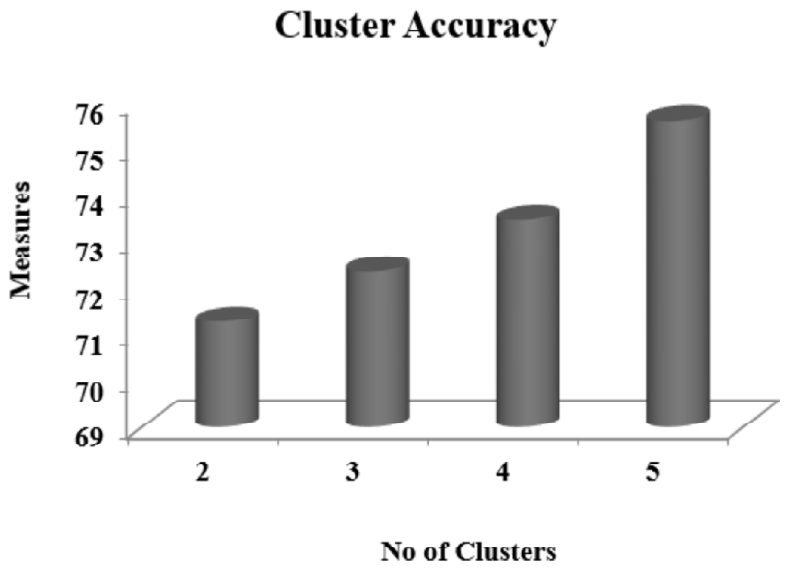

Figure 5: Graph for proposed study cluster accuracy

6.1. Comparative Analysis

When compared our work with the existing methods, we found that, our proposed study has given a better result. In our proposed study Possibilistic fuzzy c-means technique was compared to an existing fuzzy c-means algorithm to prove that our proposed technique yields a better result.

Table 5
Comparison for Proposed and Existing Time value Measures

Threshold	*Time for Proposed Study*	*Time for Existing study*
3	165487	176521
4	154785	165556
5	151236	175462
6	146321	164589

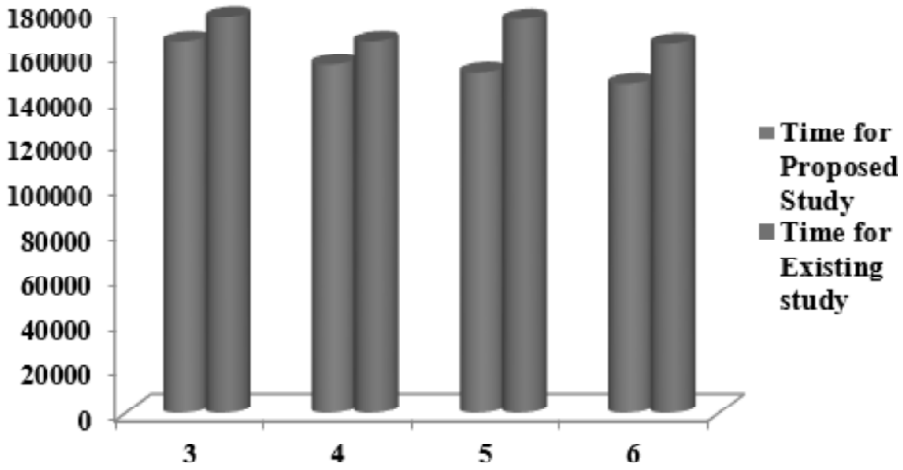

Figure 6: Graph for comparison of proposed and existing method

Here we have compared time and memory based on the threshold value. The threshold values are 3, 4, 5, and 6 based on this value, the existing FCM used memory space for the process such as 176521, 165556, 175462 and 164589. This result shows this existing FCM has taken a maximum no of memory space, but when we compare this result to our proposed PFCM technique had taken a minimum memory space such as 165487, 154785, 151236 and 146321.

Table 6
Comparison for Proposed and Existing Clustering accuracy Measures

Threshold	*Time for Proposed Study*	*Time for Existing study*
2	71.26	66.32
3	72.35	68.26
4	73.45	70.35
5	75.61	71.21

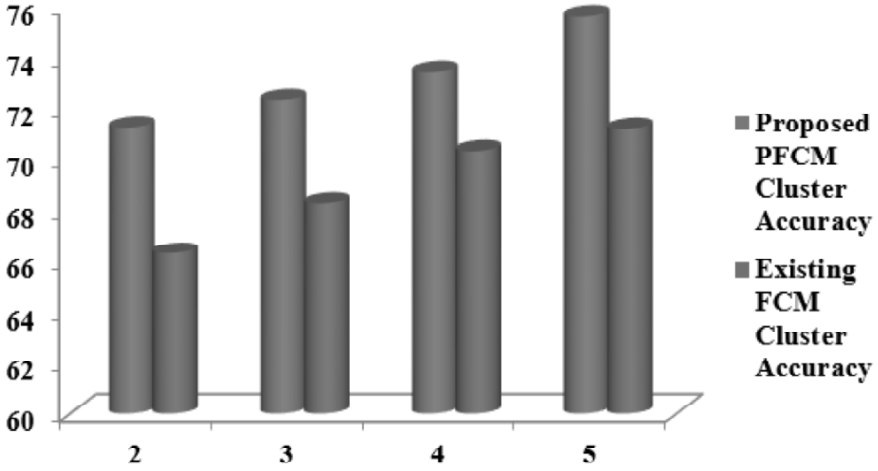

Figure 7: Graph Comparison for Proposed and Existing Clustering accuracy Measures

Here we have compared a Clustering accuracy Measures based on the clusters. The clusters are 2, 3, 4 and 5, based on this value, the existing FCM obtained clustering for the process such as 66.32, 68.26, 70.35 and 71.21. When we compared this existing FCM clustering accuracy to that of proposed PFCM, the latter gave a better accuracy such as 71.26, 72.35, 73.45 and 75.61.

7. CONCLUSION

The innovative Possibilistic fuzzy c-means based web usage mining with three diverse phases such as preprocessing, behavioral pattern identification and gain & loss calculation is elegantly presented in this study. The preprocessing phase involve to clean an unwanted data and it reduced the time. Also it can identify a user session with the aid of IP address. It is followed by the pattern identification phase to identify behavioral patterns with the aid of Possibilistic fuzzy c-means algorithm. Subsequently, the gain and the loss of the web users determined and based on the gain (the users find what they are looking for) and the loss (the users do not find what they are looking for) the web directory modified with the help of the Relevance Feedback technique which in turn reduce the loss of the end users. The efficiency of this study measured with time, memory, gain & loss and cluster accuracy. Additionally, we have compared our proposed PFCM web usage mining technique to the existing FCM clustering technique. It revealed that our proposed PFCM gave a better result.

REFRENCES

[1] Malarvizhi and Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study", *International Journal of Computer Trends and Technology*, Vol. 4, No. 8, , 2013, pp. 2940-2945.

[2] Joshila Grace, Maheswari and Dhinaharan Nagamalai, "Analysis of Web Logs and Web User In Web Mining", *International Journal of Network Security & Its Applications,* Vol. 3, No. 1, pp. 99-110, 2011.

[3] Abdul-Aziz and Rashid Al-Azmi, "Data, Text and Web Mining for Business Intelligence: A Survey", *International Journal of Data Mining & Knowledge Management Process*, Vol. 3, No. 2, , 2013, pp. 1-21.

[4] R. Jain and Purohit, "Page Ranking Algorithms for Web Mining", *International Journal of Computer Applications*, Vol. 13, No. 5, , 2011, pp. 22-25.

[5] P. Kamde and S. Algur , "A Survey on Web Multimedia Mining", *The International Journal of Multimedia & Its Applications*, Vol. 3, No. 3,, 2011, pp. 72-84.

[6] Poongothai, Parimala and Sathiyabama, "Efficient Web Usage Mining with Clustering", *International Journal of Computer Science Issues*, Vol. 8, No. 6, 2011, pp. 203-209.

[7] A. Rastogi and S. Gupta, "Web Mining: A Comparative Study", *International Journal of Computational Engineering Research*, Vol. 2, No.2, 2012, pp. 325-331.

[8] P. Mehtaa, B. Parekh, K. Modi, and P. Solanki, "Web Personalization Using Web Mining: Concept and Research Issue", *International Journal of Information and Education Technology*, Vol. 2, No. 5, 2012, pp. 510-512.

[9] Manda Jaya Sindhu, Madhavi Latha, Samson Deva Kumar and Suresh Angadi, "Multimedia Retrieval Using Web Mining", *International Journal of Recent Technology and Engineering,* Vol. 2, No. 1, pp. 106-108, 2013.

[10] S.K. Pani, D. Mohapatra, and B.K. Ratha, "Integration of Web mining and web crawler: Relevance and State of Art", *International Journal on Computer Science and Engineering* , Vol. 2, No. 3, , 2010, pp. 772-776.

[11] Jeyalatha, Sivaramakrishnan, Vijayakumar and Balakrishnan, "Web Mining Functions in an Academic Search Application", *Informatica Economical*, Vol. 13, No. 3, 2009, pp. 132-139.

[12] S. Tiwari, "A Web Usage Mining Framework for Business Intelligence", *International Journal of Electronics Communication and Computer Technology*, Vol. 1, No.1, 2011, pp. 19-22.

[13] R. Chuchra, B. Mehta, and S. Kaur, "Use of web Mining in Network Security", *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, No. 4, 2013, pp. 164-168.

[14] Abdelhakim Herrouz, Chabane Khentout and Mahieddine Djoudi, "Overview of Web Content Mining Tools", The International Journal of Engineering and Science, Vol. 2, No. 6, pp. 106-110, 2013.

[15] Bhaiyalal Birla and Sachin Patel, "An Implementation on Web Log Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, No. 2, pp. 68-73, 2014

[16] Shaheen Parveen and Ajay Kushwaha, "An Approach of Deep Web Mining For Data Extraction", International Journal of Engineering Science & Advanced Technology, Vol. 2, No. 6, pp. 1653 – 1656, 2012.

[17] Sarika Pabalkar, "Web Text Mining for news by Classification", International Journal of Advanced Research in Computer and Communication Engineering , Vol. 1, No. 6, pp. 387-391, 2012.

[18] Sridevi and Umarani, "A Survey of Semantic based Solutions to Web Mining", *International Journal of Emerging Trends & Technology in Computer Science*, Vol. 1, No. 2, 2012, pp. 50- 57.

[19] S. Ramulu, S. Kumar, S. Reddy, "A Study of Semantic Web Mining: Integrating Domain Knowledge into Web Mining", *International Journal of Soft Computing and Engineering*, Vol. 2, No. 3, 2012, pp. 522-524.

[20] M. Eshaghi, Gawali, "Web Usage Mining Based on Complex Structure of XML for Web IDS", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 2, No. 5, 2013, pp. 323-326.

[21] P.Y. Yin, and Y.M. Guo, "Optimization of Multi-Criteria Website Structure Based on Enhanced Tabu Search and Web Usage Mining", *Applied Mathematics and Computation*, Vol. 219, No. 24, 2013, pp. 11082-11095.

[22] S. Matthews, M. Gongora, A. Hopgood, and S. Ahmadi, "Web Usage Mining with Evolutionary Extraction of Temporal Fuzzy Association Rules", *Knowledge-Based Systems*, Vol. 54, No. 0, 2013, pp. 66-72.

[23] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J.M. Pérez, and I. Perona, "Web Usage And Content Mining to Extract Knowledge For Modelling The Users of The Bidasoa Turismo Website and to Adapt it", *Expert Systems with Applications*, Vol. 40, No. 1, 2013, pp. 7478-7491.

[24] D. Pierrakos and G. Paliouras, "Personalizing Web Directories with the Aid of Web Usage Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 9, 2010, pp. 1331-1344.

[25] L. Akritidis, D. Katsaros, and P. Bozanis, "Effective Rank Aggregation For Meta Searching", *The Journal of Systems and Software*, Vol. 84, No. 2, 2011, pp. 130-143.