Web Usage Based Analysis of Web Pages Using RapidMiner

M.SANTHANAKUMAR, C.CHRISTOPHER COLUMBUS
Department of Computer Science and Engineering
PSN College of Engineering and Technology
Melathediyoor, Tirunelveli, Tamilnadu
INDIA
santhanakumar@psncet.ac.in, columbus@psncet.ac.in

Abstract: - In recent times, due to the rapid usage of World Wide Web, websites are the information provider to the Internet users. Storing and retrieving the information from the web is always a challenging task. Web mining, the term is defined as extract needed information to the users from the Web. Here, the information provided by the Web is not only the exact information of user needs but also suggest the information associated to the exact one. Web mining is classified into three sub tasks such as, Web Content, Web Structure and Web Usage Mining. This paper, introduces the applications and the mining process of data mining tool (open source) Rapidminer. Here, the proposed work analyzes the usage of web pages (i.e. Browsing behavior of user) using two different clustering algorithms such as k-means, which is incorporated in the tool and Fuzzy c means (FCM) clustering using RapidMiner. The results will show operational background of FCM clustering and k-means clustering algorithm based on the cluster centroid.

Keywords: Web Mining, Web Usage Mining, k-means, FCM, RapidMiner

1 Introduction

Data mining or “Knowledge Discovery in Databases” is defined as extracting the user’s needed and useful information from the massive dataset. Massive dataset has its own challenge of storing, processing and capturing of data from the dataset. Due to the huge information flow through the Internet, people search and collect the required information from the Internet. Web mining carries out the search by not only providing the exact information but also the related information to the appropriate word set search by the user. People can access the massive amount of shared information through the Internet. Web is the only source for providing information to the users via hyperlinks. Web content mining, Web structure mining and Web usage mining are the types of web mining [1].Web Content Mining is the process of extracting information (i.e. Text, Audio, Video, Image, etc…) based on the keyword given by the user. Information Retrieval (IR) and Natural Language Processing (NLP) are the technologies used in web content mining.

The extraction of information from the web based on the relationship and structure of the web pages is referred as web structure mining. Web usage mining is the process of applying Data Mining techniques to the discovery of usage patterns from the web data, targeted towards various applications. The following Figure 1 shows the structure of web mining (i.e.,) it describes the types of web mining.

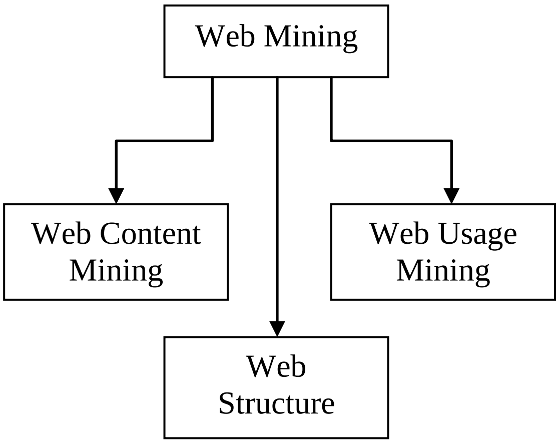


Figure 1: Structure of Web Mining

Web Usage Mining is one of the techniques which play an important role in the personalization of web pages. To perform the analysis of web access information, first the web usage dataset is collected from the internet and pre processing the dataset like filtering, noise removal, etc., For the collection of web usage dataset, the web usage data’s were gathered from different levels such as Server level, Client level and Proxy level and also from different resources through the web browsers and web server interaction using the HTTP protocol (Hyper Text Transfer Protocol) [2]. The request information sent by the user via protocol to the web server which is recorded in one file named as log file.

```
131.112.168.40 - - [01/Jul/1995:01:32:26 -
0400] "GET /images/NASA-logosmall.gif
HTTP/1.0" 200 786
```

Figure 2: Single Web Server Log entry

The structure of web server log file is given in Figure 2 which describes the single log file entry in the web server. It has,

- 131.112.168.40 (%h): denotes the IP Address of the client that sends request to Server.
- (%l): “hyphen” indicates the remote log name of the user.
- (%u): “hyphen” indicates the user id of the user sending request.
- [01/Jul/1995:01:32:26 -0400] (%t): indicates the date and time which the request was received. It is in the format of [Date/Month/Year: hour: minute: second zone]
- "GET/images/NASA-logosmall.gif HTTP/1.0"(\"%r\"): denotes requested line sent by the client in double quotes.
- 200 (%>s): indicates the status code send by the server to client.
- 786 (%b): the last part that denotes the object size returned to the client.

There are two kinds of log file format namely Common Log format (CLF) and Extended Common Log Format (ECLF) [2].

1.1 Data Collection

Data collection is defined as the collection of web access log information which is carried out from Server side, Client side and Proxy server side [3].

1.1.1 Server side collection

The Data collection from the server side, the log file contains all the information of request made by the client’s. Normally web log files are plain text and it is independent from the server. The common log file format [3] is given as follows,

“IPAddress Username Password date & timestamp URL version Status-code Bytes send”

In the extended common log file format, the server has the information like referred pages and user agent. Referred pages have the details of referral link to reach the URL and the user agent which describes the version of the browser software used for searching the information.

1.1.2 Client side collection

In the client side collection, the browsing behaviour of the users is recorded by the web browsers. For collecting the history of the user behaviour, remote agents were implemented with Java or JavaScript. It is considered as more reliable than server side collection because it overcomes both the caching and session identification problems [4].

1.1.3 Proxy Server side collection

Web usage data is collected from the proxy server which acts as an intermediate server between the web browser and web server. Here, the web access log file details are same as server side collection’s log file. Additionally, it records the information as request sent by the browser and response from the web server [3]. Proxy caching is used to reduce the loading time and network traffic load at server side as well as in client side.

1.1.4 Cookies

A cookie is a unique ID generated by the web server that is copied as a small file on client machine; server also records it for identifying the client later. Whenever the same client access the same website again from same machine, the browser reads that Unique ID and send it to the server. Thus, the web server can easily identify its user with the help of Unique ID (Cookie) that was assigned to the user (client) during the last time visit [4].

1.2 Pre-Processing

After the web usage dataset collection, perform pre processing on the dataset. Because, the data collected from the web is normally diverse, heterogeneous and unstructured. Therefore, it is necessary to do the pre-processing like filtering unnecessary and irrelevant data, predicting and filling the missing values, removing noise, resolving inconsistence before applying the algorithm [4]. Data pre-processing consists of the following processes such as, Data cleaning, User identification, Session identification and Path completion [3] is shown in the Figure 3.Web Usage Pre-processing is a difficult task due to the incompleteness of the available data.

1.2.1 Data Cleaning

Data cleaning is the process of removing unwanted data from the Web log files such as the gif, jpeg, video, audio, css etc., Also it has the HTTP status code which is less than 200 and greater than 400.To remove the irrelevant data and noise from the log file, the following steps are carried out.

1. Start
2. Open the log file
3. Split the log file
    a. Use space as the delimiter to separate the line component

  b. Separate fields like IP, Access time, date, method, referrer, agent, URL. Remove open and closed brackets in the time field
  c. From the referred field, remove “ ”

4. Close the file
5. Remove the irrelevant data such as .jpg, .gif, .css, .wav, .png, etc., from the log files what doesn't need for analysis.

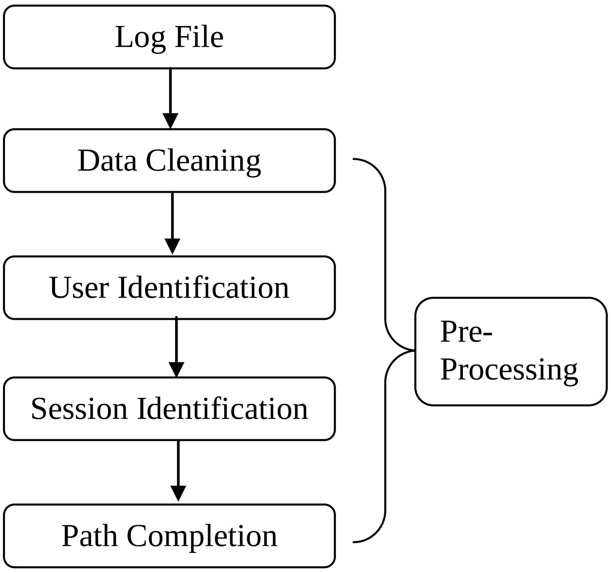


Figure 3: Pre-processing

1.2.2 User Identification

The new user can be identified based on the IP address assigned and the Web agent used by the user. Unfortunately, if both are same, the new user can be identified based on the referral pages what they used for searching the information [5]. User’s requests may not be directed towards prescribed Web pages. Sometimes it may refer some reference pages to reach the target page. Figure 4 shows the user access details of the web access log file.

1.2.3 Session Identification

To find the number of sessions based on the login and the logout time of the user access, session identification is used which finds the various user sessions from the Web access log file [6]. It also finds the number of pages visited by the user in a single session.

1.2.4 Path Completion

It is the process of identifying the reference visited to access the Web page [3]. This helps to find whether the web pages are directly accessed or accessed through any reference pages. At stated earlier not the entire request can access the corresponding Web page directly. Path completion, which helps to acquire the complete user access path. The uncompleted access path of every user session which is recognized based on the user session identification [7].

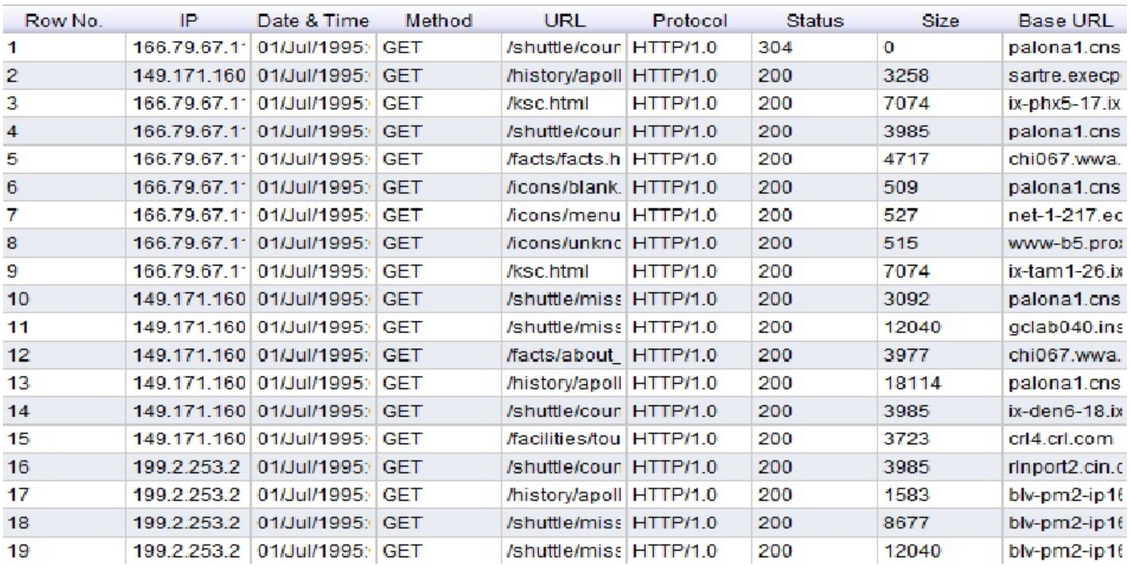
The pre-processing helps to remove the unwanted click-streams of user from the log file and also it reduces the original log file by 40-50% [7]. Initially the input dataset contains more than 10,48,576 entries in the log file. After preprocessing, the data entries are reduced to less than 10,000. Table 1 gives the details of the data (i.e.,) after removing the unwanted data from the Web log file. It shows the result of log file after preprocessing was carried out.

This paper gives the detailed performance analysis of two centroid based clustering algorithms, such as K-means and FCM. For performance analysis, these two clustering algorithms are applied onto the pre-processed dataset. Remaining section of this work is structured as follows: Section 2 describes related works of Web Usage Mining using RapidMiner tool and Section 3 gives the details about the tool RapidMiner. In Section 4, the different similarity measures such as User similarity and Session similarity are discussed. Section 5 and 6 illustrates the K-means and FCM clustering algorithms. In section 7, results and discussion are described. Section 8 describes the conclusion and future work.

2 Related Works

Web Usage Mining is one of the mining techniques used for effective personalization of Web Pages. There are number of machine learning methods available, such as classification, clustering etc... to analyze and personalize the web usage data. Before applying these machine learning methods, the data should undergone the preprocessing methods like Data Cleaning, User Identification, Session Identification and Path completion [3], [5] - [7] to improve the quality of Input file. However, Attribute Identification, Attribute Selection and Outlier Detection are some other preprocessing techniques can be applied by [8] using the RapidMiner tool. After preprocessing there are two types of similarities (i.e. User Similarity, Session Similarity) can be calculated [9] - [10]. Based on these two similarities, Web Usage data are clustered using k-means clustering [11] and FCM clustering [12] - [13] algorithms. Here, the process of web mining is carried out using by RapidMiner tool. Similarly, the RapidMiner tool is used for E-commerce Client – Server Architecture [14] and Distributed computer framework [15]. In ref. [16], the RapidMiner and Weka tools are analyzed for the usage of Educational data mining system. In [17], a methodology for student management system based on the attributes such as status, sex, nationality, etc... are analyzed using RapidMiner tool.

Table 1: Log file after pre-processing



Row No.	IP	Date & Time	Method	URL	Protocol	Status	Size	Base URL
1	166.79.67.1	01/Jul/1995:	GET	/shuttle/cour	HTTP/1.0	304	0	palona1.cns
2	149.171.160	01/Jul/1995:	GET	/history/apoll	HTTP/1.0	200	3258	sartre.execp
3	166.79.67.1	01/Jul/1995:	GET	/ksc.html	HTTP/1.0	200	7074	ix-phx5-17.ix
4	166.79.67.1	01/Jul/1995:	GET	/shuttle/cour	HTTP/1.0	200	3985	palona1.cns
5	166.79.67.1	01/Jul/1995:	GET	/facts/facts.h	HTTP/1.0	200	4717	chi067.wwa.
6	166.79.67.1	01/Jul/1995:	GET	/icons/blank.	HTTP/1.0	200	509	palona1.cns
7	166.79.67.1	01/Jul/1995:	GET	/icons/menu	HTTP/1.0	200	527	net-1-217.ec
8	166.79.67.1	01/Jul/1995:	GET	/icons/unknc	HTTP/1.0	200	515	www-b5.pro:
9	166.79.67.1	01/Jul/1995:	GET	/ksc.html	HTTP/1.0	200	7074	ix-tam1-26.ix
10	149.171.160	01/Jul/1995:	GET	/shuttle/miss	HTTP/1.0	200	3092	palona1.cns
11	149.171.160	01/Jul/1995:	GET	/shuttle/miss	HTTP/1.0	200	12040	gclab040.ins
12	149.171.160	01/Jul/1995:	GET	/facts/about_	HTTP/1.0	200	3977	chi067.wwa.
13	149.171.160	01/Jul/1995:	GET	/history/apoll	HTTP/1.0	200	18114	palona1.cns
14	149.171.160	01/Jul/1995:	GET	/shuttle/cour	HTTP/1.0	200	3985	ix-den6-18.ix
15	149.171.160	01/Jul/1995:	GET	/facilities/tou	HTTP/1.0	200	3723	crl4.crl.com
16	199.2.253.2	01/Jul/1995:	GET	/shuttle/cour	HTTP/1.0	200	3985	rlnport2.cin.c
17	199.2.253.2	01/Jul/1995:	GET	/history/apoll	HTTP/1.0	200	1583	blv-pm2-ip1
18	199.2.253.2	01/Jul/1995:	GET	/shuttle/miss	HTTP/1.0	200	8677	blv-pm2-ip1
19	199.2.253.2	01/Jul/1995:	GET	/shuttle/miss	HTTP/1.0	200	12040	blv-pm2-ip1

Kalpesh Adhatrao et.al. applied ID3 (Iterative Dichotomiser 3) in RapidMiner to generate Decision trees and C4.5 Classification algorithm on students dataset, to calculate the individual performance of students [18]. In [19], frequent visit of user visits into the websites are analyzed using Association rule mining (FP-Growth) which is associated with RapidMiner tool. Hilda Kosorus et.al. compared the functionalities of R, Weka and RapidMiner tool on to the dataset sensor data for structural health monitoring [20].

3 Rapidminer Tool

Recently, there are number of Data mining, statistical computing tools are developed and applied successfully on various data to analyze and monitoring the process of it [20]. RapidMiner project was started in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Group of Katharina Morik at the Dortmund University of Technology [1]. RapidMiner is one of the Data mining tools used to analyze the web accessed information. It is used for research, education, rapid prototyping, application development, and industrial applications [21]. It is an Open Source licensed application, which includes data cleaning, data transformation, optimization, validation and visualization. The visualization contains viewing the analyzed data in the form of scatter plot, Bar, Pie chart, etc… It also includes the various clustering and classification algorithms to do the analytical process. One of the main feature of this tool that, it will analyze data without any program coding, however if anybody wants to analyze the data with their own coding then it can also included in the tool. Different types of datasets can be imported in the RapidMiner tool such as, excel, csv, xml, arff, access etc… Since 2007, RapidMiner has been heavily extended and become one of the most important data mining and data analysis tool [1].

For this analysis process the access log file is collected from the Internet in the time period of 01-07-1995 to 28-07-1995. Following Table 2 gives the details of user's access information of that log file.

From Table 2 there are 10,939 failed requests are recorded. Table 3 denotes the error types of requests made by the user. Figure 5 shows the proposed architecture of this work.

4 Similarity Measures

Similarity measures are to be performed previously inorder to solve the pattern reorganization problem such as Clustering, Classification and Information retrieval process. The term "Similarity / Distance measure" is defined as the distance between the pair of objects. It is applicable to compare two probability density functions which are then reviewed and categorized in both semantic and syntactic relationships [22].

There are number of distance measures proposed and applied, such as Cosine similarity, Dice Similarity, Jaccard Similarity, Euclidean Distance, Manhattan Distance etc… Sometimes similarity is often attained in terms of dissimilarity or distance [23]. From the similarity measures, a tiny distance between objects are referred as high degree of similarity data and bulky distance refers to low degree of similarity data.

Table 2: Web access log file details

Hits	
Total Hits	1,886,571
Visitor Hits	1,886,571
Spider Hits	0
Average Hits per Day	67,377
Average Hits per Visitor	11.63
Cached Requests	132,360
Failed Requests	10,939
Page Views	
Total Page Views	611,295
Average Page Views per Day	21,831
Average Page Views per Visitor	3.77
Visitors	
Total Visitors	162,264
Average Visitors per Day	5,795
Total Unique IPs	81,758
Bandwidth	
Total Bandwidth	35.93 GB
Visitor Bandwidth	35.93 GB
Spider Bandwidth	0 B
Average Bandwidth per Day	1.28 GB
Average Bandwidth per Hit	19.97 KB
Average Bandwidth per Visitor	

In our proposed work, the similarities are measured based on the following criteria such as User similarity measurement, Session similarity measurement.

Table 3: Error types

S.No	Error	Hits
1	404 Not Found	10,805
2	500 Internal Server Error	62
3	403 Forbidden	53
4	501 Not Implemented	14
5	400 Bad Request	5
	Total	**10,939**

4.1 User Similarity Measurement

Computing and Comparing of user profiles are very essential task to be performed for personalizing web pages [9]. User similarity is the process of identifying the similar access of the Web pages done by the different users. As mentioned earlier the new user can be identified based on the IP address and the agent, what the users used for accessing the Web pages. Here the user similarity measure is calculated between the users and the web pages visited by them based on the mixed Euclidean distance. Mixed Euclidean distance can be calculated by the following equation

$$(n)(n-1)/2 \tag{1}$$

Figure 6 shows the user similarity measurement of the web access log file given as an input.

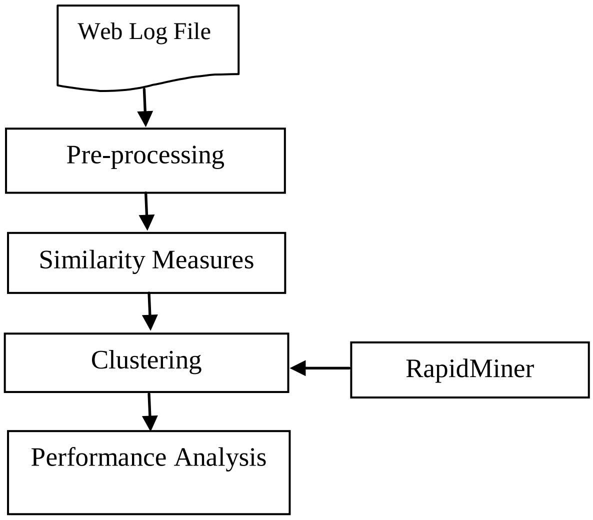


Figure 5: Proposed Work Architecture

4.2 Session Similarity Measurement

The web access log data contains the details of every request of the server, and the sequence of sessions [10]. Session similarity is defined as the similar search done by the user between two web sessions. Consider n number of sessions $S = \{s_{1,}s_{2,\ldots\ldots}s_n\}$ accessed the m number of Web pages $P = \{p_{1,}p_{2\ldots}p_m\}$ during the period of time. In this section we are going to find out the similarity of web pages accessed by the user during n number of Sessions. Figure 7 shows the session similarity measurement of a Web access log file. The requests made by the user not always directly reach the exact page. Sometimes, it refers some other pages (i.e. referral pages) to reach the target page.

HTTP referrer pages are the header files used to identify the address of the webpage or link to the resource being requested [24].

Normally, the referral pages are used for statistical and promotional purpose. The following figure 8

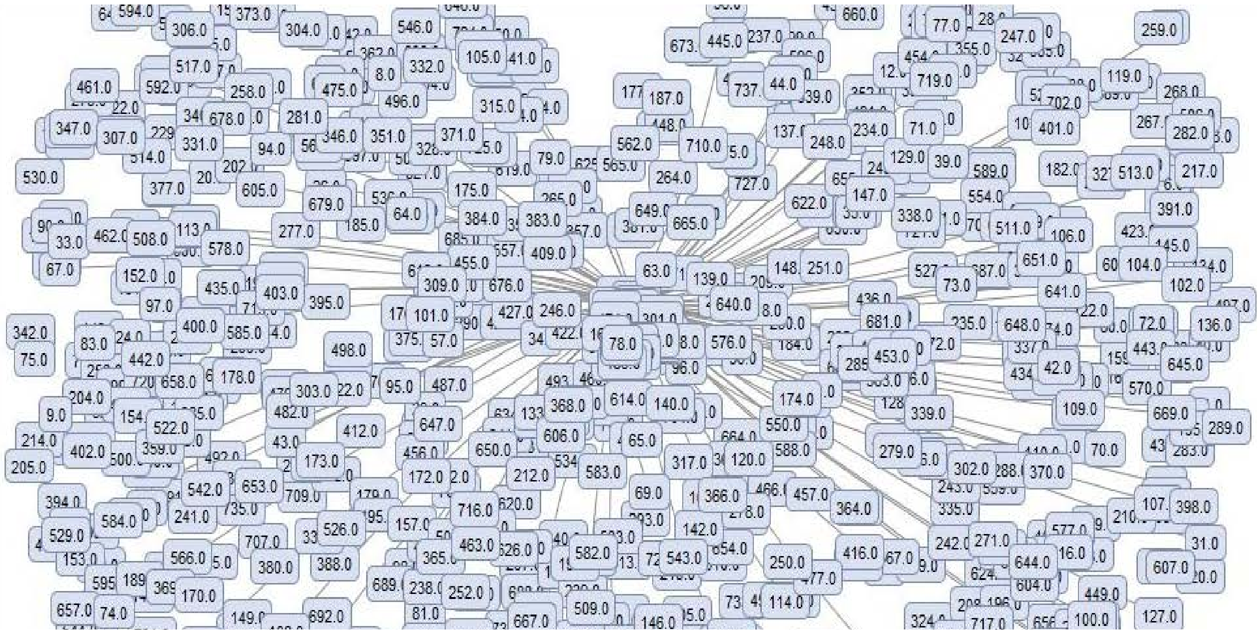


Figure 6: User Similarity Measurement

shows the number of referral pages used to access the Web page.

5 K-Means Clustering

Clustering is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters. There are many clustering algorithms such as *k*-means, *k*-medians, DBSCAN, Hierarchical clustering and *X*-means are available to do the clustering process. To analyze the web access log by any algorithm it must be preprocessed.

After finding the User similarity and Session similarity the clustering algorithm will be apply on to the data. Clustering the data is carried out based on the similar search made by the User. Here using *k*-means algorithm to group the data based on the similarity search done by the user. *k*-means clustering is an exclusive clustering algorithm i.e. each object is assigned to precisely one of a set of cluster. Objects in one cluster are similar to each other.

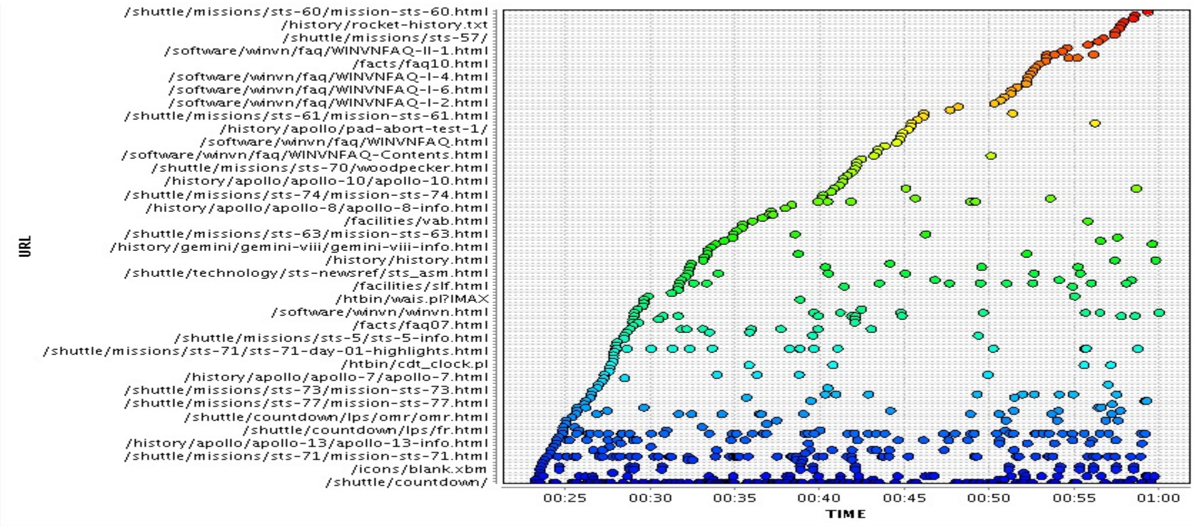


Figure 7 Session Similarity Measurement

The similarity between objects is based on a measure of the distance between them. Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. Clustering is a technique for extracting information from unlabelled data. Clustering is very useful in many different scenarios e.g. in a marketing application, user may be interested in finding clusters of customers with similar buying behavior [11].

In *k*-means clustering, centroid (center of cluster) has to be find out first. The centroid is calculated using the Euclidean distance and it is calculated by equation 1. Sometimes the centroid is one of the points in the cluster. Then, assign the *k* value for clusters are needed to be process. Generally small integer value can be assigned for *k* value. Then select *k* objects in random manner and use this as the initial set of *k* centroids. Allocate each object to the cluster, which is nearest to the centroid and recalculate the centroids of the *k* clusters. Repeat calculating centroids until the centroid may be an optimal. Figure 9 shows the clustering diagram for the given web access log file.

6 Fuzzy c Means Clustering

Fuzzy c-means (FCM) is a method of clustering which allows one part of data belonging to two or more clusters. This method is developed by Dunn in 1973 and improved by Bezdek in 1981. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature of analysis, clustering and classifier design. FCM is broadly applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis [12]. The main objective of this FCM is to cluster n number of objects into c clusters [13]. This algorithm works by assigning relationship to each data point equivalent to each cluster centre on the basis of distance between the data point and the Cluster. More the data is close to the cluster centre it becomes the member of that cluster. Obviously, summation of relationship of each data point should be equal to one.

Assume that the sample collection is $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in P^s$ (where n is the number of patterns and s is the dimension of attributes) and the goal is to divide it into *c* group with cluster centre

c_i $(i = 1, 2, \ldots, c)$ and minimize the objective function, where to find the cluster centre (c_i)

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (2)$$

The distance between two vectors are find out based on Euclidean Distance e_{ij} represents distance between the i^{th} cluster centre and j^{th} data point.

$$e_{ij} = \| c_i - x_j \| \qquad (3)$$

To find the membership matrix U_{ij} consider the following

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{e_{ij}}{d_{kj}} \right)^{2/(m-1)}} \qquad (4)$$

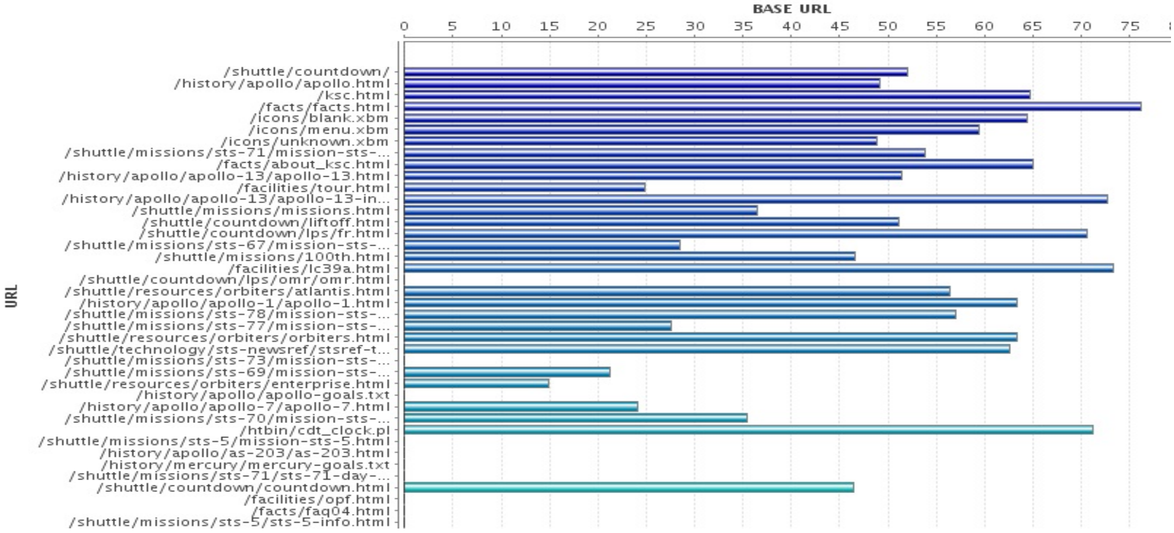


Figure 8: Referral Pages

Repeat calculating u_{ij} until getting the less dissimilarity values. The following equation is used to find the dissimilarity matrix cost function

$$J(U, c_1, \ldots.., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} e_{ij}^{2} \quad (5)$$

7 Results and Discussion

The main objective of this work is to analyze the web usage data by applying machine learning techniques such as K-means and FCM clustering algorithms using RapidMiner tool. To perform the analysis, web access log data has been collected through Internet. The file contains the sequence of user access details. In both clustering algorithms, initially the cluster centre randomly chosen [13] based on the cluster centroid and the clustering of data can be evaluated.

Table 4: Centroid Comparison of K-means and FCM

	Cluster 0		Cluster 1	
	FCM	K-means	FCM	K-means
Cluster Centroids	0.215971	0.138374	0.163095	0.126122
	0.204783	3.138650	0.761130	2.960204
	0.799673	0.348911	0.210909	0.271837
	0.743030	2.000000	0.757697	3.037143
	0.459701	0.839419	0.437498	0.789592

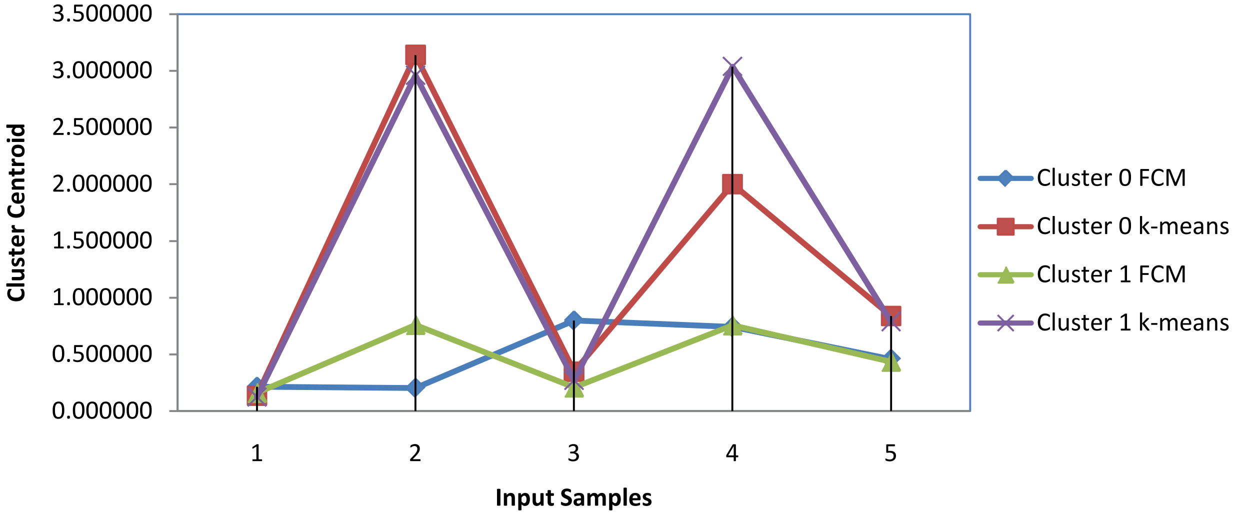


Figure 9 Performance analysis of K-means and FCM

In this work, the user's activities are characterized as user similarity and session similarity. Based on the user and session similarities the two clusters are formed. The experimental results are tabulated in Table 4, which gives the difference between K means and FCM clustering algorithms. In Table 4, cluster 0 and cluster 1 represents the cluster centroid of user similarity and session similarity measurements, respectively. Based on Table 4 the graph is plotted between the datasets and cluster centroid achieve in both K-means and FCM clustering algorithm. With the help of the graph, it is clearly identify the working background of both K-means and FCM clustering algorithms.

8 Conclusion And Future Work

In this work, the analysis of web usage data by applying two different clustering algorithms such as K-means and Fuzzy C means in web usage based dataset using the tool RapidMiner was performed.

In this work, the important steps such as preprocessing, similarity measurement and clustering methods are used for improve the input efficiency, find out user and session similarities and grouping the similar data respectively. The related data are grouped based on the cluster centroid and also the experimental result shows the performances of both clustering algorithms. This work mainly spotlight on grouping the similar data based on two similarities such as user similarity and session similarity and may useful to the web users to achieve the better access of information through internet. This work can be extended by extracting the information based on the IP address and it may be give clear idea about the websites for individual visitors.

Acknowledgement

The first author thank to University Grants Commission for providing financial support for his research work under the grant Rajiv Gandhi National Fellowship from the academic year 2013-14.

References

[1] Nico Schlitter, and Jorg Lassig, Distributed Data Analytics using RapidMiner and BOINC Zittau/Gorlitz Proceedings of the 4th rapidminer community meeting and conference (RCOMM 2013), 2013, pp 81-97.

[2] Jae Jeung Rho, Byeong-Joon Moon, Yoon-Jeong Kim and Yoon-Jeong Kim, Internet Customer Segmentation Using Web Log Data South Korea Journal of Business & Economics Research, November, Volume 2, Number 11, 2004, pp 59-74.

[3] Chitraa V. and Dr.Antony Selvdoss Davamani, A Survey on Pre-processing Methods for Web Usage Data, International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010, pp 78-83.

[4] Rajesh Shukla, Sanjay Silakari and P K Chande, Web Personalization Systems and Web Usage Mining: A Review, International Journal of Computer Applications, Volume72, No.21, 2013, pp 6-13.

[5] Mona S. Kamat, Bakal J. W. and Madhu Nashipudi, Optimization of Web Preprocessing in Web Usage Mining, International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering, Volume-2, Issue-6, 2013, pp.167-173.

[6] Naga Lakshmi, Raja Sekhara Rao and Sai Satyanarayana Reddy, An Overview of Pre-processing on Web Log Data for Web Usage Analysis, International Journal of Innovative Technology and Exploring Engineering, Volume-2, Issue-4, 2013, pp 274-279.

[7] Dipa Dixit and Kiruthika M, Preprocessing of Web Logs, International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010, pp. 2447-2452.

[8] Jianchao Han, Rodriguez, J.C. and Beheshti, M, Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner, IEEE Second International Conference on Future Generation Communication and Networking, Vol.3, 2008, pp 96-99.

[9] Giovanna Castellano, Maria Fanelli A, Corrado Mencar and Alessandra Torsello M, Similarity-based Fuzzy clustering for user profiling, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, 2007.

[10] K.Duraiswamy and Valli Mayil V, Similarity Matrix Based Session Clustering by Sequence Alignment Using Dynamic programming, Journal of Computer and Information Science, Vol.1, No.3, 2008, pp 66-72.

[11] RapidMiner – User Manual 5.0, Retrieved March10, 2014 from http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdnacdn.com/wpcontent/uploads/2013/10/rapidminer-5.0-manual-english_v1.0.pdf.

[12] Tejwant Singh, Manish Mahajan, Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering, Volume4, Issue 5, 2014, pp 89-93.

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

[14] Daniel Hunyadi, Rapid Miner E-Commerce, 12th WSEAS International Conference on Automatic Control, Modelling & Simulation, 2010, pp 316-321.

[15] Alexander Arimond, Christian Kofler, and Faisal Shafait, Distributed Pattern Recognition in RapidMiner, German Research Center for Artificial Intelligence, RapidMiner Community Meeting and Conference, Dortmund, Germany, 2010.

[16] Anduela Lile, Analyzing E-Learning Systems Using Educational Data Mining Techniques, Mediterranean Journal of Social Sciences, Vol. 2, No.3, 2011, pp 403-419.

[17] Jovica Krstevski, Dragan Mihajlov and Ivan Chorbev, Student Data Analysis with RapidMiner, Springer Chorbev ICT Innovations 2011, 2012, pp 19-28.

[18] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, Predicting Students' Performance using ID3 and C4.5 Classification Algorithms, International Journal of Data Mining & Knowledge Management

Process, Vol.3, No.5, September 2013, pp 39-52.

[19] Shiva Asadianfam and Masoud Mohammadi, Identify Navigational Patterns of Web Users, International Journal of Computer-Aided technologies (IJCAx), Vol.1, No.1, April 2014, pp 1-8.

[20] Hilda Kosorus, Jurgen Honigl and Josef Kung, Using R, Weka and RapidMiner in Time Series Analysis of sensor Data for Structural Health Monitoring, 22nd International Workshop on Database and Expert Systems Applications, 2011, pp 306-310.

[21] Jedrzej Potoniec and Agnieszka Lawrynowicz, RMonto: Ontological Extension to RapidMiner, 10th International Semantic Web Conference, 2011.

[22] Sung-Hyuk Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, International Journal of Mathematical Models and Methods in Applied Sciences, Issue 4, Volume 1, 2007, pp 300-307.

[23] Anna Huang, Similarity Measures for Text Document Clustering, NZCSRSC, Christchurch, New Zealand, 2008.

[24] ”HTTP referrer” Retrieved June 11 2014 from http://en.wikipedia.org/wiki/HTTP_referer.