iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VII **Month of publication:** July 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Implementation of The K-Means Clustering Algorithm to Analyze the User Interest by Analyzing the University Web Log Servers

Anmol Kaur[1], Raman Maini[2]
[1]*Student, Department of Computer Engineering Punjabi University, Patiala Punjab, India,*
[2]*Professor, Department of Computer Engineering Punjabi University, Patiala Punjab, India*

Abstract: ***Web Usage mining is considered as one of the very important category of the web data mining, which manages the extraction of useful and interesting data from the web log documents. Web utilization has turned out to be the essential part among the young people. The teenage group peoples are inexorably open and shared the online and social culture, which enables them to get data and keep up kinships and connections. But, however they are the very crucial young stage, which sometimes can lead to risky and immature decisions. To keep check on them we analyze the user behavior to predict their interests. The web documents are collected from the university computer center. In this work, K-means clustering algorithm is used. The clusters are formed according to the frequency of the sites visited, and then the analysis is done to judge the interest of the student. It has been find out that most of the student's interest falls in the category of the Information technology.***
Keywords: Web Usage Mining, K-Means clustering algorithm, web logs, user behavior, Mean square.

I. INTRODUCTION

In today's era of science and technology the World Wide Web has become a necessary medium to disseminate the information each and everywhere. Due to its giant nature it is scope is very vast. The wide level has resulted into abundant amount of information is available for all those who use internet for their different purposes [2]. In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task. 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.

The knowledge can be spread through World Wide Web. Internet has become an important part of our today's busy schedule. It has turned out the manner of working business, education handling the organization etc. The Web is large collection of information which is huge and dynamic in nature. That's why the complexity also increases to handle this abundant data. Like students want to examine the answer about the topics of study, business mind people like to analyze the customer's requirements. Everyone wants techniques to meet their needs. Mining can be implemented to find the Data Mining tools to find the required knowledge from internet. This gathered information is apply to obtain more command and observe the information make forecast, what would the right option and the fair appeal to move go ahead [3].

Web mining is mainly used for the extraction of the valuable data from the documents of the web. It basically includes the following these three categories named web structures mining, Web content mining and web usage mining. Web usage mining is considered as the widely used category and significant type of web data mining which is used to discover the useful data concealed in the browsing pattern of the web [9].

It is defined as the step by step approach which consists of stages defined processing of the data, discovery of the patterns and then analyzes the discovered patterns. The logs are preprocessed to remove the irrelevant and unwanted data. After this the various data mining techniques are used to find the interesting patterns which were hidden earlier. Then, the final stage is the analysis of the patterns to validate the related patterns [1].

The first step is preprocessing step, there is cleaning of the data and then it is divided into the various sets of the transactions of the user when they visit the different sites. Some other factors related to the knowledge about the content of site and the structure of site also included in the preprocessing to improve the data of the transactions [5] [13]. Second step is the pattern discovery step includes the statistical database and machine learning operations are executed to acquire the hidden patterns which reflects the user behavior

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

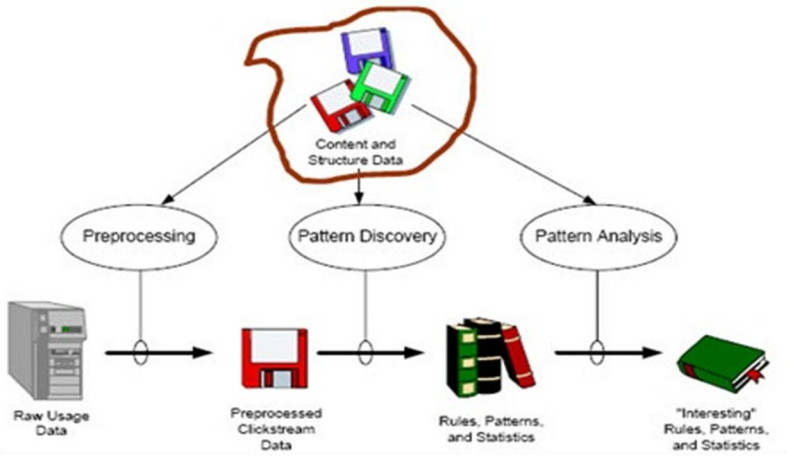


Fig 1: the web usage mining process [17]

The final step includes the patterns which were discovered and the processing of the statics, filtering which results in such type of models that are used as the input source to various applications like recommendation engines, tools for visualization, and the tools used for making the reports [8].

A. Remote host: Is the remote hostname or its IP Address [3].
B. Log name: Is the remote log name of the client;
C. Username: Is the username with which the client has verified himself,
D. Date: Is the date and time of the demand, demand: is the correct demand line as it came from the customer,
E. Status: Is the HTTP status code come back to the customer, and
F. Bytes : Is the substance length of the report exchanged [5].

Authenticated User	Bytes	Date and Time
Remote IP Addres	Modes of Request	Number of hits on Page
Remote URL	Request	Request URL

Fig 2: practical outlines of web log documents [3]

II. LITERATURE REVIEW

A. Neelima and Sireesha Rodda[2]: Implemented these three phases namely Data cleaning, User identification, Session identification**.** Depending upon the frequency of users visiting each page mining is performed. By finding the session of the user we can analyze the user behavior by the time spend on a particular page.
B. V.Anitha and P.Isakki [1]: Gives a consideration on Web use mining to anticipate the conduct of web clients in view of web server log documents. Clients utilizing pages, a regular get to ways and continuous get to pages, connections are put away in web server log records. A Web log alongside the uniqueness of the client catches their perusing conduct on a site and talking about with respect to the conduct from investigation of various calculations and diverse techniques.
C. Ashwini Ladekar [3]: The paper examines about client behavior with the help of usage mining innovation, web log information which is the wellspring of data from the framework the algorithm used is the apriori calculation to find the interests. The paper likewise presents about how the estimation of client conduct is done in light of the investigation of web logs.
D. Anurag kumar and Vaishali [4]: Focused on the client Behaviors utilizing web server log document expectation utilizing web server log record, click streams record and client data. Clients utilizing website pages, oftentimes went by hyperlinks, every

now and again got to site pages, connections are put away log records of the server. The logs of web alongside the uniqueness of the client catches the user's perusing conduct on the site and talking about with respect to the Conduct from investigation of various calculations and diverse techniques.

E. *Virendra R. Rathod and Govind V.Patel [7]*: Behavior utilizing web log document forecast utilizing records of the log, records of the click streams and client data. There are two distinctive bunching methods, specifically Fuzzy C-Means Clustering calculations and Markov shows are explored to foresee the website page which can be gotten to later on in view of the past activity of programs conduct.

F. *Amit Pratap Singh [6]***:** A web log document contains noisy information, through preprocessing step; we wipe out unessential information so this progression is important in web mining. In the wake of preprocessing example disclosure and investigation stage assumes an imperative part to distinguish criminal movement and foresee the speculated client conduct.

G. *Zakaria Suliman Zubi and Mussab Saleh Riani [5]:* Concentrate on the utilization of web mining strategies to arrange website pages sort as indicated by client visits. This arrangement pushes us to comprehend the client conduct. Additionally we will utilize some grouping and affiliation governs procedures for finding the potential information from the perusing designs**.**

III. IMPLEMENTATION

From the research gaps identified in the literature review, it has been concluded that the problem of finding the behavior or interest by analyzing the web server log files of the student is a strenuous task. Sometimes students are confused to infer about their scrutiny. From the following research one can find out their interests, after inspecting the web server log files. There is a big challenge to keep eye on the students to check what they are surfing whole day on the internet and find their behavior. There is benefit for the network administrator to block some of the sites if the student is visiting undesirable sites. For example if the student interest falls in hacking, then the network administrator can unrent the permissions to visit the particular sites.

The proposed framework aims to find the student behavior with the help of the web log records. The data is collected from the university computer center. In this recommendation engine, the customer asks for the page and connections are stored in the web log documents various fields like username, time and date, sites, categories are selected from the information of the logs. Utilizing web links which are frequently visited by the user, though the use of this information we can find or recognize the actual interest of the user and recognize which sites are regularly visited by the particular user with in the college hours and also check which site is commonly used by all the students and belong it to which category whether it belongs to social media, education, portals etc. This system is used to foresee the interests of the university students. To keep check on them we analyze the user behavior to predict their interests. The web documents are collected from the university computer center. In this work, K-means clustering algorithm is used. The clusters are formed according to the frequency of the sites visited, and then the analysis is done to judge the interest of the student. It has been find out that most of the student's interest falls in the category of the Information technology.

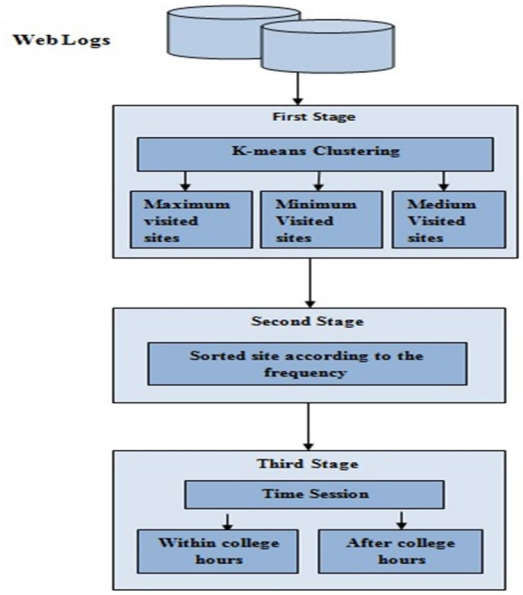


Fig 3: Steps to implement the algorithm [10]

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

IV. ALGORITM USED AND TOOL USED

In this work, the K-Means Clustering algorithm is used and the tool used is R tool to find the behavior of the student with the help of session whether the student using the internet in the college duration or after the college duration.

A. K-Means Clustering Algorithm

K- Means Clustering Algorithm is defined as the easiest and the simple unsupervised learning algorithm which is used to solve the problems of the clustering. The whole process is simple to categorize the given data into the number of clusters [11]. The basic idea is to identify the K centroids, basically one centroid for the each cluster. As the different location of centroid leads in variation of the results, so these should be chosen very carefully. It is better that they should be placed at some distances. Then choose that point which belonged to the data set and relate it to that centroid which is nearest to it [14]. When all the points are grouped, there is completion of step 1.Now calculate again the new K centroids defined as the barrycenter. The same process is followed to generate the new clusters. As a result there is generation of loop and stops when the possibility to generate the new center is almost negligible. The algorithm uses the minimizing objective function. The Objective function is defined as [12]:

$$J(v) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\|x_i - v_j\|\right)^2 \qquad 4.1$$

Where

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j. 'c_i' is the number of data points in i^{th} cluster. 'c' is the number of cluster centers.

B. Steps for the algorithm

Let $X = \{x_1,x_2,x_3,\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers

1. Select the arbitrarily 'c' centers of the given cluster.
2. The gap between the each point of data and the center of the given cluster is computed.
3. Allocate the points in such a way that whose gap is least from the center of the cluster.
4. The recalculation of the cluster center which is formed new can be done by using the equation

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i \qquad 4.2$$

5. Now, calculate again the centers of the clusters which are newly formed.
6. When there is no choice of reassignment of the data points, apart from this, step3 would be repeated.

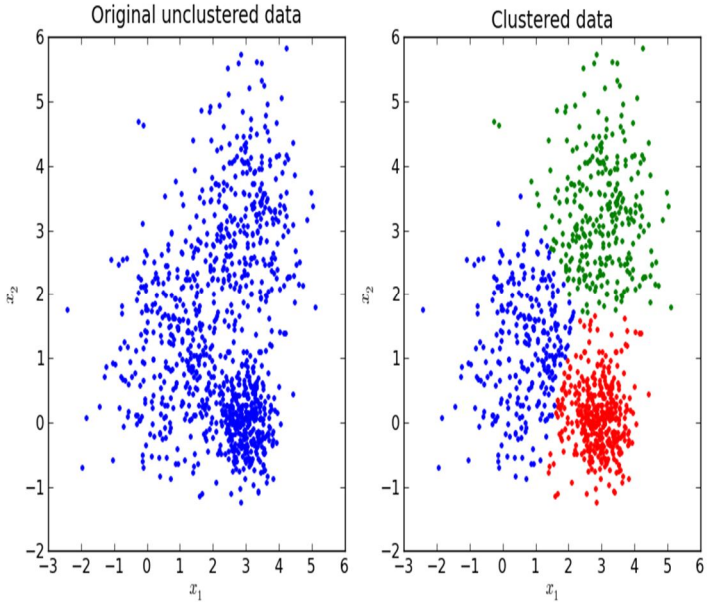


Fig 4: K-means Clustering Algorithm [15]

C. Tool Used: R-Tool

R is a device for measurements and information displaying. It is graceful, flexible and highly expressive syntax which works with the data. R has more capabilities like extraordinary graphical abilities. It is available free under the GNU General Public License and pre-compiled versions are provided for the various operating systems like windows, Mac and LINUX. Some important features of R are following [16]:

R is simple, well furnished and productive conditions, loops and the various facilities of the input and the output.

R assigns the good suite for the operators to calculate the arrays, vectors, lists and the matrices.

It has better facility for handling the data and better facility.

Here RStudio is used for the implementation. RStudio is the free of cost and the open source integrated development environment (IDE) for the R. It has the capabilities for the graphical mode and for the statistical computing.

1) Advantages of R-Tool

a) It is free and open source software, anyone can easily use.

b) It is most widely used for statistical analysis.

c) R can well import the data from import the data from other tools likes SAS, SPSS, Microsoft Excel, My SQL, Microsoft Access etc.

d) It has the capability to make graphical output in JPEG, PNG, PDF and also tabular output in LATEX HTML.

2) Disadvantages of R-tool

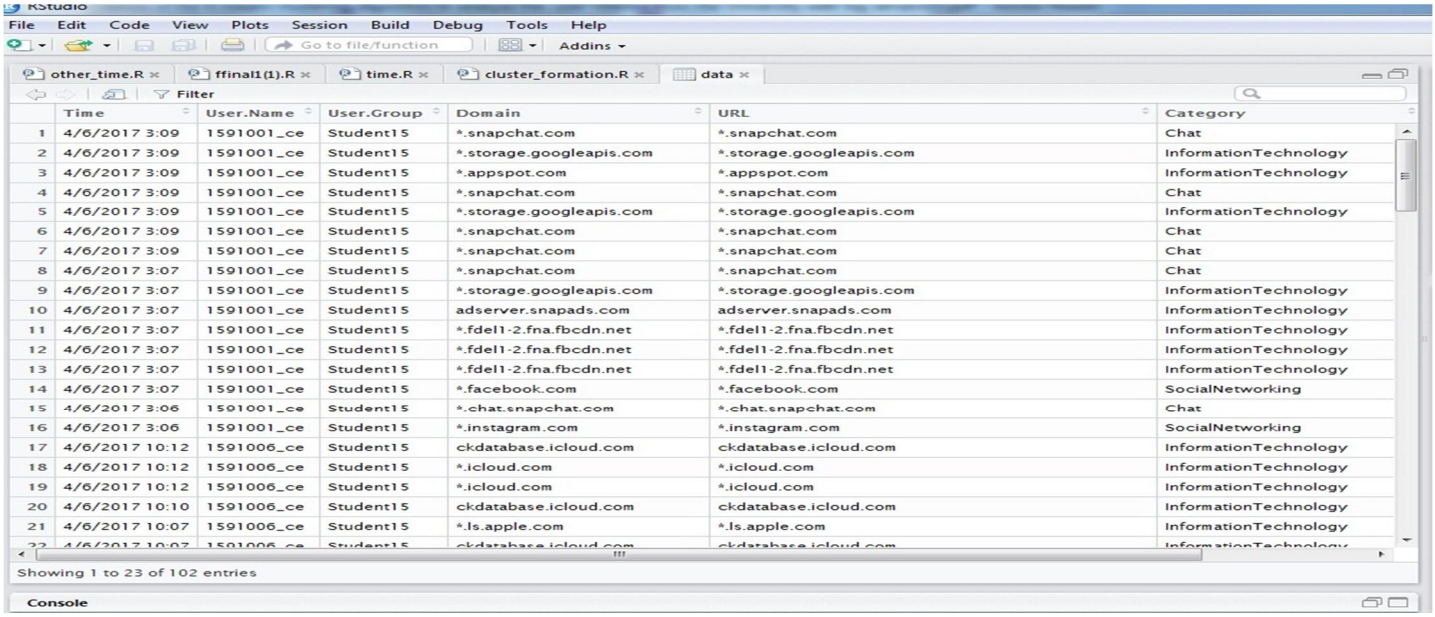
a) Consumes more memory due to large number of commands.

b) Sometimes document is sketchy and pithy and invulnerable to the non analyst.

V. RESULTS AND DISCUSSIONS

The database i.e web server log files are collected from the university computer center. There are total 58536 entries. It is very difficult to understand the results with this huge amount of data. So we take only few entries for the better understanding. The students are engaged in browsing the internet whole day. In the whole day they browse the different sites of various categories and this information is stored in the form of log files. This helps us to analyze the client interest. K-Means clustering is used to form the clusters and time is differentiated whether in the college hours the student is using the internet or not. Then according to the maximum frequency of the particular site, we can find the interest of the student.

A. Firstly import the data from the excel sheet in the proposed tool.



	Time	User.Name	User.Group	Domain	URL	Category
1	4/6/2017 3:09	1591001_ce	Student15	*.snapchat.com	*.snapchat.com	Chat
2	4/6/2017 3:09	1591001_ce	Student15	*.storage.googleapis.com	*.storage.googleapis.com	InformationTechnology
3	4/6/2017 3:09	1591001_ce	Student15	*.appspot.com	*.appspot.com	InformationTechnology
4	4/6/2017 3:09	1591001_ce	Student15	*.snapchat.com	*.snapchat.com	Chat
5	4/6/2017 3:09	1591001_ce	Student15	*.storage.googleapis.com	*.storage.googleapis.com	InformationTechnology
6	4/6/2017 3:09	1591001_ce	Student15	*.snapchat.com	*.snapchat.com	Chat
7	4/6/2017 3:09	1591001_ce	Student15	*.snapchat.com	*.snapchat.com	Chat
8	4/6/2017 3:07	1591001_ce	Student15	*.snapchat.com	*.snapchat.com	Chat
9	4/6/2017 3:07	1591001_ce	Student15	*.storage.googleapis.com	*.storage.googleapis.com	InformationTechnology
10	4/6/2017 3:07	1591001_ce	Student15	adserver.snapads.com	adserver.snapads.com	InformationTechnology
11	4/6/2017 3:07	1591001_ce	Student15	*.fdel1-2.fna.fbcdn.net	*.fdel1-2.fna.fbcdn.net	InformationTechnology
12	4/6/2017 3:07	1591001_ce	Student15	*.fdel1-2.fna.fbcdn.net	*.fdel1-2.fna.fbcdn.net	InformationTechnology
13	4/6/2017 3:07	1591001_ce	Student15	*.fdel1-2.fna.fbcdn.net	*.fdel1-2.fna.fbcdn.net	InformationTechnology
14	4/6/2017 3:07	1591001_ce	Student15	*.facebook.com	*.facebook.com	SocialNetworking
15	4/6/2017 3:06	1591001_ce	Student15	*.chat.snapchat.com	*.chat.snapchat.com	Chat
16	4/6/2017 3:06	1591001_ce	Student15	*.instagram.com	*.instagram.com	SocialNetworking
17	4/6/2017 10:12	1591006_ce	Student15	ckdatabase.icloud.com	ckdatabase.icloud.com	InformationTechnology
18	4/6/2017 10:12	1591006_ce	Student15	*.icloud.com	*.icloud.com	InformationTechnology
19	4/6/2017 10:12	1591006_ce	Student15	*.icloud.com	*.icloud.com	InformationTechnology
20	4/6/2017 10:10	1591006_ce	Student15	ckdatabase.icloud.com	ckdatabase.icloud.com	InformationTechnology
21	4/6/2017 10:07	1591006_ce	Student15	*.ls.apple.com	*.ls.apple.com	InformationTechnology

Fig 5: Structure of Log File

B. Then the proposed algorithm is applied to form the clusters ie maximum visited sites, minimum visited sites and the medium visited sites. Here the minimum visited sites are taken for the illustration.

	index	domain	frequency
1	1	*.appspot.com	1
2	2	*.chat.snapchat.com	1
3	3	*.facebook.com	4
4	5	*.g.doubleclick.net	2
5	8	*.grammarly.com	2
6	9	*.grammarly.io	1
7	11	*.instagram.com	1
8	12	*.ls.apple.com	1
9	13	*.m.taobao.com	1
10	14	*.mail.google.com	1
11	15	*.snapchat.com	5
12	16	*.storage.googleapis.com	3
13	17	*.tanx.com	1
14	18	*.urbanairship.com	1
15	20	abtest.mistat.xiaomi.com	1
16	21	adserver.snapads.com	1
17	22	api.sec.miui.com	1
18	23	cdn.content.prod.cms.msn.com	3
19	24	ckdatabase.icloud.com	3
20	25	configuration.apple.com	1
21	26	data.mistat.xiaomi.com	4
22	27	logupdate.avlyun.sec.miui.com	1

Showing 1 to 23 of 28 entries

Fig 6: Minimum Visited sites

C. K-Means Clustering graph according to the frequency of the user's visited sites.

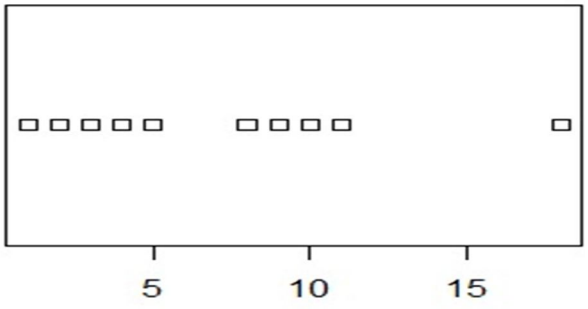


Fig 7 : Density of the Frequency of the user's visited sites

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

D. The students visited sites are sorted according to the frequency of the sites.

	username	site	catagory	ferquency	Cluster
1	1591035_ce	*.fdel1-2.fna.fbcdn.net	InformationTechnology	8	max_ferquency_cluster
2	1591035_ce	*.google.com	SearchEngines	5	max_ferquency_cluster
3	1591035_ce	*.facebook.com	SocialNetworking	3	max_ferquency_cluster
4	1591035_ce	*.wns.windows.com	InformationTechnology	3	max_ferquency_cluster
5	1591035_ce	cdn.content.prod.cms.msn.com	Portals	3	max_ferquency_cluster
6	1591035_ce	*.grammarly.com	EducationAndReferenceMaterial	2	max_ferquency_cluster
7	1591035_ce	*.grammarly.io	InformationTechnology	1	max_ferquency_cluster
8	1591035_ce	*.mail.google.com	gmailallow	1	max_ferquency_cluster
9	1591009_ce	*.googleapis.com	Portals	6	min_ferquency_cluster
10	1591009_ce	*.google.com	SearchEngines	3	min_ferquency_cluster
11	1591009_ce	*.m.taobao.com	Shopping	1	min_ferquency_cluster
12	1591009_ce	*.tanx.com	Shopping	1	min_ferquency_cluster
13	1591009_ce	*.urbanairship.com	InformationTechnology	1	min_ferquency_cluster
14	1591031_ce	*.google.com	SearchEngines	5	min_ferquency_cluster
15	1591031_ce	*.wns.windows.com	InformationTechnology	4	min_ferquency_cluster
16	1591031_ce	*.g.doubleclick.net	Advertisements	1	min_ferquency_cluster
17	1591055_ce	*.google.com	SearchEngines	3	min_ferquency_cluster
18	1591055_ce	*.g.doubleclick.net	Advertisements	1	min_ferquency_cluster
19	1591055_ce	*.wns.windows.com	InformationTechnology	1	min_ferquency_cluster
20	1591055_ce	mail.google.com	gmailallow	1	min_ferquency_cluster
21	1591001_ce	*.snapchat.com	Chat	5	medium_ferquency_cluster
22	1591001_ce	*.fdel1-2.fna.fbcdn.net	InformationTechnology	3	medium_ferquency_cluster

Showing 1 to 23 of 131 entries

Fig 8: Sorting of the students in descending order

E. Students who use the internet in the college duration are finding out along with the duration, category and the frequency.

	x.uuname	x.site_Seen	x.catgg	x.Hours2	freq
1	1591006_ce	4/6/2017 10:07	InformationTechnology	10:07	11
2	1591006_ce	4/6/2017 10:10	InformationTechnology	10:10	1
3	1591006_ce	4/6/2017 10:12	InformationTechnology	10:12	3
4	1591007_ce	4/6/2017 10:27	BusinessAndEconomy	10:27	4
5	1591007_ce	4/6/2017 10:27	InformationTechnology	10:27	2
6	1591007_ce	4/6/2017 10:27	Portals	10:27	2
7	1591007_ce	4/6/2017 10:28	BusinessAndEconomy	10:28	3
8	1591007_ce	4/6/2017 10:29	InformationTechnology	10:29	1
9	1591007_ce	4/6/2017 10:29	SearchEngines	10:29	2
10	1591007_ce	4/6/2017 10:31	InformationTechnology	10:31	1
11	1591007_ce	4/6/2017 10:31	Portals	10:31	1
12	1591007_ce	4/6/2017 10:31	SearchEngines	10:31	1

Fig 9: Students who use internet in college hours

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

F. Then the final step to find the interest of the particular user in the college duration.

	Username	Interest
1	1591006_ce	InformationTechnology
2	1591007_ce	BusinessAndEconomy

Fig 10: Interest of the students who use the internet in the college hours

G. User Interest after the college duration.

	Username	Interest
1	1591001_ce	InformationTechnology
2	1591009_ce	Portals
3	1591031_ce	SearchEngines
4	1591035_ce	InformationTechnology
5	1591055_ce	SearchEngines

Fig 11: Interest of the students who use the internet other than the college hours

Now days, the young stage people are always ready to involve in the combined social activities where they can maintain the kinships and the relationships.To keep check on them the web logs from the university computer center has been collected. The collected log contains the User name, Time, Domain, Category, URL's and IP address of each user. To ensure the privacy of the student the actual roll numbers are converted into the fictitious roll numbers. It has been concluded that firstly the clusters are formed according to the frequency of the sites that are visited by the students. The users with the most visited sited are find out. To find the interest it is important that which sites are commonly visited by all the students. So, here by applying the K-Means algorithm is used to find all the clusters. There is a heavy density on the plotted graph where there is maximum number of users. After the analysis it has been find out that the "*.google.com" is highly visited site by all the students under the category of Search Engine. The sites are sorted in their respective clusters according to the frequency of there usage. The main objective to sort the sites is to find the each user maximum and minimum visited site. We can predict the interest of the user by analyzing the maximum frequency visited site of each user. But, to check what the student is visiting in the college hours, by taking the time constraint then the students are differentiated according to the sites visited in the college hours and after the college hours. Each user category is find to judge their behavior.

Generally the students are involved into the Information technology. After generalizing this student's particular interest can be easil y find out. Here the students which are visiting the other category sites except the information technology and the educational sites i n the college hours, the network administrator can block these sites. There is a big advantage that the network administrator can bloc k all the other category sites rather than the educational sites during the session of the exams. This can lead to increase the higher ed ucational competition among the students because that time only the educational sites are activated. This way student's chance to wa ste their precious time during the exams can be avoided. Hence the students have no other choice to visit the other irrelevant site. Th is is helpful for the students to know about their actual interest. The network administrator can verify about how many students are r eally interested in the education field and how many are interested in other fields. For example, during the college duration's one stu dent found in the Business and Economy. So we can easily predict the actual interest of the student. He can choose further business subject to explore his actual talent. This proposed work is offline. Hybrid framework can be formed in future which includes the co mbination of both the offline and the online mode can be used.

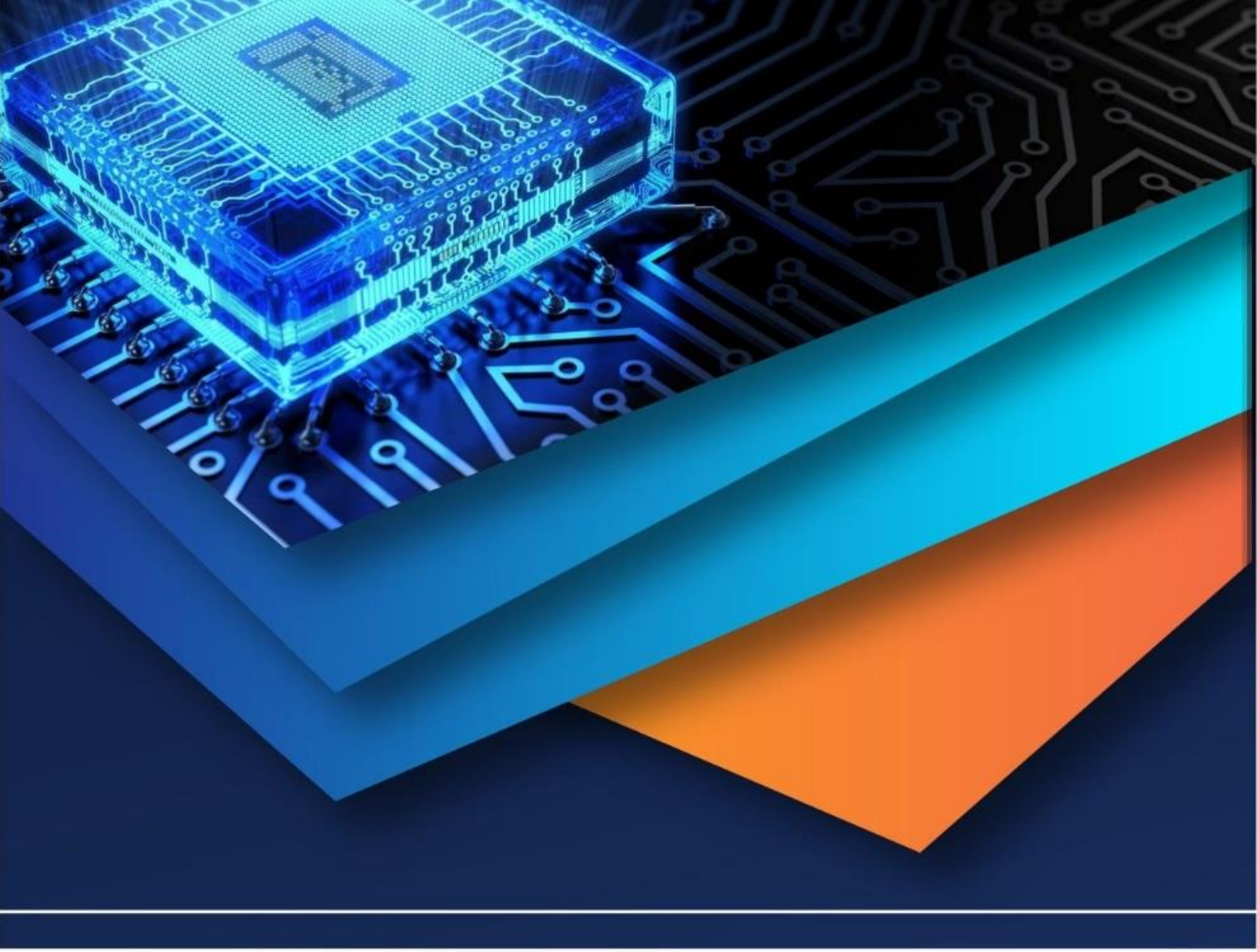
VI. CONCLUSION

In this work, the web usage mining is used. Various steps of web usage mining are discussed i.e how to clean the data and find the patterns. K-Means clustering is used to find the clusters of sites that are maximum visited, medium visited and minimum visited. There is a big challenge to keep eye on the students to check what they are surfing whole day on the internet and find their behavior. So to find the interest of the user, the web server logs are collected from the university computer center. Student's interest can be

judged on the basis that whether they use the internet within the college hours or after the college hours. We can also check the time limits of the students. There is suggestion to the network administrator to block those sites and generate the report automatically if the user is not browsing the relevant content, for example if the student continuously watching the entertainment sites during the college hours. Mostly youngster's hide their browsing history through the use of passwords on their devices. So, there is big advantage that parents can easily check the details of the browsing behavior of their children. They can make the decisions whether their children is going in the right direction or not.

REFERENCES

[1] V.Anitha and P.Isakki Devi, "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining", 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), ppno:1-4,2016.
[2] G. Neelima and Sireesha Rodda, "Predicting user behavior through Sessions using the Web log mining", 2016 International Conference on Advances in Human Machine Interaction (HMI), ppno:1-5, 2016.
[3] Ashwini Ladekar, Dhanashree, Raikar, Pooja Pawar,"Web Log Based Analysis of User's Browsing Behavior", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 11, 2014.
[4] Anurag kumar, Vaishali Ahirwar, Ravi Kumar Singh, "A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining", International Journal Of Engineering And Computer Science, Volume 6 Issue 2,ppno: 20233-2026,2017.
[5] ZAKARIA SULIZAKARIA ZUBI and SULIMAN MUSSAB SALEH EL RIANI, "Applying Web Mining Application for User Behavior Understanding", 1st WSEAS International Conference on Acoustics, Speech and Audio Processing (ASAP),2013
[6] Amit Pratap Singh,Dr. R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May –June 2014.
[7] Virendra R. Rathod and GOVIND V.PATEL, "Prediction of User Behavior using Web log in Web Usage Mining", International Journal of Computer Applications (0975 – 8887), Volume 139 Issue No.8, April 2016
[8] Ananthi.J, "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Volume 5, Issue 3, 2014.
[9] Neetu Anand, "Effective Prediction of Kid'sBehaviour Based on Internet Use", International Journal of Information and Computation Technology, Volume 4,Issue 2,2014.
[10] Avadh Kishor Singh, Ajeet Kumar and Ashish K. Maurya, "Association Rule Mining for Web Usage Data to Improve Websites", IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), ppno: 1599 – 1602,2014.
[11] Hiral Y. Modi and Meera Narvekar, "Enhancement of Online Web Recommendation System Using a Hybrid Clustering and Pattern Matching Approach", International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015), ppno: 1-6,2015.
[12] linHuaXu and HongLiu, "Web User Clustering Analysis based on KMeans Algorithm", International Conference on Information, Networking and Automation (ICINA),ppno: 6-9,2010.
[13] Gupta, Ashika, Rakhi Arora, Ranjana Sikarwar, and Neha Saxena,"Web Usage Mining Using Improved Frequent Pattern Tree Algorithms", ppno: 573 – 578,2014.
[14] https://www.google.co.in/search?q=k+means+clustering&source=lnms&tbm=isch&sa=X&sqi=2&ved=0ahUKEwiL1ceQlLvUAhWJKY8KHaoCxEQ_AUICCgD&biw=1366&bih=657
[15] http://analyticstrainings.com/?p=101
[16] https://www.google.co.in/search?q=web+usage+mining&hl=en&site=webhp&source=lnms&tbm=isch&sa=X&sqi=2&ved=0ahUKEwi_uezQl7vUAhUBOo8KHeeXDFsQ_AUIBigB&biw=1366&bih=657



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY