Journal of Theoretical and Applied Information Technology
31st October 2016. Vol.92. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT WEB USAGE MINING ALGORITHM BASED ON LOG FILE DATA

[1]TAWFIQ A. AL-ASDI, [2]AHMED J. OBAID

[1] Professor at Computer programing Dept., Babylon University, College of IT , Babel, IRAQ

[2]PhD Student, Computer programing Dept., Babylon University, College of IT, Babel, IRAQ

E-mail: [1] tawfiqasadi@itnet.uobabylon.edu.iq, [2] ahmedj.aljanaby@uokufa.edu.iq

ABSTRACT

Information on Internet and specially on website environment is increasing rapidly day by day and become very huge, this information play an important role for discovering various knowledge in the Web. Web Usage Mining one of the Web Mining algorithm categories that concern with discover and analysis useful information regard to link prediction, users' navigation, customers' behavior, site reorganization, web personalization and frequent access patterns from large web data that logs by Web server side and stored in standard text log file format called log file or Web usage data, this data can also be collected from an organization's database such as NASA. Web Usage Mining is a process of applying Data mining techniques and application to analyze and discover interesting knowledge from the Web. There are several existing research works on log file mining, some concern with web site structure, traversal pattern mining, association rule mining, Web page classification, and general statistics such as amount of time spent on a page. In this paper we will focus on mining the different segments content of Web log data entries in order to discover the hidden information and interesting browsing contents from it, then applying clustering algorithm to find similar groups of Web sites that have common browsing contents.

Keywords: *Web Mining, Web Usage Mining, Log File Analysis, Clustering, K-means, System Monitoring.*

1. INTRODUCTION

The massive growth of the amount of data and information on World Wide Web (WWW) and enormous Web pages created make mining and analysis useful information a practical challenges. World Wide Web (WWW) consist of billions of interconnected Web pages which are published by millions of authors on the world. Web page is a document that suitable to view by World Wide Web (WWW) through Web browser. Web document contain a data of all types such as structured tables, unstructured text, semi-structured content and multimedia content such as (images, audios and videos) [1].

The process used to extract and mine useful information and discovering knowledge from Web document by use Data Mining (DM) techniques is called Web mining. Web Mining is a multidisciplinary field include Data Mining (DM), machine learning, neural networks, information retrieval, statistics, and databases [2]. Web mining include wide domain of application that intent to discovering and extracting hidden information in data that stored in the Web.

The main task of the Web mining techniques is to discover and retrieve interested information from huge data set contain huge web data and its store in file called log file. Web data contain variant types of information and it's include web log data, web structure data and user profiles data. Web mining is divided into three categories web content mining, web structure mining and web usage mining [3].

Web Content Mining can be consider is the task of extracting useful and interested information from contents of web documents. Content data is the collection of web pages designed by use web language, there are many languages can be used in this manner like HTML, PHP, ASP…etc. While some Web documents can be designed by used Content Management System (CMS) like Joomla, Word press, Vivo…etc. Many techniques can employed here like text mining, Multimedia mining in order to discover or extraction similar web pages content.

Web Structure Mining is the task of using graph theory to analysis and understanding the connection structure of web site. Web structure can be divided into two classes: Extracting hyperlink patterns in the web and mining document contents.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Web Usage Mining is the application of data mining techniques to extract interesting patterns from web usage data, Usage data capture the activities of web users along with their browsing behavior at a website. Our paper deal with this kind of data in order to extract interested browsed contents and clusters Web sites Directories based on it.

Usage data can be collected at different sources levels such as server level, proxy level and user level. Web serve log is an important source for performing Web Usage Mining due its records the browsing behavior of users and their request contents. Log files can be stored in various formats such as common log or extended log [4].

2. PROBLEM DEFINITION

With explosive growth of the Web and number of users and Web sources increased significantly every day, practical data management issues such as clustering should be addressed and analyzed [5] . Clustering can be done on the Web either in users' session or in Web sources. An optimal clustering concept should satisfy compactness and separation concepts. Web sources clustered with respect to their certain characteristic or parameters such as their content, structure or their popularity. Thus Clustering on Web can be one of the following:

Web User Clustering, is a technique that groups of users that exhibiting similar browsing contents, and this can help to understanding user interesting contents and improve Web services. **Web Document Clustering**, by clustering Web documents based on its related content such as topics and this can improve search engines works and information retrieval. Finally, **Web Objects Clustering,** which can grouping relative content to better understanding and to offer user query results. New algorithms are needed to handle the combine types of data that available on the Web Sites to find meaningful clustering [6].

In this paper we concern to group similar Web sites browsing contents by classified it into four classes are text, images, Audio and videos and take into consideration the preprocessing issues to analysis valid entries content from huge log file data. Finally we will get group of similar Web sites (clusters) that have common browsing contents such cluster1 combine group of Web sites they has most actions in image, action include Post, Sharing, Downloading and simple hit to recognize from other Web sites where users interesting to seek only Audios and Videos files.

3. RELATED WORK

Discovering valuable knowledge and interested patterns from huge data has become a challenging to many research works. Growth in Web size, Different usage of World Wide Web, browsing contents in many web pages and others activities can be consider a practical issues for most of Web site administrators. Scale of data in the log file growth rapidly to several megabyte in one day thus may be exceed any available conventional database, therefore there is a need to successful way to analysis overall log data to extract useful information. Most of research work concern with pre-processing and discovering similar frequent pattern accesses, we start describing clustering algorithm from (Zamir et al, 1997, [7]) where mention a clustering method based on Word intersection to group similar Web Documents that share common Words. (Haveliwala et al, 2000, [8]) present a graph partition algorithm to cluster Web documents even in presence high dimensional space for feature representation. Another approach proposed by (Hammouda and Kamel, 2004, [9]) that combine two concepts are document similarity and histogram representation, they used a similarity among Web documents then using histogram based method for validate similarity inside each cluster. (K.Suresh et al, 2011, [10]), present approach for clustering transactions that own common characteristics and made users. (K. Poongothai et al, 2001, [11]) optimize a method for preferring similar user behaviors then construct users' profiles by analysis access entries in log file for various user, this way can help in e-commerce business sits. (Shaily et. al, 2013, [12]) used clustering algorithm for discovering similar user own common patterns by use association rule concept. Our contribution in this paper we are mining log file by applying NLP concepts and represent all contents of that log file, this log file include access events and interaction with 77 Web Sites that hosted in the Web server, then applying representative clustering algorithm to grouping similar Web Sites based on user browsing content , This method can give us an indicator what are the most visited content in every web site, most site visited, groups of users they have highest activity, most browsing fields and many other knowledge's. Data set used in this research works has collected from KUFA University, Central of Research Works in internet, Web Developing Dept., March 2016 Usage Data.

Journal of Theoretical and Applied Information Technology
31st October 2016. Vol.92. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. WEB SERVER LOG FILE

A Web log file, is a file in which Web server log information about user access to Web sites contents, then for each user request there is a particular entry in log file recognize it. Log file can differ in types and content structure in this section an overview of log file will be introduced.

4.1 Log File Types

There are four types of log files based on its information content and types of information provided by it Table 1 represent the classes of Web server log file with an example of each class [13]:

- **Access log file**: Data of all incoming request and information about client of server.
- **Error log file**: list of internal error whenever occurred in the case of that request it's not enable by server.
- **Agent log file**: provide information about user's browser.
- **Referrer log file**: provide information about link and redirects visitor site.

Table 1: Log file Types

T	Type	Format example
1	***Access***	120.236.0.14 -20011-12-12
2	***Error***	[Fri Sep 09 10:42:29.902022 2011] [core:error] [pid 35708:tid 4328636416] [client 72.15.99.187] File does not exist: /usr/local/apache2/htdocs/favicon.ico
3	***Agent***	Internet Explorer/5.0(win 7;)
4	***Referrer***	http://myblaze.sez.html>/library/lectures/news.gif

4.2 Log File Contents

Web log files contain variant information as we said based on its type and source, but in general there are common information in most common log files as follow [14]:

- **User IP address**: An IP address for users.
- **Date and Time**: information about the date and time for user request.
- **Mode of request**: Typically GET, POST, HEAD.
- **Remote Host IP address**: Used to determine unique host on internet.
- **Requested URL**: Requested file or web page by user.
- **States**: States code return by server to client such as 200, 404, and 203.
- **Bytes:** content of document or file has been transferred.
- **Agent type**: agent that used by user.
- **Remote URL**: traverse path of the user.

5. PROPOSED ALGORITHM

The proposed approach for Web Usage Mining are shown in the Figure 1, which contain two main phases are Pre-processing, Patterns discovery and analysis phase.

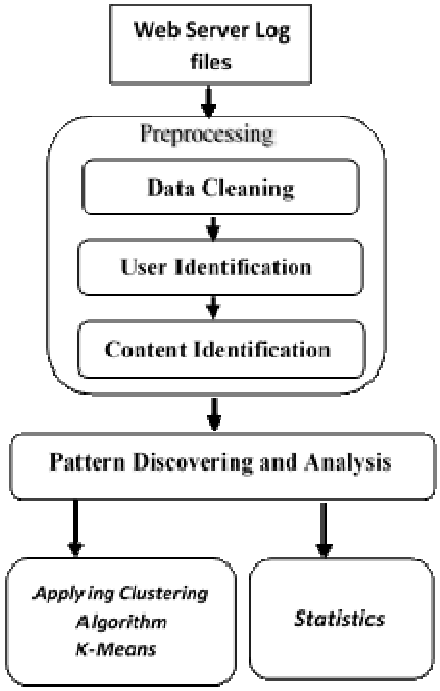

Figure 1 Proposed Diagram

5.1 Pre-Processing

In this phase Log files are combined and transformed into suitable form, this include splitting text information into useful parts and identified each portion from current text after that we are eliminate outliers and irrelevant information. This process include the following steps:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.1.1 Data Cleaning

In this step we eliminate irrelevant entries such as entries with failure Server status messages, redirect links and public access to home page due its common in all users, entries that include some other files like CSS, JavaScript, icons, Gif, etc. and other entries that does not has multimedia requests, this step also called filtering step, only desired entries contents will be consider for applying next step successfully.

5.1.2 User Identification

In this step we are recognize every user by take into consideration his own IP address, there are many issue regard to might many users might using the same machine, in this issue we are interesting just to identify valid user request instead of that request made by frequent users or certain user. By default we eliminate duplicated IP addresses that request same multimedia content, this step include identifying host IP address that assign to single host on the internet to be accessed by public users.

5.1.3 Content Identification

Content can be varied depends on users' navigation and desired content type, there are two challenges in this step, first, when user access to certain page that include desired content, user may download certain Web page in Web site, then hit will counted in log file for every contents correlated with this page if its related to user or not, second, there are some common access multimedia contents such as banner, common icons, buttons etc. will download and recorded as hits by many users when access to different Web page [15]. To solve these issues we are using an index table, in this table all multimedia content named in it, then the result produced by previous steps will filtering according to this table, for example if we have an image with name of NBA.jpg then all entries that contain this image will be collected and so on for other, this scenario can provide to us what are the main accesses media content, Web sites that has most access, types of media accesses that distinguish sites from others and finally group of Web sites that has most accesses in images, text files, audios and videos contents. The following algorithm used for cleaning and recognize the entries in log file as shown in the Figure 2.

Log file cleaning and recognize algorithm

Read *Log file in data base*
For *each record in log file*
Read *field (status code)*
If *field = Valid code then save record*
Read *field (requested file)*
If *field (requested file) ≠* ***public*** *then save record*
End for*.*

Figure 2 log file cleaning algorithm

5.2 Pattern discovering and analysis

In this phase we are employed mining algorithms depend on our interest, many techniques can be applied here but we are consider only clustering and general Statistics as follow:

5.2.1 Clustering

Is a technique applied to group items with similar properties, There are many kinds of clustering can be applied here like, user group having similar browsing behavior , Web sites pages group that has similar content, group of users they are visit similar Web sites etc. , always this is depend on what type of application used for .

The most appropriate algorithm can be used here is Representative based clustering algorithm, in this algorithm given a data set with n points in d-dimensional space , given the number of desired clusters k . the goal of this algorithm is to partition data set into k-clusters by represent each cluster with representative points that summarize cluster and assign each point to the appropriate cluster one of the most usage algorithm in this case is K-means algorithm, Figure3 represent K-means clustering algorithm that used in this Step along with normalization process [12], normalization can give better clustering result due to the various ranges in our data point in each space:

K-means clustering algorithm

Step	***Description***
Step1:	*Normalize the values in each space*
Step2*:*	*Setup the initial centroid of the k clusters*
Step3*:*	*for each record, find the nearest cluster centroid, cluster Centroid owns a subset of representative points "Record" thereby representing partitions of the data set.*
Step4*:*	*for each k cluster, find cluster centroid and update the location of each cluster center to the new value of centroid.*
Step5*:*	*repeat step 3-5 until no change in cluster centroids.*

Figure 3 K-mean algorithm steps

Journal of Theoretical and Applied Information Technology
31st October 2016. Vol.92. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.2.2 Statistics

Statistics is critical task to any Web site Administrator to know useful information regard to his own Web site contents such as discover total number of hits to every user, host, page and certain content. Hits can be counted based on each valid line in log file and these lines can include browsing, downloading and posting activities has done by different users.

The result of this phase transformed into useful tasks such as:

- Re-design web sites to include similar pages or contents.
- Improve access by grouping users that have similar behaviors.
- Most access sites, pages, content can be monitor and checked.
- Group users that have exceed bandwidth limitation.
- Monitor user's activities towards particular contents.

6. EXPERIMENTS AND RESULT

To applying proposed model in real time log file data the following steps illustrate our experiment procedure described as follow:

6.1 Parsing Log File

There are many attributes in raw log file such as User IP address, Date and time, Request Mode, Remote host IP address, URL request, Status code, Content name, Content type, Byte transferred, User Agent. One of the main issue encountered when dealing with log files is the amount of data need to be processed [16]. We consider interested attributes which has been extracted from log file, some other attributes are not consider in the preprocessing phase due to space restriction and not related to our mining process. Mining process applied to the result of Table2.

We are using assignment table for recognize different Sites, take into consideration Log file may contain many uncompleted users transactions and server errors entries, assignment table can recognize all variance content media that browsed and visited by different users based on their looking for in many Web sites. Web sites distinguished by assign DNS (Domain Name System) that assign for every particular one to recognize it through log file entries.

Table 2: Preprocessing log file

User IP	Date	Mode of request	R. Host	R. File	File type	Status code	Byte transfer
109.x.x.x	31/Mar/2016	GET	En	179372882	jpg	200	31219
141.x.x.x	31/Mar/2016	GET	Journals	Article	Text	200	31219
68.x.x.x	31/Mar/2016	GET	Journals	Article	Text	200	14485
37.x.x.x	31/Mar/2016	GET	reg_eve	Puplic	N.A	200	11236
157.x.x.x	31/Mar/2016	GET	Journals	Article	Text	200	16494
37.x.x.x	31/Mar/2016	GET	Ar	729798129	Jpg	200	50301
37.x.x.x	31/Mar/2016	GET	Libr	Haddad	Pdf	404	613
66.x.x.x	31/Mar/2016	GET	Libr	Bidery	Pdf	200	83986

6.2 Statistics

Statistics used here to extraction useful information such as count the hits for each site depend on its content, total number of valid hits and other useful statistics has been constructed. Table3 show the total number of entries for our selected log file. The sample log file as shown in the Figure4 can give an indicator that some sits were design using CMS (Content Management Systems) and in those systems request of each content that is not identified as an multimedia content by log file entries and can be considered as an Articles, Articles has text content and can be recognize in the name of that Article in CMS Web sites while some other sites are designed by using HTML or some other Web design languages, browsed content can be identified and recognized easily by compare it to CMS Web sites.

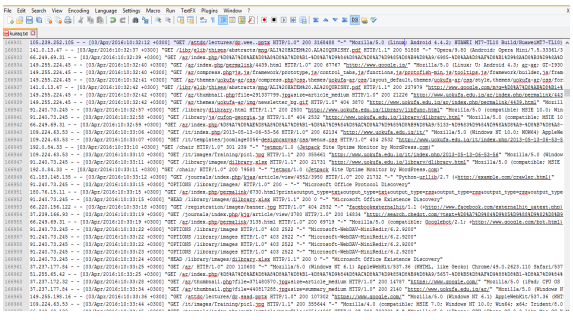

Figure 4 Sample of server log file

Following table (Table3) show statistic result for users' hits in different Web sites contents, we selected the first 10th Web sites result due to size restriction, total hits for main interface Web page not take into consideration due its visit by almost users, and it's a gate to all others Web sites. We combine video and audio hits count to be in the same field. Total number of valid Web sites are 77 exist in log file entries and hosted in our selected Web server.

Journal of Theoretical and Applied Information Technology
31st October 2016. Vol.92. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Web sites visits table

T	Total number of hits	Total number of valid hits	Total number of Directories	Site name Total number of Hits		Image file hits	Text file hits	Video / Audio hits
			Web Sites	Site name	Hits			
1	991289	750587	77 Web Sites	Site 1	507	380	120	7
2				Site2	491	15	421	55
3				Site3	358	35	321	2
4				Site4	224	2	11	211
5				Site5	524	501	22	1
6				Site6	212	177	23	12
7				Site7	364	325	14	25
8				Site8	177	31	144	2
9				Site9	487	455	32	0
10				Site10	259	242	12	5

Hits counted for every Web site can recognize the Web site filed, most browsed contents, most users' access, desired content requested, and thesis include lessons in audio, video and other structured format. In case to cluster Web sites we calculate the empirical joint probability of hits count for everyone to find the correlation of Sites to which class belongs by apply our proposed algorithm on result data.

6.3 Applying Clustering Algorithm

To find groups of similar Web sites and recognize everyone to which cluster will be assign we applying K-means clustering algorithm, K-means initially assign centroid to K cluster when user assign K value. K-means assign data point to its closest cluster then cluster centroid calculated again to update it, this process continue till no change in the final result and no data point moved from its current cluster to other clusters. Since successful application of K-means clustering algorithm depend on the parameter K, the knowledge and understanding of data analyzer on the application domain when clustering algorithm will be applying help to select good (K) value to produce meaningful clustering result in that application domain [17].

The following figure (Figure 5) show the result of applying clustering algorithm by select a proper K value to produce clusters, labeling process are used to recognize each Web site to any cluster has been assigned depend on the result on Table 3, there are some Web sites its distance bigger than other points that is because it has either equal values in tow distinct clusters or it may don't has greater values in that cluster field.

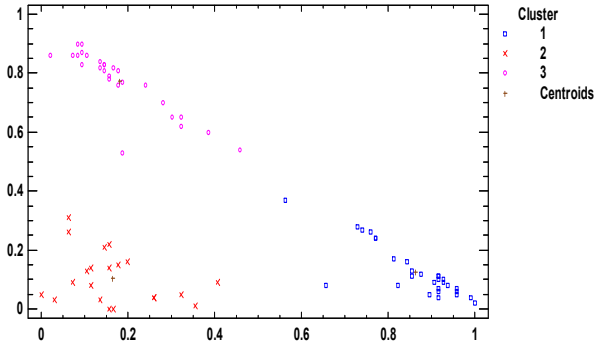

Figure 5 K-mean Clustering Result

Clustering result show there are three clusters in Figure 5, Cluster1 that combine all Web sites where users browsing, downloading, posting, etc. images content, while Cluster2 grouping Web sites that has most text files accesses events including (doc., PDF, ppt., etc.) access events, finally, Cluster3 contain Web sites where users interesting just in video and audio contents on these sites. The following table (Table 4) show the centroid points of each cluster, data set are partitioned and each point compared and assign to the cluster with the nearest centroid.

Table 4: Clusters Centroids Table

Cluster	Image	Text	Video + Audio
1	0.862428	0.125862	0.0362069
2	0.164683	0.10619	0.725238
3	0.180556	0.772222	0.0444444

7. RESULT EVALUATION

Evaluation of clustering result is a criteria which can be used to measure the goodness of the clustering algorithm, final results and the structure of proposed model used. Clustering algorithms are differ based on the type of data and application domain. In this section we discuss our selected evaluation measures, explanation of final result, limitation of the research work and some statistics result discussed in following:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

7.1 Evaluation Measures

Before applying evaluation measures we would like to see how our selected algorithm separated final clusters, we calculate Cluster distance. To calculate the distances between clusters, first task the centroid point is detected then the distance values is calculated which is equal to distance between clusters centroid points. Table 5 show the distance values between centroids of final clusters.

Table 5: Distances between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0.0000	0.9603	0.9208
Cluster2	0.9603	0.0000	0.9520
Cluster3	0.9208	0.9520	0.0000

There many evaluation measures can be used to measure the performance of clustering result, and also to evaluate selected algorithm, a proper clustering algorithm depend on the highest values can be found in every evaluation process. Most common evaluation measures used described as follow [18, 19]:

- **TP Rate (True Positive Rate)**

  Also called Sensitivity or Recall, TP rate measures the fraction of correctly labeled point's pairs to all points' pairs in same cluster or portion.

- **FP Rate (False Positive Rate)**

  FP measure the intrusion connection of points that were incorrectly clustered or classified (less values here refer to good result).

- **Precision**

  Precision measures the fraction of correctly clustered pairs of points compared to all other points' pairs with same cluster.

- **F-Measure**

  F-measure can be defined as the harmonic mean of the precision and recall scores for each cluster. Thus it tries to balance precision and recall in all clusters.

Table 6: Result of Evaluation Measures

	TP Rate	FP Rate	Precision	Recall	F-Measure	No. Of Sites in Cluster
Cluster 1	0.963	0.06	0.897	0.963	0.929	27
Cluster 2	0.931	0	1	0.931	0.966	29
Cluster 3	0.905	0.036	0.905	0.905	0.0905	21
Weighted Avg.	0.935	0.031	0.938	0.935	0.952	77

7.2 Result Explanation

Table 6, shown the number of Web sites has been assign to each cluster, for example in Cluster1 the total number of Web sites has been assign to it is 27 Web sites, this mean there is a group of 27 Web sites where most users browsed only images contents from it, while in Cluster 2 there are 29 Web sites where their users interesting only on text files contents, after analyzed Web sites we find most of these Web sites contain E-Libraries, University book store, thesis, lectures, and others text files format. Cluster 3 include audio and video files, Cluster3 Web sites contain most of visual media contents such as festivals, conferences, meeting, lessons, congress, and students parties events in audio and video contents. Finally, we got clusters that separated and partitioned Web sites based on users' browsing and interesting behaviors, that give to us a brief perception about what are the type of contents can be found in every Web site, and why Web sites differ from other in access rate.

7.3 Research Work Limitations

There many limitation in any Web application, such as type of algorithm used, not all clustering algorithm used for all dataset and this required studding and correct selecting features from dataset then select appropriate algorithm. In Web Usage Mining where Mining algorithm applied on log file contents, the first limitation is how to collect log file data source, this data contain many private information regarding to user communities like an IP address, agents, and others and this information cannot be accessed only by authorized users. So, most of Web application are developed for certain organizations to understand users' requirements and business works. Other limitation also is the analysis model has been used to analysis the log file content, this file contain text lines format and need a proper

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

way to analysis all segments content. We assume and based on our studding on selected log file, multimedia contents can be separated and Web sites can be recognized depend on the requested file from Web collections, log file format and contents are differ from each other based on Web sites structures, contents, type of technology used in it and server configuration. There is a way to categorized every Web sites and cluster it to a desired group to be clear why those Web sites browsed, what their patterns can recognized from others, structure contents and others perception.

7.4 Statistics

Other statistics can be calculated and visualized based on requirements, we illustrate in this paper some of statistics which can be visualized by using data visualization software Such as Rapid Miner, NCSS, STATISTICA and many others to understand the user events and Web sites data traffic, Table 6 illustrate the invalid entries in our selected log file. There are many options to calculate information about total users, total file requested, total sessions, and many others depend on application required.

Table 6: Invalid entries in Web Server Log File

T	Errors	Description	Hits Count
1	404	File Not Found	55760
2	500	Internal Server Error	4303
3	403	Forbidden	2422
4	400	Bad Request	733

Other statistics can be visualized, we selected to visualized duration time has been spent by user, in Figure 6, the most users had been spent less that 60 minutes when browsing Web sites contents.

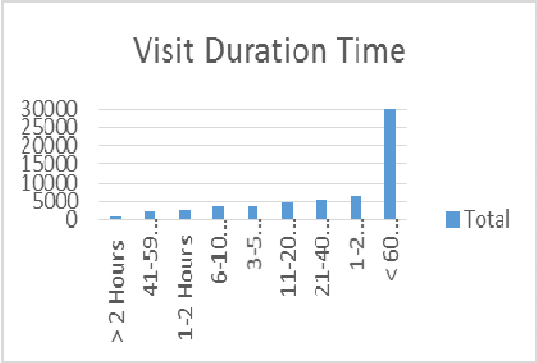

Figure 6 Users Visit Duration Time

8. FUTURE WORK

For future work, we plan to investigate a novel categorization algorithm to classifying Web pages and type of contents available in Web pages, this can help us to categorize user based on desired classes and clustering similar users based on type of interesting contents such as group of users' interest in Computer Science contents.

9. ACKNOWLEDGEMENT

This work and data analysis result has been supported by Information Technology Research and Development Center (ITRDC), KUFA University, IRAQ and College of Information Technology (CIT, Babylon University, IRAQ).

10. CONCLUSION

Web Mining is an extension of Data mining techniques in Web data. The fast growth of information and data available on World Wide Web and the big events done on these contents by communities, lead many organizations, companies, and others agencies take in consideration to analyses and understand many Web data for several requirement addressed by their community, and to develop an enhance the services and requirements. Web Usage Mining is the most filed used by Web application and its concern with the data that available in log file. The main challenges in log file analysis is a pre-processing step, the result of preprocessing lead to discover many underlying knowledge from huge data exist in it. Web Usage Mining has been applied to several Web application especially in E-commerce and business services.

In this paper we selected a log file from academic center, to analysis their Web sites and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

provide a way to understand and percept contents, structure and services provided by these Web sites then grouping similar Web sites by used clustering algorithm. Many log file analysis tools available today give limited analysis capabilities on the result such as summery statistics, frequency count, Server monitoring and others static capabilities such as for example most visited files, pages, viewing time, server performance and others. In order to perform another type of analysis this required to develop different kind of analysis and preprocessing procedure that used for analysis the deep content in that available in log file.

Thus we discuss how we analyzed the request type that made by users in this work, then extracted final result for each valid request line. After that we use our proposed model to identify type of media that requested from every Web site, this way can help in recognize and group Web sites that has common content patterns. Many Web application and services are developed also for government centers for controlling and monitoring their works and user communities, this paper present a key concept for clustering Web sites which can be address as a new looking for recent work on Web resources, and provide deep analysis to understand Web sites directions as well as their user communities behavior, we found in this work many analysis can help to support many information related many research work can be establish in future. Finally, the result has been evaluated by use most common clustering evaluation measures to verify the final result of our research work.

REFRENCES:

[1] Bing Liu, "Web Data Mining, Exploring Hyperlinks, Contents and Usage data", 2nd edition, Springer New York, ISBN: 9783642194597, 2011, PP: 1-14.

[2] Sandhya, Mala chaturvedi, 2013. "A Survey on Web Mining Algorithms", International Journal of Engineering and Science, Vol.2, No.3, 2013 pp. 25-30.

[3] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD: SIGKDD Exploration, Newsletter of Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM Vol.2, No.1, 2000.

[4] Nanhay Singh, Achine Jain and Ram S. Raw, "Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques", International Journal of Data Mining and Knowledge Management process, Vol.3, No.4, 2013, PP: 137-147.

[5] Athena Vakali, George Pallis, Lefteris Angelis, "Clustering Web Information Sources"; In Web Data management practices: Emerging Techniques and Technologies, IDEA group publishing, ISBN: 1599042282, 2007, pp. 34–55.

[6] Monika Yadav, Pradeep Mittal, "Web Mining: An Introduction", International Journal of Computer Science and Software Engineering, Vol.3, No.3, 2013, pp. 683-688.

[7] Zamir O., Etzioni O., Madanim O., and Karp R. M., "Fast and intuitive clustering of Web Documents", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997, pp. 287-290.

[8] Haveliwala T., Gionis A., and Indyk P., "Scalable technique for Clustering the Web", Proceeding of WebDB, 2000.

[9] Hammouda K. M., and Kamel M. S., "Efficient phrase-based Document Indexing for Web Document Clustering", IEEE Transaction on Knowledge and Data Engineering, Vol.18, No.10, 2004, pp. 1279-1296.

[10] K. Suresh, R. Madana M., A. Rama MohanReddy, "Improved FCM Algorithm For Clustering on Web Usage Mining", International Journal of Computer Science Issue, Vol.8, No.1, 2011, pp. 42-45.

[11] K. Poongothai, M. Parimala, S. Sathiyabama, "Efficinet Web Usage Mining with Clustering", International Journal of Computer Science Issue, Vol.8, No.6, 2011, pp. 203-208.

[12] Shaily G., L., Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery", International Journal of Data Mining Techniques and Application, Vol.2, No.1, 2013, pp. 141-150.

[13] Risto Vaarandi, "A Data Clustering Algorithm for Mining Patterns from Event Logs", Workshop on IP Operations and Management, IEEE, 2003, ISBN: 0780381998.

[14] L. K. Joshila Grace, V. Masheswari, Dhinaharan Nagamalai, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security and It's Application, Vol.3, No.1, 2011, pp. 99-110.

[15] M. H. Abd Wahab, M. N. Haji Mohd, M. F. M. Mohsin, "Discovering Web Server Log Patterns Using Generalized Association Rules Algorithm ",New Advance Technologies, Aleksandar Lazinica (Ed.), InTech, pp. 177-194, 2010, ISBN: 9789533070674.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[16] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P.Lim, "Research issues in web data mining", in Data Warehousing and Knowledge Discovery ", 1999, pp. 303–312.

[17] Nong Ye, "K-means Clustering and Density Based lustering", Data Mining: Theories, Algorithms and Examples, CRC press, 2014, pp. 153-166.

[18] Mohammed J. Zaki, Wagner Meira JR., "Clustering Validation", in: Data Mining and Analysis: Fundamental Concepts and Algorithms, CAMBRIDGE University Press, pp. 425-463, 2014, ISBN: 9780521766333.

[19] L. Pandeeswari, K. Rajeswari, "K-means Clustering and Naïve Bayes Classifier for Categorization of Diabetes Patients", International Journal of Innovative Science, Engineering and Technology ", Vol.2, No.1, pp. 179–185.