International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 16 (2016) pp 8797-8805
© Research India Publications. http://www.ripublication.com

Discovering similar user navigation behavior in Web log data

Tawfiq A. Al-asadi[1] and Ahmed J. Obaid[2]

[1]Department of Computer Science, Director, College of IT, Babylon University, Iraq.
[2]Department of Computer Science, Faculty of Education, KUFA University , Iraq.
E-mail: ahmedj.aljanaby@uokufa.edu.iq

Abstarct

With the growth of World Wide Web and large number Hosts are join continuously to the internet, huge number of access events to Web sites pages were recorded by Servers in log files , many users share, send, post and download lot of things from Web Sites, this manner can be difficult to many organization and Agents in order to monitor and control that, the recorded information and type of analysis used to extract useful knowldege and understanding it become a practical challenges to many researchers. Log files can provided many events information regard to Clients activities, server activities and so on. Many organization employee many log files analysis tools to predict, analysis and monitor users behavior towards site contents .In this paper we proposed algorithms to analysis hidden information contents in Log files and discovering patterns by identified users along with them navigation behaviors then clustering similar users based on different interesting log file content for many Web sites that hosted in Web server. Find statistics for every part in log file command line which are not present in many log files analysis tools are supported here and finally discovering frequent Web sites-Users and user's activities towards those Web sites

Keywords: Web Usage Mining ; pattern mining; Clustering; Log file analysis; Web log data.

INTRODUCTION

Information on internet and especially on Web sites increasing rapidly day by day, Web Sites play an important role in this manner where authenticated users, users are always uploads, downloads, browsed many contents according to them needs and interest. Web Server provide a way to browse Web Sites by assigning an IP address or DNS to identify it in addition hosted in it , Server record every events in the form of log file. The process of discovering hidden information from Web log file is called Web Mining. The aim of it is to obtain information about navigational behavior and retrieve useful information from very large raw data, can be represented by several millions of event records in log file. Web log data contains different kinds of information and including web document, web structure and user profiles. Web mining classified into three categories depend on which part of Web to be mined [1, 2]. Categories are Web Structure Mining, Web Content Mining and Web Usage Mining, **Figure 1** illustrate the categories of Web Mining algorithms.

Web Structure Mining is the task for discovering knowledge from the structure of hyperlinks within Web pages and given useful information for the relationship among Web pages [3, 4]. The clustering process can play important role here by grouping the Web pages based on their structure, pages can represented by nodes and their links as edges among these nodes, clustering process can be done here based on graph representation and understanding structure of Web pages and its related to other pages in other Web Site Pages . Link structure in Web pages can be classified into two types: First, hyperlinks that connect different parts in the same page (Intra). Second, hyperlinks that connect two or more different pages (Inter). The other role can be applied here by identify trustworthy pages and their hub pages for a given subject. Trustworthy pages contain important information and supported by several links referred to it that means these pages are highly referenced. Hub pages contain many links to trustworthy pages that can give a role for clustering Pages based on trustworthy pages. Web Structure Mining can be employed to efficiently improve information retrieval and document classification tasks [5].

Web Content Mining is the task of discovering different kinds of information contents and improving efficient mechanisms to organize and grouping (clustering) multimedia content to the search engines for accessing these contents by using keywords, categories, related contents etc. Multimedia contents on Web pages are varied such as structured content (i.e. XML documents), Semi-structured (i.e. HTML pages), Unstructured content (i.e. plaint text), other related contents Images, Audios, Videos which are added to those pages or linked to other hosted Sites. Recently there are some challenges appear regard to that in the case of many Web sites were designed by using not only HTML language , other Languages and systems were invited here such as Content Management Systems (CMS) etc. and the plait texts here are encrypted and stored in an SQL data bases and users events were recorded as visited articles and in this case need to combine web mining algorithms in case to mining clustering and extracted useful information from user behaviors and contents related. Web CMS is responsible for storing, control and management data and other component in long-term uses. CMS consist of repository used to store and preserve various component and use various databases to store it. Repository in CMS contain two categories, the first one comprises source files as well as CMS configuration files , these files contain information about type of content, metadata, users and group of users along with them access data , profiles and preferences. The second repository contain databases where content and files will be processed through CMS and inherit databases and Tables that are constructed for recall and process the content [6]. Many researches has been done in Web Content Mining, including text mining and its issues such as: topic discovery, association pattern discovery, Web pages classification and Web document clustering. Other important body of work by discovering knowledge from images in the field of image processing. Other research including Latent Semantic Indexing (LSI) which ties to analyzing structure of elements in document collection, another important role looking for find the position of words in document for solving the document categorization problems and extracting patterns or rules. Topic detection and tracking also addressed as Web Content Mining [7, 8].

Web Usage Mining is the task to discovering the interesting patterns from Web Usage Data. Interesting patterns include information about user access patterns along with various types of request have been made by single or many users. The aim of Web Usage Mining is to understand the browsing and navigation through Web pages to enhance many things such as: the quality of commercial services, allotment Web portals [9] or improve Web structure and Web Server performance [10]. Web Usage Mining can be defined as the extraction of useful user patterns from Web server access logs files based on data mining techniques. Sources of log files include Web server, Client server, proxy server and application servers [11, 12]. By found more than one source place that store the navigation patterns and users accesses that make the mining process more difficult. The best and reliable result can be obtain from the log file that has all three types of log file. Web page accesses that were cached in proxy servers or in client side does not contain records on server side. Proxy server provide additional information however the requested page are missing in the client side, that lead to problem for collecting information from client side. Most of Web mining algorithms work based on Server side log data, commonly used mining algorithms are association rule mining, sequence mining, clustering [13].

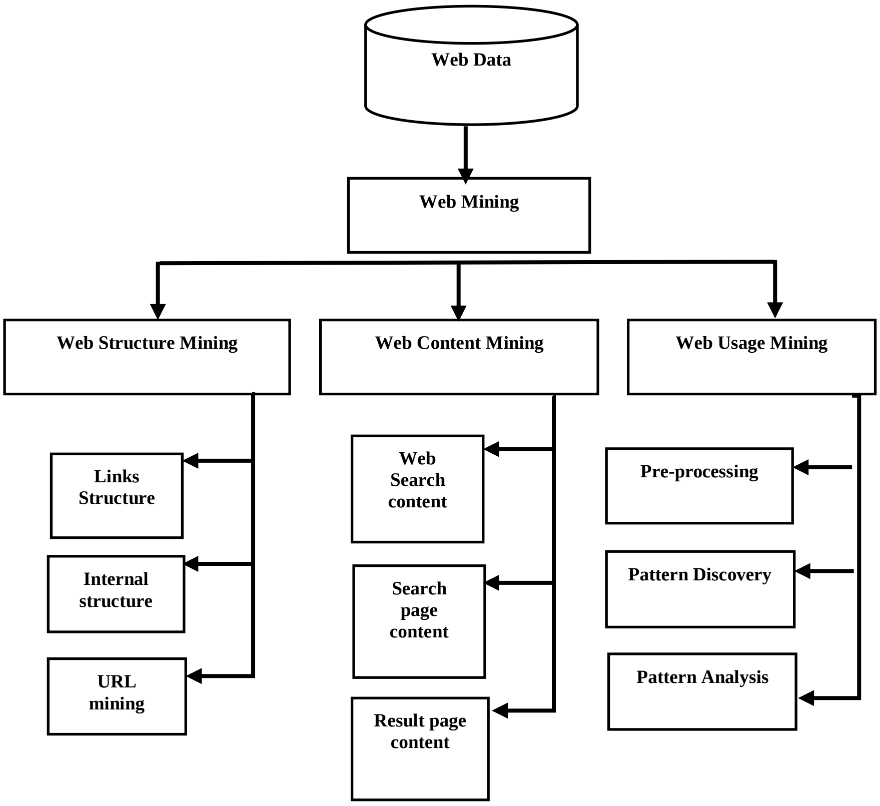


Figure 1. Web Mining Categories

The organization of the paper is as follow: section 2 illustrate the related work, section 3 discuss the log file types, format and parameters, section 4 show the Web mining phases and preprocessing steps for our log file, section 5 contain the proposed model and algorithm, section 6 our analysis result and finally conclusion for our works.

RELATED WORK

In the field of Web usage mining there are several data mining techniques have been used in order to discover interesting knowledge based on looking and focused approach. The information gained by these techniques can be used in many areas such as reconstructing Web sites, prediction next visited pages, group similar users , recommendation systems and so on. Clustering is the data mining process that group together similar items having similar properties. The clustering may include group of similar users, pages, references sites etc. Discovering group of similar users in user communities have been discussed in [20]. While in [23, 24, 25, and 26] authors used Association rule mining for discovering Web pages accessed directly by other pages. Web Usage Mining is presented in many approaches along with applying data mining techniques for discovering information. In [27] where Association rule used for discovering relation among pages, also used for detect the association among group of users with particular interest. Frequent path traversal and patterns topology of paths used with WAP–tree for representing and saving efficient patterns, others such as in [28, 29 and 30] they used Web usage mining and Meta data for discovering terrorist and attacks Web sites.

SERVER LOG FILE ANALYSIS

Your paper's figs must be without background fill color, no border fig and no border legend, no vertical line, no horizontal

Server Log file types

Web server log files are plain text files and independent from the server, generally there are four types of server logs based on types of information recorded which illustrate in **Table 1** :

- Access log file
- Agent log file
- Error log file
- Referrer log file

Table 1. Format Types of Web server log files

Log file types	Actions	Format	Extracted knowledge
Access log file	1. Records all users request processed by server. 2. Record information about users.	[Wed Oct 11 14:32:52 2000] [error] [Client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test.	Users' profiles. Frequent patterns. Bandwidth usage.
Agent log file	1. User browsers. 2. Browsers version.	"Mozilla/4.0 (compatible; MSIE 4.01; Windows NT)"	Agent version. Operating system used.
Error log file	List of errors for users request made by server.	[Wed Oct 11 14:32:52 2000] [error] [client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test	Types of errors. Generated errors IP address. Date and time of error occurred.
Referrer log file	1.Information about link. 2.Redirects visitor to Site.	"http://www.google.com/search?q=keyword", "/page.html"	Browser used. Keywords. Redirect link content.

Server Log File Format

There are three types of log file format as follow:

Common log file format

Is used by most of the web servers. The format of this log file is standardized and can be analyzed by web analysis program, the sample format of this type is shown below.

```
127.0.0.1 user-identifier frank
[10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif
HTTP/1.0" 200 2326
```

Combined log file format

Is same as common log file format but there are additional information present here, these information are "referral part, user-agent part and cookie prt", the sample format of this type as bellow.

```
127.0.0.1 user-identifier frank
[10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif
HTTP/1.0" 200 2326
```

Multiple access logs

Is consider the combination of the previous two types (common log and combined log) file format, in this type of log file format multiple directories can be can be created for access logs, the sample format of this type as shown below.

```
LogFormat "%h %l %u %t \"%r\" %>s %b"
common
CustomLog logs/access_log common
CustomLog logs/referer_log "%{Referer}i -> %U"
CustomLog logs/agent_log "%{User-agent}i"
```

Server Log File Parameters

Log files contain various parameters and can be very useful for recognized users browsing attributes, many attributes can be added or enabled depend on server configuration and user privacy agreement, some of cookies and private information can be used but in general there are common parameters an be found in log files. Below will illustrate in **TABLE 2**, the list of some parameters useful for analysis processes.

Table 2. Parameters of log file

T	Parameter name	Description
1	User name (IP address)	Identify users Who visited Website by its IP address.
2	Time stamp	Date and time when user browsed and spend time.
3	Request	Exact request line by user

4	Status code	Code sent by server after each user request
5	Bytes	Content length of document transferred
6	User agent	The browser that user used to send request
7	Request type	Method used by user to send request GET , POST

There are several works have been done on log files each work deal with particular issue of mining and task, this paper focus on identifying users and then extract knowledge about user behaviors to grouping similar users based on them browsing activities, our contribution here in case to analysis log file we select KUFA university Apache HTTP server version 1.1 main web server log file, as we mention above this log file its standardize text file format and we are applying text mining techniques to tokenize and extract interesting information from that log file.

PHASES OF WEB USAGE MINING

In order to extract knowledge from log file, several problems exist when extract useful information from that log file and also there are many outlier records need to be eliminate from it in this case we are applying general phases of Web Usage Mining to analysis and understand the extracted and valid information. The general phases of Web Usage Mining as follow:

Phase 1: Preprocessing

Preprocessing phase include some activities can be applied on log file for cleaning, identifying users, valid URL path and also eliminate outliers from log file, tasks on preprocessing phase as follow [13]:

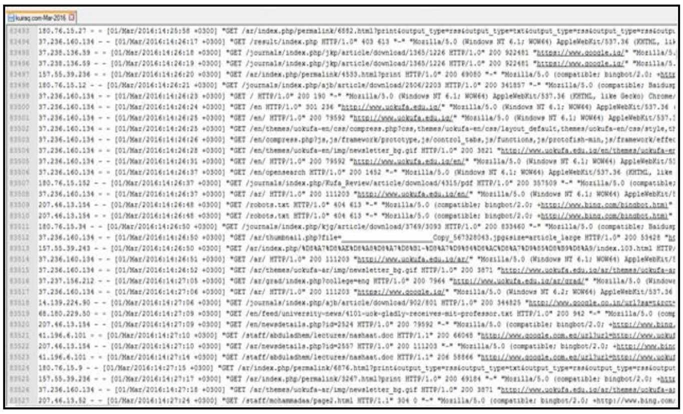
Data Cleaning

log file contain several records are irrelevant to our work like redirect path to other Sites, entries belong to top/bottom frames and records contain server error message. Error message identified through the status code that has been sent by server when user request particular content, server status code can be vary and valid status codes are show in **table 3**.

Table 3. HTTP server status codes

T	Code Syntax	Status code	Description
1	1XX	100	COUNTINUE
		101	SWITCHING PROTOCOL
		102	PROCESSING
2	2XX	200	OK
		201	CREATED
		202	ACCEPTED
		203	NON-AUTHORITATIVE INFORMATION
3	3XX	301	MOVED PERMANENTLY
		302	FOUND
		303	SEE OTHER
		304	NOT MODIFIED
4	4XX	400	BAD REQUEST
		401	AUTHORIZATION REQUIRED
		402	PAYMENT REQUIRED
		404	NOT FOUND
5	5XX	500	INTERNAL SERVER ERROR
		501	METHOD NOT IMPLEMENTED
		502	BAD GATEWAY
		503	SERVICE UNAVAILABLE
		504	GATEWAY TIME OUT

Status code show the success and failures users request, records with status code less than 200 and greater than 299 are considered failure records and eliminated from log file entries. Data cleaning also include eliminated records that browsed irrelevant paths such as CSS content, main site paths, gif, icons and maps etc. by checked suffix part of URL. **FIGURE 2** represent portion of KUFA university Main Web server (Linux server) log file format, in that server DNS were assigned to Host IP address to identify Web site that browsed by several users, We are consider to eliminate the records that browsed Main page due its common in many records because its contain links to all web sites in our server.



```
180.76.15.27 - - [01/Mar/2016:14:25:58 +0300] "GET /ar/index.php/permalink/6882.html?print&output_type=rss&output_type=txt&output_type=rss&output_type=rss&outpu
37.236.160.134 - - [01/Mar/2016:14:26:17 +0300] "GET /result/index.php HTTP/1.0" 403 613 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, lik
37.238.136.59 - - [01/Mar/2016:14:26:18 +0300] "GET /journals/index.php/jkp/article/download/1365/1226 HTTP/1.0" 200 922481 "https://www.google.iq/" "Mozilla/5.
37.238.136.59 - - [01/Mar/2016:14:26:19 +0300] "GET /journals/index.php/jkp/article/download/1365/1226 HTTP/1.0" 200 922481 "https://www.google.iq/" "Mozilla/5.
157.55.39.236 - - [01/Mar/2016:14:26:20 +0300] "GET /ar/index.php/permalink/4533.html?print HTTP/1.0" 200 69080 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http
180.76.15.12 - - [01/Mar/2016:14:26:21 +0300] "GET /journals/index.php/ajb/article/download/2806/2203 HTTP/1.0" 200 341857 "-" "Mozilla/5.0 (compatible; Baidusp
37.236.160.134 - - [01/Mar/2016:14:26:23 +0300] "GET / HTTP/1.0" 200 190 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/
37.236.160.134 - - [01/Mar/2016:14:26:24 +0300] "GET /en HTTP/1.0" 301 236 "http://www.uokufa.edu.iq/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (
37.236.160.134 - - [01/Mar/2016:14:26:25 +0300] "GET /en/ HTTP/1.0" 200 79592 "http://www.uokufa.edu.iq/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.3
37.236.160.134 - - [01/Mar/2016:14:26:25 +0300] "GET /en/themes/uokufa-en/css/compress.php?css,themes/uokufa-en/css/layout_default,themes/uokufa-en/css/style,th
37.236.160.134 - - [01/Mar/2016:14:26:26 +0300] "GET /en/compress.php?js,js/framework/prototype,js/control_tabs,js/functions,js/protofish-min,js/framework/effec
37.236.160.134 - - [01/Mar/2016:14:26:28 +0300] "GET /en/themes/uokufa-en/img/newsletter_bg.gif HTTP/1.0" 200 3821 "http://www.uokufa.edu.iq/en/themes/uokufa-en
37.236.160.134 - - [01/Mar/2016:14:26:31 +0300] "GET /en/ HTTP/1.0" 200 79592 "http://www.uokufa.edu.iq/en/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/53
37.236.160.134 - - [01/Mar/2016:14:26:37 +0300] "GET /en/opensearch HTTP/1.0" 200 1452 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like
180.76.15.152 - - [01/Mar/2016:14:26:37 +0300] "GET /journals/index.php/Kufa_Review/article/download/4315/pdf HTTP/1.0" 200 357509 "-" "Mozilla/5.0 (compatible;
37.236.160.134 - - [01/Mar/2016:14:26:37 +0300] "GET /ar/ HTTP/1.0" 200 111203 "http://www.uokufa.edu.iq/en/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/5
207.46.13.154 - - [01/Mar/2016:14:26:48 +0300] "GET /robots.txt HTTP/1.0" 404 613 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
207.46.13.154 - - [01/Mar/2016:14:26:48 +0300] "GET /robots.txt HTTP/1.0" 404 613 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
180.76.15.34 - - [01/Mar/2016:14:26:50 +0300] "GET /journals/index.php/kjg/article/download/3769/3093 HTTP/1.0" 200 833460 "-" "Mozilla/5.0 (compatible; Baidusp
37.236.160.134 - - [01/Mar/2016:14:26:50 +0300] "GET /ar/thumbnail.php?file=____________________Copy_567328043.jpg&size=article_large HTTP/1.0" 200 53428 "ht
157.55.39.243 - - [01/Mar/2016:14:26:50 +0300] "GET /ar/index.php/%D8%A7%D8%AE%D8%A8%D8%A7%D8%B1-%D8%A7%D9%84%D8%AC%D8%A7%D9%85%D8%B9%D8%A9/index.103.html HTTP/
37.236.160.134 - - [01/Mar/2016:14:26:51 +0300] "GET /ar/ HTTP/1.0" 200 111203 "http://www.uokufa.edu.iq/ar/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/5
37.236.160.134 - - [01/Mar/2016:14:26:52 +0300] "GET /ar/themes/uokufa-ar/img/newsletter_bg.gif HTTP/1.0" 200 3871 "http://www.uokufa.edu.iq/ar/themes/uokufa-ar
37.237.156.212 - - [01/Mar/2016:14:27:05 +0300] "GET /ar/grad/index.php?college=eng HTTP/1.0" 200 7964 "http://www.uokufa.edu.iq/ar/grad/" "Mozilla/5.0 (Window
37.236.160.134 - - [01/Mar/2016:14:27:06 +0300] "GET /ar/ HTTP/1.0" 200 111203 "https://www.google.iq/" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36
14.139.224.90 - - [01/Mar/2016:14:27:06 +0300] "GET /journals/index.php/ajb/article/download/902/801 HTTP/1.0" 200 344825 "http://www.google.co.in/url?sa=t&rct=
68.180.229.50 - - [01/Mar/2016:14:27:09 +0300] "GET /en/feed/university-news/4101-uok-gladly-receives-mit-professor.txt HTTP/1.0" 200 942 "-" "Mozilla/5.0 (comp
207.46.13.154 - - [01/Mar/2016:14:27:09 +0300] "GET /en/newsdetails.php?id=2524 HTTP/1.0" 200 79592 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.
41.196.6.101 - - [01/Mar/2016:14:27:10 +0300] "GET /staff/abduladhem/lectures/nashaat.doc HTTP/1.1" 200 66048 "http://www.google.com.eg/url?url=http://www.uokuf
207.46.13.154 - - [01/Mar/2016:14:27:10 +0300] "GET /ar/newsdetails.php?id=2557 HTTP/1.0" 200 111203 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing
41.196.6.101 - - [01/Mar/2016:14:27:14 +0300] "GET /staff/abduladhem/lectures/nashaat.doc HTTP/1.1" 206 58866 "http://www.google.com.eg/url?url=http://www.uokuf
180.76.15.9 - - [01/Mar/2016:14:27:15 +0300] "GET /ar/index.php/permalink/6876.html?print&output_type=rss&output_type=txt&output_type=rss&output_type=rss&output
157.55.39.236 - - [01/Mar/2016:14:27:17 +0300] "GET /ar/index.php/permalink/3247.html?print HTTP/1.0" 200 69184 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http
37.236.160.134 - - [01/Mar/2016:14:27:18 +0300] "GET /ar/themes/uokufa-ar/img/newsletter_bg.gif HTTP/1.0" 200 3871 "http://www.uokufa.edu.iq/ar/themes/uokufa-ar
207.46.13.52 - - [01/Mar/2016:14:27:24 +0300] "GET /staff/mohammadea/page2.html HTTP/1.1" 304 0 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/
```

Figure 2. Portion of Web Server Log file format

The result of this step produce the valid entries in log file, next step used to identifying unique users and distinguish users that belong to same IP address. The following algorithm in **Figure 3** used for eliminated irrelevant entries in log file data.

Data Cleaning Algorithm

```
Input: Web Server Log file data
Output: Log file data
Step1: Read log file record from (Web Server Log File).
Step2: IF    (log File Record) .URL == (gif, Css, Main.php, index.php )
          AND (Status code < 200 ^ Status code > 209)
```

```
            Remove from log file.
        End IF.
Step3: Repeat Step 1 and Step 2 until EOF (Web Server Log file).
Step4: Stop and Save file in Data base.
END
```

Figure 3. Data Cleaning Algorithm Steps

User Identification

Web Usage Mining does not required knowledge for user's identifying, there is a need to distinguish among different user's behavior. Server logs record of multiple sessions for user may visit Web site frequently. By absent authentication mechanisms in many Web Server some Web site used Cookies in Client-side, Due to privacy content this feature may disable by users, therefore IP address alone not sufficient to identify unique users in general by assigning many sessions to map IP address [15]. In case of absent user authentication and client-side cookies the possible accurate user identifying method by combination IP addresses with User agent and referrer [13]. The following figure (**FIGURE 4**) show user-identification algorithm steps that used for identifying different users from log file browsing data

```
User Identification Algorithm

Input: log file data
Output: Unique Users Table.
Step1: Initialization
        Create Table include the following field:
        ([User ID, IP's address, Date, Time, Request, Site name, User Agent, Size)].
Step2: Read record from Log file data
Step3: User's IP addresses of tow sequential records are compared.
Step4: IF ((IP address) is not in Users Table) THEN
            Assign User ID to IP address
            Add both to Users Table
        ELSE
        IF ((IP address) is in User's Table) THEN
        Check (User Agent if same) then Add it with Same User ID
        ELSE Assign (next User ID) to IP address
            Add both to Users Table
Step5: Repeat Step2-5 until EOF (log file data)
Step6: STOP, Store Result.
END.
```

Figure 4. User-Identification Algorithm steps

Phase 2: Mining Phase

Many techniques can be applied here after preprocessing phase to extract knowledge such as association rule mining, frequent pattern mining, Classification, Clustering etc.

Discovering and analysis users patterns

We are focus on clustering technique in case to extract knowledge about similar User's behaviors based on browsing Web sites characteristics. This technique help in many aspects for understanding similar user's interested content and Web sites contents, frequent User's- Site browsing content, Effects of Site content to Users and other indicators related to this work. **Table 4** illustrate information gained after pre-processing steps, based on this result we build appropriate Model for applying clustering algorithm to group of similar User's navigation behaviors.

Table 4. Identifying Unique user's navigation

User Id	IP address	Date	Time	Request	Site Name	Agent
User1	109.127.95.50	21/Oct/2015	15:01:12	GET	AR	"Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
User2	141.8.143.183	21/Oct/2015	15:01:15	GET	journals	"Mozilla/5.0
User3	37.236.136.3	21/Oct/2015	15:01:18	GET	AR	"Mozilla/5.0 (Windows NT 6.1)
User3	37.236.136.3	21/Oct/2015	15:01:26	GET	journals	"Mozilla/5.0 (Windows NT 6.1)
User1	109.127.95.50	21/Oct/2015	15:01:20	GET	conf	"Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
User4	157.55.39.119	21/Oct/2015	15:01:15	GET	journals	"Mozilla/5.0
User5	66.249.69.39	21/Oct/2015	15:01:26	GET	AR	"Mozilla/5.0
User5	66.249.69.31	21/Oct/2015	20:45:36	GET	Libr	"Mozilla/5.0
User5	66.249.69.31	21/Oct/2015	20:45:39	GET	AR	"Mozilla/5.0
User3	37.237.208.165	21/Oct/2015	20:45:42	GET	Libr	"Mozilla/5.0 (Windows NT 6.1)

Usage Data pre-processing result is a set of **M** Web sites views, **W** = {W_1, W_2, W_3 … W_m}, and a set of (**N**) user transactions, **T** = {t_1, t_2, t_3… t_n} where each (**t**$_i$) is a subset of W. For data mining tasks such as association rule mining and clustering the ordering of Web site views is not relevant, we represent each user transactions as a vector over M dimensional space of Web sites views. In most Web Usage mining algorithm and collaborative filtering applications weights were used to construct profiles of similar users. Weights may be user rating, spend time on that page and either binary representing the presence or absence of that user from page view, product view and Site view. In our situation we are deal with this cases by eliminated the records that visit main page due it's consider the gate for other Web sites links and consider the User spent time for each transaction have been made by a particular user. Spend time threshold used here to distinguish the users that browsed Sites for viewing Site component from others who search for particular content. Valid user's transactions treated to build Users-Web sites visit matrix, the following Table (**Table 5**) represent the occurrence of users based on valid transactions to construct Users-Web site visit matrix.

Table 5. User-Web site occurrences visit matrix

User Id	AR	Journals	Conf	Libr	Art	Busin	Comm	Educ	Gelog
User1	103	55	18	0	0	0	2	30	0
User2	20	29	33	1	0	355	240	0	2
User3	15	0	0	2	1	412	133	5	1
User4	207	72	51	0	1	4	1	28	2
User5	1	2	0	78	95	22	1	0	31
User6	1	3	3	1	0	107	77	1	5
User7	192	73	71	1	2	1	1	0	0
User8	1	1	0	0	0	0	0	55	109
User9	1	0	0	96	5	77	26	1	0
User10	7	84	31	0	0	0	11	2	0
User11	1	1	0	155	2	182	191	1	0
User12	98	65	24	1	1	0	0	14	0

Due to space of real Table we are show only small part, above Table show for example user1, user2, user4, user7 and user 12 are more interested in AR Web site while user5, user9 and user11 more interested for thesis and reading books from Library Web site. User –Web site Visit matrix produce many visit fractions for this purpose we consider the occurrences of all users to be within similar scale.

Cluster analysis and grouping similar users

Many data mining techniques can be applied in this manner to deal with fractions of User-Visit matric, for some data mining algorithm different range values lead to a tendency and immoderate influence for variables on the final result in order to scale the effect of it [16]. Normalization work well in this manner for small values close to (0.0) and higher ones to (1.0). MIN-MAX Normalization one of simplest and most used by scaling the difference by the range. The MIN-MAX formula is given as in Equation (1).

$$x_i = \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} \qquad (1)$$

Then by applying threshold the new value updated as in Equation (2)

$$X_i = \begin{bmatrix} 0 & if & x < T \\ 1 & if & x \geq T \end{bmatrix} \qquad (2)$$

Then new **Table 6** after applying Equation (1) and (2) represent a user's Web site visit matrix, clustering can be applied for the enhanced matrix to find groups of similar users based on browsing and navigation patterns. Given the mapping of user transactions into multi-dimensional space as enhanced vectors of Web sites visit as in **Table 6**, standard Hierarchal clustering algorithm can efficient employed here to take the similarity of groups of users members with respect to many Web sites visit patterns in the manner to form each possible number of groups n that have similar behaviors. Many clustering algorithm have been applied here some algorithms consider click stream to cluster dynamic users behaviors by using Mixture Models, this process can be too complex to be modeled by using basic probability distribution because each user may show different behaviors correspond to different tasks, different task reflect different distribution periodically in such application such as dynamic Web sites. Mixture Markov Models were applied in [17, 18] to cluster users based on similarities in navigation behaviors.

PROPOSED MODEL

In order to discover similar user navigation behavior, log data need to be preprocesses, eliminate non-relevant data then applying data minig techniques on result data. When applying data mining techniques on web data this is called Web data minig, Web mining inlcude several tasks based on problem found and interesting result. **Figure 4** show our proposed model for discovering hidden information in Log file data for varios users activities.

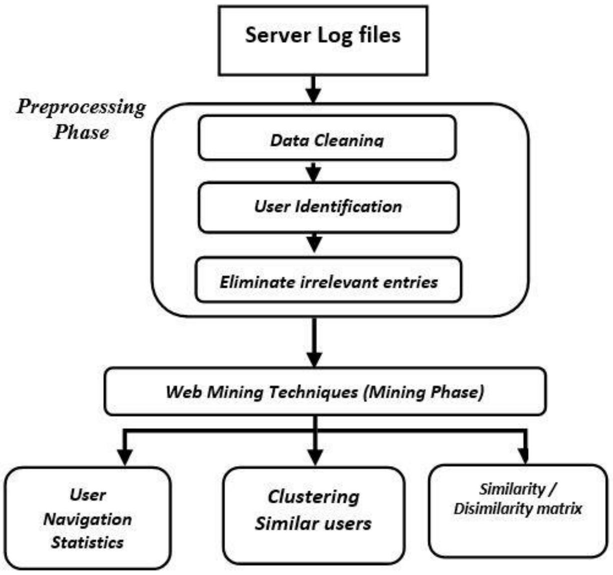

Server Log files
Preprocessing Phase
Data Cleaning
User Identification
Eliminate irrelevant entries
Web Mining Techniques (Mining Phase)
User Navigation Statistics
Clustering Similar users
Similarity / Disimilarity matrix

Figure 5. Proposed Diagram

After applying equation (1, 2) for the result Table 6, users with small number of visiting count were eliminate in each Web site, users with high visit count pick higher values and were grouped into users-Web site interest. Table 7 show Users-Visit intersect in each Web site

Table 6. User-Web sites Intersection matrix

User Id	AR	Journals	Conf	Libr	Art	Busin	Comm	Educ	Gelog
User1	1	1	0	0	0	0	0	1	0
User2	0	1	1	0	0	1	1	0	0
User3	0	0	0	0	0	1	1	0	0
User4	1	1	1	0	0	0	0	1	0
User5	0	0	0	1	1	0	0	0	1
User6	0	0	0	0	0	0	1	0	0
User7	1	1	1	0	0	0	0	0	0
User8	0	0	0	0	0	0	0	1	1
User9	0	0	0	1	0	0	0	0	0
User10	0	1	1	0	0	0	0	0	0
User11	0	0	0	1	0	1	1	0	0
User12	1	1	1	0	0	0	0	0	0

above **Table 6** show users with 0 values are (non-interest/non-visited) Web sites by corresponding users, while Web sites with 1 values refer to users are more interesting to visit and browsed contentf from those Web sites, from the above result we can infer for example user1 and user4 are more interest to Web site 1and 2, while user2 and user3 are interest to Web site 6 and 7.

Incase to find the similarities among users, many binary similarities measures can be applied here, in [19] list similarity and distance measures were applied in binary data. The goal of this measures is to find similarities among data points in our scenario we are consider this data as a dynamic because the behaviors of users may change through time and based on them interest, in this case we are applying 2 scenarios as follow :

- The first one is by using Cluster Identification Algorithm (CIA) which can be visualize grouping similar users in our Result Matrix and identified similar users by calculating the cells intersection ratio among them, this process yield blocks of similar users that share similar Web-Sites browsed and eliminated not browsed content Table 6 consider to applying CIA algorithm.
- Second scenario is by using distance measure , we are consider each user is a particular case and its browsed Websites differ from others, in order to find similarities among users so we are arranged Websites based navigation orders for example the result in **Table 4** and **Table 5** are combined together to form a vectors for users, we use character-Based coded to represent Web sites names to be simple for comparing, Users with small hits occurrences were not considered, user behaviors can discovered through continuously visited sites by users, relative frequency were calculated here for each user, as in Equation (3),

$$F(U_i) = \frac{m_i}{\sum M_i} \geq \ T \qquad (3)$$

Where users i = 1……N , F(U$_i$) relative frequency for user i, mi is hits count of user (i) in particular Web site J , Mi is total number of hits for user (i), finally T selected threshold.

Minimum values are discarded that does not satisfying the selected threshold, user's vectors result as follow:

Table 7. User-Web Sites Navigation behaviors

User Id	AR	Journals	Conf	Libr	Art	Busin	Comm	Educ	Gelog
User1	1	1	0	0	0	0	0	1	0
User2	0	0	0	0	0	1	1	0	0
User3	0	0	0	0	0	1	1	0	0
User4	1	1	1	0	0	0	0	0	0
User5	0	0	0	1	1	0	0	0	1
User6	0	0	0	0	0	1	1	0	0
User7	1	1	1	0	0	0	0	0	0
User8	0	0	0	0	0	0	0	1	1
User9	0	0	0	1	0	1	1	0	0
User10	0	1	1	0	0	0	0	0	0
User11	0	0	0	1	0	1	1	0	0
User12	1	1	1	0	0	0	0	0	0

Then find similarities among users, similar users are grouped together to form new cluster following **Table 8** show similarity matrix among users:

Table 8. User-Web Sites Navigation behaviors

User Id	User 1	User 2	User 3	User 4	User 5	User 6	User 7	**User 8**	**User 9**	**User 10**
User1	0	0	0	0.50	0	0	0.50	0.25	0	0.33
User2		0	1	0	0	1	0	0	0.67	0.67
User3			0	0	0	1	0	0	0.67	0
User4				0	0	0	1	0	0	0.67
User5					0	0	0	0.25	0.20	0
User6						0	0	0	0.67	0
User7							0	0	0	0.67
User8								0	0	0
User9									0	0
User10										0

Jaccard and Bray-Curtis for similarity/dissimilarity measures were applied , from **Table 7** many hierarchal algorithms can be applied for clustering similar users such as Single Linkage and Complete Linkage the result of **Table 7** after applying clustering algorithm has ben shown in **Table 8**. To find similarities among users Equation (4) used and result compare with distance measures used in Equation (5).

$S(U_A, U_B) = \dfrac{a}{a+b+c}$	(4)
$D(U_A, U_B) = \dfrac{\sum \lvert X_{ij} - X_{ik} \rvert}{\sum (X_{ij} + X_{ik})}$	(5)

Table 9. Users - Similarity values

ID	Cluster-name	Users
1	Cluster1	user1,user4,user7,user10,user12
2	Cluster2	user2,user3,user6,user9,user10,user11
3	Cluster3/4	user5 /user8

CONCLUSION

This paper focuse on discovering the hidden information from main server general log file, main server contain combination for all Web sites access information that hosted on it in text format, this file include navigation activities for many Web sites in order to understand the behaviors of users towards those sites not for single Web site, the contribution of the paper is to extract information from huge log file and consider novel approaches to deal and analysis users patterns, then extracted useful information for valid sessions after that clustering approach has been applied to grouping similar users navigations behaviors, this can give as indicators frequent users interest towards different Web sites content, monitor users activities for particular Web site, consume bandwidth for each user during selected period, monitor Web sites visits and browsed content and many others activities for future works.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Kosala and Blockeel: “Web mining research: A survey,” SIGKDD : SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000.

[2] 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.

[3] J. Hou and Y. Zhang: “Effectively finding relevant web pages from linkage Information.” IEEE Trans. Knowledge Data Eng., Vol. 15, No. 4, pp. 940-951, 2003.

[4] H. Han and R. Elmasri: “Learning rules for conceptual structure on the Web,” J. Intell. Inf. Syst., Vol. 22, No. 3, pp. 237-256, 2004.

[5] Renáta Iváncsy, István Vajk: “Frequent Pattern Mining in Web Log Data” , Acta Polytechnica Hungarica Vol. 3, No. 1, 2006 .

[6] Daniel MICAN, Nicolae TOMAI, Robert Ioan COROŞ: “Web Content Management Systems, a Collaborative Environment in the Information Society” , Informatica Economică vol.13, no 2/2009.

[7] Shiqun Yin, Yuhui Qiu, Chengwen Zhong, Jifu Zhou: “Study of Web Information Extraction and Classification Method”, Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007.

[8] M. A. Bayir, I. H. Toroslu, A. Cosar: “A Performance Comparison of Pattern Discovery Methods on Web Log Data” , AICCSA-06, the 4th ACS/IEEE International Conference on Computer Systems and Applications .

[9] M. Eirinaki and M. Vazirgiannis: “Web mining for web personalization” , ACM Trans. Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003.

[10] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu: “Mining access patterns efficiently from web logs” , Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, 2000, pp.396-407 .

[11] Kohavi, R: “Mining e-commerce data: The good, the bad, and the ugly” , Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, 8-13, 2001 .

[12] Anthony Scime: “Web mining: Application and techniques” , IDEA , chapter 19 , pp 373-376.

[13] R. Cooley, B. Mobasher, and J. Srivastava: “Data preparation for mining world wide web browsing patterns” , Knowledge and Information Systems, Vol. 1, No. 1, pp. 5-32, 1999 .

[14] L.K. Joshila Grace, V. Maheswari, and Dhinaharan Nagamalai: “Web Log Data Analysis and Mining” , Proc CCSIT-2011, Springer CCIS, Vol 133, pp 459-469, Jan 2011 .

[15] Bing Liu: “Web data mining, exploring Hyperlinks, Contents and Usage Data” , Second edition, Springer , pp-550-557, 2011 .

[16] Daniel T. Larose, Chantal D. Larose: “Data Mining and predictive analysis” , Second edition, WILEY , PP-28-35., 2015 .

[17] Cadez, I., D. Heckerman, C. Meek, P. Smyth, S. White: “Model-based clustering and visualization of navigation patterns on a web site” , Data Mining and Knowledge Discovery,7(4): p. 399-424, 2003 .

[18] Ypma, A., T. Heskes: “Automatic categorization of web pages and user clustering with mixtures of hidden Markov models” , In Proceedings of Mining Web Data for Discovering Usage Patterns and Profiles ,WEBKDD-2002, 2003.

[19] Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert: "A Survey of Binary Similarity and Distance Measures" Systemics, Cybernetics And Informatics , Volume 8 - Number 1, 2010.

[20] Paliouras, G., C. Papatheodorou, V. Karkaletsis, C. Spyropoulos: "Discovering user communities on the Internet using unsupervised machine Learning techniques" Interacting with Computers, 14(6): p. 761-791, 2002.

[21] Chen H., Fan H., Chau M., Zeng D.: "MetaSpider: meta-searching and categorization on the web" Journal of the American Society for Information Science and Technology, 52 (13), 1134–1147, 2001.

[22] Chung W.: "Visualizing E-Business stakeholders on the web: a methodology and experimental results " International Journal of Electronic Business, 2001.

[23] J. Punin, M. Krishnamoorthy, M. Zaki: "Web usage mining: Languages and algorithms ", in Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2001.

[24] P. Batista, M. ario, J. Silva: "Mining web access logs of an on-line newspaper", 2002.

[25] O. R. Zaiane, M. Xin, J. Han: "Discovering web access patterns andtrends by applying olap and data mining technology on web logs", in ADL '98: Proceedings of the Advances in Digital Libraries Conference.Washington, DC, USA: IEEE Computer Society, pp. 1-19, 1998.

[26] J. F. F. M. V. M. Li Shen, Ling Cheng, T. Steinberg: "Mining the most interesting web access associations", in WebNet 2000-World Conferenceon the WWW and Internet, pp. 489-494, 2000.

[27] M. Eirinaki, M. Vazirgiannis: "Web mining for web personalization", ACM Trans. Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003.

[28] X. Lin, C. Liu, Y. Zhang, X. Zhou: "Efficiently computing frequent tree-like topology patterns in a web environment", in TOOLS '99: Proceedings of the 31st International Conference on Technology of Object-Oriented Language and Systems. Washington, DC, USA: IEEE Computer Society, p. 440, 1999.

[29] X. A. Nanopoulos, Y. Manolopoulos: "Finding generalized path patterns for web log data mining", in ADBIS-DASFAA '00: Proceedings of the East-European Conference on Advances in Databases and Information Systems Held Jointly with International Conference on Database Systems for Advanced Applications. London, UK: Springer-Verlag, pp. 215-228 , 2000..