Using Web Logs Dataset via Web Mining for User Behavior Understanding

Zakaria Suliman Zubi , Mussab Saleh El Raiani

Abstract— Web usage mining focuses on the discovering of potential knowledge from the browsing patterns of the users. It leads us to find the correlation between pages in the analysis stage. The primary data source used in web usage mining is the server log-files (web-logs). Browsing web pages by the user leaves a lot of information in the log-file. Analyzing log-files information drives us to understand the behavior of the user. Web-logs include web server access logs and application server logs. Web-log is an essential part for web mining to extract the usage patterns and study the visiting characteristics of user. Our proposal focus on the use of web mining techniques to classify web pages type according to user visits. This classification helps us to understand the user behavior. Also we will use some classification and association rule techniques for discovering the potential knowledge from the browsing patterns.

Keywords— Web mining; web usage mining; classification, association rule; log-file, web-log, user behavior.

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data; it includes web content mining, web structure mining and web usage mining.

“Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications.” Note that web usage mining differs from web structure mining and web content mining,

In that web usage mining reflects the behavior of humans as they interact with the Internet. Because of this, web usage mining is of intense interest for e-marketing and e-commerce professionals. Analysis of user behavior can provide insights leading to customization and personalization of a user’s web Experience.[22]

Using standard data mining techniques such as clustering and association rules , aparticular user may be associated with other users exhibiting similar behavior patterns and preferences. Then this user may be offered specialized links and sales opportunities tailored to his or her own preferences, based on information provided by the clustering or association rule algorithms. For example, the e-vendor may provide a choice of items to the user based on items the user has already browsed. “Customers who bought this book also bought . . . ” (from Amazon.com, arguably the world leaderin applied web usage mining). Recommendation making is one of the most common

applications of knowledge gained through web usage mining.[23]

Web Usage Mining is a special type of web mining tool, which can discover the knowledge in the hidden browsing patterns and analyses the visiting characteristics of the user. It is a complete process that includes various stages of data mining cycle, including Data Preprocessing, Pattern Discovery & Pattern Analysis [6]. Initially, the web log is preprocessed to clean, integrate and transform into a common log. Later, Data mining techniques are applied to discover the interesting characteristics in the hidden patterns. Pattern Analysis is the final stage of web usage mining which can validate interested patterns from the output of pattern discovery [1,2].

In the pre-processing stage, the data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site [8]. Other sources of knowledge such as the site content or structure may also be used in pre-processing or to enhance user transaction data.

In the pattern discovery stage, statistical, database, and machine learning operations are performed to obtain hidden patterns reflecting the typical behavior of users, as well as summary statistics on Web resources, sessions, and users [5]. In the final stage of the process, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools, and Web analytics and report generation tools figure 1 summarize the

Zakaria Suliman Zubi- He is an Associate Professor since 2010. Works currently at Sirte University, Faculty Of Science, Computer Science Department Sirte, P.O Box 727, Libya,. Email : {zszubi@yahoo.com}.

Mussab Saleh El Raiani – He is a postgraduate student at the Libyan Academy, Computer Science Department, Misurata, Libya. Email: {msb_rio@yahoo.com}

process [3,4].

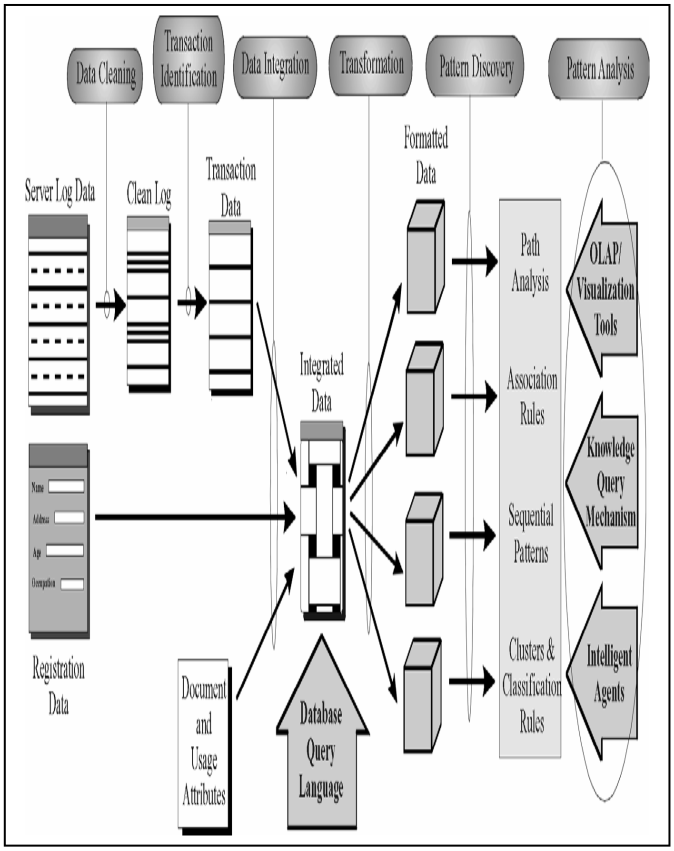


Fig. 1: the web usage mining process

Before we begin we must familiarize ourselves with the types of data forms available for the analysis of user behavior. Web usage information takes the form of web server log files, or web logs. For each request from a user's browser to a web server, a response is generated automatically, called a web log file, log file, or web log (not to be confused with blogs, of course, which are essentially web journals, sometimes, called web logs).

II. WEB SERVER LOG FILES

Before we begin we must familiarize ourselves with the types of data forms available for the analysis of user behavior. Web usage information takes the form of web server log files, or web logs. For each request from a user's browser to a web server, a response is generated automatically, called a web log file, log file, or web log (not to be confused with blogs, of course, which are essentially web journals, sometimes called web logs).

This response takes the form of a simple single-line transaction record that is appended to an ASCII text file on the web server. This text file may be comma-delimited, space-delimited, or tab-delimited.[23]

A standard log-file has the following format:

[**remotehost**; logname; username; date; request; status code; bytes] where:

- ***remotehost***: This field consists of the Internet IP address of the remote host making the request,such as 141.243.1.172". If the remote host name is available through a DNS lookup,this name is provided, such as "wpbfl2-45.gate.net." to obtain the domain name of the remote host rather than the IP address,the server must submit a request, using the Internet domain name system (DNS) to resolve (i.e., translate) the IP address into a host name. Since humans prefer to work with domain names and computers are most efficient with IP addresses, the DNS system provides an important interface between humans and computers.

- ***logname***: This field is used to store the authenticated client user name, if it is required. The logname field was designed to contain the authenticated user name information that a client needs to provide to gain access to directories that are password protected. If no such information is provided, the field defaults to a hyphen.

- ***username***: is the username with which the user has authenticated himself,

- ***date/time***: field format:[DD:HH:MM:SS] where DD represents the day of the month and HH:MM:SS represents the 24-hour time, However,it is more common for the date/time field to follow the following format: "DD/Mon/YYYY:HH:MM:SS offset," where the offset is a positive or negative constant indicating in hours how far ahead of or behind the local server is from Greenwich MeanTime (GMT). For example, a date/time field of "09/Jun/1988:03:27:00 -0500" indicates that a request was made to a server at 3:27 a.m. on June 9, 1988, and the server is 5 hours behind GMT.

- ***HTTP request:*** The HTTP request field consists of the information that the client's browser has requested from the web server.

The entire HTTP request field is contained within quotation marks. Essentially, this field may be partitioned into four areas:

(1) the request method.

(2) the uniform resource identifier (URI).
(3) the header.
(4) the protocol.

The most common request method is GET, which represents a request to retrieve data that are identified by the URI.

Status code: The status code field provides a three-digit response from the web server to the client's browser, indicating the status of the request. A sample of the possible status codes that a web server could send follows.

1- Successful transmission (200 series)

Indicates that the request from the client was received, understood, and completed.

200: success
201: created
202: accepted
204: no content

2- Redirection (300 series)

Indicates that further action is required to complete the client's request.

301: moved permanently
302: moved temporarily
303: not modified
304: use cached document

3- Client error (400 series)

Indicates that the client's request cannot be fulfilled, due to incorrect syntax or a missing file.

400: bad request
401: unauthorized
403: forbidden
404: not found

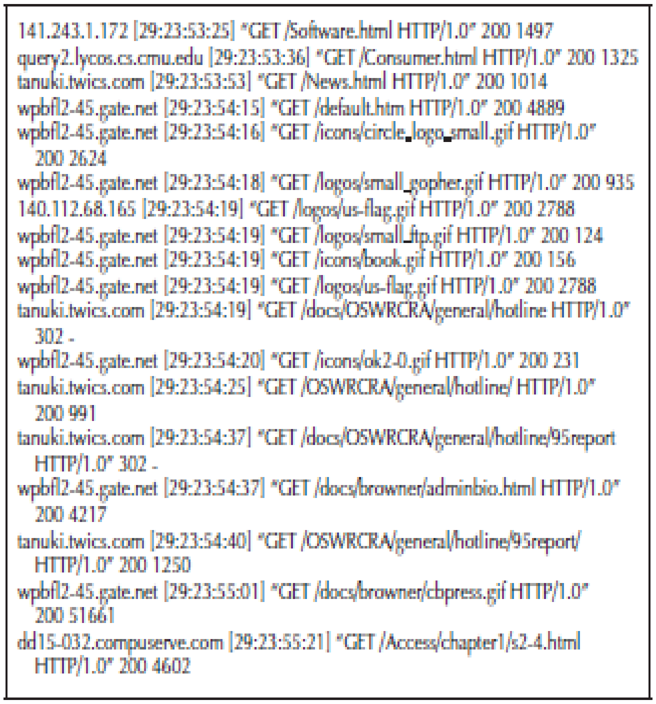
4-Server error (500 series)

Indicates that the web server failed to fulfill what was apparently a valid request.

500: internal server error
501: not implemented
502: bad gateway
503: service unavailable.

- ***Bytes***: The transfer volume field indicates the size of the file (web page, graphics file, etc.),in bytes, sent by the web server to the client's browser. Only GET requests that have been completed successfully (Status = 200) will have a positive value in the transfer volume field. Otherwise, the field will consist of a hyphen or a value of zero. This field is useful for helping to monitor the network traffic, the load carried by the network throughout the 24-hour cycle is the content-length of the document transferred[24].

The following figure is a fragment of a common log-file:



```
141.243.1.172 [29:23:53:25] "GET /Software.html HTTP/1.0" 200 1497
query2.lycos.cs.cmu.edu [29:23:53:36] "GET /Consumer.html HTTP/1.0" 200 1325
tanuki.twics.com [29:23:53:53] "GET /News.html HTTP/1.0" 200 1014
wpbfl2-45.gate.net [29:23:54:15] "GET /default.htm HTTP/1.0" 200 4889
wpbfl2-45.gate.net [29:23:54:16] "GET /icons/circle_logo_small.gif HTTP/1.0"
  200 2624
wpbfl2-45.gate.net [29:23:54:18] "GET /logos/small_gopher.gif HTTP/1.0" 200 935
140.112.68.165 [29:23:54:19] "GET /logos/us-flag.gif HTTP/1.0" 200 2788
wpbfl2-45.gate.net [29:23:54:19] "GET /logos/small_ftp.gif HTTP/1.0" 200 124
wpbfl2-45.gate.net [29:23:54:19] "GET /icons/book.gif HTTP/1.0" 200 156
wpbfl2-45.gate.net [29:23:54:19] "GET /logos/us-flag.gif HTTP/1.0" 200 2788
tanuki.twics.com [29:23:54:19] "GET /docs/OSWRCRA/general/hotline HTTP/1.0"
  302 -
wpbfl2-45.gate.net [29:23:54:20] "GET /icons/ok2-0.gif HTTP/1.0" 200 231
tanuki.twics.com [29:23:54:25] "GET /OSWRCRA/general/hotline/ HTTP/1.0"
  200 991
tanuki.twics.com [29:23:54:37] "GET /docs/OSWRCRA/general/hotline/95report
  HTTP/1.0" 302 -
wpbfl2-45.gate.net [29:23:54:37] "GET /docs/browner/adminbio.html HTTP/1.0"
  200 4217
tanuki.twics.com [29:23:54:40] "GET /OSWRCRA/general/hotline/95report/
  HTTP/1.0" 200 1250
wpbfl2-45.gate.net [29:23:55:01] "GET /docs/browner/cbpress.gif HTTP/1.0"
  200 51661
dd15-032.compuserve.com [29:23:55:21] "GET /Access/chapter1/s2-4.html
  HTTP/1.0" 200 4602
```

Fig. 2: Portion of a typical server log

III. Phases of Web Usage Mining

Web Usage Mining has been defined as the application of data mining techniques to large Web data repositories in order to extract usage patterns, namely the visitor behavior [7]. As further step, pattern discovery and patter analysis allow for profiling users and their preferences. For that statistical methods play a fundamental role. It is definitively possible to identify suitable attributes and main features characterizing a typology of users, thus providing a Web personalization [11]. Statistical methods could be classification, association rules and clustering analysis.

A. Phase 1: Preprocessing:

Main task of user profiling is to organize data such to collect consistent information about the users. A distinction can be made between explicit information and implicit information: the former is derived from the data provided by the user filling any type of application form, whereas the latter can be induced by the log-files, the cookies as well as any other method of tracking the web navigation [18]. The data can be collected at two levels. First, at the Server level, data are memorized into log-files, recording all the activities of the Server such for instance; the Web page selections. One problem is that some information cannot be captured by the Server, as is the case of the caching where users visualizes pages that are already loaded in the computer memory[10].

The cookies are very useful for the tracking activity as in fact they keep memory of the Web navigation in a text file sent to the user browser [9]. One of the drawbacks is imputed to the privacy, indeed the user could cancel the cookies from the own computer so that the link with the Server can be lost. At the Client level, instead, the so-called agents are suitable applications to memorize huge data sets recording the browser activity. Such applications can be included directly in the browser tool box but the analysis cannot be complete as the browser cannot catch the whole Web navigation.

Both implicit and explicit information can be jointly considered although they keep a fundamental distinction [19].

Whereas data for the explicit information is collected in a proper form for the statistical analysis, data for the implicit information need a pre-processing process. This includes the following activities with their brief description.

Several problems exist during the preprocessing phase where log-files are transformed into a form that is suitable for mining. The following are preprocessing tasks that have been identified:

1) <u>Data Cleaning:</u> The log-file is first examined to remove irrelevant entries such as those that represent multimedia data and scripts or uninteresting entries such as those that belong to top/bottom frames.

2) <u>Page view Identification:</u> Identification of page views is heavily dependent on the intra-page structure of the site, as well as on the page contents and the underlying site do-main knowledge. each page view can be viewed as a collection of Web objects or resources representing a specific “user event,”.

3) <u>User Identification</u>: Since several users may share a single machine name, certain heuristics are used to identify users; we use the phrase user activity record to refer to the sequence of logged activities belonging to the same user. In general, the user identification procedure could be used to identify users:

   a. *Sort the web log file by ID address and then by time stamp.*

   b. *For each distinct ID address, identify each agent as belonging to a different user.*

   c. *For each user identified in step 2, apply path information garnered from the referrer field and the site topology to determine whether this behavior is more likely the result of two or more users.*

   d. *To identify each user, combine the user identification information from steps 1 to 3 with available cookie and registration information.[5]*

4) <u>Session identification</u> :aims to split the page access of each user into separated sessions. defines the number of times the user has accessed a web page and time out defines a time limit for the access of particular web page for more than 30 minutes if more the session will be divided in more than one session.

<u>Session Identification Procedure</u>

The session identification procedure may be summarized as follows:

1) For each distinct user identified in the preceding section, assign a unique session ID.

2) Define the timeout threshold *t.*

3) For each user, perform the following:

   a. *Find the time difference between every two consecutive web log entries.*

   b. *If this difference exceeds the threshold t, assign a new session ID to the later entry.*

4) Sort the entries by session ID.[5]

<u>Sessionization heuristics fall into two basic categories:</u>

Time-oriented heuristics apply either global or local time-out estimates to distinguish between consecutive sessions.

Example:

<u>User:pc1</u>

Session1

05:10	*http://www.databaseanswers.org*	*pc1*	*Education*
05:22	*http://database.firstnormalform.htm*	*pc1*	*Education*

Session2

07:43	*http://www.aljazeera.net/news/arabic*	*Pc1*	*News*
07:44	*http://www.aljazeera.net/news*	*Pc1*	*News*
07:50	*http://www.complete-review.com/*	*Pc1*	*Social*

<u>User2:pc2</u>

Session1

05:13	*http://www.dotnetperls.com*	*Pc2*	*Education*
05:17	*http://www.dotnetperls.com/data*	*Pc2*	*Education*
05:26	*http://www.java2s.com*	*Pc3*	*Education*
05:30	*http://www.java2s.com/.htm*	*Pc3*	*Education*

Session2

Fig. 3: Example of sessionization with a time-oriented heuristic

- <u>Path Completion</u>: Not all page views seen by the user are recorded in the web server log. For example, many people use the “Back” button on their browsers to return to a page viewed previously.

When this happens, the browser returns to a page that was previously cached locally rather than accessing the web server again. This leads to “holes,” missing pages, in the web server’s record of the user’s path through the Web site.Knowledge of site topology must be applied to complete these paths, in a process known as path completion.

Once the missing pages have been identified, they are inserted into the session file along with an estimate of the duration spent on the missing page. These duration estimates may be classified according to whether the missing page is a navigation page, with a shorter duration estimate, or a content page, with a longer duration estimate [23].

- Transaction Identification: Once a session of completed paths is determined, page references must be grouped into logical units representing Web transactions before any mining can be carried out.

These operations can be crucial for a correct identification of the initial data set. It depends on the state of the dataset and the purposes of the analysis. Pre-processing is an important step of Web Mining, as it is fundamental for identifying uniquely sessions and/or users [16]. Some situations represent an obstacle to the correct identification of users and/or sessions.

- Single IP address – more sessions: It is not evident that to one single IP address does not correspond necessarily one session and one user. Indeed, if the provider uses a Proxy Server then to a single IP address it can correspond more requests of different users;

- More sessions – single IP address: In order to secure the privacy of navigators some providers or applications assign the same IP address to more requests (also from the same user), thus it is impossible to consider a single IP address to a single user;

- More IP addresses – single User: An user can connect to the same Web site by more internet points, thus allowing for tracking the Web site paths of each connection;

- More browsers – single User: It is sufficient that the user navigation is through another browser so that it is considered as a new user from the Agent [17].

In this work, we assumed that each address/agent/operating system as a single user, also we added the login information to the log file to get user name

B. Phase 2: Mining

There are four main mining techniques that can be applied to Web access logs to extract knowledge, but we will focus on algorithms based on association rule mining (ARM) and classification. Here are the four techniques:

- Sequential-pattern-mining-based: Allows the discovery of temporally ordered Web access patterns.

- Association-rule-mining-based: Finds correlations among Web pages type.

- Clustering-based: Groups users with similar characteristics.

- Classification-based: Groups users into predefined classes based on their characteristics.

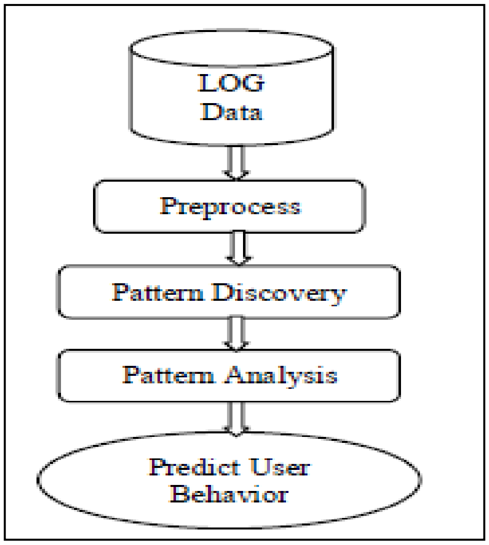


Fig. 4: Web Log Mining System Structure

C. Phase 3: Applying Mining Results

The last phase of WUM involves the analysis and translation of mining results into useful actionable tasks such as the following:

- Re-design Websites so that correlated pages type are found together.

- Improve access time by perfecting pages types frequently accessed sequentially.

- Improve caching by storing pages types frequently revisited.

- Enhance surfing experience by relocating pages in such a way that users need not visit unnecessary pages types to get to their desired pages.

IV. Web logs

A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. These log-files can be stored in various formats one of these formats is shown in figure 5.

However, the site usage data recorded by server logs may not be completely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not stored in a server log. Also, any essential information will not be available in a server log sometimes [15]. The Web server logs can also store other kinds of usage information such as cookies and query data in separate logs.

Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. Cookies rely on implicit user cooperation and thus have raised growing concerns regarding user privacy. Query data is also usually generated by online visitors while searching for sites types relevant to their information needs [11]. Besides usage data, the server side also provides content data, structure information and Web page meta-information (such as the size of a file and its last modified time).

Time	URL	User	Type
05:10	http://www.databaseanswers.org	pc1	Education
05:13	http://www.dotnetperls.com	Pc2	Education
05:17	http://www.dotnetperls.com/data	Pc2	Education
05:22	http://database.firstnormalform.htm	pc1	Education
05:26	http://www.java2s.com	Pc3	Education
05:30	http://www.java2s.com/.htm	Pc3	Education
07:30	http://www.ferryhalim.com/orisinal/	Pc2	Entertainment
07:33	http://www.ferryhalim.com/oris.htm	pc2	Entertainment
07:40	http://www.complete-review.com/	Pc3	Social
07:42	http://www.complete.com/main.html	Pc3	Social
07:43	http://www.aljazeera.net/news/arabic	Pc1	News
07:44	http://www.aljazeera.net/news	Pc1	News
07:50	http://www.aljazeera.net/news/pages	Pc1	News

Fig. 5: Simulated web server logs.

V. Classification Techniques

Classification is the task of mapping a data item into one of several predefined classes [13]. In the web domain, the user may develop a profile which belongs to a particular class or category. This requires extraction and selection of features that best describe the properties of the given class or category. Classification can be done by using supervised learning algorithms such as decision trees, naive Bayesian classifiers, k-nearest neighbor classifiers, and Support Vector Machines.

Classification techniques play an important role in Web analytics applications for modeling the users according to various predefined metrics. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category [1, 5]. This requires extraction and selection of features that best describe the properties of a given class or category. We will focus on k-nearest neighbor (K-NN) which was considered as a predictive technique for classification models. Whereas; k represents a number of similar cases or the number of items in the group.

In K-NN technique the training data in the model when a new case or instance is presented to the model, the algorithm looks at all data to find a subset of cases that one most similar to it and uses them to predict the outcome. Using classification on server logs may lead to the discovery of interesting rules such as: 30% of users who placed an online order in /Product/Music are in the 18-25 age groups and live on the West Coast.

VI. Association Rule Mining Techniques

Association rule mining discovery and statistical correlation analysis can find groups of web pages types that are commonly accessed together (Association rule mining can be used to discover correlation between pages types found in a web log), This, in turn, enables Web sites to organize the site content more professionally [12].

Most common approaches to association discovery are based on the Apriori algorithm. This algorithm finds groups of items (page-views appearing in the preprocessed log) occurring frequently together in many transactions (i.e., satisfying a user specified minimum support threshold). Such groups of pages types are referred to as frequent datasets. Association rules which satisfy a minimum confidence threshold are then generated from the frequent datasets.

For example, association rule discovery using the Apriori algorithm [14] (or one of its variants) may disclose a correlation between users who visited a page containing electronic products to those who access a page about sporting equipment. Away from being applicable for business and marketing applications, the presence or absence of such rules can help Web designers to restructure their Web site. The association rules may also serve as a heuristic for prefetching documents in order to reduce user-perceived latency when loading a page from a remote site.

Discovers the correlations between pages types that are most often referenced together in a single server session provide the information to answer such questions as:

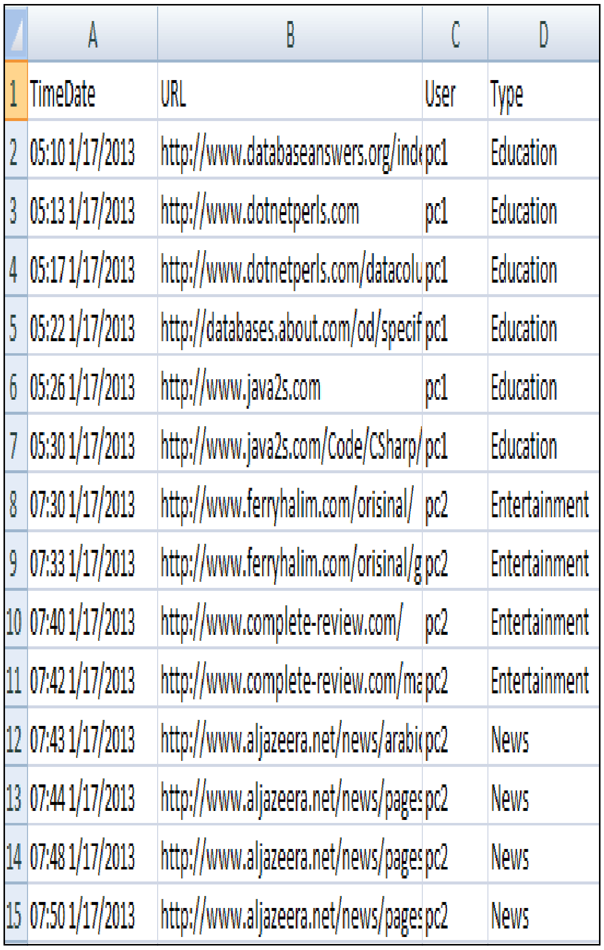
- What are the set of pages types frequently accessed together by web users?

- What page type will be fetched next?

VII. Implementation of Experiments & Results of using Association Rules

A. Prepare the simulated log-file

We convert the log files into a flat Excel file with extension xls, whereas, shown in the table:

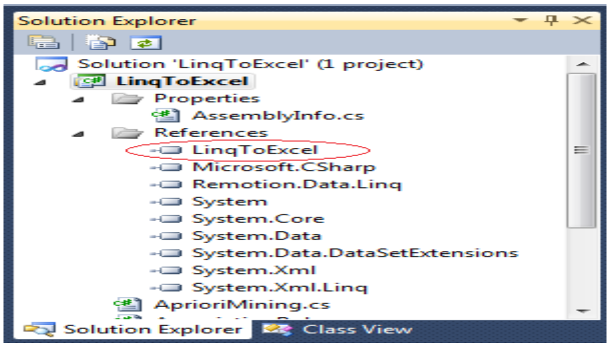


	A	B	C	D
1	TimeDate	URL	User	Type
2	05:10 1/17/2013	http://www.databaseanswers.org/inde	pc1	Education
3	05:13 1/17/2013	http://www.dotnetperls.com	pc1	Education
4	05:17 1/17/2013	http://www.dotnetperls.com/datacolu	pc1	Education
5	05:22 1/17/2013	http://databases.about.com/od/specif	pc1	Education
6	05:26 1/17/2013	http://www.java2s.com	pc1	Education
7	05:30 1/17/2013	http://www.java2s.com/Code/CSharp/	pc1	Education
8	07:30 1/17/2013	http://www.ferryhalim.com/orisinal/	pc2	Entertainment
9	07:33 1/17/2013	http://www.ferryhalim.com/orisinal/g	pc2	Entertainment
10	07:40 1/17/2013	http://www.complete-review.com/	pc2	Entertainment
11	07:42 1/17/2013	http://www.complete-review.com/ma	pc2	Entertainment
12	07:43 1/17/2013	http://www.aljazeera.net/news/arabic	pc2	News
13	07:44 1/17/2013	http://www.aljazeera.net/news/pages	pc2	News
14	07:48 1/17/2013	http://www.aljazeera.net/news/pages	pc2	News
15	07:50 1/17/2013	http://www.aljazeera.net/news/pages	pc2	News

Fig. 6. Simulated excel web server log

B. Connecting C# to log-file table

We used C# programming language to import log-file table to the application by using Linq to Excel .Net library .

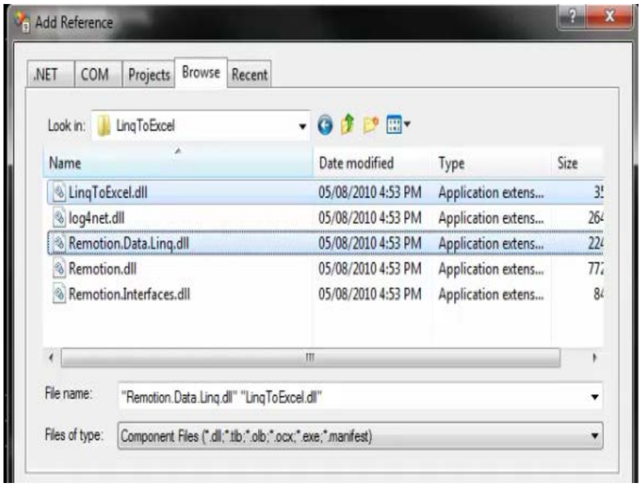


```
Solution 'LinqToExcel' (1 project)
  LinqToExcel
    Properties
      AssemblyInfo.cs
    References
      LinqToExcel
      Microsoft.CSharp
      Remotion.Data.Linq
      System
      System.Core
      System.Data
      System.Data.DataSetExtensions
      System.Xml
      System.Xml.Linq
    AprioriMining.cs
```

Fig.7. Adding library to at the solution explorer

C. Browsing to downloaded library by adding

- LinqToExcel.dll
- Remotion.Data.Linq.dll



Look in: LinqToExcel

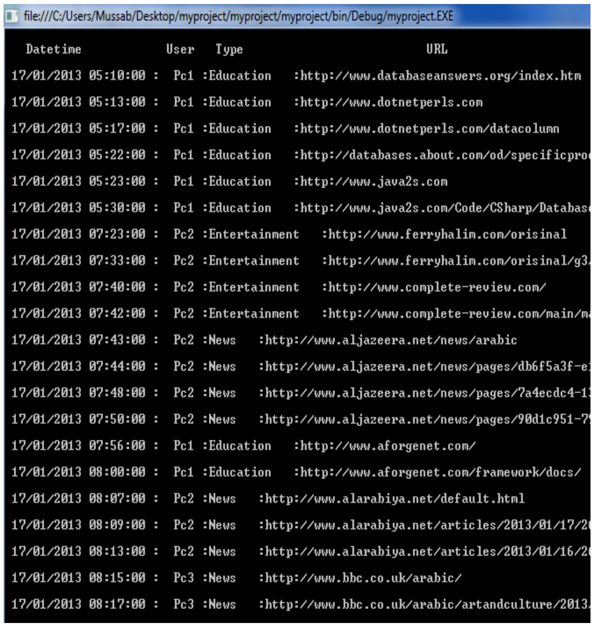
Name	Date modified	Type	Size
LinqToExcel.dll	05/08/2010 4:53 PM	Application extens...	3!
log4net.dll	05/08/2010 4:53 PM	Application extens...	26
Remotion.Data.Linq.dll	05/08/2010 4:53 PM	Application extens...	22
Remotion.dll	05/08/2010 4:53 PM	Application extens...	77
Remotion.Interfaces.dll	05/08/2010 4:53 PM	Application extens...	8

File name: "Remotion.Data.Linq.dll" "LinqToExcel.dll"

Files of type: Component Files (*.dll;*.tlb;*.olb;*.ocx;*.exe;*.manifest)

Fig 8 browsing to library references

D. Import log-file database to the application.

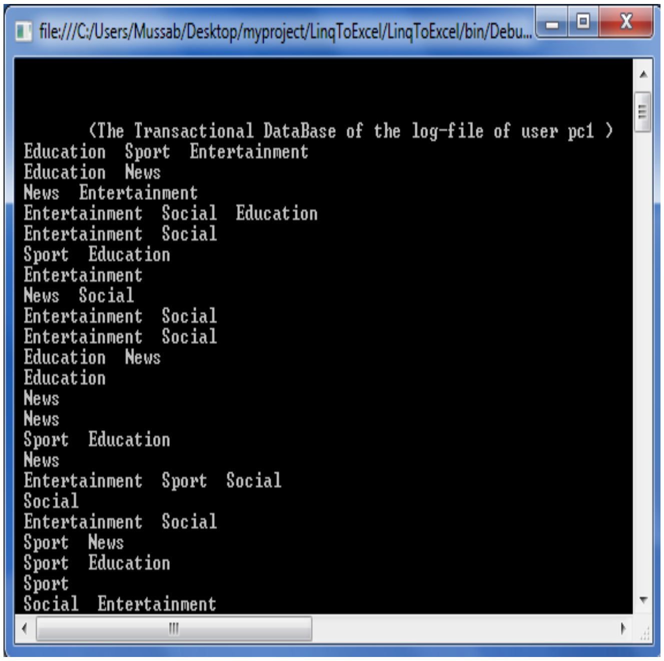


```
file:///C:/Users/Mussab/Desktop/myproject/myproject/myproject/bin/Debug/myproject.EXE

 Datetime            User  Type                     URL
17/01/2013 05:10:00 :  Pc1 :Education   :http://www.databaseanswers.org/index.htm
17/01/2013 05:13:00 :  Pc1 :Education   :http://www.dotnetperls.com
17/01/2013 05:17:00 :  Pc1 :Education   :http://www.dotnetperls.com/datacolumn
17/01/2013 05:22:00 :  Pc1 :Education   :http://databases.about.com/od/specificpro
17/01/2013 05:23:00 :  Pc1 :Education   :http://www.java2s.com
17/01/2013 05:30:00 :  Pc1 :Education   :http://www.java2s.com/Code/CSharp/Databas
17/01/2013 07:23:00 :  Pc2 :Entertainment   :http://www.ferryhalim.com/orisinal
17/01/2013 07:33:00 :  Pc2 :Entertainment   :http://www.ferryhalim.com/orisinal/g3
17/01/2013 07:40:00 :  Pc2 :Entertainment   :http://www.complete-review.com/
17/01/2013 07:42:00 :  Pc2 :Entertainment   :http://www.complete-review.com/main/m
17/01/2013 07:43:00 :  Pc2 :News   :http://www.aljazeera.net/news/arabic
17/01/2013 07:44:00 :  Pc2 :News   :http://www.aljazeera.net/news/pages/db6f5a3f-e
17/01/2013 07:48:00 :  Pc2 :News   :http://www.aljazeera.net/news/pages/7a4ecdc4-1
17/01/2013 07:50:00 :  Pc2 :News   :http://www.aljazeera.net/news/pages/90d1c951-7
17/01/2013 07:56:00 :  Pc1 :Education   :http://www.aforgenet.com/
17/01/2013 08:00:00 :  Pc1 :Education   :http://www.aforgenet.com/framework/docs/
17/01/2013 08:07:00 :  Pc2 :News   :http://www.alarabiya.net/default.html
17/01/2013 08:09:00 :  Pc2 :News   :http://www.alarabiya.net/articles/2013/01/17/2
17/01/2013 08:13:00 :  Pc2 :News   :http://www.alarabiya.net/articles/2013/01/16/2
17/01/2013 08:15:00 :  Pc3 :News   :http://www.bbc.co.uk/arabic/
17/01/2013 08:17:00 :  Pc3 :News   :http://www.bbc.co.uk/arabic/artandculture/2013
```

Fig. 10. Loading web server log

E. Extract the transactional database of web sever log for every user where every transaction represnts a session.

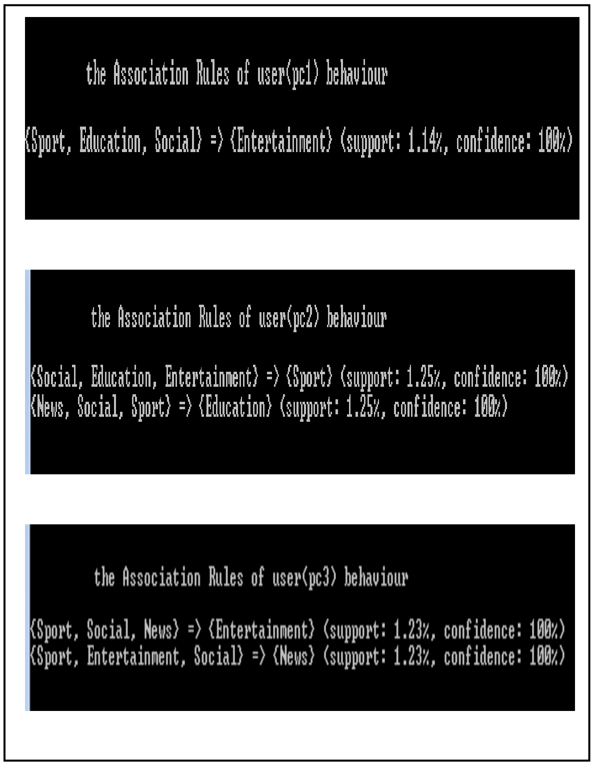

file:///C:/Users/Mussab/Desktop/myproject/LinqToExcel/LinqToExcel/bin/Debu...

```
        (The Transactional DataBase of the log-file of user pc1 )
Education  Sport  Entertainment
Education  News
News  Entertainment
Entertainment  Social  Education
Entertainment  Social
Sport  Education
Entertainment
News  Social
Entertainment  Social
Entertainment  Social
Education  News
Education
News
News
Sport  Education
News
Entertainment  Sport  Social
Social
Entertainment  Social
Sport  News
Sport  Education
Sport
Social  Entertainment
```

Fig. 11,Sample of Transaction database for user pc1

F. find the association rules of user behavior

The below figure illustrates the Rule Mining results after applying the Apriori algorithm to the transactional database of the identified user.



```
        the Association Rules of user(pc1) behaviour

{Sport, Education, Social} => {Entertainment} (support: 1.14%, confidence: 100%)
```

```
        the Association Rules of user(pc2) behaviour

{Social, Education, Entertainment} => {Sport} (support: 1.25%, confidence: 100%)
{News, Social, Sport} => {Education} (support: 1.25%, confidence: 100%)
```

```
        the Association Rules of user(pc3) behaviour

{Sport, Social, News} => {Entertainment} (support: 1.23%, confidence: 100%)
{Sport, Entertainment, Social} => {News} (support: 1.23%, confidence: 100%)
```

Fig. 12: Samples of the association rules of users (pc1,pc2,pc3) behavior

VIII. CONTRIBUTION OF THIS WORK

In this work we extracted a potential knowledge from the log-files while browsing patterns via users accessing web pages types. We used classification techniques to classify web pages that have been visited by users. For instance, if the user visited many pages randomly these pages can be classified in many profiling classes such as sports, economic, political, etc. Suppose that this user visited sports pages 3 times, economic, political pages one time, and then we can say that this user has a sporty behavior. Association rule mining is used as well to find the correlation between these web pages.

IX. CONCLUSION

This paper has tried to provide the important of the rapidly growing area of Web Usage mining. With the growth of Web-based applications, particularly, there is a significant interest in analyzing Web usage data to better understand Web usage data, and apply the knowledge to better serve users. The web data source used in web usage mining is the server log-files (web-logs). It contained all the information that the user leaves when accessing the web pages. Web logs are preprocessed to clean, integrate and transform into a common log. Later, Data mining techniques are applied to discover the interesting characteristics in the hidden patterns. Pattern Analysis is applied which can validate interested patterns from the output of pattern discovery.

We used statistical methods such as classification, association rule mining discovery and statistical correlation analysis which can find groups of web pages types that are commonly accessed together. Classification is used to map the data item into one of several the predefined classes. The class will belongs into one category such as sport or politics or education or..etc. In this paper, the selecting criteria of the features that best describe the properties of a given class or category is conducted via k-nearest neighbor (K-NN) algorithm.

Association rule mining can be used to discover correlation between sites types found in a web log. This, in turn, enables web sites to organize the site content more efficiently [10, 11]. Web data are a real source to analyze the user behavior in the web. An important step is preprocessing of the web data in the web log-files It also allow us to find the unknown patterns that must be validated or rejected by an expert in a wide professionalism after investigation to help us personalize any web site.

ACKNOWLEDGMENT

This work has been fully supported and funded by the Libyan Higher Education Ministry throughout Sirte University.

REFERENCES

[1] Abraham and V. Ramos. Web usage mining using artificial ant colony clustering and genetic programming. In Procs. Int. Conf. CEC03 on Evolutionary Computation, pages 1384–1391. IEEE Press, 2003.

[2] Cyrus Shahabi, Amir M Zarkesh, Jafar Adibi, and Vishal Shah. Knowledge discovery from users web-page navigation. In Workshop on Research Issues in Data Engineering, Birmingham, England, 2005.

[3] J. D. Velásquez, H. Yasuda, T. Aoki and R. Weber, A new similarity measure to understand visitor behavior in a web site, IEICE Transactions on Information and Systems, Special Issues on Information Processing Technology for web utilization, E87-D(2): 389-396,2004.

[4] J. g. Liu, h. h. Huang. Web Ming for Electronic Business Application, Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, Chengdu, China, 2003:872~876.

[5] J. Srivastava, R. Cooley, M. Deshpande, P. N. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations,2003, 1(2):1~12.

[6] Kun-lung Wu, Philip S Yu, and Allen Ballman. Speed- tracer: A web usage mining and analysis tool. IBM Systems Journal, 37(1), 2006.

[7] M. Eirinaki, M. Vazirgiannis. Web mining for web personalization. ACM Transactions on Internet Technology(TOIT), 2003(1): 1~27.

[8] M.S. Chen, J. Han, and P.S. Yu. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6):866-883, 2008.

[9] Olfa Nasraoui, Raghu Krishnapuram, and Anupam Joshi. Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In Eighth International World Wide Web Conference, Toronto, Canada, 2007.

[10] O. R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In Advances in Digital Libraries, pages 19{29, Santa Barbara, CA, 2007.

[11] Punin, J., Krishnamoorthy, M., & Zaki, M. (2001). LOGML - Log Markup Language for Web Usage Mining. Paper presented at the Proceedings of the WEBKDD Workshop 2001: Mining Log Data Across All Customer TouchPoints (with SIGKDD01), San Francisco, USA.

[12] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of the 20th VLDB Confer- ence, pages 487{499, Santiago, Chile, 2008.

[13] Robert Cooley, Bamshad Mobasher, and Jaideep Sri-vastava. Web mining: Information and pattern discovery on the world wide web. In International Conference on Tools with Arti_cial Intelligence, pages 558-567, Newport Beach, 2009. IEEE.

[14] Srivastava, J. et al. (2000). Web usage mining: Discovery and applications of usage patterns from Web data, ACM SIGKDD Explorations, 1(2).

[15] S. W. Changchien, T. Lu. Mining association rules procedure to support on-line recommendation by customers and products fragmentation. Expert Systems with Applications, 2008(20): 325~335.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery: An overview. In Fayyad U., Piatetsky-Shapiro G., Smyth P.,Uthurusamy R, editors, Advances in Konwledge Discovery and Data Mining, AAAI/MIT Press, 2007:1~34.

[17] Zakaria Suliman Zubi and Marim Aboajela Emsaed. 2010. Sequence mining in DNA chips data for diagnosing cancer patients. In Proceedings of the 10th WSEAS international conference on Applied computer science (ACS'10), Hamido Fujita and Jun Sasaki (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 139-151.

[18] Zakaria Suliman Zubi. 2009. Using some web content mining techniques for Arabic text classification. In Proceedings of the 8th WSEAS international conference on Data networks, communications, computers (DNCOCO'09), Manoj Jha, Charles Long, Nikos Mastorakis, and Cornelia Aida Bulucea (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 73-84.

[19] Zakaria Suliman Zubi and Rema Asheibani Saad. 2011. Using some data mining techniques for early diagnosis of lung cancer. In Proceedings of the 10th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases (AIKED'11), Zoran Bojkovic, Janusz Kacprzyk, Nikos Mastorakis, Valeri Mladenov, and Roberto Revetria (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin.

[20] M. Baglioni, U. Ferrara, A. Romei, S.Ruggieri, and F. Turini, Preprocessing and mining web log data for web personalization, Advances in Artificial Intelligence, Volume 2829,Springer, Berlin, 2003.

[21] L. Catledge and J. Pitkow, Characterizing browsing strategies in the world wide web,Computer Networks and ISDN Systems, 27: 1065– 1073, 1995.

[22] Zdravko Markov and Daniel T .Larose ,Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage, 2007 John Wiley & Sons.

[23] David Gourley, Brian Totty, Marjorie Sayer and Anshu Aggarwal, HTTP: The Definitive Guide.

Zakaria Suliman Zubi-- born in Benghazi Libya, in 1969. He received his Ph.D. in Computer Science in 2002 from Debrecen University in Hungary; he is an Associate Professor since 2010. He is a reviewer of many scientific journals such as Word Scientific and Engineering Academy and Society (WSEAS) , Journal of Software Engineering and Applications (JSEA), Member of the International Association of Engineers (IAENG), Journal of Engineering and Technology Research (JETR) , World Academy of Science Engineering and Technology (WASET) journal, an Associate Editor in the Journal of the WSEAS Transactions on Information Science and Applications and more local journals in Libya. He is a member of the Association for Computing Machinery society (ACM), a member of IEEE society, a member of the Word Scientific and Engineering Academy and Society (WSEAS). He published as authors and a co-author in many researches and technical reports in local and international journals and conference proceedings.