



# Discovering Valuable Patterns through Internet Web-Log Access Analysis

Navin Kumar

Ph: (410) 455-8673

Aryya Gangopadhyay

(410) 455-2620

Department of Information Systems University of Maryland Baltimore County 1000 Hilltop Circle, Baltimore, MD  
21250{navin1.gangopad}@umbc.edu

## ABSTRACT

This study outlines the usage of data mining techniques to analyze web log files. It shows the direction in which web mining can be performed to unearth concealed information in huge access log data. An attempt has been made to give an overview on how to derive association rules from web server data mining. Paper also discusses implementation of OLAP technology to perform web usage analysis.

**Keywords.** Web mining, web log analysis, usage analysis, association rules, OLAP cubes

## 1.0 INTRODUCTION

It is difficult to visualize the present market situation without taking World Wide Web (WWW) into consideration. Everyday new complexities are being faced in web site designing and site navigations. When the market is moving into internet, it is also a major concern for business industries to observe customer's interests. All these problems together seek for a rich-content and easily accessible site. To provide a better web designing, navigation through the web pages becomes an important input for analytical purposes. Here, Web mining refers to discovery and analysis of useful information from the World Wide Web. It also focuses its attention on user accesses data. Once the interesting patterns are recognized from web access log data, it will help the company in restructuring and better management of the web site, giving more effectiveness to it. Important point is that the web servers register a web log entry for every single access they get in which they save the URL requested, the IP address from which the request originated, and a timestamp, and with the rapid progress of WWW technology, and the ever growing popularity of the WWW, a huge number of Web access log records are being collected [3]. Frequently visited sites would easily end up with repository of hundreds of megabytes of log data. Here comes the issue of data mining considering colossal files of raw web log data where retrieving significant and useful information is a nontrivial task [3].

This study is an attempt to analyze web logs to dig information about session identification, user navigations, web usage analysis, and association rules between various pages. In order to show data mining techniques with examples, *a hypothetical company* (www.OnlineBookStore.com) is considered in this paper. The behavior of the web page readers is imprinted in the web server log files. Analyzing and exploring regularities in this behavior will significantly improve system performance, enhance the quality and delivery of Internet information services to the end user, and identify population of potential customers for electronic commerce. Thus, by observing people using collections of data, data mining will bring considerable contribution to company's web site designers.

## 2.0 ABOUT THE COMPANY

This hypothetical company hosts a website for online book shopping (*Site map is shown in appendix*) There are some web pages to provide the information about the company itself. Different categories of books include computers, children, sports, and fictions. A site navigation tree is presented below. Different categories of books are managed in their respective

directories e.g. computer books are kept in /Computer directory. /Home.html is the root file and all other files can be reached from this root node. Category files e.g. Sports.asp is accessible from SpecialOffer.asp. Moreover, users cannot directly access two files e.g. to reach Java2.asp from Java1.asp, user will navigate through /Computer/Computer.asp.

### Note:

1. Every web page name is followed by a letter (A,B,C,...) in the site map. These letters have been used for derivation of association rules.
2. Access log data are randomly generated for analytical purposes, and so are **not** the actual data.

## 3.0 STEPS FOLLOWED IN WEB MINING

- Convert server log files into relational table format
- Data Preprocessing and Cleaning
- Session Identification
- Path Completion Analysis
- Transaction Identification
- Discovery of Association Rules
- OLAP Technology Implementation
- Characterization and Comparison (Web usage analysis)

### ACCESS LOG FILE in text format:

```
130.85.253.114 — [05/Apr/2002:11:34:55 +0100] "GET /Home.html
HTTP/1.0" 200 2048
130.85.253.114 — [05/Apr/2002:11:35:37 +0100] "GET /Category.html
HTTP/1.0" 200 1536
130.85.253.114 — [05/Apr/2002:11:36:25 +0100] "GET /Computer/
Computer.asp HTTP/1.0" 200 2048
.....
.....
```

The access log file is stored on server in ASCII format.

### Log file contains following information about each navigation:

- IP Address of the computer the request is coming from
- User ID of the user who generated the request (if assigned any)
- Date and time of the request
- user action (GET or POST)
- URL of requested page
- Name and version of the protocol
- Status code (or error code) of the request e.g.
  - 200: successfully received
  - 400: bad request
  - 505: HTTP version not supported
- Size of the page in bytes

Table2: Identification of session from log files

Session Number	Session IP Address	Start Row number	End Row Number	Access Path
1	130.85.253.114	1	5	A-D-F-M-N
2	130.85.253.114	6	10	A-L-B-H-A3
3	207.46.230.220	11	13	B-H-A4
4	207.70.7.168	14	20	A-D-G-Q-T-R-W
5	209.96.148.192	21	22	C-E
6	144.92.104.37	23	33	A-E-L-F-O-P-I-A6-H-H-A2
7	170.248.128.30	34	37	D-I-A5-A7
8	199.171.55.3	38	40	E-J-K
9	198.82.162.11	41	45	A-D-I-A7-F

### 3.1 Step 1 - Convert The Raw Data From Access Log Files Into Relational Data Format.

A script is run to extract and put ASCII data into respective columns in a table. Table 1 (appendix) presents web usage data restructured in relational format. (*IE denotes Internet Explorer*)

### 3.2 Step 2 - Data Preprocessing and Cleaning

Data preprocessing is performed on relational data. At this stage, preprocessing and cleaning processes have been discussed taking discovery of association rules into consideration. Step 8 continues further with discussion on data preprocessing for OLAP implementation.

#### a) Data reduction: Remove undesired fields

Data set is reduced by avoiding some attributes which do not directly influence the web mining results. Hence, from current log data, *attributes User ID, User Action, and File Size are omitted.*

#### b) Data cleaning:

(i) HTTP protocol requires individual connection for every file. Hence even request for an ASP page may lead to several log entries on graphics/scripts written in log file. Concentrating on html/asp pages, records having file path ending with gif, jpeg, jpg, bmp are omitted. Here *row number 5,20,22,24,33,43,45, 46 and 54 are removed.*

(ii) Only successful requests are filtered.

If status code=200 → request successful; record is retained.

If status code!=200 → request unsuccessful; record is removed.

Here *row number 24, 37, and 56 are removed from the log data table.* These records are stored in “**error\_info**” table to keep track of various errors occurring in web pages.

(iii) After removing error records, *status code and HTTP protocol fields are discarded requiring no attention for association rules derivation.* Table 2 (appendix) represents data after cleaning and preprocessing.

### 3.3 Step 3 - Session Identification

Discovery of association analysis from web log data starts with *transaction identification.*

*Transaction can be identified by “a collection of user clicks to a single Web server during a user session”.* A criterion for identifying server session is implemented by examining if client has not surfed through the site for a reasonable long time.

**Method used to identify unique user sessions** If the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session.

Let's assume that the **Timeout** (the session becomes inactive if not used for this specific time) is **30 minutes**.

Each log record *l* will contain this information:

*l.ip*: client IP address

*l.uid*: user id of the client

*l.url*: URL accessed

*l.time*: access time

*Session can be identified by comparing all the tuples as described :*

$$S = \langle ip_i, uid_i, \{(l'_1.url, l'_1.time), \dots, (l'_m.url, l'_m.time)\} \rangle$$

where, for,  $1 \leq k \leq (m-1), l'_k \in L, l'_k.ip = ip_i, l'_k.uid = uid_i, l'_{k+1}.ip = ip_i, l'_{k+1}.uid = uid_i$ , and,  $(l'_k.time - l'_{k+1}.time) \leq W (= 30 \text{ Minutes})$

By looking at table 2, we find that all IP addresses except 130.85.253.114 have unique user sessions. IP address 130.85.252.114 has two user sessions because time difference between row number five and six is (05/Apr/2002:13:14:37-05/Apr/2002:11:42:23) greater than 30 minutes. Hence 130.85.253.114 has two user sessions, one from row number 1 to 5, and another from row number 6 to 10. A summary **user session table** is shown below:

### 3.4 Step 4 - Path Completion Analysis

Even after unique user sessions are successfully identified, another problem remains if there are important accesses that are not recorded in the access log. We refer this problem as *path completion*. An approach is used here to identify missing web pages from the log file. **This approach is explained as follows:**

If a page request is not directly linked to the last requested page, it is assumed that the page is already in user's recent request history, and so user backtracked with the “back” button available on browser. This in effect calls cached version of the pages until a new page was requested. Site map can be used here to identify the missing pages. If more than one page in the user's history contains a link to the requested page, assumption follows that the page closest to the previously requested page is the source of the new request. Missing page references inferred through this approach are added to the user session file.

**Session number 1:** Access path is A-D-F-M-N. There is no direct path from M to N. Hence, the user must have backtracked to reach N i.e. user would have used page F to reach page N. hence complete path would be A-D-F-M-F-N.

**Session number 2:** Access path is A-L-B-H-A3. Again there is no direct path to reach B from L. That means user has once backtracked to reach B from L via A. Hence complete path would be A-L-A-B-H-A3.

**Session number 3:** Access path is B-H-A4. It does not require any further modification in the access path.

**Session number 4:** Access path is A-D-G-Q-T-R-W. There is no direct path from R to W, so user must have backtracked to G first and then has navigated to W. hence complete access path is A-D-G-Q-T-Q-G-R-V.

**Session number 5:** Access path is C-E. This is a complete path considering the fact that the site has common header with links to all first level pages.

**Session number 6:** Access path is A-E-L-F-O-P-I-A6-H-H-A2. There is no direct path from L to F suggesting backtracking. Navigation from A6 to H suggests I as the middle page. Page H appears twice indicating removal of one page reference from access path. Hence final access path would be A-E-L-E-A-F-O-F-P-F-A-I-A6-I-A-H-A2.

**Session number 7:** Access path is D-I-A5-A7. Final path is D-I-A5-I-A7.

**Session number 8:** Access path is E-J-K. Final path is E-J-E-K.

**Session number 9:** Access path is A-D-I-A7-F. To navigation from A7

Table 3: Identification of access path (path completion) from session file

Session Number	Session IP Address	Start Row number	Access Path
End Row Number			
1	130.85.253.114	1	5
2	130.85.253.114	6	10
3	207.46.230.220	11	13
4	207.70.7.168	14	20
5	209.96.148.192	21	22
6	144.92.104.37	23	33
7	170.248.128.30	34	37
8	199.171.55.3	38	40
9	198.82.162.11	41	45

Table 4: Transaction table from session table (using Maximal Forward reference)

Session Number	Access Path	Transaction
1	A-D-F-M-F-N	A-D-F-M, A-D-F-N
2	A-L-A-B-H-A3	A-L, A-B-H-A3
3	B-H-A4	B-H-A4
4	A-D-G-Q-T-Q-G-R-W	A-D-G-Q-T, A-D-G-R-W
5	C-E	C-E
6	A-E-L-E-A-F-O-F-P-F-A-I-A6-I-A-H-A2	A-E-L, A-F-O, A-F-P, A-I-A6, A-H-A2
7	D-I-A5-I-A7	D-I-A5, D-I-A7
8	E-J-E-K	E-J, E-K
9	A-D-I-A7-I-D-F	A-D-I-A7, A-D-F

to F, there are two possibilities, reaching either from page A or page D. Because page D is the closest to the requested page, the final path would be A-D-I-A7-I-D-F.

### 3.5 Step 5 - Transaction Identification by Maximal Forward Reference

This approach divides large access paths into smaller ones to identify smaller transactions. *This technique works as follows:*

Users are apt to travel objects back and forth in accordance with the links and icons provided. This senses possibility of two types of references: backward and forward references. A *backward* reference is the revisit of previously visited resource; on the other end, a *forward* reference is the visit

of a new resource in user session path. Transaction is defined as the set of pages in the path from the first page in a user session up to the page before a backward reference is made. When backward references occur, a forward reference path terminates. New transaction starts with next forward reference. This resulting forward reference path is termed as a *maximal forward reference*.

Considering session data from table 4, for session number 1, access path is A-B-F-M-F-N. According to the rule, first transaction will end at A-B-F-M, and so another transaction will be A-B-F-N. Table 5 presents all transactions.

Each of the transactions represents a basket and each resource an item. *Apriori algorithm* is now applied to derive association rules.

Table 5

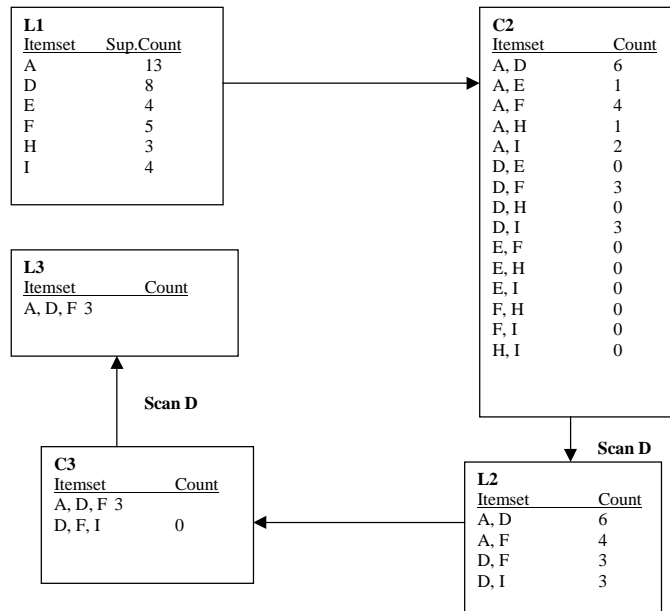
Transaction ID	Items	Transaction ID	Items	Transaction ID	Items
100	A-D-F-M	106	A-D-G-R-W	112	A-H-A2
101	A-D-F-N	107	C-E	113	D-I-A5
102	A-L	108	A-E-L	114	D-I-A7
103	A-B-H-A3	109	A-F-O	115	E-J
104	B-H-A4	110	A-F-P	116	E-K
105	A-D-G-Q-T	111	A-I-A6	117	A-D-I-A7
				118	A-D-F

Total number of transactions, N = 19  
Let's assume that *minimum support level* = 15% = 2.85 ~ 3  
*minimum confidence level* = 50%

Itemset	Support count	Itemset	Support count	Itemset	Support count	Itemset	Support count
A	13	J	1	S	0	A2	1
B	1	K	1	T	1	A3	1
C	1	L	2	U	0	A4	1
D	8	M	1	V	0	A5	1
E	4	N	1	W	1	A6	1
F	5	O	1	X	0	A7	2
G	2	P	1	Y	0		
H	3	Q	1	Z	0		
I	4	R	1	A1	0		

### 3.6 Step 6 - Discovery Of Association Rules From Transaction Data

Because the **minimum support is 3**, the next picture shows frequent 1-itemset **L1** and subsequent analysis.



Hence, Apriori algorithm provides these frequent item sets (resource sets) for the transaction data: {A}, {D}, {E}, {F}, {H}, {I}, {A, D}, {A, F}, {D, F}, {D, I}, {A, D, F}

#### Strong association rules for 2-itemsets:

Minimum confidence = 50%

**Rule 1:**  $\forall x \in \text{transaction}, \text{contains}(X, A) \Rightarrow \text{contains}(X, D)$

$$\text{Confidence} = \frac{\text{Support}(A, D)}{\text{Support}(A)} = \frac{6}{13} = 46.15\%$$

Rule **cannot** be established.

**Rule** 2:

$\forall x \in \text{transaction}, \text{contains}(X, D) \Rightarrow \text{contains}(X, A)$

Confidence=75% i.e. this rule is a **strong** rule. It implies that 75% of the users navigating page D (/Category.html) also are visiting page A (/Home.html).

**Rule** 3:

$\forall x \in \text{transaction}, \text{contains}(X, A) \Rightarrow \text{contains}(X, F)$

Confidence=30.76%, rule **cannot** be established.

**Rule 4:**  $\forall x \in \text{transaction}, \text{contains}(X, F) \Rightarrow \text{contains}(X, A)$

Confidence=80% i.e. this rule is a **strong** rule.

**Rule** 5:

$\forall x \in \text{transaction}, \text{contains}(X, D) \Rightarrow \text{contains}(X, F)$

Confidence=37.5%, rule **cannot** be established.

**Rule** 6:

$\forall x \in \text{transaction}, \text{contains}(X, F) \Rightarrow \text{contains}(X, D)$

Confidence=60%, rule is established as **strong** rule.

**Rule** 7:

$\forall x \in \text{transaction}, \text{contains}(X, D) \Rightarrow \text{contains}(X, I)$

Confidence=37.5%, this rule **cannot** be established

**Rule** 8:

$\forall x \in \text{transaction}, \text{contains}(X, I) \Rightarrow \text{contains}(X, D)$

Confidence=75%, association rule is **strong**.

#### Strong association rules for 3-itemsets:

**Rule** 1:

$\forall x \in \text{transaction}, \text{contains}(X, A) \wedge \text{contains}(X, D) \Rightarrow \text{contains}(X, F)$

$$\text{Confidence} = \frac{\text{Support}(A, D, F)}{\text{Support}(A, D)} = \frac{3}{6} = 50\%$$

**Strong** association rule is established. It infers that 50% of the time, if the client visits both /Home.html and /Category.html, he/she will also visit /Computer/Computer.asp.

**Rule** 2:

$\forall x \in \text{transaction}, \text{contains}(X, A) \wedge \text{contains}(X, F) \Rightarrow \text{contains}(X, D)$

Confidence=75%, so this is a **strong** association rule.

**Rule** 3:

$\forall x \in \text{transaction}, \text{contains}(X, D) \wedge \text{contains}(X, F) \Rightarrow \text{contains}(X, A)$

Confidence=100%, so %, it qualifies as **strong** association rule.

### 3.7 Step 7 - OLAP Technology Implementation

#### Data Preprocessing:

To construct multidimensional cube technology, we first need to convert raw data into various dimensions with some defined to facilitate generalization and specialization.

- Access time is represented by **TIME** dimension with **schema hierarchy**.

*second<minute<hour<day<month<year<All*

- URL** stores file structure by server domain (if multiple servers running simultaneously), directory (where the file resides), file name, extension (.asp, .html, .cgi etc.). **Schema hierarchy:**

*file extension<file<directory<server domain<All*

- Client IP Address** defines organization name and domain with **operation-derived hierarchy**.

*Organization name<domain name<All*

- Time Spent** is difference in access time for the current page and the next page. A **set-grouping hierarchy** would then be:

If t = time spent (in seconds),  
*{very\_short\_stay, short\_stay, moderate\_stay, long\_stay}  $\subset$  all (Time Spent)*

*{0<t<30}  $\subset$  very\_short\_stay*

*{60<t<60}  $\subset$  short\_stay*

*{60<t<180}  $\subset$  moderate\_stay*

*{t>180}  $\subset$  long\_stay*

- Range hierarchy** on **File\_Size** (in bytes):

*{tiny, small, medium, large, huge}  $\subset$  all (File Size)*

*{0<File\_size<1000}  $\subset$  tiny*

*{1000<File\_size<2000}  $\subset$  small*

$\{2000 < \text{File\_size} < 4000\} \subset \text{medium}$

$\{4000 < \text{File\_size} < 5000\} \subset \text{large}$

$\{\text{File\_size} > 5000\} \subset \text{huge}$

• **Set-grouping hierarchy for Type\_of\_Resource:**

$\{\text{script, images}\} \subset \text{all (Type\_Of\_Resource)}$

$\{\text{asp, html}\} \subset \text{script}$

$\{\text{jpg, jpeg, gif, bmp}\} \subset \text{images}$

• **Browser type**

$\text{Browser\_version} < \text{Type\_of\_browser} < \text{all}$ .

We can construct OLAP cubes with abovementioned dimensions. OLAP operations drill-down, roll-up, slice and dice, can be performed to view and analyze web log data from different angles. One such *measure* is “**number of hits**” at a particular level of granularity. Step 8 explains more on OLAP operations.

### 3.8 Step 8 - Characterization and Comparison

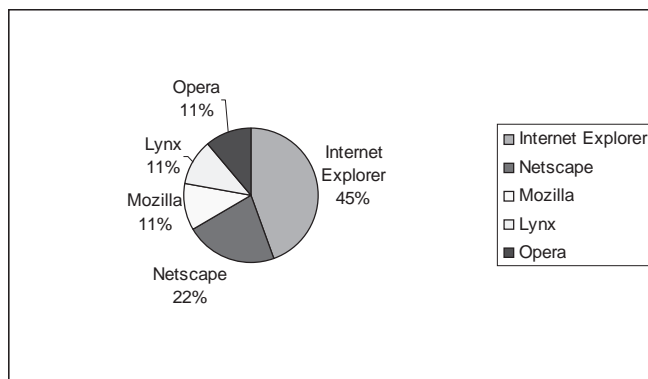
#### a) Finding “TOP N” requested URLs

Selecting top N pages is achieved by performing drill-down on URL dimension to file, and **all** to other dimensions. For this sample data, top 6 pages in decreasing order of count would be

1) /Home.html, 2) /Category.html, 3) /Sports/Sports.asp, 4) /CompanyInfo.html, and 5) /Computer/Computer.asp, and 6) /Fiction/Fiction.asp

#### b) Comparison of browser types for web site access

Web mining can be performed on the browser type to find out commonly used browsers, and more information on their compatibility for the web site can be further studied. Pie chart below presents web site usage by browser type obtained by drilling to browser dimension, and “**all**” for other dimensions.



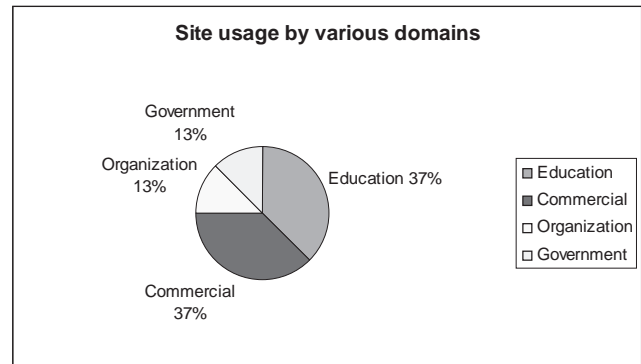
It explains that *Internet Explorer is the most frequently used browser*. Web site can be redesigned to provide maximum support for Internet Explorer and Netscape.

#### c) Comparison of various domains on web site access

The access log file is processed to convert the IP addresses into host names. E.g. www.umbc.edu has education domain.

IP Address	Host Name	Domain Type	IP Address	Host Name	Domain Type
130.85.253.114	www.umbc.edu	Education	144.92.104.37	www.wisc.edu	Education
207.46.230.220	www.microsoft.com	Commercial	170.248.128.30	www.accenture.com	Commercial
207.70.7.168	www.infotech.com	Commercial	199.171.55.3	www.sba.gov	Government
209.96.148.192	www.myvirginia.org	Organization	198.82.162.11	www.vt.edu	Education

It is easy to mine which kind of domains the web site is receiving the requests from. Pie chart explains that web site is mostly used by commercial sites (37%) and educational institutions (37%). It is obtained by drilling on IP address to ‘domain name’, and “**all**” along other dimensions.



### 4.0 CONCLUSION

As web is a one of the biggest repositories ever built, analyzing web access logs can help us understand the user behavior and would lead to better web site management by making use of Web Usage Mining. Furthermore, more information about user agents and referring resources can be collected to reveal better information. Several tools like NetTracker, SAS Webhound, Analog, and WebTrends Log Analyzer can be used to explore web log data in efficient way.

### REFERENCES

1. Cooley R., Mobasher B., and Srivastava J., Data preparation for mining World Wide Web browsing patterns, Knowledge and Information Systems, V1(1), 1999
2. Bartoloni G., Web usage mining and discovery of association rules from HTTP server logs, 2001 (<http://www.prato.linux.it/~gbartolini/html/wum.html>)
3. Jiawei Han, Man Xin, Osmar Zaiane, Discovering web access patterns and trends by applying OLAP and data mining technology on web logs, In Proceedings of the Fifth IEEE Forum on Research and Technology Advances in Digital Libraries, 1998.
4. White paper on Speed Tracer: A Web usage mining tool (<http://www.research.ibm.com/journal/sj/371/wu.html>)
5. Log File Basics (<http://slis-two.lis.fsu.edu/~log/basics-1.htm>)
6. WWW access statistics for the last 12 months (<http://www.netstore.de/stats/www2002/frames.html>)
7. Hits summary detail (<http://www.ideva.com/reports/ideva/idevahitssummary.htm>)
8. Request detail report (<http://www.ideva.com/reports/ideva/idevarequests.htm>)



## APPENDIX

### Site map

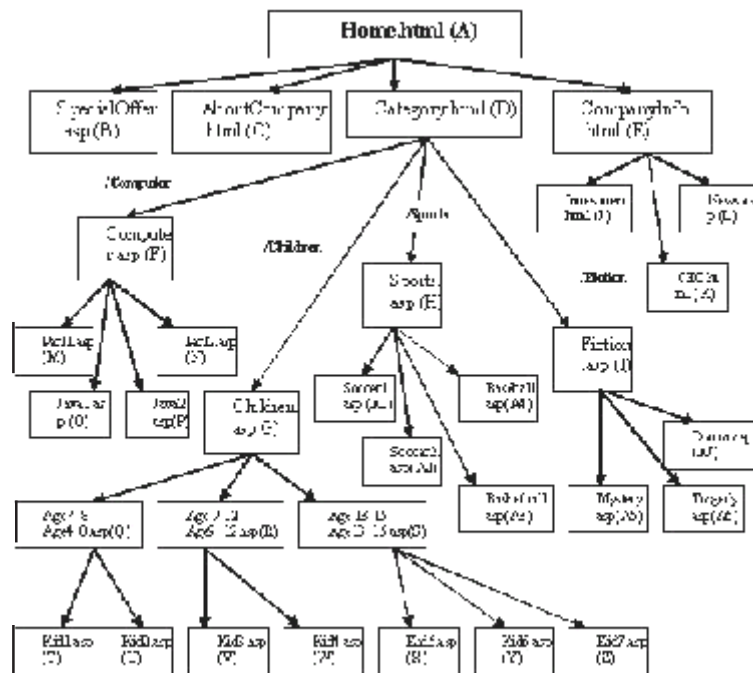


Table 1: Access log file in tabular format

Row #	Client IP Address	User ID	Access Time	User Agent	File Path	HTTP Protocol	Status Code	File Size	Browser Type
1	130.36.253.111	--	1/10/2002 11:54:53	IE5.0	/Home.html	HTTP/1.0	200	3748	IE
2	130.36.253.111	--	1/10/2002 11:55:27	IE5.0	/Category.html	HTTP/1.0	200	1154	IE
3	130.36.253.111	--	1/10/2002 11:56:45	IE5.0	/Category.html	HTTP/1.0	200	2648	IE
4	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Category.html	HTTP/1.0	200	1154	IE
5	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Category.html	HTTP/1.0	200	4054	IE
6	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Category.html	HTTP/1.0	200	1762	IE
7	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
8	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	716	IE
9	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	1762	IE
10	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
11	130.36.253.111	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	1154	IE
12	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	1154	IE
13	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
14	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
15	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
16	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
17	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
18	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
19	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
20	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
21	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
22	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
23	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
24	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
25	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
26	201.46.270.220	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
27	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
28	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
29	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
30	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
31	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
32	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
33	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
34	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
35	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
36	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
37	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
38	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
39	144.32.1.43	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
40	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
41	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
42	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
43	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
44	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
45	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
46	170.248.128.30	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
47	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
48	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
49	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
50	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
51	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
52	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
53	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
54	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
55	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE
56	199.11.155.3	--	1/10/2002 11:58:32	IE5.0	/Home.html	HTTP/1.0	200	2648	IE

Table 2: Web log data after preprocessing and cleaning

Row No.	Client Address	IP	Access Time	File Path	Browser Type
1	130.85.253.14		03/Apr/2002:11:34:35	/Home.html	IE
2	130.85.253.14		03/Apr/2002:11:35:37	/Category.html	IE
3	130.85.253.14		03/Apr/2002:11:36:25	/Computer/Computer.asp	IE
4	130.85.253.14		03/Apr/2002:11:38:12	/Computer/Per1.asp	IE
5	130.05.250.14		03/Apr/2002:11:42:23	/Computer/Per2.asp	IE
6	130.85.253.14		03/Apr/2002:13:14:57	/Home.html	Netscape 6.2
7	130.85.253.14		03/Apr/2002:13:15:55	/News.asp	Netscape 6.2
8	130.85.253.14		03/Apr/2002:13:17:13	/Special/Difer.asp	Netscape 6.2
9	130.85.253.14		03/Apr/2002:13:18:21	/Sports/Sports.asp	Netscape 6.2
10	130.85.253.14		03/Apr/2002:13:20:31	/Sports/Basketball.asp	Netscape 6.2
11	207.46.230.220		06/Apr/2002:05:41:23	/Special/Difer.asp	Opera 6.31
12	207.46.230.220		06/Apr/2002:05:43:45	/Sports/Sports.asp	Opera 6.31
13	207.46.230.220		06/Apr/2002:05:44:11	/Sports/Basketball.asp	Opera 6.31
14	207.70.7.168		07/Apr/2002:04:11:19	/Home.html	IE
15	207.70.7.168		07/Apr/2002:04:13:23	/Category.html	IE
16	207.70.7.168		07/Apr/2002:04:14:43	/Children.asp	IE
17	207.70.7.168		07/Apr/2002:04:17:17	/Children/Age4_8.asp	IE
18	207.70.7.168		07/Apr/2002:04:18:43	/Children/Kid.asp	IE
19	207.70.7.168		07/Apr/2002:04:24:31	/Children/Age9_12.asp	IE
20	207.70.7.168		07/Apr/2002:04:27:55	/Children/Kid.asp	IE
21	209.96.148.192		08/Apr/2002:19:23:49	/AboutCompany.html	Mozilla 4.0
22	209.96.148.192		08/Apr/2002:19:25:41	/CompanyInfo.html	Mozilla 4.0
23	144.92.104.37		09/Apr/2002:14:23:02	/Home.html	IE
24	144.92.104.37		09/Apr/2002:14:24:20	/CompanyInfo.html	IE
25	144.92.104.37		09/Apr/2002:14:27:33	/News.asp	IE
26	144.92.104.37		09/Apr/2002:14:28:52	/Computer/Computer.asp	IE
27	144.92.104.37		09/Apr/2002:14:29:09	/Computer/fan1.asp	IE
28	144.92.104.37		09/Apr/2002:14:32:33	/Computer/fan2.asp	IE
29	144.92.104.37		09/Apr/2002:14:35:38	/Fiction/Fiction.asp	IE
30	144.92.104.37		09/Apr/2002:14:37:50	/Fiction/George.asp	IE
31	144.92.104.37		09/Apr/2002:14:38:10	/Sports/Sports.asp	IE
32	144.92.104.37		09/Apr/2002:14:40:57	/Sports/Sports.asp	IE
33	144.92.104.37		09/Apr/2002:14:45:41	/Sports/occ21.asp	IE
34	170.248.123.30		11/Apr/2002:07:03:54	/Category.html	Netscape 6.2
35	170.248.123.30		11/Apr/2002:07:05:41	/Fiction/Fiction.asp	Netscape 6.2
36	170.248.123.30		11/Apr/2002:07:08:19	/Fiction/Mystery.asp	Netscape 6.2
37	170.248.123.30		11/Apr/2002:07:12:11	/Fiction/Drama.asp	Netscape 6.2
38	199.171.56.3		11/Apr/2002:20:33:10	/CompanyInfo.html	Linux 3.8.3
39	199.171.56.3		11/Apr/2002:20:36:05	/Investment.html	Linux 3.8.3
40	199.171.56.3		11/Apr/2002:20:39:35	/CDD.html	Linux 3.8.3
41	198.82.152.1		11/Apr/2002:08:40:11	/Home.html	IE
42	198.82.152.1		11/Apr/2002:08:51:06	/Category.html	IE
43	198.82.152.1		11/Apr/2002:08:52:27	/Fiction/Fiction.asp	IE
44	198.82.152.1		11/Apr/2002:08:54:35	/Fiction/Drama.asp	IE
45	198.82.152.1		11/Apr/2002:08:59:46	/Computer/Computer.asp	IE

0 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

[www.igi-global.com/proceeding-paper/discovering-valuable-patterns-through-internet/32167](http://www.igi-global.com/proceeding-paper/discovering-valuable-patterns-through-internet/32167)

## Related Content

---

### Reputational Mechanisms in Consumer-to-Consumer Online Commerce

Mikhail I. Melnik (2018). *Encyclopedia of Information Science and Technology, Fourth Edition* (pp. 2833-2840).

[www.irma-international.org/chapter/reputational-mechanisms-in-consumer-to-consumer-online-commerce/183994](http://www.irma-international.org/chapter/reputational-mechanisms-in-consumer-to-consumer-online-commerce/183994)

### Towards Higher Software Quality in Very Small Entities: ISO/IEC 29110 Software Basic Profile Mapping to Testing Standards

Alena Buchalceva (2021). *International Journal of Information Technologies and Systems Approach* (pp. 79-96).

[www.irma-international.org/article/towards-higher-software-quality-in-very-small-entities/272760](http://www.irma-international.org/article/towards-higher-software-quality-in-very-small-entities/272760)

### Healthcare Technology Adoption at the Group Level

Andre L. Araujo (2015). *Encyclopedia of Information Science and Technology, Third Edition* (pp. 3396-3403).

[www.irma-international.org/chapter/healthcare-technology-adoption-at-the-group-level/112770](http://www.irma-international.org/chapter/healthcare-technology-adoption-at-the-group-level/112770)

### Dynamics in Strategic Alliances: A Theory on Interorganizational Learning and Knowledge Development

Peter Otto (2012). *International Journal of Information Technologies and Systems Approach* (pp. 74-86).

[www.irma-international.org/article/dynamics-strategic-alliances/62029](http://www.irma-international.org/article/dynamics-strategic-alliances/62029)

### Open Access

Diane Fulkerson (2015). *Encyclopedia of Information Science and Technology, Third Edition* (pp. 4878-4885).

[www.irma-international.org/chapter/open-access/112934](http://www.irma-international.org/chapter/open-access/112934)