CHAPTER: 12

FUNDAMENTALS OF WEB LOG FILES

[1]**Dr. SUDHAKAR RANJAN**
[1]School of engineering and technology
Apeejay stya university Sohna-Palwal road, Gurugram

[2]**Dr. PARIKSHIT VASISHT**
[2]School of engineering and technology
Apeejay stya university Sohna-Palwal road, Gurugram

Ch.Id:-ASU/GRF/EB/RPETHEAT/2022/Ch-12
DOI: https://doi.org/10.52458/9789391842888.2022.eb.grf.asu.ch-12

INTRODUCTION

Whenever user surfs a website every web server maintains the list of actions performed/ requested by the user into a web log files. These web logs are thereafter analyzed using some kinds of mining technique i.e. web, data mining in order to extract useful information and utilized them for future purpose.

- **User Access Report***:* The given information available in the server log, a possible approach to grouping various accesses into user sessions is to use both time stamps and agent information.
- **Path Traversal Report***:* Once user sessions are identified, 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. 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 sessions. However, there might be groups of pages not on the same traversal path but frequently visited together by users visited. forward reference.
- **Group visit report:** Frequent traversal paths identify pages that are on the same forward path in a Web presentation. These pages represent consecutive subsequences in the maximum forward paths of user *For example*

User Action GET

Table 1 Access log files

Client IP address	User id	Login count	Access Time	File path	HTTP protocol	Status code	File size
130.85.234.112	abb	1	33:34:11 /5/april/2006	A-B	1.0	200	2048
130.23.233.111	asd	1	34:23:11 /5/april/2006	A-C-F-N-S-T	1.0	200	2056
123.34.123.23	asdd	1	40:23:11	A-D-K	1.0	200	2056

			/5/april/2006				
123.34.125.67	sde	1	50:24:11 /5/april/2006	A-D-K-Q	1.0	200	2056
135.45.123.56	erd	1	53:25:12 /5/april/2006	A-D-J-K	1.0	200	2056

SITE MAP

Every website contains number of the web pages and these web pages are linked with other web pages. The graphical representation of the complete website is called site map. It illustrates the various link associated with web pages. Let us suppose, every web page of website is termed as A,B,C,… see figure1. Access log data are randomly generated for analytical purposes, and so are not the actual data. Site map shown in figure1 can be assumed to be engineering College web site.

A. College home.html

B. Specialofficer.asp

C. About college.html

D. Studentinfo.html

E. Computer info.html

F. Computerdeptt.html

G. Spoorts.html

H. Electronics.html

I. Mechanical.html

J. Mtech.html

K. Btech.html

M. Mca.html

N. Faculty.html

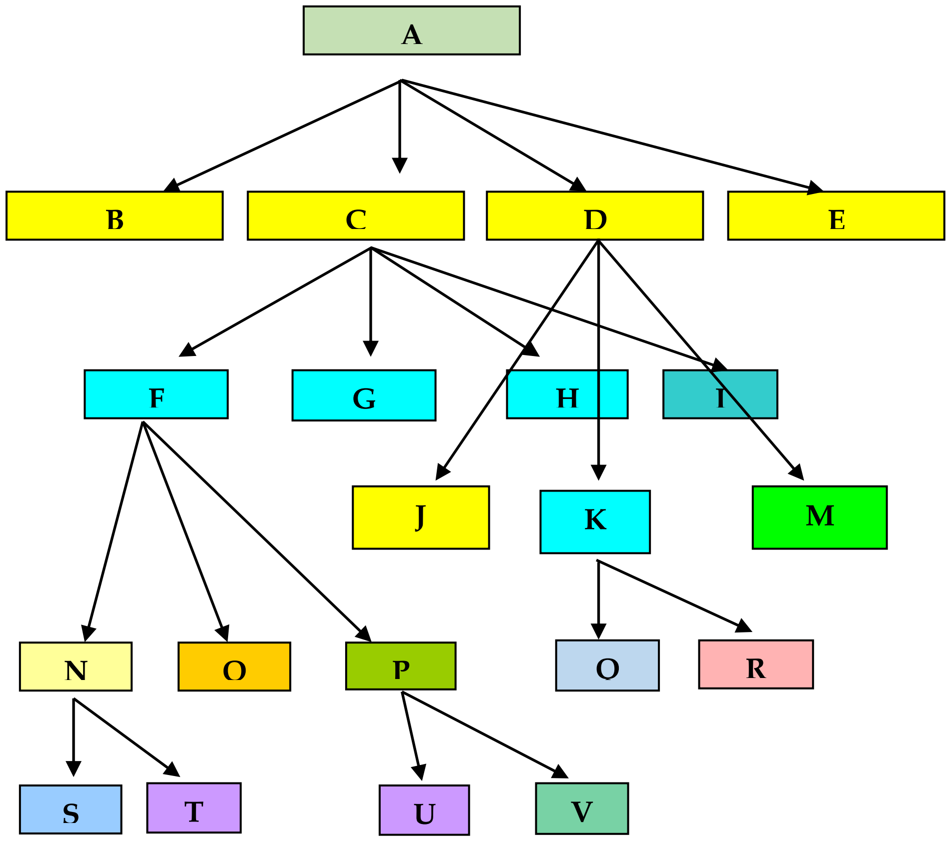

Fig.1: Site Map

Each web page has been depicted with the help of A, B, C, D…in figure 3.1 web page A depicts the home page of the web site which has links with the other sites B, C, D, E

- C has links with the sites F, G, H, I.

A. F has links with the sites N, O, P.

B. N has links with S, T.

C. O has no links.

D. P has links with U, V.

E. G, H, I has no links.

- B, E has no links.
- D has links with J, K, and M.

A. J and M has no links.

B. K has links with Q and R.

C. Q, R has no links.

In order to traverse a particular web page from the home page A say V. The path followed is shown in the figure 2. Starting from the home page A moving to C there after C moving to F and F moving to P and P moving to V. The over all path will be A-C-F-P-V

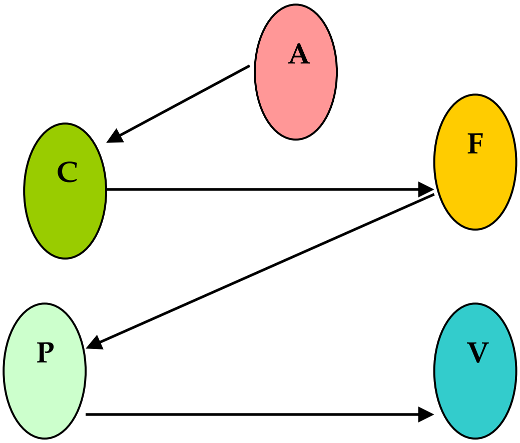

Fig.2 Path Traverse A-V

STEPS FOLLOWED IN ANALYZING THE WEB LOG FILES

The basic steps followed in this work for analyzing the web log are shown in Figure 3.

1. Logging
2. Log Files into Relational Data Format
3. Data Processing and cleaning
4. Session Identification
5. Transaction Identification

6. Association Rule (Apriori Algorithm)
7. Frequent User-access-report, Frequent Path–traversal-report, Frequent Group-visit-report

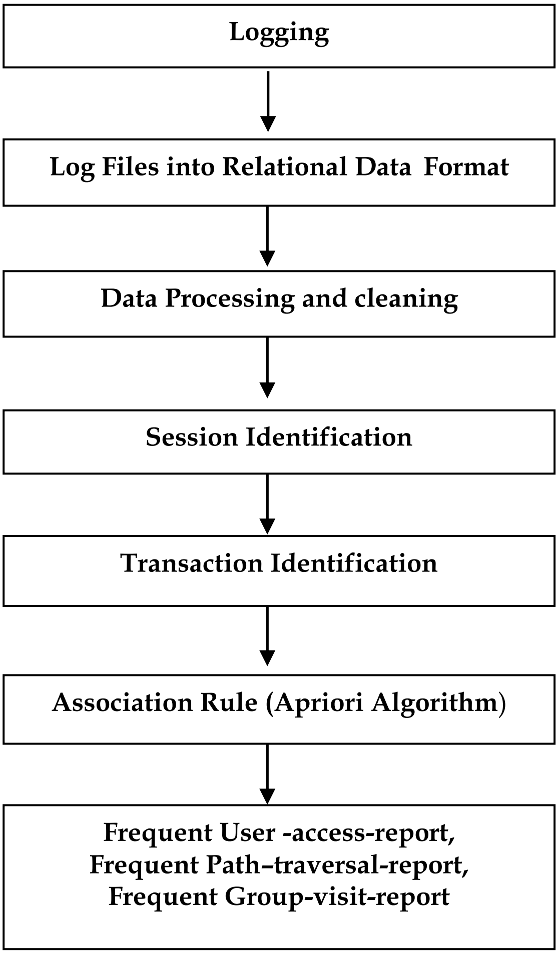
Logging
Log Files into Relational Data Format
Data Processing and cleaning
Session Identification
Transaction Identification
Association Rule (Apriori Algorithm)
Frequent User -access-report,
Frequent Path–traversal-report,
Frequent Group-visit-report

Fig.3. Flow Diagram of "A Frame Work for analyzing web-log files using data mining Technique

CONCLUSION

In this chapter discussed the basic concepts of web log files, access file along with site map, path traverse and step to analyzing web log files of web site with example.

REFERENCE

1. *White paper on Speed Tracer: A Web usage mining tool (http://www.research.ibm.com/journal/sj/371/wu.html)*
2. *Bamshad Mobasher, Namit Jain, Eui-Hong Han, Jaideep Srivastava (1996). Web mining: pattern discovery from World Wide Web Transactions.*
3. *Robert Cooley, Bamshad Mobasher, Jaideep Srivastava (1997). Grouping Web page references into transactions for mining World Wide Web browsing patterns.*
4. *Robert Cooley, Bamshad Mobasher, Jaideep Srivastava (1997). Web Mining: information and pattern discovery on the World Wide Web.*
5. *Masseglia, P. Poncelet, M. Teisseire (1999). Using data mining techniques on Web access logs to dynamically improve Hypertext structure.*
6. *Robert Cooley, Bamshad Mobasher, Jaideep Srivastava (1999). Data preparation for mining World Wide Web browsing patterns.*
7. *Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang- Ning Tan (2000). WebUsage Mining: Discovery and applications of usage patterns from Web data.*
8. *Sudhakar Ranjan, Komal Kumar Bhatia "Query interface integrator for domain Specific Hidden web" International Journal of Computer Engineering and Applications (IJCEA) ISSN 2321-3469, Vol. IV, Issue I/III, Oct.-Dec. 2013.*
9. *Sudhakar Ranjan, Komal Kumar Bhatia "Indexing for Vertical Search Engine: Cost Sensitive"International Journal of Emerging Technology & Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal), Volume 3, Issue 10, October 2013.*
10. *.Sudhakar Ranjan, Komal Kumar Bhatia"Web Log for Domain Specific Hidden Web", International Journal of Computer Engineering and Applications (IJCEA) ISSN 2321-3469, Vol VII, Issue I, July 2014.*
11. *Sudhakar Ranjan, Komal Kumar Bhatia "Indexing for Domain Specific Hidden Web" International Journal of Computer Engineering and Applications (IJCEA) 2321-3469, Vol VII, Issue I, July 2014.*
12. *Sudhakar Ranjan, Komal Kumar Bhatia "Transaction in Hidden Web" International Journal of Computer Engineering and Applications (IJCEA) 2321-3469, Vol VIII, Issue II, Nov 2014.*
13. *.Sudhakar Ranjan, Komal Bhatia: Design a Least Cost Vertical Search Engine based on DSHWC" selected for publication in IJIRR ,2017*