

Preprocessing in Web Usage mining

Marathe Dagadu Mitharam

ABSTRACT - Web usage mining to discover history for login user to web based application. Web usage mining is the process of data mining techniques. Web usage mining to extract useful information from server log files. It is an automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web sites.

Goal - Analysis for user interaction to various website

Web usage mining consists following sections.

1) Pre-processing

2) Pattern discovery

3) Pattern Analysis

In this paper describes First phase in detail.

----- ◆ -----

1) INTRODUCTION:-

The process may involve pre-processing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data mining operations. Collectively, we refer to this process as *data Collection*. After data collection we can do preprocessing section.

Preprocessing include the fusion and synchronization of data from multiple log files, data cleaning, pageview identification, user identification, session identification (or sessionization), episode identification, and the integration of

clickstream data with other data sources such as content or semantic information.

2) DATA PRE-PROCESSING:-

Fig. 1 it shows the process of web usages mining. In this section to be discuss about the pre-processing section in brief. Pre-processing section depends on web log files or various raw log files. Web usages mining process incomplete without using preprocessing section. We now examine some of the essential tasks in pre-processing. Data preprocessing depends on server log file.

Format of Server Log File:-

IP Address	rfc931	authuser	Date and time of request	request	status	bytes	referer	user agent
128.101.35.92	-	-	[09/Mar/2002:00:03:18 -0600]	"GET /~harum/ HTTP/1.0"	200	3014	http://www.cs.umn.edu/	Mozilla/4.7 [en] (X11; i; SunOS 5.8 sun4u)

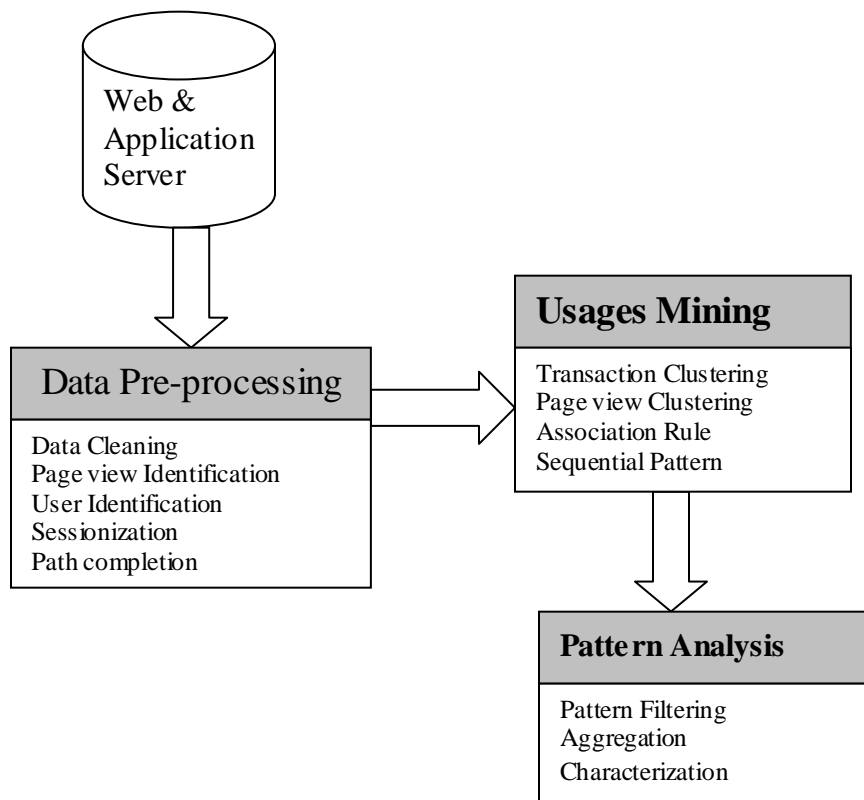


Fig. 1 Web Usages mining process

2.1) Data Fusion and Cleaning:-

In some cases, multiple servers are used to reduce the load on any particular server. Data fusion refers to the merging of log files from several Web and application servers. A user comes from multiple Web or application servers then data fusion merge data and solved various user identification session etc.

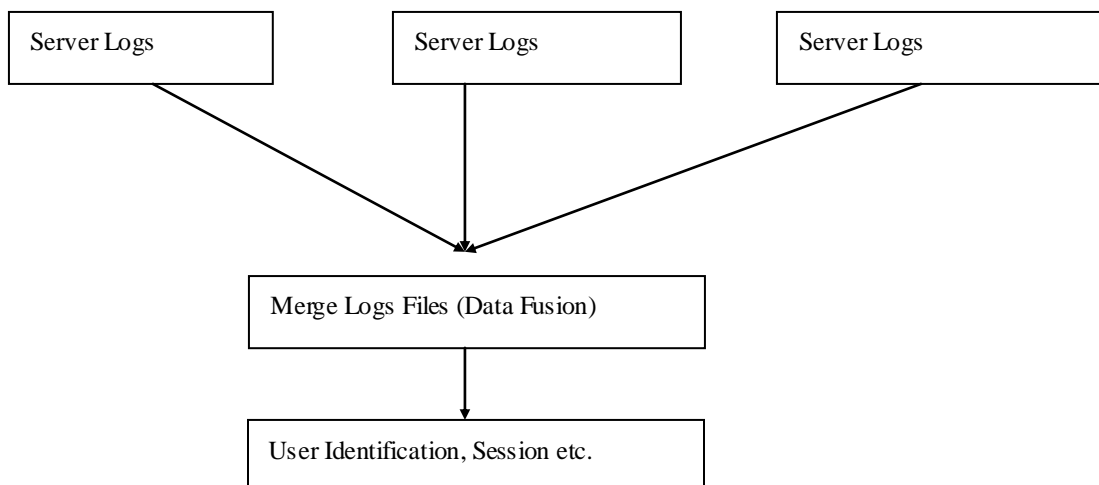


Fig. 2 Data Fusion


Data cleaning is mostly use to removing extraneous references to embedded objects that may not be important for the purpose of analysis, including references to style files, graphics, or sound files. Some information should not provide useful information in analysis or data mining tasks then Data cleaning is used. Remove erroneous references.

2.2) User Identification:-

In web usages mining does not require knowledge about a user history because the users visit or request given more than one time to the server. If we visit more that one time, then it generate multiple sessions for each user. It is also known as User activity Records. User Identify by using IP address and User Agent in log files. Client request to server then it generate log files at that time client also send user agent to server.

Consider, for instance, the example of Fig. 3 depicts a portion of a partly preprocessed log file (the time stamps are given as hours and minutes only).

The combination of IP + AGENT we can find out users. In Fig. 3 shows IP address and User agent then we judge user identifications.



TIME	IP	URL	REFF	AGENT
0:04	192.168.100.101	A	-	IE5; Win2k
0:10	192.168.100.101	B	A	IE5; Win2k
0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp
0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp
0:28	192.168.100.101	C	B	IE5; Win2k
0:33	192.168.100.101	D	C	IE5; Win2k
0:35	192.168.100.102	D	C	IE6;Xp

Fig. 3 Log file

Fig. User 1

TIME	IP	URL	REFF	AGENT
0:04	192.168.100.101	A	-	IE5; Win2k
0:10	192.168.100.101	B	A	IE5; Win2k
0:28	192.168.100.101	C	B	IE5; Win2k
0:33	192.168.100.101	D	C	IE5; Win2k

Fig. User 2

TIME	IP	URL	REFF	AGENT
0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp
0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp
0:35	192.168.100.102	D	C	IE6;Xp

In Fig.3 shows 192.168.100.101 this IP visit more than one time as well as 192.168.100.102 also visits. Then we judge or find user depends for IP and User Agent.

In the above (User 1 & User 2) we can find out users as per IP address and User Agent.

2.3) SESSION:-

A session is a sequence of page views by a single user during a single visit. A Session is the process of User activity record of each user in the log files. Session it shows single user visiting to web pages. In the ASP or ASP.Net session object is used, in this session object is used single user login status manipulation purpose. Same think should use in web usage mining to find how many sessions create a single user login to website. Session is partitioned after user identification.

Session captures in two way :- 1) Time oriented

2) Structure oriented

Time Oriented:- Time oriented is depends on the Time stamps or date and time of request in the server log file. In the time oriented session there are two types i) The difference between First request and last request is ≤ 30 minutes. ii) The difference between First request and next request is ≤ 10 . Using these two points we judge time oriented sessions.

In the above Fig. User2 first request given 0:12 and last request given 0:35, then difference between ≤ 30 minutes and difference between every request is ≤ 10 minutes then it's called as one session. Suppose that chart extend or request given then generate for different

output.

e.g.

TIME	IP	URL	REFF	AGENT
0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp
0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp
0:35	192.168.100.102	D	C	IE6;Xp
0:45	192.168.100.102	E	D	IE6;Xp
0:49	192.168.100.102	F	C	IE6;Xp
0:55	192.168.100.102	G	F	IE6;Xp

Fig. 4 log file

In the above Fig. 4 shows log file then we find out the session by using time oriented then it's generate two sessions.

Session 1

TIME	IP	URL	REFF	AGENT
0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp
0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp
0:35	192.168.100.102	D	C	IE6;Xp

Session 2 show as follow

Session 2

TIME	IP	URL	REFF	AGENT
0:45	192.168.100.102	E	D	IE6;Xp
0:49	192.168.100.102	F	C	IE6;Xp
0:55	192.168.100.102	G	F	IE6;Xp

Structure Oriented: Structure oriented capture in the referrer fields of the server logs. Structure oriented depends on Referrer fields is currently open or that user currently login referrer. Means it's belonging to more than one "open" constructed session.

e.g.

TIME	IP	URL	REFF	AGENT
0:04	192.168.100.101	A	-	IE5; Win2k
0:10	192.168.100.101	B	A	IE5; Win2k
0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp
0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp
0:48	192.168.100.101	C	B	IE5; Win2k
0:52	192.168.100.101	D	C	IE5; Win2k
0:58	192.168.100.102	D	C	IE6;Xp

Fig. 5 Login status for 101 and 102 IP

In the above fig.5 shows structure oriented session means 192.168.100.102 this user session is open for time stamp 0:12 to 0:25 and 0:58. It is consider as one session.

TIME	IP	URL	REFF	AGENT
0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp

Session 1

0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp
0:58	192.168.100.102	D	C	IE6;Xp

Fig. 6 Example of session with the structure oriented.

Using time oriented it generate two session as below, because the difference between first and last request is >30 minutes.

Session 1

0:12	192.168.100.102	A	-	IE6;Xp
0:15	192.168.100.102	B	A	IE6;Xp
0:20	192.168.100.102	C	B	IE6;Xp
0:25	192.168.100.102	D	C	IE6;Xp

Session 2

0:58	192.168.100.102	D	C	IE6;Xp
------	-----------------	---	---	--------

2.3) PATH COMPLETION:-

Path completion it is also preprocessing task. After completion sessions we start path completion, because that user how web pages visited that should be confirmed using path completion phase. Path completion is depends on mostly URL and REFF fields in server log file. It is also graph model. Graph model represents some relation defined on Web pages (or web), and each tree of the graph represents a web site. . Each node in the tree represents a web page (html document), and edges between trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site.

In the path completion Missing Reference this method also used. Missing Reference means the user backtrack should not be stored in server log file. It cached in client side.

e.g.	URL	Reff
	A	--
	B	A
	D	B
	E	D
	F	E
	B	C

Then we draw the structure of visiting in Fig. 6

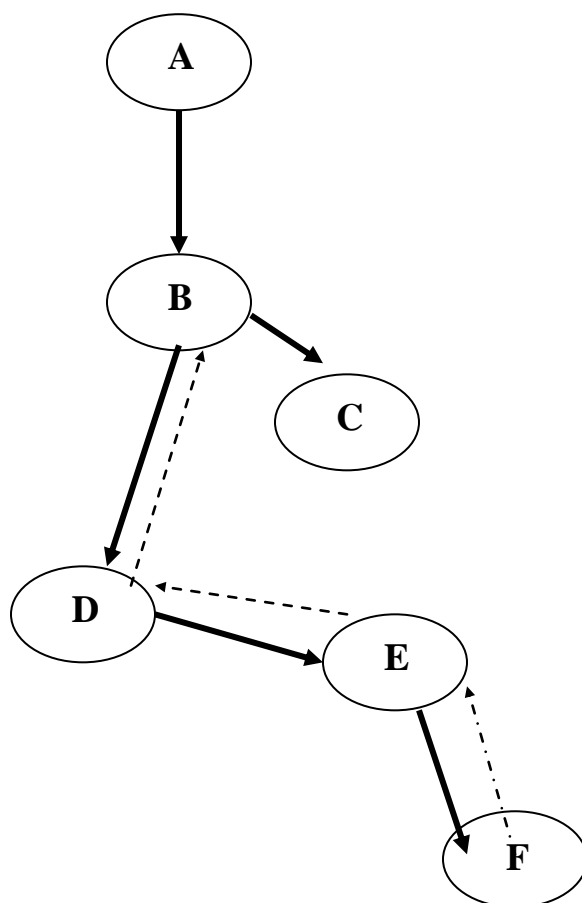


Fig. 7 Web site Structure

In above Fig. 7 Shows the visiting web pages as per server log file. The dotted arrow shows back track means the click on back button this information not store in server log file. This information stored in only client side. It is known as **Missing Reference**.

URL	Reff
A	--
B	A
D	B
E	D
F	E
B	C

In the above chart shows A to F Web pages visiting as linking by linking but in last B to C at that time F to B visiting as a back track then this information not store in server log file. At that time user click on back button and this information store only a client side.

3) CONCLUSION:-

This paper has attempted to for the purpose of web usage mining. The proposed methods were successfully tested on the log files. If we want to check User, Session and Path completion then refer this paper. The results which were obtained after the analysis were satisfactory and contained valuable information about the Log Files.

4) REFERNECES:-

- 1) Web data mining – Bing Liu
- 2) PPT for Web usage mining- Bing Liu
- 3) Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD, Jan 2000.
- 4) Jaideep Srivastava Paper

- 4) WCA. Web characterization terminology & definitions.
- 5) <http://www.w3.org/1999/05/WCA-terms/>. Vigente
al
19/11/2005

Author Name:- Dagadu Mitharam Marathe
R.C.Patel A.C.S. College, Shirpur,
Maharashtra (INDIA).
At/post- Thalner Tal-Shirpur Dist- Dhule(MS) India