Effective Cleaning of Educational Web Site Usage Patterns and Predicting their Next Visit

Harish Kumar
PhD Scholar, Mewar University, Chittorgarh

Anil Kumar Solanki, Ph.D
Director,

ABSTRACT
Universities with web education rely on web usage analysis to obtain students behavior for education promotion. The Internet is an information gateway and as a medium for business and education industry. Finding hidden information from Web log data is called Web usage mining. The aim of discovering similar patterns in Web log data is to obtain information about the navigational behavior of the users. Web usage mining, from the data mining aspect, is the task of applying data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of users navigating on the Web. Web usage mining aim is to find out useful information from the educational weblogs. These useful data pattern are used to analyze behavior of user. The focus of this paper is to generate a cleaning algorithm and provide an overview how to use frequent pattern mining techniques for discovering different types of patterns in a Web log. In this paper we premeditated different pattern for web pages, general statics, activity statics, visitor's statics, browser used

Keywords
Web mining, Web usage, Web logs, Navigational behavior.

1. INTRODUCTION
The Internet grows at an amazing rate as an information gateway and as a medium for online education industry. Students use Internet to find out the required information because the Internet is an infinite source of data that can come either from the Web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information daily collected by all the servers around the world [1] [2].Web based higher education is growing very quickly. The number of students taking advantage of online education is increasing by about 25% percent every year in India. But Registration at a college and University are growing by only 1.5% a year. So we need to be storing their personal information and their web navigation patterns in web log data. Various search engines, and other large educational document repositories *(e.g.* customer support databases, product specification databases, press release archives, news story archives, *etc.)* are growing so rapidly that it is difficult and costly to categorize every document. The information collection through data mining has allowed E-education Applications to make more revenues by being able to better use of the internet that helps students to make more decisions. Web-based applications for electronic commerce, online education, news etc., are becoming common practice and widespread. So many servers manage their cookies for distinguishing server address. User Navigation pattern are in the form of web logs .These Navigation patterns are refined and resized and modeled as a new format. This method is known as "Loginizing"[3].Web log mining has been applied to the e- education system, the main purpose of which is to get the access and interaction information of students through the access of server log files. Various data mining techniques are used to discover the hidden information in the Web. Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for mining useful information or knowledge from data [3]. Web mining does not only mean applying data mining techniques to the data stored in the Web. Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. Users leave navigation traces, which can be pulled up as a basis for a user behavior analysis. In the field of web applications similar analyses have been successfully executed by methods of Web Usage Mining. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, web user behavior and high-performance computing, to deliver advanced business intelligence and web discovery solutions [4]. The focus of this

paper is to provide an overview how to use pattern mining techniques for discovering different types of patterns in an e-Educational Web log. The three web patterns to be discussed are frequent item sets, sequences and tree patterns.

2. WEB MINING

Web mining may be classified into three categories, namely weblog mining, web content mining, and web structure mining.

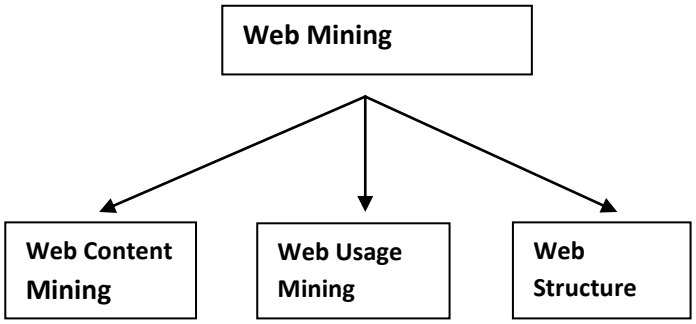


Figure1: Categorization of Web Data mining

Web content mining (WCM) is to find useful information in the content of web pages[4] e.g. free Semi-structured data such as HTML code, pictures, and various unloaded files. Web structure mining (WSM) is use to generating a structural summary about the web site and web pages. Web structure mining tries to discover the link structure of the hyperlinks at the inter document level. Web content mining mainly focuses on the structure of inner document, Web usage mining (WUM) is applied to the data generated by visits to a web site, especially those contained in web log files. I only highlighted and discussed research issues involved in web usage data mining. Web usage mining (WUM) or web log mining, users' behavior or interests is revealed by applying data mining techniques on web. Three main sources of web log file are

1. Client Log File,
2. Proxy Log File
3. Server Log File.

A log file contains the following field

- The client's host name or its IP address,
- The client id (generally empty and represented by a \-")
- The user login (if applicable),
- The date and time of the request,
- The operation type (GET, POST, HEAD, etc.),
- The requested resource name,
- The request status,
- The requested page size,
- The user agent (a string identifying the browser and the operating system used),and
- The referrer of the request which is the URL of the Web page containing the link that the user followed to get to the current page.

User behavior can be best analyzed from client log file because log files collected from client logs are much reliable and accurate then server log file and proxy log file. An extended log file contains a sequence of lines containing ASCII characters terminated by either the sequence LF or CRLF. Log file generators should follow the line termination convention for the platform on which they are executed. Analyzers should accept either form. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by whitespace; the use of tab characters for this purpose is encouraged. If a field is unused in a particular entry dash "-" marks the omitted field. Directives record information about the logging process itself. Lines beginning with the # character contain directives. The following directives are defined:

Version: <*integer*>.<*integer*>

The version of the extended log file format used. This draft defines version 1.0.

Fields: [<*specifier*>...]

Specifies the fields recorded in the log.

Software: *string*

Identifies the software which generated the log.

Start-Date: <*date*> <*time*>

The date and time at which the log was started.

End-Date :< *date*> <*time*>

The date and time at which the log was finished.

Date:<*date*> <*time*>

The date and time at which the entry was added.

Remark: <*text*>

Comment information. Data recorded in this field should be

Ignored by analysis tools.

Sample web log format is as follows.

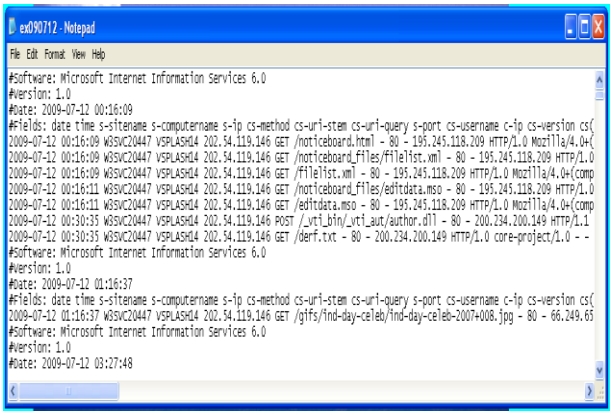


Figure 2: Categorization of Web log Mining

User profile was derived from the analysis of web log file and meta data of page contents. For discovery and analysis of usage patterns from the available data, it is necessary to perform three steps: Preprocessing, Pattern Discovery, Pattern Analysis.

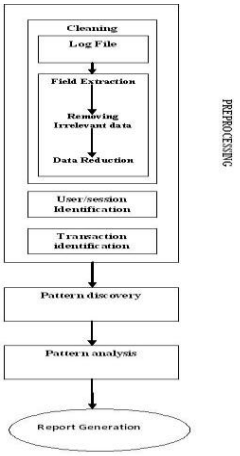


Figure 3: Basic steps of Web log Mining

3. DATA CLEANING ALGORITHEM

Data cleaning is the first preprocessing step. In this step all redundant, irrelevant and bulky noise data are eliminated from the web log files. When we request an educational web page which contains additional Web information like images or script files, on that time web navigator generate several requests. Data cleaning is depends on the web site i.e. if a site contains more irrelevant data then requires more cleaning steps otherwise simple cleaning is sufficient for web analysis. My proposed algorithm is worked on USP identification. USP identification algorithm has three phases

1. User identification phase.
2. Session Identification phase.
3. Path completion and identification phase.

Data cleaning is the first preprocessing step. In this step all redundant, irrelevant and bulky noise data are eliminated from the web log files. When we request an educational web page which contains additional Web information like images or script files, on that time web navigator generate several requests. If these requests are still present when the data mining step is performed, uninteresting patterns like “Page, Img1, Img2, Img3" may be found, making the pattern analysis step more complex [3] [4]. Also, Web robots, generally, have a predefined (programmed) behavior and the analysts are not interested in mining these requests.

It involves irrelevant reference to embedded objects. That may not be important for purpose of analysis, including references to style files, graphics or sound files. We filter out documents that are not requested directly by users. These are image requests in the log that are retrieved automatically after accessing requests to a document containing links to these files.

1. Collect log file data from weblog folder of the www root (FTP path of a web site).
2. Swap the data into text file.
3. Arrange the text file according to the date of created.
4. Pass the text file one by one in function clean(textfile.log)
5. Call Function clean(“weblogfile.log”, “ string”)
6. Main Function
7. Call Function removeLineFromFile(String file, String lineToRemove) {
8. If filename not found
9. Exit
10. Else
11. Read the file
12. Create a temp file for deletion
13. Pass the string S to Function isSubstring
14. Check whether the string is present or not
15. IF present
16. Call MAIN FUNCTION CALLING
17. Else
18. Handel the exception and EXIT from main Function
19. Function isSubstring (String s1, String s2)
20. {

Match the string

If found

Then

Pass to main Function

Else

Exit

}

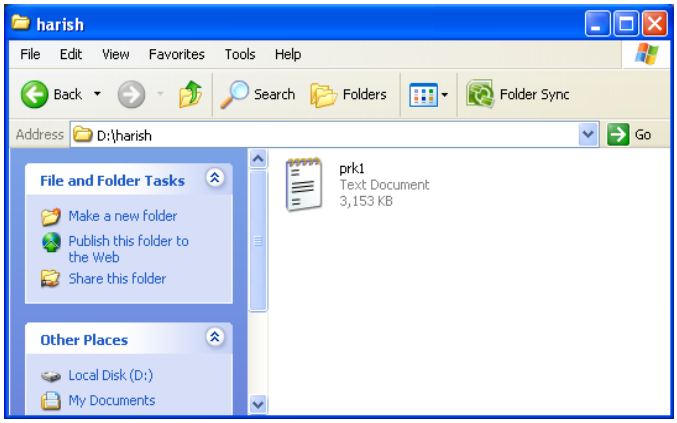
21. MAIN FUNCTION CALLING(STRING S[ ])

22. Remove the irrelevant data with desired extension(like .gif,.css etc)

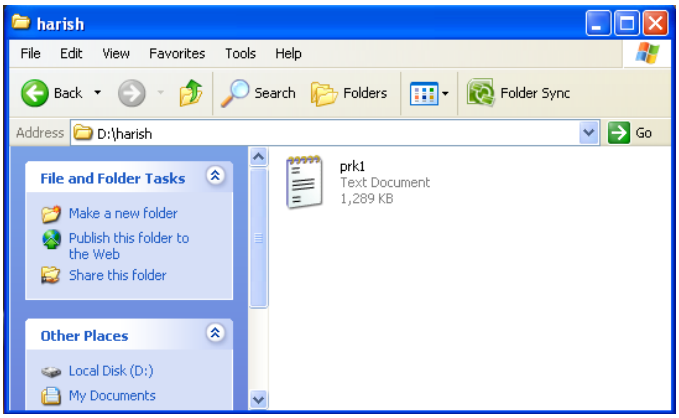
23. }

Log Set	Step No	Website Name	Duration	Size before Filtering	Size after Filtering
LS 1	1	www.krishnacollege.ac.in	1-01-12 To 2-01012	3153 Kb	1.1 Mb
LS 1	2	www.krishnacollege.ac.in	1-01-12 To 2-01012	1.1 Mb	786 Kb
LS 1	2	www.krishnacollege.ac.in	1-01-12 To 2-01012	786Kb	355Kb

The function complexity is much lesser then the other cleaning algorithms. Before Compiling size of web log file is 3153Kb. So this code helps to clean the irrelevant field from the log file. Clean data is use for designing Web Graph. This method increased the average length of the request data reduction for dataset.



After 1 step cleaning size is reduce up to 1,289 Kb.



After 2 step of cleaning size is approximate 1.1 Mb.

4. RELATED WORK

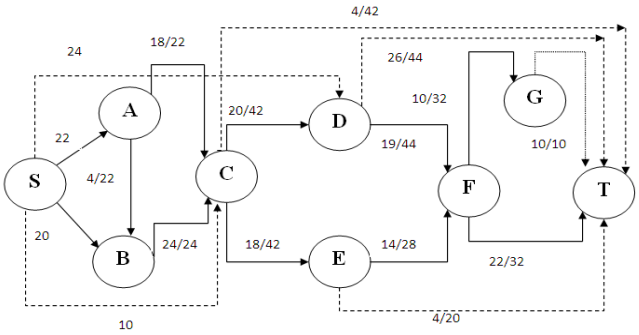
The pages and hyperlinks of the World-Wide Web may be viewed as nodes and arcs in a directed graph. The relationship between sites and pages indicated by these hyperlinks gives rise to what is called a *Web graph.* When it is viewed as a purely mathematical object, each page forms a node in this graph and each hyperlink forms a directed edge from one node to another. Generally user visit a web site in sequential nature means user visit first home page then second page and then third and then finish his work with this user leaves his navigation marks on a server. These navigation marks are called navigation pattern that can be used to decide the next likely web page request based on significantly statistical correlations. If that sequence is occurring very frequently then this sequence indicated most likely traversal pattern. If this pattern occurs sequentially, Markov chains have been used to represent navigation pattern of the web site [5][7]. Important properties of Markov Chain:

- Markov Chain is successful in sequence matching generation. Markov model is depending on previous state.
- Markov Chain model is Generative.
- Markov Chain is a discrete – time stochastic process.

The log file records the entry for each navigation click and can be processed into the specific manner of time ordered session. Every navigation session includes a *trail* that the user followed through the space, which we take without loss of generality to be a web site This trail can be as a sequence of clicks that has taken no longer than a given time span, or one such that the time between clicks in the sequence is no longer than a given short time span [2][8]. Markov chain present state is depending on previous state. If a web site contains more navigation pattern ("Interesting Pattern") and high supporting threshold is assign to it and less interesting patterns are ignored. This concludes that at various level of web site a different threshold value is assigned. Suppose the navigation patterns for above cleaned web logs with their frequency of use are as follows.

Pattern Name	**Frequency of occurrence**
S A B C D T	**4**
S E F G T	**8**
S B C E F T	**4**
S A C D T	**4**
S B C D T	**6**
S A C E T	**14**
S B C T	**4**
S D F G T	**2**
S D FT	**10**
S D T	**12**
S B C D F T	**6**
S E F T	**2**

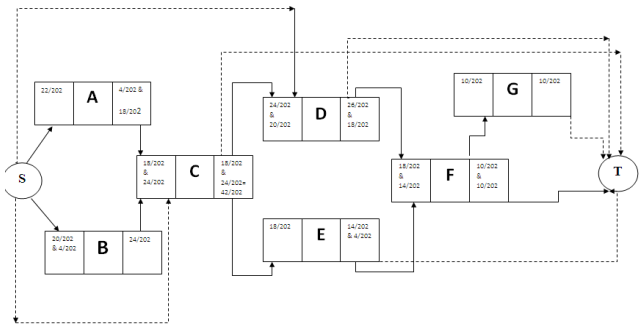
This pattern is similar to [3] pattern and first order Markov chain model for the same is as follows.



The above graph is represented by matrix.

	A	B	C	D	E	F	G
A	(S-A/B)⁴ (S-A/C)¹⁴ (S-A/C)² FREQ=3	B: A,B/C B: 0 B: 0 1((A,B/C),0,0)	C:B,C/D C:A,C/E C:A,C/D 3((B,C/D),(A,C/E),(A,C/D))	D:C,D/T D: 0 D: C,D/T 2((C,D/T),0,(C,D/T))	E: 0 E:C,E/T E: 0 1(ω,(C,E/T), ω)	F: 0 F: 0 F: 0 0(ω, ω, ω)	G: 0 G: 0 G: 0 0(ω, ω, ω)
B	x	(S-B/C)⁴ (S-B/C)⁴ (S-B/C)⁴ (S-B/C)² FREQ=4	C:B,C/E C:B,C/D C:B,C/D C:B,C/T- ω 4((B,C/E),(B,C/D), (B,C/D), (B,C/F),	D: 0 D:C,D/T- ω D: C,D/F D: 0 2(0,(C,D/F),(C,D/F), ω)	E:C,E/F E: 0 E: 0 E: 0 1((C,E/F), 0,ω, ω)	F: E,F/T- ω F: 0 F: D,F/T- ω F: 0 2((E,F/T),ω , D,F/T, ω)	G: 0 G: 0 G: 0 G: 0 0(ω, ω, ω, ω)
C	x	x	x	x	x	x	x
D	x	x	x	(S-D/F)² (S-D/F)¹² (S-D/T)¹² FREQ=3	E: 0 E: 0 E: 0 0	F: F,G/T F: D,F/T- ω F: 0 2((F,G/T),(D,F/T), ω)	G: G,T/ ω G: 0 G: 0 1((G,T/ ω), ω ,ω)
E	x	x	x	x	(S-E/F)⁴ (S-E/F)² FREQ=2	F: E,F/G F: E,F/T- ω 2((E,F/G),(E,F/T))	G: G,T/ ω G: ω 1((G,T/ ω), ω)
F	x	x	x	x	x	x	x
G	x	x	x	x	x	x	x
	22	24	42	44	28	32	10

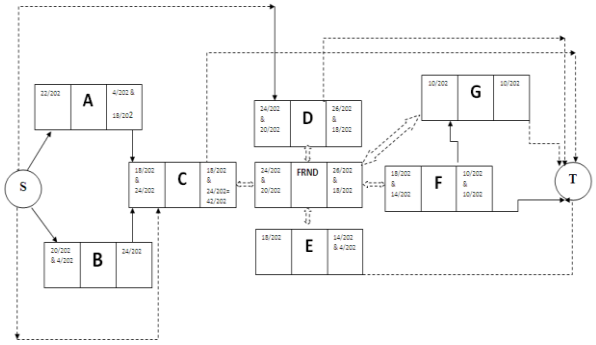
Above matrix indicates that every node having number of forward links and backward links. If we plot the graph with forward and backward link .If we download any web page this shows all the forward links of that particular page. This indicates that we need to be drawing a graph which shows all the forward and backward link of a page or say node.



4.1 ADVANCE MODEL

State C is not accurately showing his actual probability. The accuracy of changing probability from a state can be increased

by separating the in paths to it [10]. For solve this accuracy problem state cloning is required or need to find out another improved method for accuracy increasing. Cloning increase the data structure overhead because it is required to be created for those nodes having problem of accuracy. But the friend node can easy solve this problem. This is similar to friend function but does not have own importance.



In this model all the states are connected with this, the probability estimation for the node C-D is accurate and approx 0.099 which is equal to first order Markov chain. So this model can be successfully implemented. Now need to find out the new pattern with probability matrix and from their occurrence.

5. BENEFITS OF CURRENT MODEL

Benefits of this model are as follows.

1. Decrease overhead.
2. Precise Probability calculation.
3. Adopted by dynamic model.
4. Accurate results.
5. Predict the next pattern on the basis of probability ratio.

6. REFRENCES

[1] Ajith Abraham, "Business Intelligence from Web Usage Mining" Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375-390.

[2] Jos´e Borges, Mark Levene "An Average Linear Time Algorithm for Web Usage Mining" Sept 2003.

[3] Kumar Harish, Solanki A.K "Adaptive Markov Chain For Next Page Access Prediction" Vol. 9 No. 7 JUL 2011 Aug 25, 2011 by IJCSIS .

[4] Hengshan Wang, Cheng Yang, Hua Zeng " Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", *Communications of the IIMA, Volume 6 Issue 2*

[5] Jaideep Srivastava_y, Robert Cooleyz, Mukund Deshpande, Pang-Ning Tan "Web Usage Mining: Discovery and Applications of UsagePatterns from Web Data" Volume 1 Issue 2-Page13

[6] Alice Marques, Orlando Belo "Discovering Student web Usage Profiles Using Markov Chains" The Electronic Journal of e-Learning Volume 9 Issue 1 2011, (pp63-74)

[7] Ji He,Man Lan, Chew-Lim Tan,Sam-Yuan Sung, Hwee-BoonLow, "Initialization of Cluster refinement algorithms: a review and comparative study", Proceeding of International Joint Conference on Neural Networks[C].Budapest,2004.

[8] Renata Ivancsy, Ferenc Kovacs "Clustering Techniques Utilized in Web Usage Mining" International Conference on Artificial Intelligence, Knowledge Engineering and Data

[9] Bhawna.N and Suresh. J "Generating a New Model for Predicting the Next Accessed Web Page in Web Usage Mining" Third International Conference on Emerging Trends in Engineering and Technology, ICETET.2010.56

[10] Bindu Madhuri, Dr. Anand Chandulal.J, Ramya. K, Phanidra.M "Analysis of Users' Web Navigation Behavior using GRPA with Variable Length Markov Chains" IJDKP.2011.1201.

[11] Renata Iváncsy, István Vajk: A time- and memory-efficient frequent itemset discovering algorithm for association rule mining. IJCAT 27(4): 270-280 (2006).

AUTHOR PROFILE

Harish Kumar has completed his M.Tech (IT) from Guru Gobind Singh Indraprastha University, Delhi. He is currently pursuing his PhD from Mewar University, Chittorgarh.

PROF. Anil Kumar Solanki has obtained his PhD in Computer Science & Engineering from Bundelkhand University.He has published various paper in National and International journal. He has published various papers in international conferences, books and journals.