ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Identification of Frequent Navigation Pattern Using Web Usage Mining

[I]Latika Tamrakar, [II]S. M. Ghosh
[I]Kalyan PG College, Bhilai, India
[II]R.C.E.T., Bhilai, India

Abstract
With the explosive growth of knowledge sources available on the World Wide, it has become more important to find the useful information from these huge amounts of available data. At the same time, in the number of websites presents a challenging task for web designer to organize the contents of the websites to provide the needs of web users. These problems can be solved by path analysis using web user navigation patterns. In addition, web designers can improve the design and organization of websites based on the obtained solutions. Our approach is to give an overview about the discovery of association rules from Web logs data coming from HTTP servers.

Keywords
Pre-processing, log files, Apriori Algorithm, Frequent pattern.

I. Introduction

Web usage mining, from the data mining side, is the job of applying data mining techniques to discover usage patterns form Web data in order to identify with and better serve the needs of users navigating on the Web. As every data mining task, the process of Web usage mining also consists of three main steps (i) pre-processing, (ii) pattern discovery and (iii) pattern analysis.

Pattern discovery means applying the introduction frequent pattern discovery methods to the log data. For this reason the data have to be transformed in the pre-processing phase such that the output of the alteration can be used as the input of the algorithms and drawing conclusions. The input of the process is the log data.

The data has to be pre-processed in order to have the appropriate input for the mining algorithms. The different methods need different input formats, thus the pre-processing phase can provide three types of output data. The frequent patterns discovery phase is irrelevant. Also the duplicates of the same pages are omitted, and the pages are ordered in a predefined order[6].

1. Web Data

In Web Usage Mining, data can be collected in server logs, browser logs, proxy logs, or obtained from an organization's database. There are many kinds of data that can be used in Web Mining.

1. Content: The visible data in the Web pages or the information which was meant to be imparted tc the users. A major part of it includes text and graphics (images).
2. Structure: Data which describes the organization of the website, it is divided into two types. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page.
3. Usage: Data that describes the usage patterns of Web pages, such as IP addresses, page references, and the data and time of accesses and various other information depending on the log format.

2. Data Sources

The data sources used in Web Usage Mining may include web data repositories like:

1. Web Server Logs – These are logs which maintain a history of page requests. The W3C maintains a standard format for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request data/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log.
2. Proxy Server Logs - A Web proxy is a caching mechanism which lies between client browsers and web server. It helps to reduce the load time of web pages as well as the network traffic load at the server and client site
3. Browser Logs – Various browsers like Mozilla, Internet Explorer etc. can be modified or various JavaScript and Java applets can be used to collect client side data.

3. Web usage Mining Process

We used different web server log analysers like Web Expert Lite 6.1 and Analog6.0 to analyse various sample web server logs obtained.

The key information obtained was:

Total Hits, Visitor Hits, Average Hits per Day, Average Hits per Visitor, Failed Requests, Page Views Total Page Views, Average Page Views per Day, Average page Views per Visitor, Visitors Total Visitors Average Visitors per Day, Total Unique IPs, Bandwidth, Total Bandwidth, Visitor Bandwidth, Average Bandwidth per Day, Average Bandwidth per Hit, Average Bandwidth per Visitor.

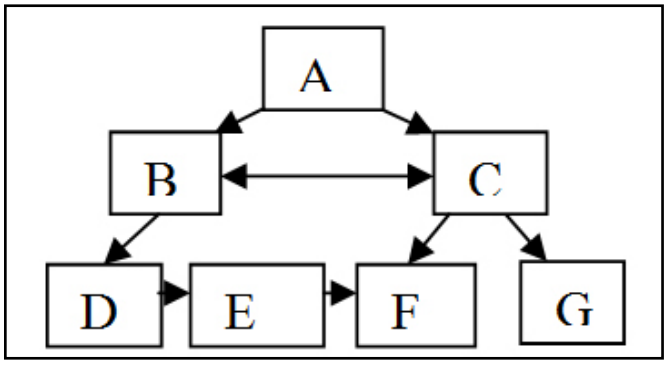

Fig. 1: Structure of Web site

4. Web Log Data

The fields available for the identification of the user, apart the uid field (which most of the times is empty), are

- IP address (or hostname if DNS lookups are performed).
- User Agent
- Referring URL
- Method
- Protocol
- Path
- Agent
- Date
- Time

II. Methodology

The proposed work is a complete pre-processing methodology that allows the analyst to transform any collection of web server log files into structured collected of tables in relational database model. The log file of a Website of the same organization is pre-processed to apprehend the behaviours of the users that navigate in a transparent way. Afterwards, this file is cleaned by removing all unnecessary requests, such as implicit requests for the objects embedded in the web pages and the requests generated by non-human clients of the Web site (i.e. Web robots). Then, the remaining requests are grouped by user, user sessions, page views, and visits, Finally, the cleaned and transformed collections of requests are saved onto a relational database model. We have provided filters to filter the unwanted, irrelevant, and unused data. Analyst can select the log files from different Web servers and decide what entries he/she is interested (HTML, PDF, and TEXT).

1. The pre-processing phase

In Web Usage Mining, with the term pre-processing phase I intend a set of operations that process the available sources of information (HTTP server and auxiliary ones) and lead to the creation of an ad-hoc formatted dataset to be used for knowledge discovery through the application of data mining techniques such as association rules, sequential patterns, clustering.

i. Usage Pre-processing

Usage pre-processing is arguably the most difficult task in the Web Usage mining process due to the incompleteness of the available data. Unless a client side tracking mechanism is used, only the IP address, agent and server side click-stream are available to identify users and server sessions. Some of the typically encountered problems are:

Single IP address/ Multiple Sever

- Multiple IP address/ Single Server
- Request from a user to one of several IP addresses
- Multiple IP address/Single
- Multiple Agent/Singe Users

ii. Content pre-processing

Content pre-processing consists of converting the text, image, scripts, and other files such as multimedia into forms that are useful for the Web Usage mining process. Often, this consists of performing content mining such as classification or clustering. While applying data mining to the content of Web sites is an interesting area of research in its own right, in the context of Web Usage mining the content of a site can be used to filter the input to, or output the pattern discovery algorithms.

2. The different Pre-processing Phases

The pre-processing phase of HTTP server information is made up of the following four steps.

- Data cleaning;
- User session identification;
- Path completion (at this point we have the user session file);
- Transaction identification (transaction file creation).

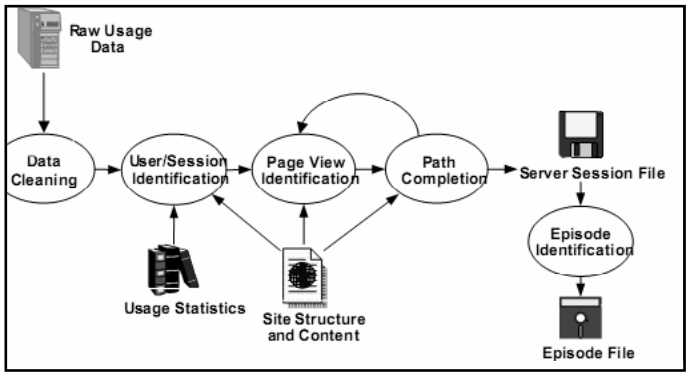

Fig. 2 : Pre-processing of Web Usage Data

i. Data Cleaning

Well, it is probably surprising for the end user to know that most of the information stored in an HTTP server log file is useless for the majority of KDD cases. On an average size Web server access log files easily reach tens of 2megabytes per day, causing the analysis process to be really slow and inefficient without an initial cleaning task. Just think that every time a Web browser downloads an HTML document on the Internet, the images included are requested as well, each of those accesses are stored in the log file of the server.

ii. Session identification

In most cases, the log file provides only the computer address (name or IP) and the user agent. For Web sites requiring user registration, the log file also contains the user login that can be used for the user identification. Sometimes the user login is not available, then each IP is considered as a user, although it is possible that an IP address can be used by several users.

iii. Path completion

This is a pretty complex task, because it involves the use of referring URLs and auxiliary information (site topology in particular). There are several reasons why some references are missing in the path that leads to the request of a particular resource; one of the major reasons is that cache mechanisms are applied both by clients and proxies.

iv. Transaction identification

The identification of transactions varies from case to case, depending on the Web Usage Mining technique that we want to use.

3. The A-priory Algorithm

Apriori is classic algorithm for learning association rules. Apriori algorithm is operated on database containing transactions (for example, details of a website frequentation). Other algorithms are designed for finding association rules in data having no transaction, or having no timestamps (DNA sequencing).

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

As it is common in association rule mining, given a set of item sets (for example, each listing of individual items purchased, sets of retail transactions), this algorithm finds subsets which are common to at least a minimum number C of the item sets. Apriori uses a bottom up approach, where frequent subsets are extended one item at a time (the step is known as candidate generation), and groups of these candidates are tested against the data. When no further successful extension are found then the algorithm terminates.
Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length h from item sets of length h-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent h-length item sets. After that, it scans the transaction database to determine frequent item among the candidates.

i. Algorithm

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is generally decomposed into two sub problems.

- Find those item sets whose occurrences go beyond a predefined threshold in the database.
- Generate association rules from those large item sets with the constraints of minimal confidence.

Suppose one of the large item sets is Lh = {I1, I2, …..Ik}, association rules with this item sets are generated in the following way the first rule is {I1, I2, …..Ih - 1}=? {Ih}. By checking the confidence this rule can be determined as interesting or not. Then, other rules are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them[4]. This process iterates until the antecedent becomes empty.
Find frequent set Lh – 1
Join step
Ch is generated by joining Lh – 1 with itself
Prune step.
Any (h-1) – itemset that is not frequent cannot be a subset of a frequent k- itemset, hence should be removed.
Where
(Ch Candidate itemset of size h)
(Lh frequent itemset of size h)

ii. Apriori Pseudo code

$Apriori\ (T, \varepsilon)$
L_1
$\leftarrow large\ 1$
$- itemsets\ that\ appear\ in\ more\ than\ \varepsilon\ transactions$
$k \leftarrow 2$
$while\ L_{k-1} \neq$
$C_k \leftarrow gnerate\ (L_{k-1})$
$for\ transactions\ t \in T$
$C_t \leftarrow Subset\ (C_k, t)$
$for\ candidates\ c \in C_t$
$count\ [c] \leftarrow count\ [c] + 1$
$L_k \leftarrow \{c \in C_k\ count\ [c] \geq \varepsilon\}$
$k = k + 1$
$return \bigcup_k L_k$

III. Results and Discussion (Experimental)

Table 1: Web Log Report

IP	METHOD	Protocol	PATH	AGENT	OS	DATE	Time
127.0.0.1	GET	HTTP1.1	/A.asp	Internet Explorer	Windows XP	Fri May 11	21:22:31
127.0.0.1	GET	HTTP1.1	/C.asp	Internet Explorer	Windows XP	Fri May 11	21:22:36
127.0.0.1	GET	HTTP1.1	/G. asp	Internet Explorer	Windows XP	Fri May 11	21:22:37
127.0.0.1	GET	HTTP1.1	/A. asp	Internet Explorer	Windows XP	Fri May 11	21:25:51
…….	…….	…….	…….	…….	…….	…….	…….
…….	…….	…….	…….	…….	…….	…….	…….

Description Various activities of users, which they perform at the time of accessing the web pages, are stored in Web Log Report. The various activities can the IP Address of user, Agent name of user, the page name, time and date of accessing, method, protocol etc.

Frequent Path Report

Table 2 . Frequent Path Report

FREQUENT PATH REPORT
/A.ASP:/B. ASP:/C.ASP:/D.ASP:/F.ASP:/G.ASP:
FREQUENT BINARY PATH
/A.ASP → /B.ASP /A.ASP → /C.ASP /B.ASP → /C.ASP /B.ASP → /D.ASP /C.ASP → /F.ASP /C.ASP → /G.ASP
FREQUENT CUBE PATH
/A.ASP → /B.ASP → /C.ASP /A.ASP → /B.ASP → /D.ASP /A.ASP → /C.ASP → /F.ASP /A.ASP → /C.ASP → /G.ASP /B.ASP → /C.ASP → /F.ASP /B.ASP → /C.ASP → /G.ASP
FREQUENT SQUARE PATH
/A.ASP → /B.ASP → /C.ASP → /F.ASP /A.ASP → /B.ASP → /C.ASP → /G.ASP

Description A-priori algorithm is applied on table no. 2 for getting most frequent accessed pages of users. The pages are called as most frequent, if the minimum numbers of users, who access the page or association of pages, are more than minimum support. Here minimum support is taken as 40% users access the page or association of pages, then the pages are called as most frequent pages.

IV. Conclusions

For the discovery of most frequent associated pages. Association Mining Rule (Apriori algorithm) is used, so that most frequent navigation pages can be retrieved. Pattern analyser can use these patterns for performing some important applications like system Improvement by page caching, site modification, page personalization, website restructuring etc.

References

[1] *Kim, Wooju. Song, Young U. Hong, June S. "Web enabled expert system using hyperlink based inference". Export system with application 2004 pp l-13.*
[2] *Michele Facca, Federico. Luca Lanzi, Pier. "Mining interesting knowledge from web loghas a survey". Data & Knowledge Engineering. Accepted for publication 2004.*
[3] *Hsu, Jeffrey. "Data mining trends and developments The Key Data Mining Technologies and Application for the 21st Century". Proc of ISECON 2002.*
[4] *K. Pazhani Kumar , S. Arumugaperumal "Association Rule Mining and Medical Application: A Detailed Survey", October 2013*
[5] *Chakrabarti, Soumen. "Mining the web discovering Knowledge from hypertext data". San Francisco, CA. Morgan Kaufmann Publishers An imprint of Elsevier Science 2003 pp 1-13.*
[6] *Kousalya, Suguna, SaravananImproving, "the Efficiency of Web Usage Mining Using K-Apriori and FP-Growth Algorithm ",March-2013*
[7] *Arotaritei, Dragos. Mitra, Sushmita. "Web mining a survey in the fuzzy framework". Fuzzy Sets and Systems vol. 148, 2004 pp 5-19.*
[8] *Larsen, Jan. Lars Hansen, Kai. Szymkowiak Have, Anna. Christiansen, Torben. Kolenda, Thomas. "Webmining learning from the World Wide Web". Computational Statistics & Data Analysis. 38. 2002 pp 517-532.*
[9] *Eirinaki, Magdalini. Vazirrgiannis, Michalis. " Web Mining for Web[19]. Gatetrade net. Information on gatetrade net and some of their solutions Marketplace Personalization". ACM Transaction on Internet Technology vol. 3, no. 1, 2003. pp 1-27.*
[10] *De Young Colin G. Spence, lan. "Profiling information technology users en route to dynamic personalization". Computers in Human Behavior. Vol. 20. 2004. Pp 55-65.*
[11] *Rakesh Agrawal, Sakti Ghosh, Tomasz lmielinski, Bala lyer, and Arun Swami, An Interval Classi_er for Database Mining Applications" , VLDB-92, Vancouver, British Columbia, 1992, pp 560-573.*

Author's Profile

Latika Tamrakar has completed her MCA in 2004. She is perusing MPhil(CS) from Dr. C.V. Raman University, Kota ,Bilaspur. C.G.,India.

Prof. Samarendra Mohan Ghosh received his PhD from Chhattisgarh Swami Vivekanda Technical University, India in Software Engineering. He is Professor in Rungta College of Engineering and Technology, Chhattisgarh teaching Data Mining, Software Engineering, ERP and Business Intelligence related topics