RESEARCH ARTICLE OPEN ACCESS

An Efficient Path Completion Technique for Web Log Mining

Shanta H Biradar
Information Science and Engineering Department
Sir M visvesvaraya Institute Of Technology
Bangalore-562157
India

ABSTRACT

World Wide Web is a huge repository of web pages and links. It provides abundance information for the internet users. The growth of web is tremendous as approximately one million pages are added daily. Users' accesses are recorded in web logs.Because of the tremendous usage of web, the web log files are growing at a faster rate and the size is becoming huge. Web data mining is the application of data mining techniques in web data. Extraction of user behavior is an important work in web mining. Web Usage mining applies mining techniques in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling , prefetching, creating attractive web sites etc., preprocessing, pattern discovery and pattern analysis. Web log data is usually noisy and ambiguous and preprocessing is an important process before mining. For efficient mining process the transactions are to be constructed accurately which is an important task of preprocessing. This paper describes about the accomplishment of path completion, finding content path set, and travel path set which shows user interest.

Keywords:- Preprocessing, Web Usage, Path Completion, Travel Path Set, Content Path Set.

I. INTRODUCTION

In this internet era web sites on the internet are useful source of information in day to day activities. So there is a rapid development of World Wide Web in its volume of traffic and the size and complexity of web sites. As per August 2010 Web Server survey by Netcraft there are 213,458,815 active sites. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user's visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval etc., web mining has become one of the important areas in computer and information science.

Web usage mining is the process of applying data mining technology to the web log data and is the pattern of extracting user's interests from their traversals of web pages. Web servers accumulate data about user's interactions in log files whenever requests for resources are received. Log files records information such as client IP address, URL requested etc., in different formats such as Common Log format, Extended Common Log format(ECLF) which is issued by Apache and Internet Information Server. Most of the servers use ECLF format.

Web usage mining consists of three main steps: data Preprocessing, Knowledge Extraction and analysis of extracted results. Preprocessing is an important step because of the complex nature of the Web architecture which takes 80% in mining process. The raw data is pretreated to get reliable sessions for efficient mining. It includes the domain dependent tasks of data cleaning, user identification, session identification, and path completion and construction of transactions. Data cleaning is the task of removing irrelevant records that are not necessary for mining. User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. Path completion is used to fill missing page references in a session. Classification of transactions are used to know the users interest ad navigational behavior. The second step in web usage mining is knowledge extraction in which data mining algorithms like association rule mining techniques, clustering, classification etc., are

applied in preprocessed data. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge. [6] Knowledge query mechanism such as SQL is the most common method of pattern analysis.

This paper focuses on path completion process which is used to append lost pages and construction of transactions in preprocessing stage. In this study a referrer-based method is proposed to efficiently construct the reliable transactions in data preprocessing. The paper is organized into six sections, in the first section Introduction of the concept is given, in the second section Literature survey for path completion is discussed, in the third section an overview of Data preprocessing phase is discussed, in the fourth section path completion technique is discussed, in the fifth section Transaction identification and analysis is Discussed and in the last section experimental results are shown followed by conclusion.

II. LITERATURE SURVEY

Chungsheng Zhang and Liyan Zhuang [1] suggested that reconstruction of accurate user sessions from logs is a challenging task as the HTTP protocol is stateless and connectionless Path completion is an important activity in preprocessing phase, as many patterns can be discovered and analyzed after forming the complete and accurate path. Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah [3] uses the link sequence information for prediction user links, D. W. Chueng, B. kao, and J. W. Lee [4] analyses the web pages visited by users and performs topic spotting.

In user session identification and path completion methods, the most common methods are timeout, maximal forward reference and reference length methods. Many authors have implemented path completion phase with different parameters, The method proposed by Cooley. R., Mobasher, B. & Srivastava, J.[2], assumes that the amount of time a user spends on a page depends whether the page is an auxiliary or content page. Z. Chen, A. Fu, J. Tang and F. Tung [10][11], they defined each session as the set of pages from the first page in a request sequence to the final page before a backward reference is made. Yan LI, Boqin FENG, Qinjiao MAO [9] they have implemented path completion algorithm using three steps. 1) The incomplete access path is identified, and path combination is conducted. 2) The content and auxiliary pages are identified by using the Maximal Forward References (MFR) and Reference Length (RL) algorithms. 3) The complete path for each user session is acquired by using referrer information and the reference length of some pages of this complete path is modified by using Average Reference Length Auxiliary Pages. G. Arumugam, S. Suguna [5] proposed User Session Identification Algorithm (USIDALG) containing two modules for the activities related to User Identification, and Session Identification.

Cooley, R., Mobasher, B. & Srivastava, J. [2] and Z. Chen, A. Fu, J. Tang and F. Tung [10] suggest that to process l
backward references among L logs the time complexity is O (N/2 * l) where N is the number of pages on the server. The complete path generation process fails to generate a correct path when the pages are referred from some other servers. So, level of accuracy is reduced. Cooley, R., Mobasher, B. & Srivastava, J. [2], there is no algorithm for generating a complete set of the USS. Yan LI, Boqin FENG, Qinjiao MAO [9], using SbSfxminer and Absfxminer the complete set of the USS is generated with the time complexity of •i=1 MFRS|MFRSi| G. Arumugam, S. Suguna [5] the performances are analyzed on parameters like 1. Generating a complete path 2. Time complexity to generate the complete set of User Session Sequence (USS) 3. Accuracy in generating a complete set of the USS. They find the Time complexity for generating the complete USS by applying the formula O (l* log (n/2)) /(No. of search / sec). In all the approaches the pages designed using cms are not considered , so this area requires focus and we
contrive to work on this.

III. AN OVERVIEW OF DATA PREPROCESSING FOR UNIVERSITY WEBSITE ACCESS DOMAIN (UWAD)

This phase is used to clean and process the data for making it available for analysis. Our preprocessing phase comprises
various steps like, Data Cleaning, User Identification, Session Identification, Path Completion, Generating Site structure (Site Map, Mapping Page Number and Name) and Generating Academic Events. A detail description of steps is given by
Nirali Honest, Dr. Bankim Patel, Dr. Atul Patel [7] [8]. Figure 1 shows the architecture of the data preprocessing phase for
UWAD.

Fig. 1
Architecture of Preprocessing phase UWAD

Website Structure considered in our work is based on a website designed using Content Management System, which shares two characteristics, 1) The pages are generated with unique identifiers and 2) The pages may have logical names apart from actual name. These characteristics are important to understand before carrying out path completion. The website structure we consider is shown in Figure 2.

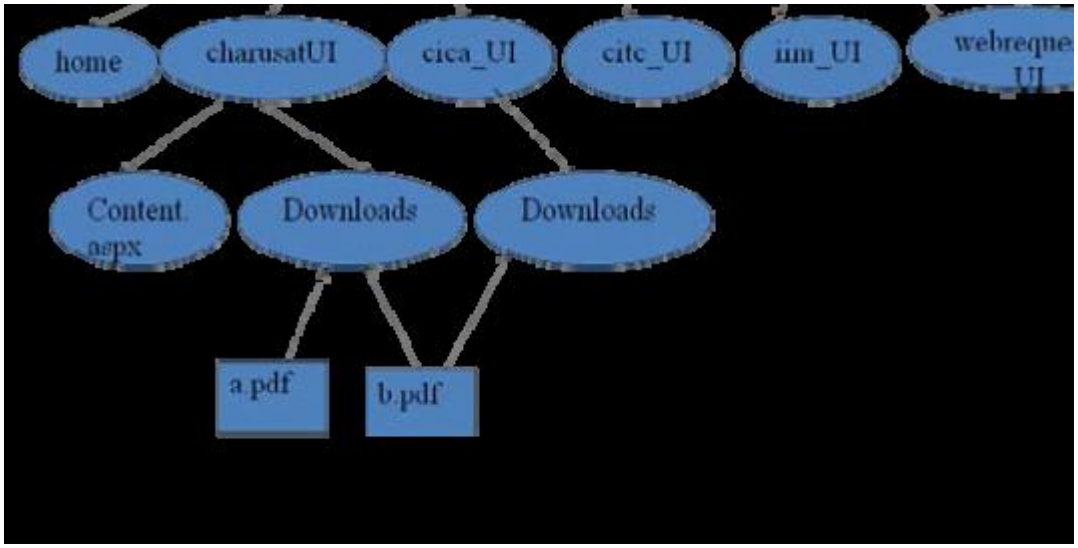


Fig. 2 Website structure

IV. USER AND SESSION IDENTIFICATION

The log file after cleaning is considered as Web Usage Log Set
WULS = {UIP, Date, Method, URI, Version, Status, Bytes, ReferrerURL, BrowserOS }. The next important and complex step is unique user identification. The complexity is due to the local cache and proxy servers. To overcome this cookies are used. But users may disable cookies. [8] Another solution is to collect registration data from users. But users neglect to give their information due to privacy concerns. So majority of records does not contain any information in the user-id and authentication fields. The fields which are useful to find unique users and sessions are

- IP address
- User agent
- Referrer URL

A. User Identification.

Users are identified by using these fields as follows.

- If two records has different IP address they are distinguished as two different users else if both IP address are same then User agent field is checked.
- If the browser and operating system information in user agent field is different in two records then they are identified as different users.

B. Session Identification.

After users are identified the next step is identification of sessions. A session is a sequence of activities made by one user during one visit to the site. The goal of session identification is to divide the page accesses of each user into individual sessions. These sessions are used as data vectors in various classification, prediction, clustering into groups and other tasks.

After identifying users if both IP address and Browser OS are same the referrer url field is checked. If URL in the referrer URL field in current record is not accessed previously or if referrer url field is empty then it is considered as a new user session. Reconstruction of accurate user sessions from server access logs is a challenge task since the access log protocol[HTTP protocol] is stateless and connectionless [3]. There are three heuristics available to identify sessions from users. Two are based on time and one based on the navigation of users through the web pages.

Time Oriented Heuristics: The simplest methods are time oriented in which one method based on total session time and other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [2] to 24 hours [10] while default time is 30 minutes by R.Cooley [9]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes the second entry is assumed as a new session.

Navigation Oriented Heuristics: This method uses web topology in graph format. It considers webpage connectivity, however it is not necessary to have hyperlink between consecutive page requests. But the proposed method is not following this heuristics due to the complexity of web topology.
From WULS , the set of user sessions are extracted as USS= {USID, (URI_1, $ReferURI_1$, $Date_1$).....(URI_k, $ReferURI_k$,$Date_k$))} where $1 \leq k \leq n$ and n denotes the amount of records in WULS. Every record in WULS must belong to a session and every record in WULS can belong to one user session only.
After grouping the records into sessions the path completion step follows.

V. PATH COMPLETION

Due to proxy servers and cached versions of the pages used by the client using 'Back' , the sessions identified have many missed pages. So path completion step is carried out to identify missing pages. Path Set is the incomplete accessed pages in a user session. It is extracted from every user session set.

A. Path Combination and Completion

Path set(PS) is access path of every USID identified from USS. It is defined as PS = (USID,(URI1,Date1,

RLength1),… (URIk, Datek, RLengthk)) where Rlength is computed for every record in data cleaning stage. After identifying path for each USID path combination is done if two consecutive pages are same. In the user session if any of the URL specified in the Referrer URL is not equal to the URL in the previous record then that URL in the Referrer Url field of current record is inserted into this session and thus path completion is obtained. The next step is to determine the reference length of new appended pages during path completion and modify the reference length of adjacent ones. Since the assumed pages are normally considered as auxiliary pages the length is determined by the average reference length of auxiliary pages. The reference length of adjacent pages is also adjusted.

VI. TRANSACTIONS IDENTIFICATION

The goal of transactions identification is to create meaningful clusters of references for each user [8]. Transaction identification is done by merges or divides approaches. To find out the user's travel pattern and user's interests, two kinds of transactions are defined. i.e., travel path transactions and content only transactions. The travel path is a combination of auxiliary and content pages accessed by a user. The content only transactions are only content pages which are used in mining to discover user's interest and cluster users visiting the same web site. There are three divide approaches as follows.

Transaction identification by Reference Length: Reference Length approach is based on the fact that depending upon the time taken a user spends on a page correlates to whether the page should be classified as auxiliary or content pages for that user.

Transaction identification by Maximal Forward Reference: This approach is based on the forward references in a path of pages accessed by a user. A forward reference is defined to be a page not already in the set of pages for current transaction and a backward reference is defined as a page that is already contained in the set of pages for the current transaction. A new transaction is started when the next forward reference is made. In this the last page in maximal forward reference are content pages and the pages leading to forward reference is treated as auxiliary pages.

Transaction Identification by Time Window: The time window approach partitions a user session into time intervals no larger than a specified parameter. If W is the time window then (Datem.time - Date1.time) $\leq$ W where m is the last page in a session.In the proposed method a combination of all methods are used and Content Path Set and Travel-path transactions are identified. First by using Maximal Forward Reference the paths in a session is split into forward reference paths. Travel paths of a user session are found. For example for a user access path 11-12-13-12-14, two travel paths will be identified by MFR algorithm are 11-12-13 and 11-12-14 where 11, 12 are the unique identification for each record in User Session Set. Travel Path Set is defined as the set of user travel paths, the member of TPS includes travel paths, the member of TPS includes travel paths having same USID, defined as

TPS = <USIDi, TPi1, TPi2, …,TPin > Where TP is the travel path is a group of URIs which are arranged according to the access time, a travel path including k URIs is defined as TP = {URI1, URI2,..., URIk}.

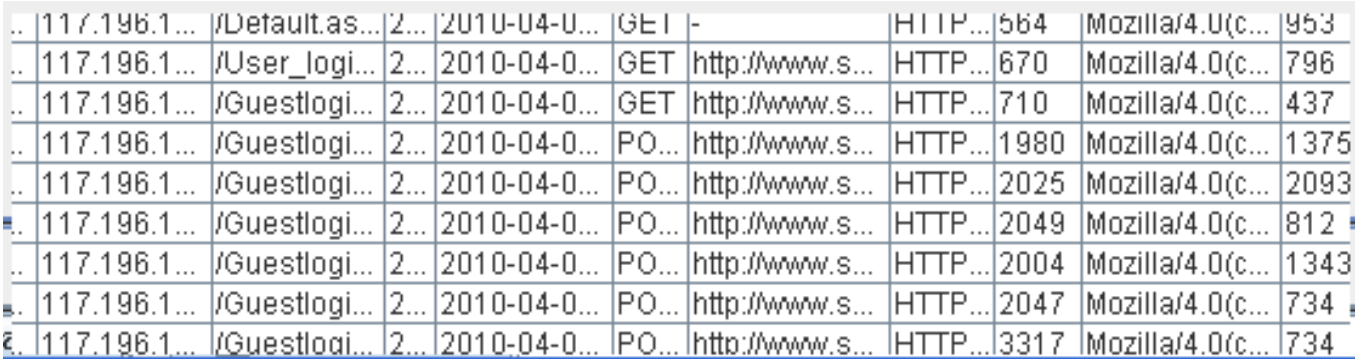
Reference Length algorithm is used to distinguish content pages from auxiliary pages. The algorithm depends on the time spent on viewing a page. A page is identified as content page if it exceeds a cut-off time or as auxiliary page if it is less than cutoff time. Cutoff time is calculated using a formula $t = -\lambda . Ln\ r$ where r is the percentage of content pages in the log found from the site . Normally the last page in every travel path is identified as content pages and leading pages are auxiliary pages. λ is the mean reference length of all pages in the log. In this the last record is ignored since last pages are normally considered as content pages. But it may be auxiliary pages also. To solve this issue the third algorithm Transactions by Time Window is used. In this a default time is fixed for each session and divided the path into transactions. The time difference between the first and last page access is calculated. That is considered as total time of transaction. Then the difference between Time Window and calculated total time is calculated. If the difference is less than cut off time it is considered as auxiliary page or as content page.

From the above algorithms Content transactions are identified. Content Path Set (CPS) is

the set of content pages, used for mining, corresponding to each user session, is written as CPS = < USIDi, CPi1, CPi2…CPik > where k is the number of content pages for the ith user session.

VII. EXPERIMENT

The sample log of our college web site is retrieved for a period of 20 days in the month of April in the year 2010. Initially there are 750 records in the log file. Data Cleaning is done by removing unsuccessful, extra records and entries from robot entry are cleaned. Finally a log of 332 records are obtained. A sample of the cleaned log file is given below



..	117.196.1...	/Default.as...	2...	2010-04-0...	GET	-	HTTP...	564	Mozilla/4.0(c...	953
..	117.196.1...	/User_logi...	2...	2010-04-0...	GET	http://www.s...	HTTP...	670	Mozilla/4.0(c...	796
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	GET	http://www.s...	HTTP...	710	Mozilla/4.0(c...	437
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	PO...	http://www.s...	HTTP...	1980	Mozilla/4.0(c...	1375
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	PO...	http://www.s...	HTTP...	2025	Mozilla/4.0(c...	2093
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	PO...	http://www.s...	HTTP...	2049	Mozilla/4.0(c...	812
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	PO...	http://www.s...	HTTP...	2004	Mozilla/4.0(c...	1343
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	PO...	http://www.s...	HTTP...	2047	Mozilla/4.0(c...	734
..	117.196.1...	/Guestlogi...	2...	2010-04-0...	PO...	http://www.s...	HTTP...	3317	Mozilla/4.0(c...	734

Users and sessions are identified after finding reference length of all records. There are 20 users identified by comparing IP address and Browser and Operating Systems and 28 sessions are identified by using referrer url method and divided with a time limit of 30 minutes.

No of UserId	ClientIp	No Of UserSessionId
1	117.196.160.55	1
4	117.196.161.21	8
5	122.164.14.149	5
6	122.164.49.135	10
4	122.164.61.151	4

After sessions are identified path set is framed and missing paths are inserted and transactions are found. Travel path transactions and content transactions are identified from the sessions by using the proposed method. The differences between the sets found with Time Window and without Time Window are also evaluated.

VIII. CONCLUSION

A data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. It has undergone arious steps such as data cleaning, user identification, session identification, path completion and transaction identification. Different from other implementations records are cleaned effectively by removing robot entries. The reference length is computed by considering the byte transfer rate. Apart from using Maximal forward reference (MFR) and reference length(RL) algorithm Time Window concept is also combined to find content pages. Travel path transactions are constructed to know the navigational behavior of users. Content page set is used for analyzing users and so that modification of sites can be done. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found.

REFERENCES

[1] Chungsheng Zhang and Liyan Zhuang , “New Path Filling Method on Data Preprocessing in Web Mining ,“, Computer

and Information Science Journal , August 2008.

[2] Cooley, R., Mobasher, B. & Srivastava J., ,"Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information Systems, I , Page(s): 5-32, 1999.

[3] Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah, "Kowledge Discovery from Users Web-Page navigation", , IEEE RIDE 1997.

[4] D. W. Chueng, B. kao, and J. W. Lee, "Discovering user Access patterns on the World-Wide Web", Proc. First Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-97).

[5] G. Arumugam, S. Suguna," Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs", Network and Service Security, 2009.

[6] Cyrus Shahabi, Amir M.Zarkessh, Jafar Abidi and Vishal Shah "Knowledge discovery from users Web page navigation, ", In. Workshop on Research Issues in Data Engineering, Birmingham, England,1997.

[7] Article "Preprocessing phase for University Website Access Domain", International Journal of Scientific & Engineering Research, (IJSER) –ISSN : 2229-5518, 4, No.6, June 2013.

[8] Nirali Honest, Dr. Bankim Patel, Dr. Atul Patel," Sessionization Process for the Pages Designed with the Concept of CMS", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X,Volume 3, Issue 9, September 2013 .

[9] Yan LI, Boqin FENG, Qinjiao MAO,"Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology,2008.

[10] Z. Chen, A. Fu, J. Tang and F. Tung, "Optimal algorithms for finding user web access sessions", Journal of World Wide Web: Internet and Information Systems, Vol. 6, Page(s): 259-279, 2003. Springer.

[11] Z. Chen, A.Fu, R.H. Fowler & C. Wang, "Efficient Web Mining of Frequent Traversal Patterns", in Anthony Acime, Web Mining: Applications and Techniques, Page(s): 322-338, Idea Group Publishing, August 2004.