WEB LOG PREPROCESSING BASED ON PARTIAL ANCESTRAL GRAPH TECHNIQUE FOR SESSION CONSTRUCTION

S.Chitra[1], Dr.B.Kalpana[2]

[1]Assistant Professor, Postgraduate and Research Department of Computer Science, Government Arts College, Coimbatore, Tamilnadu, India.
[2]Associate Professor, Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, Tamilnadu,India
[1]chitra.sivakumar@ymail.com
[2]kalpanabsekar@yahoo.com

ABSTRACT

Web access log analysis is to analyze the patterns of web site usage and the features of users behavior. It is the fact that the normal Log data is very noisy and unclear and it is vital to preprocess the log data for efficient web usage mining process. Preprocessing comprises of three phases which includes data cleaning, user identification and session construction. Session construction is very vital and numerous real world problems can be modeled as traversals on graph and mining from these traversals would provide the requirement for preprocessing phase. On the other hand, the traversals on unweighted graph have been taken into consideration in existing works. This paper oversimplifies this to the case where vertices of graph are given weights to reflect their significance. The proposed method constructs sessions as a Partial Ancestral Graph which contains pages with calculated weights. This will help site administrators to find the interesting pages for users and to redesign their web pages. After weighting each page according to browsing time a PAG structure is constructed for each user session. Existing system in which there is a problem of learning with the latent variables of the data and the problem can be overcome by the proposed method.

KEYWORDS

Web Usage Mining, Partial Ancestral Graph (PAG), Session Construction, Directed Acyclic Graph (DAG), Preprocessing, Robots Cleaning

1. INTRODUCTION

In this present internet World Wide Web sites on the internet are the source of useful information. As a result there is a huge improvement in its volume of traffic, the size and difficulty of web sites. World Wide Web grows rapidly day by day. So researchers are paying more and more attention on the effectiveness of services obtainable to the users over the internet. Web usage mining is an active technique which is used in this field of research. It is also called as web log mining in which the data mining techniques are applied to web access log. A web access log is a

Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, CRYPSIS, ICAIT, ICDIP, ITCSE, CS & IT 07,
pp. 473–482, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2446

time series record of user's requests each of which is sent to a web server at any time a user sent a request. Due to different server setting parameters, many types of web logs are there, but typically the log files share the same basic information such as client IP address, request time, requested URL, HTTP status code, referrer etc.

Web usage mining extracts regularities of user access behavior as patterns, which are defined by combinations, orders or structures of the pages accessed by the internet. Web usage mining consists of three main steps:

- Collection of Web data
- Data Preprocessing
- Knowledge Extraction
- Analysis of Extracted Results

The process in which Collection of data is the first step in which it represents the activities or clickstreams recorded in the Web server log. Preprocessing is an important step since the Web architecture is very complex in nature and 80% of the mining process is done at this phase.

Administrators of the web sites have to know about the users background and their basic needs. For this statistical analysis such as Google Analytics are used to examine the logs in terms of page views, page exit ratio, visit duration etc. With the help of this statistical analysis administrators can know about frequently accessed page, average view time and so on. But there are few drawbacks in statistical analysis. It gives low level error report on unauthorized entry points, invalid URLs are not found properly etc. Web usage mining enables administrators to provide complete analysis than statistical methods. It extracts a lot of patterns for administrators to analyze. This paper provides a method which analyses log files and extracts access patterns containing browsing time of each page using graphs [16].

Graph and traversal are extensively used to model a number of classes of real world problems. For example, the structure of Web site can be modeled as a graph in which the vertices represent Web pages, and the edges correspond to hyperlinks between the pages [7]. Mining using graphs turns out to be a center of interest. Traversals on the graphs are the models of User navigations on the Web site [14]. Once a graph and its traversals are specified, important information can be discovered. Frequent substructure pattern mining is an emerging data mining problem with many scientific and commercial applications [15]. This paper provides a new version to the previous works by considering weights attached to the vertices of graph. Such vertex weight may reflect the importance of vertex. For example, each Web page may have different consequence which reflects the value of its contents.

In existing graph of DAG, not all the variables can be measured, that variables values are called as observed variables and all other variables are called as latent variables.But in proposed method both the latent variable and the observed variable is measured.The remainder of this paper is organized as follows. The next section presents some basic of web log server. Section 3 provides the main steps of web usage mining. Section 4 describes about PAG construction and some experimental results are illustrated in Section 5. Finally, the conclusions are drawn in Section 6.

2. RELATED WORKS

Various commercial available web server log analysis tools are not designed for high traffic web servers and provide less relationship analysis of data relationships among accessed files which is essential to fully utilize the data gathered in the server logs [3]. The statistical analysis introduces a set of parameters to describe user's access behaviors. With those parameters it becomes easy for administrators to define concrete goals for organizing their web sites and improve the sites according to the goals. But the drawback in this analysis is that the results are independent from page to page. Since user's behavior is expected to be different dependent on length of browsing time, the calculation of accurate browsing time is more important [5].

A labeled graph is a tuple $G = (V, E, \varphi)$, where V is the set of vertices, E is the set of edges and $\varphi: V \rightarrow L$ is a labeling function with L a finite set of labels [9]. For an edge $(u, v) \in E, u$ is the parent of v and v is the child of u. If there is a set of vertices $\{u_1, \ldots, u_n\} \subseteq$ V such that$(u_1, u_2) \in E, \ldots, (u_{n-1}, u_n) \in E$, $\{u_1, \ldots, u_n\}$ is called a path, u_1 is an ancestor of un and u_n is a descendant of u_1. There is a cycle in the graph if a path can be found from a vertex to itself. An edge $(u, v) \in E$ of the graph is said to be a transitive edge if besides the edge(u, v), there also exists another path from u to v in G. A labeled graph is without cycles. Let $D = \{D_1, \ldots, D_n\}$ be a set of labeled PAGs and $\epsilon \geq 0$ be an absolute frequency threshold. PAG algorithm specifies that a PAG P is a frequent embedded sub-PAG of D if it is embedded in at least ϵ PAGs of D.

Prediction of users interest is the most important aspect of web usage mining. For this frequency and order of visited pages are considered. But Time spent on web pages is more important factor which is estimated from the log information and it is used as an indicator in other fields such as information retrieval, human-computer interaction (HCI) and E-Learning [2].

Duration time is the time that a user spends on reading a page in a session. Let P_i and P_{i+1} are two adjacent pages in a session. The timestamp field of P_i is Ti, and of P_{i+1} is Ti+1. Suppose T_3 is the loading time of P_i, and T_4 is the loading time ancillary files. By subtracting the time required for loading P_i and the ancillary files from the time difference between the requests of P_i and that of P_{i+1}, the duration time of P_i can be calculated [4].

The browsing time of an accessed page equals the difference between the access time of the next and present page. But with a more careful analysis, this difference includes not only user's browsing time, but also the time consumed by transferring the data over internet, launching the applications to play the audio or video files on the web page and so on. The user's real browsing time is difficult to be determined; it depends on the content of the page, the real-time network transfer rate, user's actions and computer's specifications and so on [13].

All of these works attempt mainly to find the exact browsing time of users so that web administrators can understand the interest of their users in web pages. In the proposed method a more accurate browsing time is found and creation of sessions as graphs depending on the time accessed.

3. PREPROCESSING

The quality of session construction significantly affects the whole performance of a web usage mining system. To improve the quality log data should be reliable. Preprocessing is a vital phase before mining to select the reliable data. Data Cleaning, user identification, sessions construction are the steps in preprocessing.

3.1 Data Cleaning

Data Cleaning enables to filter out useless data which reduce the log file size to use less storage space and to facilitate upcoming tasks [8]. It is the first step in data preprocessing. The log format used in this method is Extended Common Log Format with the fields as follows: "IP address, username, password, date/timestamp, URL, version, status-code, bytes-sent, referrer-URL, user-agent".

If a user needs a particular page from server entries like gif, JPEG, etc., are also downloaded which are not helpful for further investigation are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be eradicated from log files. Thus removal process includes elimination of irrelevant records as follows:

- If the status code of all record is fewer than 200 and better than 299 then those records are eradicated.
- The cs-stem-url field is verified for its extension filename. If the filename has gif, jpg, JPEG, CSS, and so on they are eradicated.
- The records which request robots.txt are eradicated and if the time taken is incredibly little like less than 2 seconds are considered as automated programs traversal and they are also eradicated [8].
- All the records which have the name "robots.txt" in the requested resource name (URL) are recognized and straightly eradicated.

3.2 User Identification

In this step users are identified from log files. Sites needed registration stores that the user data in log records. But those sites are few and often neglected by users. IP address, referrer URL and user agent in the log record is considered for this task. Unique users are identified as follows:

- If two records has dissimilar IP address they are differentiated as two different users else if both IP address are similar then User agent field is verified.
- If the browser and operating system information in user agent field is dissimilar in two records then they are recognized as different users else if both are identical then referrer URL field is checked.
- If URL in the referrer URL field in present record is not accessed before or if URL field is blank then it is considered as a new user.

3.3 Session Identification

A user session is defined as a sequence of requests made by a single user over a certain navigation period and a user may have a single or multiple sessions during a period of time. The objective of session identification is to segregate the page accesses of each user into individual sessions. Reconstruction of precise user sessions from server access logs is a difficult task because the access log protocol (HTTP protocol) is status less and connectionless. There are two simple methods for session identification. One is 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 [12] to 24 hours [8] at the same time as default time is 30 minutes by R. Cooley [4]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it goes over 10 minutes the second entry is understood as a new session. The third method based on navigation of users through web pages. But this is accomplished by using site topology which is not used in our method.

4. PAG CONSTRUCTION

In the proposed method sessions are modeled as a graph. Graph mining extracts users access patterns as a graph structure like the web sites link structure. To make efficient analysis when users handle more pages at the same time using tab browsers graph mining gives excellent results. Vertices are represented as web pages and edges are represented as hyperlink between pages. A graph is represented as a tuple of vertices, edges which connect the vertices [13]. User navigations are given as traversals in a graph. Each traversal can be represented as a sequence of vertices, or equivalently as a sequence of edges. PAG construction phase has following tasks.

4.1 Calculation of Browsing Time

The first task is to calculate browsing time of each page. For this the timestamp fields of the records are considered. Real Browsing time is very difficult to calculate since it depends on network transfer rate, user's actions, and computer specifications and so on. Browsing Time and Request Time recorded in log are abbreviated as BT andRT. Browsing time BT_p of page 'p' is equal to the period of time with the time difference between the RT_p of the request which include 'p' as a reference and another RT of the request which include 'p' as a requested page. In the log record one of the fields is bytes_sent which is the size of the web page. 'c' is the data transfer rate. So the real browsing time is assumed as

$BT_p = BT_p$' – bytes_sent / c
where BT_p' is the difference between reference and request page of 'p'.

4.2 Calculation of Weight of Pages

The second task in this method is to fix minimum and maximum browsing time for each page as BT_{min} and BT_{max} is used to calculate the weighing function which is to be used as a label in the graph. They are assumed by the administrators. The next step is to discretise the browsing time and given to each page as the weight which denotes the length of browsing time. Weighting function is calculated as follows

Wt (p, BT_p) = 0 when $BT_p \neq$ null and $BT_p < BT_{min}$

Wt (p, BT_p) = 1 when $BT_p \neq$ null and

$BT_{min} \leq BT_p \leq BT_{max}$

Wt (p, BT_p) = 2 when $BT_p \neq$ null and $BT_{max} < BT_p$ Wt (p, BT_p) = 3 when BT_p = null

If weight is '0' it is assumed as the time to browse is too short and the user simply passed the page. If weight is '1', administrators conclude it is a valid browsing time and user is interested in the content of the page. If weight is '2' the time is too long and it is assumed as if the user left the page and if the weight is '3' the page does not exists as reference page in that session. It is assumed as the end page and the user does not move from this page.

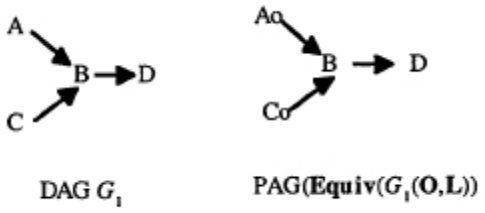
4.3 PAG Construction

A PAG (Partial Ancestral Graph) is used to represent any subset of Equiv (G (O, L))(Equivalence graph). A PAG is a extended graph consist of set of vertices O, and a set of edges between vertices were they may be following kind of edges: A ↔ B, A o─o B, A o→ B, A ←o B, A→ B or A←B. we say that the A endpoint of A→ B is "-" ; A endpoint of an A ↔ B is "<"; and then A endpoint of A o→ B is "o". The conventions of B endpoint are analogous. In addition pair of edge points may be connected by underlining. A partial Ancestral Graph for the set of directed graph G each carrying the same set of Observed variables O, contain partial information about the ancestral relation in G, namely only those ancestral relations is common to all members of G.

If G is a set of directed graph included in Equiv (G (O,L)), ψ (with vertices O) is a PAG for G if and only if

1. There is the edge between A and B in ψ if and only if every graph in G does not entail that A and B are the independent to the subset of O\{A,B}.
2. If there is an edge in ψ out of A, i.e. A → B then A is the ancestor of B in every graph in G.
3. If there is an edge in ψ into B, then in every PAG in G, B is not the ancestor of A.
4. If there is any underlining A*-*<u>B</u>*-*C in ψ then B is the ancestor of (atleast one of) A or C in every graph of G.
5. Any edge endpoint not marked in one of the above ways is left with a small circle thus o-*.

Note that the only condition (1) gives necessary and sufficient condition about the features of PAG. All the other conditions are merely necessary conditions. That means that there can be more than one PAG representing a given set G. Thus the PAG can be used to represent both the ancestor relation among the members of O common to the members of G, and the set conditional independence relation among the members of O in G. The PAG have two separate uses, they can be used in algorithm to perform fast condition and the other one is used to calculate the effect of any ideal intervention in the system. PAG also called as POIPGs.



DAG G_1 PAG(Equiv(G_1(O,L))

4.4 Pattern Extraction Phase

Once a graph and its traversals are specified, valuable information can be retrieved through graph mining. Normally they are in the form of patterns. Frequent patterns which are sub traversals occurred in a large ratio are considered for analysis. To discover PAG's i.e., sub graphs PAG mining algorithm is used which derive closed frequent sets. It replaces closed frequent PAG mining problem with the problem of closed frequent item-set mining on edges with the restriction that all the labels of the vertices in a PAG must be distinct. By the reconstruction of PAG structures from the mined closed frequent edge set, closed frequent PAG's are obtained. POIPGS extracts the embedded PAGs based on not only on parent-child relationship but also ancestor-descendant relationship of vertices.

4.5 Clustering Pattern

The last step is clustering of the mined patterns. The purpose of clustering is to group patterns which have similar page transitions. Each pattern is analyzed as different user behavior with browsing time. Weight of each page is not considered in clustering. The similarity of the patterns is to be estimated. Similarity of graphs is based on the labels of vertices and the edges. There are many clustering algorithms available to group the similar patterns. Administrators have to analyze the patterns respectively and it is time-consuming. They have to understand the meaning of each and every sub pattern to find out the problem of their web sites. If a content page has 0 weights then they have to redesign the page.

5. EXPERIMENTAL RESULTS

To confirm the usefulness and effectiveness of the proposed methodology, an experiment is carried out with the web server log of the library. The preliminary data source of the experiment is from May 28, 2006 to June 3, 2006, which size is 129MB. Experiments were carried out on a 2.8GHz Pentium IV CPU, 512MB of main memory, Windows 2000 professional, SQL Server 2000 and JDK 1.5. Table-I is the obtained results from the experiment.

Table-I The Processes and Results of Data Preprocessing in Web Usage Mining

Number of records in raw web log	Number of records after data cleaning	Number of users	Number of session construction using PAG
747890	112783	55052	57245

Table 1 show that after data cleaning, the number of log data diminished from 747890 to 112783. Four samples from the same university are obtained to evaluate the cleaning phase. From Figure-1 it is confirmed that the unwanted and irrelevant records are cleaned.

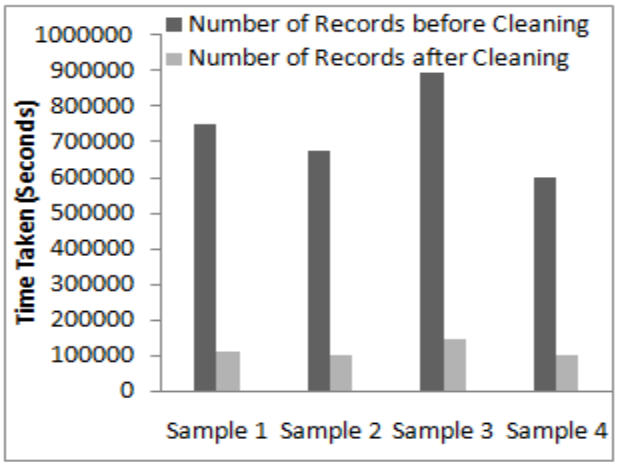


Figure-1: Data Cleaning of Sample Records

Table-II Comparison of Data Identification Using Various Methods

Various Model	Dimension of Data	Searching of Data
Dempster model	12	15.656
Whittaker model	14	4.42
PAG model	10	3.53

From Table-II, Comparing with other methods it can be observed that using the Partial Ancestral Graph the data is searched in short span of time as compared with the existing DAG method. The PAG model which as the greater flexibility in representing the data and the subset of data on the basis of conditional independence while comparing with other methods. In existing method we cannot identify the subset of the data.

6. CONCLUSION

Web log data is a collection of huge information. Many interesting patterns available in the web log data. But it is very complicated to extract the interesting patterns without preprocessing phase. Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. But understanding user's interest and their relationship in navigation is more important. For this along with statistical analysis data mining techniques is to be applied in web log data. In this paper, proposed a method to analyze web logs in detail by constructing sessions as Partial Ancestral Graphs. The graph easily searches the latent variables. The PAG is used to find the subset of the system and used to calculate the interventions of the system data. The proposed method takes advantage of both statistical analysis and web usage mining. Web site administrators follow the results and improve their web sites more easily. From the experimental results it is obvious that the proposed method successfully cleans the web log data and helps in identifying the data in short span of time.

REFERENCES

[1] Bollen, K (1989). Structural Equations with Latent Variables. Wiley, Newyork.
[2] Catlegde L. and Pitkow J., "Characterising browsing behaviors in the World Wide Web", Computer Networks and ISDN systems, 1995.
[3] Cooley, R., Mobasher, B., and Srivastava, J. (1999). "Data preparation for mining World Wide Web browsing patterns", Knowledge and Information Systems, 1999.
[4] Cooley, R., Mobasher, B., and Srivastava, J., "Web mining: Information and Pattern Discovery on the World Wide Web,", International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.
[5] Koichiro Mihara, Masahiro Terabe and Kazuo Hashimoto," A Novel web usage mining method Mining and Clustering of DAG Access Patterns Considering Page Browsing Time",2008
[6] Peter I. Hofgesang , "Methodology for Preprocessing and Evaluating the Time Spent on Web Pages", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,2006.
[7] Seong Dae Lee, Hyu Chan Park, "Mining Weighted Frequent Patterns from Path Traversals on Weighted Graph ", IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.4, April 2007.
[8] Spilipoulou M.and Mobasher B, Berendt B. "A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS Journal on Computing Spring, 2003.
[9] Suresh R.M. and Padmajavalli .R "An Overview of Data Preprocessing in Data and Web usage Mining, "IEEE, 2006.
[10] Termier, A., Tamada, Y., Numata, K., Imoto, S.,Washio, T., and Higuchi, T. (2007). DIGDAG, a first algorithm to mine closed frequent embedded sub-DAGs. In The 5th International Workshop on Mining and Learning with Graphs (MLG '07).
[11] WANG Tong,HE Pi-Lian"Find Duration Time Maximal Frequent Traversal Sequence on Web Sites", IEEE International Conference On Control and Automation , 2007.
[12] Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, IEEE, 2008.

[13] Yan Li and Boqin FENG "The Construction of Transactions for Web Usage Mining", International Conference on Computational Intelligence and Natural Computing, IEEE, 2009.
[14] Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani, M., "Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM", First International Conference on Networked Digital Technologies, Pp.224-249, 2009.
[15] Nina, S.P., Rahman, M., Bhuiyan, K.I. and Ahmed, K., "Pattern Discovery of Web Usage Mining", International Conference on Computer Technology and Development, Vol. 1, Pp.499-503, 2009.
[16] Chu-Hui Lee and Yu-Hsiang Fu, "Web Usage Mining Based on Clustering of Browsing Features", Eighth International Conference on Intelligent Systems Design and Applications, Vol. 1, Pp. 281-286, 2008.
[17] Bamshad Mobasher "Data Mining for Web Personalization,", LCNS, Springer-Verleg Berlin Heidelberg, 2007.
[18] Whittaker, J.(1990) Graphical Models in Applied Multivariate Statistics Wiley, NJ.
[19] Cochran, W.G. (1938). The omission or addition of an independent variate in multiple linear regressions. JRSS Supplement, 5, pp. 171-176.
[20] Richard Son, T.(1996). A discovery algorithm for directed cyclic graphs. Uncertainty in Artificial intelligence, proceeding, 12th conference, Morgan Kaufman, CA.

AUTHORS

Mrs. S. Chitra is an Assistant Professor of Computer Science in Government Arts College, Coimbatore. She received her Masters' degree in Computer Science from Avinashilingam University, Coimbatore. She has around 15 years of teaching experience at the post graduate and under graduate levels. Presently she is a Ph.D research scholar in Avinashilingam University. Her areas of interest are Data Mining and Web Mining. She is a life member of The Indian Science Congress Association, Kolkata.

Dr. B. Kalpana is an Associate Professor of Computer Science in Avinashilingam University, Coimbatore, Tamilnadu, India. She received her Ph. D in Computer Science from Avinashilingam University, Coimbatore. She specializes in Data mining. She has around 22 years of teaching experience at the post graduate and under graduate levels. She has published and presented papers in several refereed international journals and conferences. She is a member of the International Association of Engineers and Computer Scientists, Hongkong, Indian Association for Research in Computing Sciences (IARCS) and the Computer Society of India.