A Web Usage Mining Approach to User Navigation Pattern and Prediction in Web Log Data

Computer Science

KEYWORDS : Web usage mining; Navigation pattern; classification; weblog; clustering improved pairwise nearest neighbour.

SUJATHA PADMAKUMAR	Department of Computer Science, CMS College of Science and Commerce, Coimbatore, India.
Dr.PUNITHAVALLI	Dean, Sri Ramakrishna College of Arts and Science for women, Coimbatore, India.
Dr.RANJITH	HEAD OF Department of Computer Application Karunya University, Coimbatore, India.

ABSTRACT

Web Usage Mining (WUM) are specifically designed to carry out the user navigation pattern in web log files Organizations collect large volumes of data in their daily operations, generated automatically by web servers and collected in server access logs and this task is used for analyzing the data representing usage data and to predict the future movements. The improved pairwise nearest neighbour algorithm is used to group the potential clusters and the maximum likelihood classification algorithm is to predict users' future requests. The Experimental results the quality of clustering for user navigation pattern in web usage mining and for the prediction of users next request.

1. INTRODUCTION

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files. It mines the web log files derived from the users' interaction with the web pages during the visit to the web pages. Extraction of interesting information from Web log data has result in web mining. Web usage mining is an important technology for understanding user's behaviors on the web and is one of the favourite areas of many researchers in the recent time. The web pages can be used to identify the typical behavior of the user and to make prediction about desired pages [3].Web usage mining consists of three phases, namely pre-processing, pattern discovery and pattern analysis (fig 1).

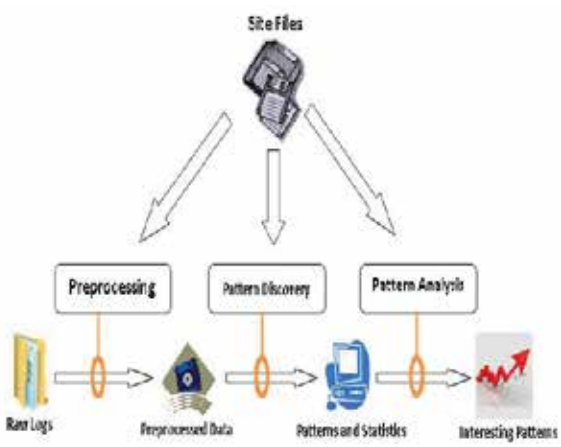

Figure: 1 Web Usage Mining Process

1.1 PREPROCESSING

Pre-processing "consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery". It always necessary to adopt a data cleaning method to eliminate the impact of the irrelevant items to the analysis result. The usage pre-processing probably is the most difficult task to the incompleteness of the available data (R.Cooley 2000).Without sufficient data it is very difficult to identify the users.

1.2 PATTERN DISCOVERY

This phase consists of different techniques derived from various fields such as statistics, machine learning, data mining, and pattern recognition etc, applied to the Web domain and to the available data. The task for discovering the patterns offer some techniques as statistical analysis, association rules, clustering, classification, sequential pattern, and dependency modelling

1.3. PATTERN ANALYSIS

Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

- Validation: to eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.
- Interpretation: the output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations.

The main objective of the proposed system is to predict the user interest by a classification process that identify the potential user from the web log data and the clustering process to group the similar interest of users' and the classification process to predict the future request.

2. RELATED WORK

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [9].Pattern discovery from web data is the key component of web mining and it converge algorithms and techniques from several research areas. Baraglia and Palmerini (2002) proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Liu and Keselj (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users" future requests and Mobasher (2003) presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et al. (2002) [10] proposed a hybrid approach for analyzing the visitor click sequences. To form highly dense clusters, it is proposed to do pair-wise nearest neighbor base clustering approach using only k– neighbors A.Anitha (2010) [7]. Maximum Likelihood classification for knowledge discovery from user's navigation patterns.

3. METHODOLOGY

3.1 WEB LOG FILES

Web log file is log file automatically created and maintained by a web server. Every "hit" to the Web site, including each view of a HTML document, image or other object, is logged. These files record the browsing behavior of site visitors. Data can be collected from multiple users on a single site. Log files are stored in various formats such as Common log [8] (fig 2) or combined log formats. This line consist the following fields.

- Client IP address
- User id ('-'if anonymous)

Volume : 3 | Issue : 4 | April 2014 • ISSN No 2277 - 8179

- Access time
- HTTP request method
- Path of the resource on the Web server
- Protocol used for the transmission
- Status code returned by the server
- Number of bytes transmitted

Client IP	Access Date and Time	Method	URL STEM	PROTOCAL	STATUS	BYTES	BROWSER
216.140.123.22-	31/May/2008 05:54:14+0400	"GET	elearning/index.html	HTTP/1.0"	200	9440	"Mozilla/4.0(compatible)"
216.140.123.22-	31/May/2008 05:54:15+0400	"GET	elearning/lessons.jpg	HTTP/1.0"	200	1164	"Mozilla/4.0(compatible)"
216.140.123.22-	31/May/2008 05:54:15+0400	"GET	elearning/lessons/style.css	HTTP/1.0"	200	842	"Mozilla/4.0(compatible)"
216.140.123.22-	31/May/2008 05:54:15+0400	"GET	elearning/lessons.jpg	HTTP/1.0"	200	11349	"Mozilla/4.0(compatible)"
216.140.123.22-	31/May/2008 05:54:15+0400	"GET	elearning/lessons/CS.jpg	HTTP/1.0"	200	319	"Mozilla/4.0(compatible)"

Figure :2 Sample web log files

3.2 DATA CLEANING

Irrelevant information which is not useful for mining purposes [7] can be removed from the HTTP server log files e.g. access performed by spiders, crawlers ,robots(these are automatic agents that surf the Web to collect and store the information e.g. search engine spiders)and files with extension name jpg, gif, css .

3.3 USER IDENTIFICATION

Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address. If the IP address of a user is same as previous entry and user agent is different then the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address [8].

3.4 SESSION IDENTIFICATION

The browsing speed is calculated as the number of viewed pages / session time. After handling the network robot entries, a series of decision rules are applied to group the users as potential and not-potential users. Given a set of training data containing valid log attributes, C4.5 classification algorithm is used to classify the users. The attributes selected are time (>30 seconds), number of pages referred in a session (Session time=30 minutes) and the access method used. The decision rule for identifying potential users is "If Session Time > 30 minutes and Number of pages accessed > 5 and Method used is POST then the classify user as "Potential" else classify as "Not- Potential". The purpose of introducing classification is to reduce the size of the log file. This reduction in size will help for efficient clustering and prediction.

3.5 CLUSTERING PROCESS

Clustering is the process of grouping the objects in such a way that , intra cluster similarity is high and inter cluster similarity is low. The pair wise nearest neighbor approach is a bottom - up hierarchical clustering technique, by which every object belongs to individual clusters initially, pair wise merging of objects is done at every step based on their similarity. Finally, resulting in a single cluster. The distance calculations [5] are replaced by similarity measure. Similarity between two transact ions are given by the ratio of, total number of unique pages referenced by them to the number of common references. For every transaction,its first k nearest access sequences are identified. Among the whole set of k -neighbors, the pair of sequences having high similarity is identified and merged. For this new cluster, the new k – neighbors are identified from 2k neighbors and its back neighbors are also updated [4]. The merging is continued until no more merging is possible. In this process , only the pair of access sequence having similarity value greater than a predefined threshold is selected for merging process. By this approach, the distant objects that are irrelevant to mining process are eliminated resulting in homogeneous access patterns.

1. Collect the access log information from web servers
2. Retrieve only IP address and URL details from access log by removing noise and filtering irrelevant details
3. Form click stream transactions and place each transaction in individual cluster

Perform pattern discovery by improved pair-wise nearest neighbor method on k neighborhood as follows:

(i) Find out similarity values between transactions
(ii) Identify first k neighbors having similarity greater than threshold, for every transaction and remove other neighbors
(iii) Cluster the pair with highest similarity
(iv) Update similarity for objects in the neighborhood of merged pair
(v) Find out new set of k neighbors from 2k neighbors of merged pair
(vi) Update the neighbors in the back list of merged pair
(vii)Repeat steps (iii) to (iv) until no more merging is possible.

3.6 PREDICTION PROCESS

Maximum likelihood classification (MLC) algorithm is a statistical decision rule that examines the probability function of a data for each of the classes and assigns the pixel to the class with the highest probability. This method is most often used in image classification and is used to classify user requests from web log data in the present study. The maximum likelihood equation used is called Mahalanobis minimum distance (MD) and is defined equation (1).

$$MD = (x\text{-}mI)^{TC}{}_{I}^{-1}(x\text{-}m_{I})$$

Where CI is the covariance matrix for the particular imagined movement considered, left or right and T stands for the transposition operator. The Mahalanobis distance is used in a minimum-distance classifier as follows: Let mR, mL be the means for the right and left imagined movement classes, and let CR, CL be the corresponding covariance matrices. A feature vector x is classified by measuring the Mahalanobis distance d from x to each of the means, and assigning x to the class for which the Mahalanobis distance is minimum. In this paper, the full covariance matrix is used to calculate the MD. The MLC algorithm is used to classify user requests into NI and I, based on the amount of time spent by them in a website. If the amount of time spent is more than 30 seconds, then they are considered as genuine users. According to [9], interested users exhibit certain access patterns; they access certain web pages for a rather long time because they need time to spend on its contents. The user who does not have interest simply accesses many pages quickly to browse content. This algorithm results in the prediction of users' next request.

4. EXPERIMENTAL RESULTS

4.1 CLUSTERING RESULT

This dataset has pre-processed web logs of the site www.microsoft.com. It records 38,000 randomly selected anonymous users of the site of which 31,501 are used for pre-processing out of which 10,481 weblog data are pre-processed and are used to find the potential user with 4000 are used for testing the ensemble model for clustering and classification. The number of visits made by the browsers in 24 hours to these 15 pages is presented in Fig 3 number of clusters found is another parameter that was used to analyze the performance of the clustering algorithm in Fig 4.

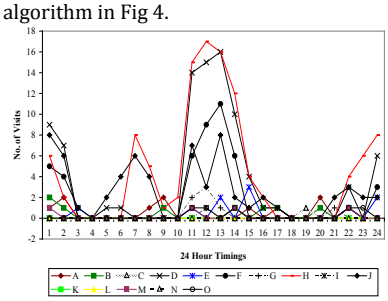

Figure: 3 Page visited details for twenty four hours.

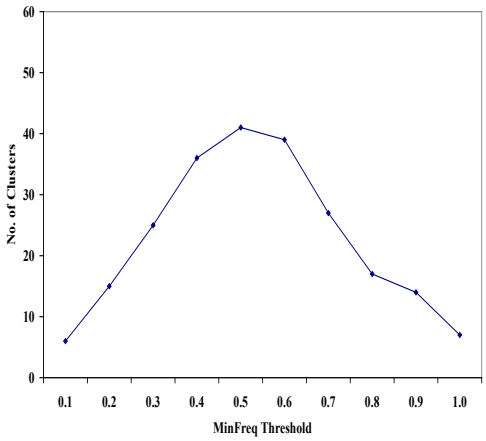

Figure: 4 Effect of Threshold on number of clusters

4.2 PREDICTION RESULT

The first set is used for generating prediction and the second set is used to evaluate the predictions. Let asnp denote the navigation pattern obtained for the active session„s" and let T be a threshold value. The prediction set is denoted as P(asnp, T) and the evaluation set is denoted as evalnp The three parameters can then be calculated using Equations (1) and (2) and the results are projected in Figures 5 for accuracy and in Fig: 6 for coverage.

$$\text{Accuracy} = \frac{|P(as_{np}, T) \cap eval_{np}|}{|P(as_{np}, T)|} \quad (1)$$

$$\text{Coverage } (P(as_{np} T)) = \frac{|P(as_{np}, T) \cap eval_{np}|}{|eval_{np}|} \quad (2)$$

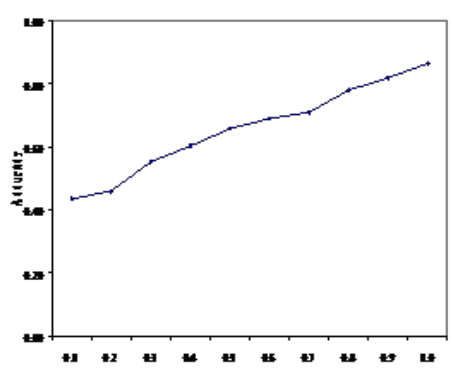

Figure: 5 Prediction Accuracy with threshold value from 0.1 to 1.0

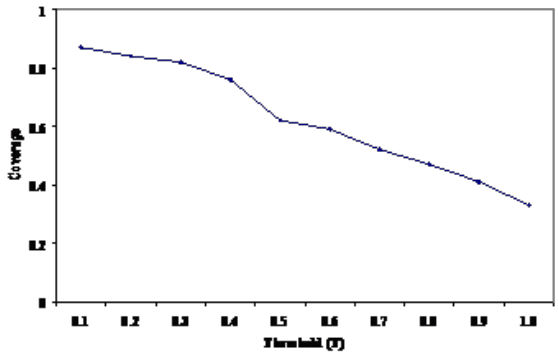

Figure: 6 Prediction Coverage with threshold value from 0.1 to 1.0

5. CONCLUSION

In this study, a usage navigation pattern prediction system was presented. The system consists of three stages. The first stage is the cleaning stage, where unwanted log entries were removed. In the second stage, cookies were identified and removed. The result was then segmented to identify potential users. The result was then segmented to identify potential users. From the potential user, Improved pairwise nearest neighbor clustering algorithm was used to discover the navigation pattern. A Maximum Likelihood classification algorithm was then used to predict future requests. The experimental results prove that the proposed amalgamation of techniques is efficient both in terms of clustering and classification. In future, the proposed work will be compared with existing systems to analyze its performance efficient. Plans in the direction of using association rules for prediction engine are also under consideration.

REFERENCE

[1] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava," Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information System,1999,pp. 1-27. | [2] Robert Cooley, Bam shad Mobasher, and Jaideep Srivastava." Grouping Web page references into transactions for mining World Wide Web browsing patterns", Knowledge and Data Engineering Workshop, New port Beach, CA.IEEE, 1997, pp.2- 9. | [3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan, "Web usage mining: Discovery and applications of usage patterns from Web data", SIGKDD Explorations, 2000, Vol.1.pp. 12-23. | [4] Pasi Franti,Olli Virmajoki, and Ville Hautamaki "Fast Agglomerative Clustering Using a k Nearest Neighbor graph",IEEE transaction on pattern analysis and machine intelligence.Vol 28,No 11. November 2006, pp 1875 -1880. | [5] Pasi Franti, Timo Kaukoranta,Day - Fann Shen and Kuo-Shu Chang "Fast and Memory Efficient implementation of exact PNN",IEEE Transaction on image processing,Vol 9,No 5,May 2000.pp 773-777. | [6] A.Anitha A New Web Usage Mining Approach for Next Page Access Prediction International Journal of Computer Applications (0975 – 8887) Volume 8–No.11, October 2010 7. | [7] Pang-Ning Tan, and Vipin Kumar, "Discovery of Web robot sessions based on their navigational patterns. Data mining and knowledge discovery", 2002, 6(1), pp. 9-35. | [8] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava,"Data Preparation for Mining World Wide Web Browsing Patterns ,", journal of knowledge and Information Systems,1999. | [9]Suneetha and Krishnamoorthi (2010) International Conference on Computing, Communications and Information Technology Applications, International Conference on Computing, communications and Information Technology Applications, (CCITA 2010), Coimbatore, India. | [10] Cooley, R., Srivastava, J. and Mobasher, B. , "Web mining: Information and pattern discovery on the world wide web,Tools with Artificial Intelligence", Ninth IEEE International Conference on In Tools with Artificial Intelligence, 1997.Proceedings., Vol. 10, pp. 0558-567. | [11] Eirinaki, M. and Vazirgiannis, M. , "Web mining for web personalization", ACM Transactions on Internet Technology (TOIT), 2003, Vol. 3, Issue 1, Pp. 1-27. | [12] Sujatha, V. and Punithavalli, 2012. Improved user navigation pattern prediction technique from web log data. Procedia Eng., 30: 92-99. |