IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 3, Issue 1 (July-Aug. 2012), PP 44-50
www.iosrjournals.org

Performance of Real Time Web Traffic Analysis Using Feed Forward Neural Networks and K-Means Algorithm

V.SarathChandra [1] *itsarath@yahoo.com,*
M.Tech 2nd Year Nimra Institute of Science and Technology,
Under the Guidance of
Nagul Shaik [2], Dr. M. Kishore Kumar[3]

Asst Professor in Dept of CSC,
Professor, M.Tech., Ph.D HOD Dept of CSC

Abstract: *Predicting of user's browsing behavior is an important technology of E-commerce application. The prediction results can be used for personalization, building proper web site, improving marketing strategy, promotion, product supply, getting marketing information, forecasting market trends, and increasing the competitive strength of enterprises etc. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining is usually an automated process whereby Web servers collect and report user access patterns in server access logs. The navigation datasets which are sequential in nature. Clustering web data is finding the groups which share common interests and behavior by analyzing the data collected in the web servers, this improves clustering on web data efficiently using proposed robust algorithm. In the proposed work a new technique to enhance the learning capabilities and reduce the computation intensity of a competitive learning multi-layered neural network using the K-means clustering algorithm. The proposed model use multi-layered network architecture with a back propagation learning mechanism to discover and analyze useful knowledge from the available Web log data.*

I. Introduction:

The goal of web usage mining is to find out the useful information from web data or web log files. The other goals are to enhance the usability of the web information and to apply the technology on the web applications, for instance, pre-fetching and caching, personalization etc. For decision management, the result of web usage mining can be used for target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc. Forecasting the users' browsing behaviors is one of web usage mining issues. In order to achieve the purpose, it is necessary to understand the customers' browsing behaviors through analyzing the web data or web log files. Predicting the most possible user's next requirement is based on the previous similar behavior. There are many advantages to implement the prediction, for example, personalization, building proper web site, improving marketing strategy, promotion, product supply, getting marketing information, forecasting market trends, and increasing the competitive strength of enterprises etc[2].The terminology of web mining was proposed by Etzioni in 1996. Web mining is applied to web pages and services of Internet in order to find and extract the available knowledge. Web mining can be categorized into three categories (as Fig 1.) which are web content mining, web structure mining and web usage mining

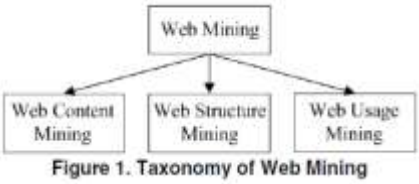

Figure 1. Taxonomy of Web Mining

Figure 1

Web content mining focuses on useful knowledge which is extracted from web pages. Web structure mining is used to analyze the links between web pages through the web structure to infer the knowledge. Web usage mining is extracting the information from web log file which is accessed by users. Lee and Fu proposed a Two Levels of Prediction Model in 2008 (as Fig. 2) . The model decreases the prediction scope using the two levels framework. The Two Levels of Prediction Model are designed by combining Markov model and

Bayesian theorem. In level one, Markov model is used to filter the most possible of categories which will be browsed by user. In the level two, Bayesian theorem is used to infer precisely the highest probability of web page.

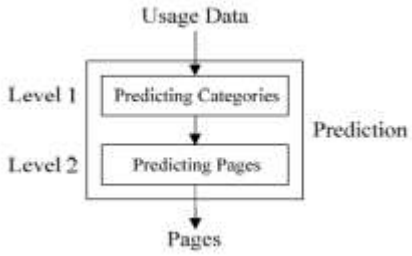

Figure 2

In level one, it is to predict the most possible user's current state (web page) of category at time *t*, which depends on user's category at time *t*-1 and time *t*-2. Bayesian theorem is used to predict the most possible web pages at a time *t* according to user's states at a time *t*-1. In the Two Levels of Prediction Model framework the similarity matrix *S* of category is established. The approach of establishing similarity matrix is to gather statistics and to analyze the users' behavior browsing which can be acquired from web log data. The enterprise proxy log is access log of the employees visiting the World Wide Web by using the proxy servers. Mining the enterprise proxy log provides a new aspect of analyzing the user behavior of surfing the web, and in addition it helps to optimal the cache strategies of proxy servers and the intranet management. In this paper, we focus on providing web recommendations. Firstly, a few features of the enterprise proxy log are presented by comparing with the web server log.

(1) Unlimited access to the web sites. Web server log is access log of the users surfing a particular site, while the enterprise proxy log has no limit to the sites which the users access. It makes that although we have a huge amount of data records, but these records are discrete and make the access patterns more hidden.
(2) Unknown to the information of the web sites. This includes the topology of the site, the classification of the pages, and the mark of the attach pages. This information has an important affect on filtering the access log to make mining algorithms correct.
(3) Diversifying the behavior motivations. The motivation of user browsing one site is usually related to the topic of this site, so we can expect to mine out some access patterns related to this topic. But that's not true while mining the enterprise proxy log for its WWW-oriented feature. We will be confronted with more access patterns and more motivations while mining the enterprise proxy log.
(4) Rapid growth of new pages in log. It's different from the web server log for its WWW-oriented feature.

II. Related Work:

S.Madria et al. [Madria 1999] gave details about how to discover interesting facts describing the connectivity in the Web subset, based on the given collection of connected web documents. The structure information obtained from the Web structure mining has the followings:

- The information about measuring the frequency of the local links in the web tuples in a web table The information about measuring the frequency of web tuples in a web table containing links within the same document
- The information measuring the frequency of web tuples in a web table that contains links that are global and the links that point towards different web sites
- The information measuring the frequency of identical web tuples that appear in the web tables.

Most of the research in Web usage mining is focused on applications using web Server Data. The only remained information after users visits a web site is the data about the path through the pages they have accessed. Most of the Web log analysis tools only use the textual data from log files. They ignore the link information, which is also very important. Web usage mining tries to discover useful information from the data extracted from the interactions of the users while surfing on the Web. It also has focus on the methods predicting user behavior while the user interacts with Web. Tasawar *et al.*, [3] proposed a hierarchical cluster based preprocessing methodology for Web Usage Mining. In Web Usage Mining (WUM), web session clustering plays a important function to categorize web users according to the user click history and similarity measure. Web session clustering according to Swarm assists in several manner for the purpose of managing the web resources efficiently like web personalization, schema modification, website alteration and web server performance. The author presents a framework for web session clustering in the preprocessing level of web usage mining. The framework will envelop the data preprocessing phase to practice the web log data and change the categorical web log data into numerical data. A session vector is determined, so that suitable similarity and

swarm optimization could be used to cluster the web log data. The hierarchical cluster based technique will improve the conventional web session method for more structured information about the user sessions. Yaxiu *et al.,* [4 put forth web usage mining based on fuzzy clustering. The World Wide Web has turn out to be the default knowledge resource for several fields of endeavor, organizations require to recognize their customers' behavior, preferences, and future requirements, but when users browsing the Web site, several factors influence their interesting, and various factor has several degree of influence, the more factors consider, the more precisely can mirror the user's interest. This paper provides the effort to cluster similar Web user, by involving two factors that the page-click number and Web browsing time that are stored in the Web log, and the various degree of influence of the two factors. The method suggested in this paper can help Web site organizations to recommend Web pages, enhance Web structure, so that can draw more customers, and improves customers' loyalty. Web usage mining based on fuzzy clustering in identifying target group is suggested by Jianxi *et al.*, [5].Data mining deals with the methods of non-trivial extraction of hidden, previously unknown, and potentially helpful data from very huge quantity of data. Web mining can be defined as the use of data mining methods to Web data. Web usage mining (WUM) is an significant kind in Web mining. Web usage mining is an essential and fast developing field of Web mining where many research has been performed previously. The author enhanced the fuzzy clustering technique to identify groups that share common interests and behaviors by examining the data collected in Web servers. Houqun *et al.*, [6] proposed an approach of multi-path segmentation clustering based on web usage mining. According to the web log of a university, this paper deals with examining and researching methods of web log mining; bringing forward a multi-path segmentation cluster technique, that segments and clusters based on the user access path to enhance efficiency.

III. Data Preprocessing:

The normal procedure of data preprocessing includes five steps data cleaning, user identification, user session identification, path completion and user transaction identification. While applying these to the enterprise proxy log, we encounter some new challenges. Web pages are becoming more and more colorful with attachments such as advertisements. However, as mentioned in section 1, we have no information about the web sites. It makes the normal data cleaning methods still have a lot of noisy pages which affects the data mining. Besides, this also disables the step of path completion for lacking information. For these reasons, the data preprocessing we used in this paper makes some modifications, which includes data cleaning, user identification, incremental filtering, user session identification and user transaction identification. These steps are shown as Fig. 2. We use the method of data cleaning according to [7] and user identification is easier because of the authentication information. Through the observation of the content pages and the attached pages, based on the feature that the attached pages are requested automatically when the related content pages are requested, we present some hypotheses. These hypotheses are as follows:

(1) Because with the feature above, the requested time of an attached page is immediately after the requested time of the related content page. We set this interval to be 1 second.

(2) An attached page usually can refer from many different pages. Although a content page can also refer from more than one other pages, but such records are much fewer in logs. So we assume that a page referred from more than 10 different pages is an attached page.

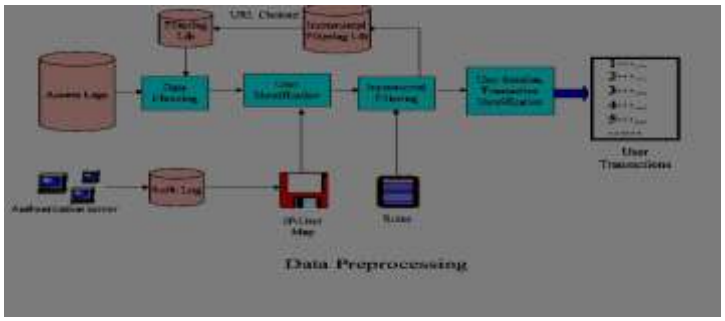

Figure 3

(3) Through our observation, the size of a content page surely is larger than 4500 bytes. Even if a few non-attached pages are smaller than 4500 bytes, they usually have too little content to attract user and can be ignored without affecting the mining results. Based on the above hypotheses, we propose a method of incremental data filtering and put those filtered pages into an incremental filtering lib. In this lib, we choice the pages referred from more than 10 different pages into the filtering lib to assist the data cleaning. The result shows that after a period of time. Filtering lib and the incremental filtering lib can gain the feature of attached pages. It helps a lot to find the advertisements or to find the naming rules of the attached pages. After incremental filtering, we apply the process of user session identification using 30 minutes as interval . A transaction is a subset of related pages

that occur within a user session, which targets at creating meaningful clusters of references for each user. Although lacking the topology of the web sites, we can construct a visit tree by using the records expressed by (URL, Referrer URL). As the Fig.4 shows, The transactions generated from Fig.3 are as follows. We express a transaction as: (user, transaction-id, {(navigational pages), (index page), (content pages)} to keep the relation of pages. *(user, transaction-id, transaction={(A,B,D), ,(H)}) (user, transaction-id, transaction={(A), C, (E, G)}) (user, transaction-id, transaction={(A, C), F, (I, J, K)})*

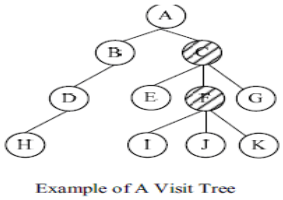

Figure 4

IV. Proposed Architecture*:*

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and Ungrateful step of the overall process. This system only describe it shortly and say that its main task is to ”clean” the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process. So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

Field Extraction*:* The log entry contains various fields which need to be separate out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most used separator character is ',' or 'space ' character. The Field Extract algorithm is given below.

Input: Log File
Output: DB
Begin
1. Open a DB connection
2. Create a table to store log data
3. Open Log File
4. Read all fields contain in Log File
5. Separate out the Attribute in the string Log
6. Extract all fields and Add into the Log Table (LT)
7. Close a DB connection and Log File
End

Data Cleaning*:* Data cleaning eliminates irrelevant or unnecessary items in the analyzed data. A web site can be accessed by millions of users. The records with failed HTTP status codes also may involve in log data. Data cleaning is usually site specific, and involves extraneous references to embedded objects that may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data [8].An algorithm for cleaning the entries of server logs is presented below

Input: Log Table (LT)
Output: Summarized Log Table (SLT)
‘*’ = access pages consist of embedded objects (i.e .jpg, .gif, etc)
‘**’ =successful status codes and requested methods (i.e 200, GET etc)
Begin
1. Read records in LT
2. For each record in LT
3. Read fields (Status code, method)

4. If Status code='**'and method= '**'Then
5. Get IP_address and URL_link
6. If suffix.URL_Link= {*.gif,*.jpg,*.css}Then
7. Remove suffix.URL_link
8. Save IP_sddress and URL_Link
End if
Else
9. Next record
End if
End

Proposed Framework: Portal Pendidikan Utusan normally known as Tutor.com and its server log consists of 19 attributes. The attributes are:-

a) Date: The date from Greenwich Mean Time (GMT x 100) is recorded for each hit. The date format is YYYY-MM-DD. The example from Fig. 1 above shows that the transaction was recorded at 2003-11-23.

b) Time: Time of transactions. The time format is HH:MM:SS. The example from Fig. 1 above shows that the transaction time was recorded at 16:00:13.

c) Client IP Address: Client IP is the number of computer who access or request the site.

d) User Authentication: Some web sites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a Website, that user's "username" is logged in the fourth field of the log file.

e) Server Name: Name of the server. In Fig. 1 the name of the server is CSLNTSVR20.

F) Server IP Address: Server IP is a static IP provided by Internet Service Provider. This IP will be a reference for access the information from the server.

g) Server Port: Server Port is a port used for data transmission. Usually, the port used is port 80.

h) Server Method (HTTP Request): The word request refers to an image, movie, sound, pdf, txt, HTML file and more. The above example in Fig. 1 indicates that folder.gif was the item accessed. It is also important to note that the full path name from the document root. The GET in front of the path name specifies the way in which the server sends the requested information. Currently, there are three formats that Web servers send information [8] in GET, POST, and Head. Most HTML files are served via GET Method while most CGI functionality is served via POST.

i) URI-Stem: URI-Stem is path from the host. It represents the structure of the websites. For examples:- /tutor/images/icons/fold.gif

j) Server URI-Query: URI-Query usually appears after sign "?". This represents the type of user request and the value usually appears in the Address Bar. For example" ?q=tawaran+biasiswa&hl=en&lr=&ie=UTF-8&oe=UTF- 8&start=20&sa=N"

k) Status: This is the status code returned by the server; by definition this will be the three digit number [2]. There are four classes of codes:

i. Success (200 Series)
ii. Redirect (300 Series)
iii. Failure (400 Series)
iv. Server Error (500 Series)

A status code of 200 means the transaction was successful. Common 300-series codes are 302, for redirect from http://www.mydomain.com to http://www.mydomain.com, and 304 for a conditional GET. This occurs when server checks if the version of the file or graphics already in cache is still the current version and directs the browser to use the cached version. The most common failure codes are 401 (failed authentication), 403 (Forbidden request to a restrict subdirectory, and the dreaded 404 (file not found) messages. In the above transmission a status is 200 means that there was a successful transmission.

a) Bytes Sent: The amount of data revisited by the server, not together the header line.

b) Bytes Received: Amount of data sent by client to the server.

c) Time Stamp: This attribute is used to determine how long a visitor spent on a given page.

d) Protocol Version: HTTP protocol being used (e.g. HTTP/1.1).

e) Host

- **T**his is either the IP address or the corresponding host name
- (www.tutor.com.my) of the remote user requesting the page.

Session Identification:

1: Calculate the browsing time of a web page by a user by finding the difference between two consecutive entries and subtract the time taken value
2: Compare the browsing time with minimum and maximum time of each web page

3: If the browsing time is less than minimum time fix the weight as „0 else if itis between minimum and maximum, then weight is fixed as „1 , if the weight exceeds maximum fix as 10 and if referrer URL is null weight is fixed as 100.
4: If the same page is visited by the user again in user s setincrement the corresponding entry.
5: Weights are stored in the matrix in the corresponding cells. The value aij represents a weight based on users browsing time in page j. until last row in users set.

K MEANS Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Proposed Hierarchical clustering:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to
[d[(r),(s)]=mind[(i),(j)]
where the minimum is over all pairs of clusters in the current clustering.
Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to [L(m) = d[(r),(s)]

Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way:d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]
If all objects are in one cluster, stop. Else, go to step 2.

V. Experimental results*:*

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operation system platform is Microsoft Windows XP Professional (SP2). The dataset is taken from real time php based real time web analyzer. Some of the results in the dynamic web statistics are given below as:

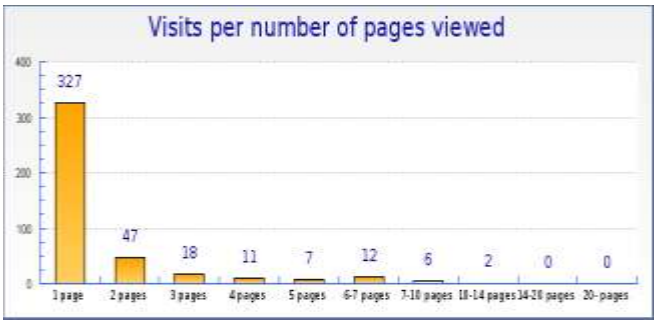

Figure 5: Number of pages viewed by the users

Figure 6

K-means Results:

Cluster centroids:

Attribute	Cluster# 0 (5)	1 (25)
Date	12/14/2011	12/12/2011
Maximum actions in one visit	7.2	9.48
Actions	124.8	101.28
Unique visitors	84.2	66.12
Visits	86.4	67.08
Actions per Visit	1.44	1.508
Avg. Visit Duration (in seconds)	36.8	25.32
Actions by Returning Visits	9.2	3.84
Unique returning visitors	6.2	2.84
Returning Visits	7	2.88
Avg. Actions per Returning Visit	1.32	1.348
Avg. Duration of a Returning Visit (in sec)	84.8	30.32
Conversions	0	0
Visits with Conversions	0	0
Conversion Rate	0	0
Revenue	0	0
Pageviews	124.8	101.28
Unique Pageviews	94.6	78.2
Downloads	0	0
Unique Downloads	0	0
Outlinks	0	0
Unique Outlinks	0	0

VI. Conclusion*:*

Data preprocessing is an important task of WUM application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. This paper presents two algorithms for field extraction and data cleaning. K means algorithm to the cleaned log file gives effective clustered results for web analysis. Not every access to the content should be taken into consideration. So this system removes accesses to irrelevant items and failed requests in data cleaning. After that necessary items remain for purpose of analysis. Speed up extraction time when users' interested information is retrieved and users' accessed pages is discovered from log data. The information in these records is sufficient to obtain session information.

References

[1] A new perspective of web usage mining: using enterprise proxy log yu zhang

[2] Two Levels of Prediction Model for User'sBrowsing Behavior1 Chu-Hui Lee, Yu-Hsiang Fu

[3] Murata, T. and K. Saito (2006). Extracting Users' Interests from Web Log Data. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06) 0-7695-2747-7/06.

[4] Pabarskaite, Z. (2002). Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining. 24th Int. Conf. information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia.

[5] Yun, L., W. Xun, et al. (2008). A Hybrid Information Filtering Algorithm Based on Distributed Web log Mining. Third 2008 International Conference on Convergence and Hybrid Information Technology 978-0-7695-3407-7/08 © 2008 IEEE DOI 10.1109/ICCIT.2008.39.

[6] Ou, J.-C., C.-H. Lee, et al. (2008). "Efficient algorithms for incremental Web log mining with dynamic thresholds." The VLDB Journal (2008) 17:827–845 DOI 10.1007/s00778-006-0043-9.

[7] Suneetha, K. R. and D. R. Krishnamoorthi (2009). "Identifying User Behavior by Analyzing Web Server Access Log File." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[8] Wahab, M. H. A., M. N. H. Mohd, et al. (2008). Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science,Engineering and Technology 48 2008.

[9] Stermsek, G., M. Strembeck, et al. (2007). A User Profile Derivation Approach based on Log-File Analysis. IKE 2007: 258-264.

[10] Yuan, F., L.-J. Wang, et al. (2003). Study on Data Preprocessing Algorithm in Web Log Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[9] Khasawneh, N. and C.-C. Chan (2006). Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 0-7695-2747-7/06 © 2006.