Web usage Mining for Exploring User Needs and Interest

Dr. V. Govindasamy
Dept. of Information Technology,
Pondicherry Engineering College,
Pondicherry, India.

V. Akila
Dept. of Computer Science and Engineering,
Pondicherry Engineering College,
Pondicherry, India.

Senthilnathan. A , Kapilaguru. A, Rajesh. V , Senthamizh Selvan. P
Dept. of Computer Science and Engineering,
Pondicherry Engineering College,
Pondicherry, India.

D. Dinesh
Dept. of Computer Science and Engineering,
Alpha College of Engineering and Technology,
Pondicherry, India.

***Abstract* - Web Usage mining helps in finding the user needs by analyzing the web server log files to make the administrators of the web sites to modify their web site to attract more number of users. This is very vital in commercial sites where the administrator has to know what his customer wants or for what the users prefer their site to other such web sites. Web usage mining is the process of getting that information about the users through the log files. This mining creates a access pattern which shows how a user uses a particular site. i.e the sequence of web pages the user has used.**

The main focus of our proposed system is to improve the efficiency of the data pre-processing technique compared to those that are in use currently. The result of the mining depends greatly on the pre processing technique. If the data has been cleansed and transferred to the database in a proper manner without any unwanted details the work of mining will be greatly reduced. Web log files are collected from the web site's administrator and are transformed to a database. Removal of unwanted details is done and user recognition and session recognition are done as parts of pre processing. Then the pre processed data is used for finding the access pattern which gives the user interest and needs. The web site can be modified according to the result of this. The result of this proposed work is calculated by comparing the modified site with the old one and a graph is plotted for various parameters such as ease of access and user satisfaction.

Keywords - Session Recognition;User recognition;Web usage mining

I. INTRODUCTION

A. Web Mining

Web Mining [1] is the extraction of appealing and useful patterns and implicit information from artifacts that are activity related to the World Wide Web. Web usage mining gives imminent help in web site maintenance , ecommerce and business intelligence applications. Extracting patterns from the log data by the use of event processing and machine learning technique is the norm of the day. This field has direct relevance in the field of business intelligence and web information system. There are three closely related fields to web mining. They are Web Content Mining, Web Structure Mining, and Web Usage Mining.

Web content mining pertains to extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining pertains to inferring knowledge from links between references . Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

B. Web Usage Mining

Web Usage Mining [2] [3] extracts navigation patterns of users by applying data mining techniques to server logs, heuristics, Web structure, Web content, user profile, registration data etc. One can examine its functions from two aspects: one is that it can give insights to Business Web site to perform navigation traffic analysis and to launch effective marketing strategies across products. The other is that it helps designer to optimize logic structure of Web site and provides more efficient customized services. Though the architectures and techniques of Web Usage Mining tools differ greatly in their implementations, we can still separate their task into three steps: pre-process, knowledge discovery and pattern analysis.

C. Preprocessing

Datasets are ambiguous, inconsistent and are noisy in nature. The data preprocessing is necessary to carry on a transformation to those databases [4]. The result is that the database will become consistent.Data used in preprocessing includes server log files, Web page content, Web page structure, user profile and registration data etc. Purpose of the phase is to offer a reliable and integrated data source to next phase that is, knowledge discovery. It should remove entries that contain noise. Such data offer nothing to the analyzing of user navigation behaviors.
In the data pretreatment work, mainly include

- data cleansing
- user identification
- session identification
- path completion.

a) Data Cleaning

Data cleaning is used to purge irrelevant items. These techniques are of importance for any type of web log analysis. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

b) User Identification and Session Identification

The task of user and session identification is to identify the different user sessions from the original web access log. User's identification is, to identify the users who access web site and identify the order in which the pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session comprises of a series of web pages user browse in a single access.

c) Path Completion

Another vital step in data preprocessing is path completion. The reasons that result in path's incompletion, are for instance, retrieving pages from local cache, agent cache, "post" technique and uses of browser's "back" button. This can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log [5] may be less than the real one. Using the local caches,proxy servers also produce difficulties for path completion because users can access the pages from the local caches or the proxy servers without leaving any record in server's access log. The result of this is the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended to the web log.

D. Objective

The aim of our proposed system is to implement an efficient data pre-processing technique and to implement a data mining algorithm which finds the access patterns of the users of a particular site. The proposed work is compared with other existing mining techniques by taking a survey in the web site modified based on the results of our project.

II. LITERATURE SURVEY

A. Web Usage Mining Solutions Based on SQL server 2000

Web usage mining [6] has to go through the procedures as follows: first source data collection, sort-out and transfer, then data pre-processing (including data cleansing, user recognition, session recognition and so on) and then mode discovery with a certain approach, and finally model analysis in combination with knowledge of certain fields. SQL Server2000 provides major tools for certain related processing such as data transfer service (DTS) and analysis services, and with SQL+ADO+VB/VC, Web usage mining is thus completed.

B. Data Transfer Service (DTS)

To transfer and integrate data from different data sources is the pre-condition and basis of data analysis and data mining. Web usage mining need to firstly sort out the .txt format web logs through format arrangement and transfer them to the data base, the function of which is provided by DTS under SQL Server2000.

C. Data Processing Service (SQL+ADO+VB+VC)

After data from data sources has been transferred to data bases, pre-processing will be done for source data which includes data cleansing, user recognition, session recognition etc. in web usage mining. The data preprocessing is made convenient with the powerful data processing function of SQL Server2000 (SQL), plus VB/VC for user-end programming and plus ADO for data connection.

D. Analysis Services

After data preprocessing, the next key step is to do data analysis and data mining. Web usage mining needs to analyze and mine the pre-processed web logs, that is, mode discovery and mode analysis. The analysis service components of SQL Server2000 include OLAP and DM, which provide tools for mode discovery and mode analysis.

III. PROPOSED WORK

A. Data Pre-processing

The data pre-processing techniques [7] [8] [9] in existence is not an efficient one. So we propose an identification strategy, based on the referred web page. This is based on user identification and session identification. At stage of Session Identification, the strategy based on preset priori threshold combined with session reconstruction is introduced. First, the initial session set is developed by the method of preset priori threshold, and then the initial session set is optimized by using session reconstruction.

The result of data preprocessing will directly affect the accuracy and reliability of the pattern detection algorithm processing results. As the first step of web log mining, data preprocessing impacts the rule and pattern produced by the data mining algorithm, which is the foundation of the entire web log mining and the key of quality assurance.

The system uses a sequential access pattern mining. The efficient sequential pattern mining algorithm is used to identify frequent sequential web access patterns. The access patterns are then stored as a Pattern-tree. This pattern tree is then used for matching and generating web links for recommendations.

The overall process of our proposed work is shown in fig 1. It includes two modules

- Data pre-processing
- Personalization

Data pre-processing includes data cleansing with user and session recognition which is performed from the web log files taken from the server. The pre-processed data is then mined, and the access pattern is found. This is useful in modifying the structure of the web site to meet the users needs.

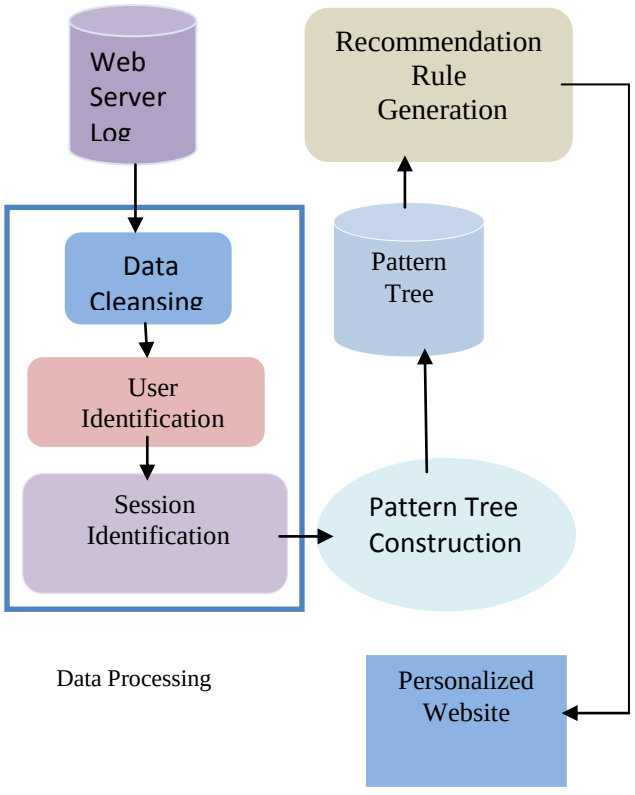

Fig. 1. Web Usage Mining for Exploring User Needs and Interest

A. Optimized User Identification

The strategy of User Identification based on the referred page refers to identifying users according to the referred page without considering the topology structure of site. The description of the algorithm is given in fig.2.

Input: N records of web log file, each record including IP, Agent.
Output: User set identified
Procedure OUI:
Step 1: *While*((i<N)
{
Step 2: *If*(U_i.IP != U_{i+l}.IP) //whether IP is the same
{ The user is a new one. }
Step 3: *Elseif*(U_i.Agent != U_{i+l}.Agent)//Check if the browser and operating system
{ The user is a new one. }
Step 4*: Elseif*(U_i.URL has been requested OR the referred page of U_i.URL is null)
{ The user is a new one.}
Step 5: *Else*
{ The user is the same one.}
i = i+l;
Return User Set.
}

Fig.2. Optimized User Identification algorithm

First of all, the algorithm judges user IP address and different IP is on behalf of the different user. If the IP address is the same, but the user's browser and operating system is different, that is a different user. When the IP address is the same and the user's browser and operating system is also the same, the method based on the referred page will be used. According to the referred page, if the page requested has been requested or the referred page is null, that is a new user. Otherwise that is the same one. The experiments prove that the algorithm significantly improves the efficiency and the accuracy of user identification.

B. Optimized Session Identification

Optimized method of Session identification is divided into two steps. First, the initial session set is developed by the method of fixed priori threshold, and then the initial session set is optimized using session reconstruction. The initial session set S is developed as follows. Given a page time threshold 't' when user stays at the page. If the time-gap of two continuous requests does not surpass 't', the two requests belong to the same session. Otherwise they belong to different session. 't' is generally set to 10 minutes. In the initial session set S, there may be such circumstances that the records in the same session are divided into different sessions or records which objectively are the different sessions are divided in the same session. So the session restructuring algorithm is used for tuning session set S and creating more realistic session set. The description of session restructuring algorithm is as follows.

In session identification, there may be practical session <L_l,"', L_i, L_j,"' ,L_n> which is divided into <L_l,"',L_{i-l}, L_i> and <L_j,. .. , L_{n-l}, L_n>.L_i and L_j belong to the same practical session. It means that the user has not turned to another topic, or the user has not left the site. Simply put, there is direct or indirect link between L_i and L_j in the topology structure of the web site. Based on the above facts, if L_i and L_j are the session border in the optimization process of session, L_i and L_j are linked into a session on two conditions. One condition is that the pattern of L_i->L_j is the user frequent accessing pattern, another is that L_j could be hyperlinked by L_i or L records in front of L_i in the Web site topology. In session identification, <L_1,"' ,L_{i-l}, L_i> and <L_j,. " , L_{n-l}, L_n> are two practical sessions. They may be divided into the same session <L_1,. .. , L_i, L_j,,··,L_n>.Although L_i and L_j are two consecutive records of the same user, being not in the same practical session shows that user has shifted to another topic. The user reaches to the page recording L_j by a certain amount of retreat or entering directly address. Based on the above facts, if L_i and L_j are internal records in the optimization process of session, L_i and L_j should be ruptured when L_j could not be hyperlinked by L_i or L records in front of L_i in the topology structure of the web site. In summary, the essence of session restructuring algorithm is union and rupture. Union the records which are divided into different sessions but objectively belong to the same session. Rupture records which are objectively in different sessions but divided in the same session. The algorithm for Optimized Session Identification is given in fig.3.

Input: The initial session set S, S={S_l,S_2,S_3 ... S_i ... S_n }, S_i is a session in S, S_i={ L_l, L_2, L_3 ... L_n }. // (1 <=i<=n)
Output: Optimized session set S'
PROCEDURE OSI:
Step 1: Enter orderly each session of S and each record L of the same session
Step 2: *Do Case*
Case (L_i and L_j are the last record(L_i) and the first record(L_j)of two consecutive sessions):
If (the pattern of L_i->L_j is the user frequent accessing Pattern)
Connect L_i and L_j
// Consolidation of two consecutive sessions
ElseIf (In the web site topology, L_j could be hyperlinked by L_i or L records in front of L_i)
Connect L_i and L_j
//Consolidation of two consecutive sessions
Else
L_i and L_j remain in the last record(S_i) and the first record(S_j)of two consecutive sessions.
Step 3: *Case* (L_i and L_j are two consecutive records of the same Session):
If (The time-gap of L_i,L_j > given threshold && In the web site topology, L_j could not be hyperlinked by L_i or L records in front of L_i)
Rupture L_i, L_j //L_i, L_j are divided into two session
Else
L_i and L_j are still two consecutive records of the same session
Step 4: Return Optimized Session Set

Fig.3 Optimized Session Identification algorithm

C. Pattern Matching Algorithm

a) Decision Trees

Decision Trees algorithm is a data mining algorithm. The Decision Trees algorithm calculates the outcome based on values in a training set. For example, a person in the age group 20-30 who makes over Rs.600,000/year and owns a home is more likely to own a car than someone in the age group of 15-19 who doesn't own a home. Based on age, income, and home ownership, the Decision Trees algorithm can calculate the odds of that person needing a car service based on the historical values.

The pre-processed data is mined to find the user pattern. Using this pattern the web site is re-structured according to the user interests and the efficiency is calculated.

IV IMPLEMENTATION

The proposed system has been implemented by SQL server 2005 Data Mining add-ins for data pre-processing and mining works. The proposed work(i.e Data pre-processing and pattern matching) have been implemented using java. The pre-processed data obtained as the result is used for data mining.

A. Performance Analysis

The web site for which the log files have been mined is modified on the basis of the results and the performance is measured by conducting some polls in the web site regarding the results of the modified site. Various parameters such as ease of use of the site, user satisfaction and the number of pages a user has to go through before reaching his required page are taken into consideration for conducting the poll. The results are tabulated and graphs are plotted against the number of users vs the result of the polls.

Different tables and graphs have plotted based on the results of the polls. The following lists them all.

a) User Satisfaction

The polls have been conducted for three days. Users were asked to vote in the poll. User satisfaction denotes which site (old one or modified new one) they feel is better. The graph is plotted in fig .4.

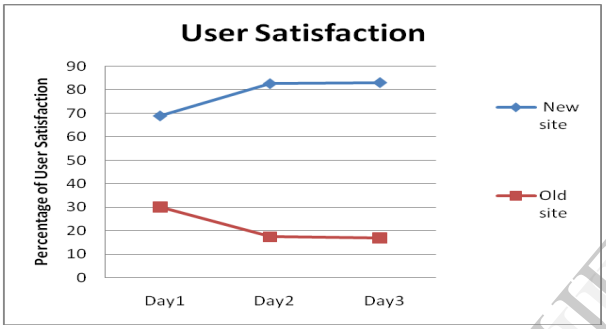

Fig. 4. User Satisfaction

b) Ease of Use

Ease of use represents which site the users have felt much easier to use. The graph is plotted in fig.5.

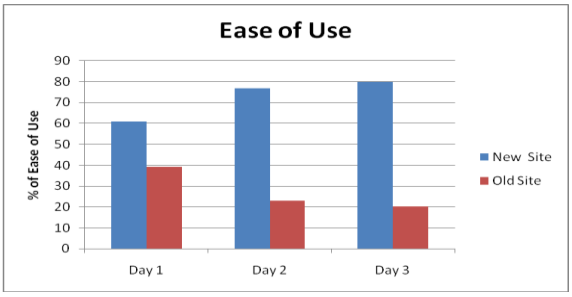

Fig. 5. Ease of Use

c) Number of Intermediate Pages

This denotes the number of web pages a user has to go through to reach his desired page. The number of pages are calculated for both old and new site and is plotted in fig.6.

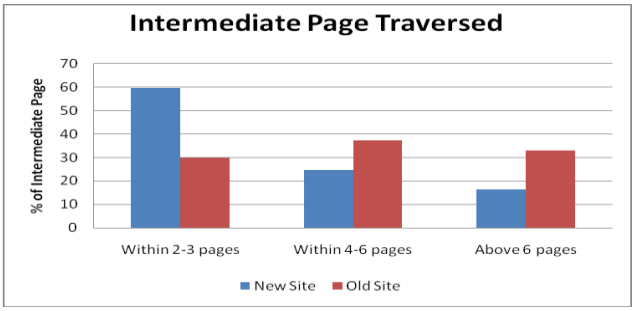

Fig. 6. Intermediate pages Traversed

V CONCLUSION

An system for web usage mining has been designed and implemented, which allows to find the users interest in a particular web site by finding the web access pattern. An efficient algorithm for data pre-preprocessing is designed with optimized user recognition and session recognition and is implemented. This project concludes that it gives web sites with improved qualities such as user satisfaction, ease of access to users etc. when compared to a web site that is not structured in an efficient way.

Internet is an ever evolving thing. New and improved techniques come every now and then. As future work we extend this data mining algorithm to be used for mining the data from proxy server logs in addition to server logs. Proxy logs are the ones that is gaining significance nowadays. Moreover, the web personalization can be done in more detailed manner to increase the efficiency of the result of this mining algorithm. We believe that this optimized data pre-processing solves the problems that were there in previous pre-processing techniques.

REFERENCES

[1] WANG Xiao-Gang, LI Yue, "Web Mining Based on User Access Patterns for Web Personalization" , ISECS International Colloquium on Computing, Communication, Control, and Management, pp.194-197, 2009.

[2] DeMinDong , "Exploration on Web Usage Mining and Its Application", Intelligent Systems and Applications, pp.1-4, 2009.

[3] Rajni Pamnani, Pramila Chawan, "Web Usage Mining: A Research Area in Web Mining", IEEE 2009.

[4] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan Mohamad Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology ,vol.2.no.12,pp.155-162, 2008.

[5] JIANG Chang-bin, Chen Li, "Web Log Data Preprocessing Based on Collaborative Filtering", Second International Workshop on Education Technology and Computer Science. ,pp.118-121, 2010.

[6] J Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology, Vol.11, no.4, March 2011

[7] Ling Zheng, HuiGui, Feng Li, "Optimized Data Preprocessing Technology for Web Log Mining", International Conference On Computer Design And Appliations ,pp.19-22,2010.

[8] Harald Genther and Manfred Glesner, "Advanced data preprocessing using fuzzy clustering techniques", Methods for Data Analysis in Classification and Control, vol.85,no. 2 , pp.155-164 , 1997.

[9] Nehal G. Karelia, Prof. Shweta Shukla, "Data Preprocessing: A Pre requisite for Web Log Files", International Journal of Engineering Research & Technology, vol.3,no.4,pp.1571-1574 , 2014