EFFICIENT USER NAVIGATION PATTERN PREDICTION TECHNIQUE FROM WEB LOG DATA

Vyas Mahesh Bharat[1], Mali Prasad Atmaram[2], Prof V. N. Nirgude[3]

[1] Student, Computer Dept. SRESCOE Kopargaon , Maharashtra, India
[2] Student, Computer Dept. SRESCOE Kopargaon, Maharashtra, India
[3] Assistant Professor, Computer Dept, SRESCOE Kopargaon, Maharashtra, India

ABSTRACT

Abstract: Web Usage Mining (WUM) is one of the most interesting areas of data mining. The main aim of WUM is to survey the web log files in order to extract users' navigation pattern .The web log files contain abstract data which needs to be processed in order to discover the meaningful data from it. Later on mining techniques are applied for clustering users, to organize frequently used data sets, for classification of users and association rule mining. This paper emphasizes on identifying user navigation pattern from web log data. The working is divided into two steps:

In first step web log data is processed. In second step the processed web log data is analyzed in order to identify the user access navigation pattern from it.

Keyword: - Web usage mining; Navigation pattern; classification; weblog; clustering; Graph partitioning.

1. INTRODUCTION

Web Usage Mining (WUM) is the automatic discovery of user access pattern from web servers. Organizations collect large volumes of data in their daily operations, generated automatically by web servers and collected in server access logs. This paper presents the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. In the first stage PUCC focuses on separating the potential users in web log data, and in the second t stage clustering process is used to group the potential users with similar interest and in the third stage the results of classification and clustering is used to predict the user future requests. The experimental results represent that the approach can improve the quality of clustering for user navigation pattern in web usage mining systems. These results can be used for predicting user's next request in the huge web sites

2. EXISTING SYSTEM

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 .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 Personalize system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et al. (2002) proposed a hybrid approach for analyzing the visitor click sequences. Jalal et al. proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge (2010) presented another user navigation pattern mining system based on the graph partitioning.

3. PROPOSED SYSTEM

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 Personalize system which provides dynamic recommendations, as a list of hypertext links, to users.

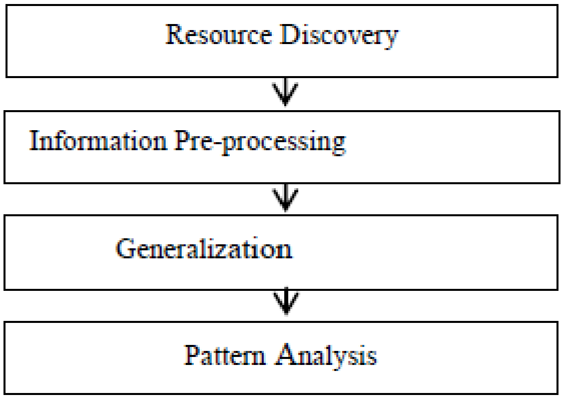


Fig 1: Process of Web Mining

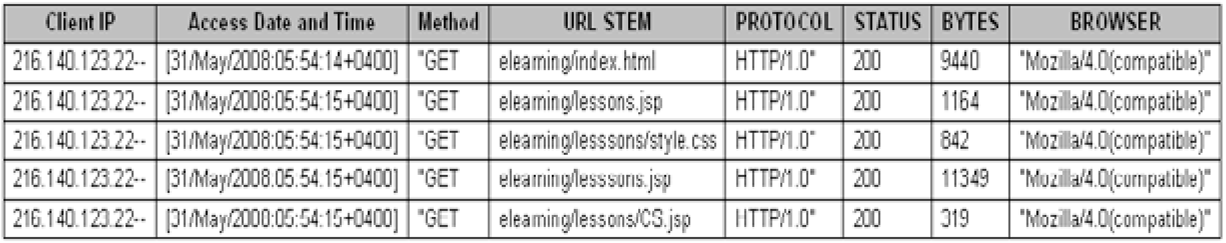
4. MODULES

1) Weblog 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. The raw web log file format is essentially one line of text for each hit to the web site. At least two log file formats exists: Common Log File format (CLF) and Extended Log File format,

1 The log file is text file. Its records are identical in format.

2 Each record in the log file represents a single HTTP request.



Client IP	Access Date and Time	Method	URL STEM	PROTOCOL	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.jsp	HTTP/1.0"	200	1164	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lesssons/style.css	HTTP/1.0"	200	842	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008.05.54.15+0400]	"GET	elearning/lesssons.jsp	HTTP/1.0"	200	11349	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2000:05:54:15+0400]	"GET	elearning/lessons/CS.jsp	HTTP/1.0"	200	319	"Mozilla/4.0(compatible)"

2) Pre-processing:

The first step of PUCC is the pre-processing of web log data, where the unformatted log data is converted into a form that can be directly applied to mining process. The pre-processing steps include cleaning, user identification and session identification. Cleaning is the process which removes all entries which will have no use during analysis or mining.

3) Identification of Potential Users:

This step of PUCC focuses on separating the potential users from others. It uses decision tree classification using C4.5 algorithm to identify interested users. They use a set of decision rules for this purpose. The algorithm worked efficiently in identifying potential users, but had the drawback that it completely ignored the entries made by network robots. Search engines normally use network robots to crawl through the web pages to collect information. The number of records created by these robots in a log file is extremely high and has a negative impact while discovering navigation pattern. This problem is solved in this paper by identifying the robot entries first before segmenting the user groups into potential and not-potential users

ALGORITHM (C4.5)
1. Selecting dataset as an input to the algorithm for processing.
2 .Selecting the classifiers
3. Calculate entropy, information gain, and gain ratio of attributes.
4. Processing the given input dataset according to the defined algorithm of C4.5 data mining.
5. According to the defined algorithm of improved C4.5 data mining processing the given input dataset.
6. The data which should be inputted to the tree generation mechanism is given by the C4.5 and improved C4.5 processors. Tree generator generates the tree for C4.5 and improved C4.5 decision tree algorithm.

5. CLUSTERING PROCESS

The k-means algorithm is used for clustering process where each cluster’s center is represented by the mean value of the objects in the cluster. Here k represents the number of clusters to be formed

Method:

- Arbitrarily choose K objects from D which is the initial cluster center.
- Repeat the same procedure
- Re-assign each object to the cluster to which the objects is most similar based on the mean value of the objects in the cluster
- Update the cluster means that is calculate the mean value of object for each cluster until no change

6. PREDICTION ENGINE

The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts user’s future requests. The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts users‟ future requests. For this purpose we use LCS algorithm. The main aim of LCS is to find the longest subsequence common to all sequences in a set of sequences. The algorithm works with two features. The first property state that if two sequences X and Y both end with the same element, then their LCS will be found by removing the last element and then finding LCS of the shortened sequence. The second property is used when the two sequences X and Y does not end with the same symbol.
Algorithm Steps: (LCS Algorithm)

1. To find the longest subsequence common to Xi and Yj , the elements Xi and Yj are compared.

2. If equal, then the sequence LCS (Xi-1, Yj-1) is extended by that element, Xi.

3. If they are not equal, then the longer of the two sequences, LCS (Xi, Yj-1), and LCS (Xi-1,Yj), is retained.

4. If they are both of the same length, but are not identical, then both are retained

7. CONCLUSIONS

As we know that web log data contains raw data. Thus preprocessing the web log data is a significant and the most basic steps in web mining. It removes the unwanted items and recognizes users with browsing information. The different patterns can be then discovered by applying the mining techniques. The different patterns can be then discovered by applying the mining techniques

REFERENCES

[1].V.SUJATHA a, PUNITHAVALLI b, a*,"IMPROVED USER NAVIGATION PATTERN PREDICTION TECHNIQUE FROM WEB LOG DATA", International Conference on Communication Technology and System Design 2011

[2].Gaurav L. Agrawal1, Prof. Hitesh Gupta," Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering

[3].Costas S. Iliopoulos and M. Sohel Rahman , "A New Efficient Algorithm for Computing the Longest Common Subsequence", {csi, sohel}@dcs.kcl.ac.uk

[4].Neetu Anand ,"Identifying the User Access Pattern in Web Log Data ", International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3536-3539

[5].Ranieri Baraglia, Paolo Palmerini," A Web Usage Mining System", CNUCE, Istituto del Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy.

[6].Haibin Liu, Vlado Kesˇelj," Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting user's future requests ",ScienceDirect

[7].BAMSHAD MOBASHER," Web usage mining can help improve the scalability, accuracy, and flexibility of recommender systems", COMMUNICATIONS OF THE ACM