A Hybrid Technique to Predict Web Page Access on the Base of CPMM

Palak Solanki, Mitesh Patel
Silver Oak College of Engineering and Technologies
Ahmedabad, Gujarat

***Abstract*- Web server stores details about user activities in web log files. This log files are then used to mine patterns for web page prediction. When Users make request from their browser to web server for the information they require. These requests are stored in web log files on web server. Data Preprocessing will help to clean and format Log files. When different web mining algorithms are applied on this filtered data from logs the frequent pattern can be mined. Further these patterns will be useful for prediction for user's next request on the base of user's country.**

***Index Terms*- Web Page Access, Eclat Algorithm, Markov Model, Confidence Pruned Markov Model, Web Navigation Pattern**

I. INTRODUCTION

Web is having huge amount of data but it is also very important part of our life nowadays. when numbers of users and web applications have increased on internet, activities have been increased between them as well. Because of the increase in their interaction data is increased also. When the data is huge in amount, we need more analysis of those data for making them more useful for users for different purposes. To understand the relationship of user and user access pattern which exist in the particular sessions we require web usage analysis.

Certain patterns are grouped together for specific reasons. If we don't know more about domain knowledge the these patterns provide very little information about those reasons. [3] The result of analysis of web access logs can help to understand the user behavior and web structure.

Large data are collected automatically by Web servers and accumulated in access log files. Examination of server access data can offer considerable and valuable information. So the providers need only a tool to analyze these logs. This is where web usage mining is helpful. By Mining data from web server log files with accurate constraints web user navigation pattern can be discovered. This data is used for personalization and recommendation.

'Web usage mining' is the process of extracting useful information from server logs. Web usage mining is the process of finding out what users are looking for on internet. [1] Web usage mining is process of finding patterns of navigation behavior from users visiting websites. There are many web pages and when you move from one to another web page a certain path is created. This path is stored inside web server in access log files. A series of web pages in a website requested by a visitor in a single visit is known as session. Web log mining is process of pattern discovery from web access logs. [2] These navigational patterns can be used for online promotion and personalization. These discovered navigation pattern is important to answer some of the questions like - Can the web site predict user's next page to visit? Can the website personalized according to the specific group of users? Is the web site efficient enough to deliver information? Can the user understand the structure of website? By analyzing web log files we can answers these questions.

II. BACKGROUND THEORY

Web Usage Mining is the discovery of user access patterns from Web servers. The figure below shows the steps of process. [3]

1.Data Preprocessing: It consists of four steps: data cleaning, user identification, session identification, path completion. The Purpose of data preprocessing is to offer structural, reliable and integrated data source to pattern discovery.

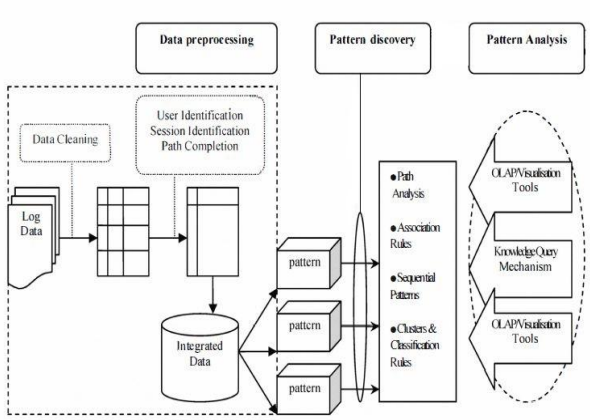

Figure 1 Architecture of web usage mining [3]

2.Pattern Discovery: It is the very important step of the Web mining. It consists the algorithms and techniques such as data mining, machine learning, statistics and pattern recognition. The techniques used for discovering patterns and rules are statistical analysis, association rules, clustering, classification, sequential pattern and dependency modeling. For characterizing, comparing, predicting or classifying data from the Web access log, the discovered knowledge is represented in the form of rules, tables, charts, graphs and other visual presentation.

3.Pattern Analysis: Pattern analysis is the Last step of the WUM. This process is used to mine the interesting rules or patterns from the output of the second step by removing the irrelative rules or patterns.

Application of Web usage Mining:

1.To keep track of previously accessed ages of user.

2.To predict the future access by identifying frequent access behavior for the users.

3.To use frequent patterns for advertising and making business decision.

4.To improve and personalize websites by analyzing users access behavior.

III. PATTERN DISCOVERY

There are three approaches used for web navigation pattern. [2] Which are:

1.Association Rule: Association Rules are able to discover related item occurring together in same transaction, and is used to find interdependency, co-relation among the pages. Such number of rules generated could be very large so two measures support and confidence is employed, which determines importance and quality of rules.

2.Sequential Pattern Mining: When a time domain is attached attributes this results into Sequential pattern. Now if we have certain user specified minimum support for sequences then the problem of mining sequential patterns is to find the maximal frequent sequences among all of them.

3.Clustering and Classification: It is an automated process of assigning a class label or mapping a user based on browsing history or on the basis of some other attribute with one of existing class. It can be done by various inductive learning algorithms like decision tree classifier, naïve Bayesian classifiers, Support Vector Machine. Web session clustering is one of the important techniques which aim to group usage sessions on the basis of some similarity measures. Mostly used clustering approaches are either partition or hierarchical.

IV. PATTERN ANALYSIS

The output of pattern discovery by is not always used in the form that people can understand so this output is transformed into the form that can easily integrate.This can be done by using some analysis methodologies and tools.

There are two approaches for the pattern analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations.[4]

WEB LOG FILES

A Web log is a file to which the Web server writes information each time a user requests a website from that particular server. The log file that resides in the web server notes the activity of the client who accesses the web.Content of Log files :

User name,Visiting Path,Path traversed,Time stamp,Last visited pageSuceess Rate,User Agent,URL,Request Type.

ACCESS LOG FILES:

All user requests are stored in the web access log.Storing the information in the access log is only the start of log management. The next step is to analyze this information to produce useful statistics. The information given below describes server

configuration to record information in web access log. Here are three log formats considered for access log entries in the case of Apache HTTP Server Version 2.4. [10]

1.Common Log Format
2.Combined Log Format
3.Conditional Logs

Markov Model:

To understand the random process where probability are involved Markov are mostly used.Now predicting a web page on the base of user's history is a random process. So to predict users navigation pattern for web site Markov model is one of the most appropriate one. For this model, the sequence of web-pages that were accessed by a user are used as input. The aim of this is to build Markov models.These models can be used to predict the next possible the web-page that the user will access.

In traditional Markov Model first it compare two sequences.One sequence is from log file that is previously used for accessing web pages and another sequence is current access.After the comparison it predict the next possible web-page that user can Another use of this model is modeling and analyzing Web access sequences.

Markov models are represented by three parameters < A, S, T >, where A is the set of all possible actions that can be performed by the user; S is the set of all possible states for which the Markov model is built; and T is a The transition probabilities represent the likelihood of moving between those states. It is Transition Probability Matrix where for ith state how many times the action j is performed is enterd in tij value.[5]

Confidence Pruned Markov Model:

In Markov Model, S represent Markov states and A represent possible access. Now one makov state can have more than one actions from that state to another state.Suppose there are two possible actions. Now for these situation probability of these two actions are different from one another.When One has more than another action then there are two possibilities:

1. If the support for this state is low, the difference in those actions probabilities is high. This shows that the state produces reliable predictions

2. If the difference between outgoing probabilities in the above situation is very less, the large training set is used to make the difference reliable.

CPMM uses statistical techniques to determine for each state, if the probability of the most frequently taken action is significantly different from the probabilities of the other actions that can be performed from this state.

The state is pruned if difference between probabilities is not significant because then the state is less accurate.The state is retained if difference between probabilities is significant. There are two steps to find the significance for most probable action from second most.First step is to compute the confidence interval around the most probable action.Second step is to check if second action's probability falls within that interval or not.If it is within the interval then the state is prunes otherwise it is kept as it is.

If p' is the probability of the most probable action, then its 100(1-α) percent confidence interval is given by $p' - z_{\alpha/2}(p'(1-p')/n)^{1/2} <= p <= p' + z_{\alpha/2}(p'(1-p')/n)^{1/2}$,where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution and n is the frequency of the Markov State. The degree of pruning in CPMM is controlled by confidence coefficient α. As the value of α decreases the size of the confidence interval increases, resulting in more pruning. If the difference in the probabilities between the two most probable actions is relatively small, the state will most likely be retained.[6]

Apriori	**Eclat**
User Prior knowledge of item sets	Maintain a transaction list for each item. so only transaction database is required.
BFS	DFS
Generate all subset for each transaction	Doesn't generate all subset for each transaction
Iterative approach : level wise search	Doesn't search complex data structure
Extra computation overhead occurs due to iterative approach	Doesn't pay extra computation overhead

Paper	Algorithm	Author	Advantage	Drawback
1	Apriori [7]	Kotiyal, B., Kumar, A., Pant, B., Goudar, R. H., Chauhan, S., et Junee, S.	It uses large item set. It can be easily paralleliz ed and It is easy to impleme nt	It is cost wise difficult to handle huge number of candidate sets . It has extra computatio n overhead due to its iterative approach.
2	FP Growth [8,9]	Pal, Monti BabuLal, and Dinesh C. Jain.	Use compact data structure and eliminate s repeated database scans.	It lacks Good candidate generation method.
3	Eclat [7]	Kotiyal, B., Kumar, A., Pant, B., Goudar, R. H., Chauhan, S., et Junee, S.	It maintains a transactio n list for each item. So only transactio n database is required. So It doesn't have extra computat ion overhead.	It doesn't generate all subset for each transaction. It doesn't search complex data structure.
4	Path Traversal Graph [2]	Yao Te Wang et Anthony J T Lee	It is appropriate for sequential patterns in incremental mining.	Cannot mine nonconsecutive browsing patterns

V. PROPOSED METHODOLOGY

In proposed methodology we combine a Frequent Pattern Mining algorithm with Prediction Model.

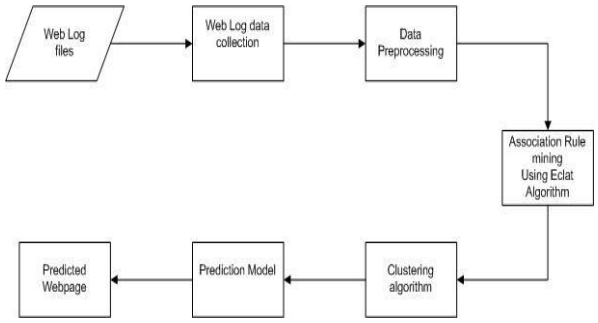

Figure 2 Proposed algorithm

PROPOSED ALGORITHM :

Step 1: Web logs are collected from web server with different parameters

Step 2 :Web Log is preprocessed by different techniques like data cleaning, user identification, session identification and path completion.

Step 3:For mining frequent pattern, Association rule mining algorithm-Eclat algorthm will be used on training sessions.

Step 4:To cluster Web sessions or user among the mined pattern K means algorithm will be used.

Step 5 :Take remaining sessions as test sessions and apply Prediction Model-Confidence Pruned Markov Model on it. Confidence Pruned Markov Model is used to predict the next web page according to discovered pattern in previous step.

Step 6 :This will result into the predicted web page according to user's request based on higher probability of web page.

VI. IMPLEMENTATION TOOL

WampServer is userd to run Php Program. Here Google Chrome browser is used.Wampserver is a local server package for Windows. WampServer is open source, free to use under the GPL license agreement, relatively simple package that automatically installs everything.

VII. RESULT

Here Figure 3 shows the comparison between original and Preprocessed web log data. Original web log have 963 entries while after Data preprocessing no of logs are reduced to 235.

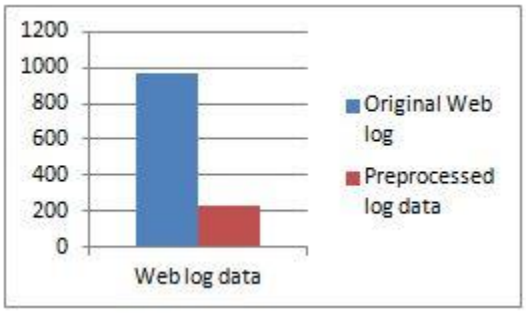

Figure 3 Comparision between original and preprocessed web log

The figure 4 shows the frequency for pattern for every session.where session is described by session ID.

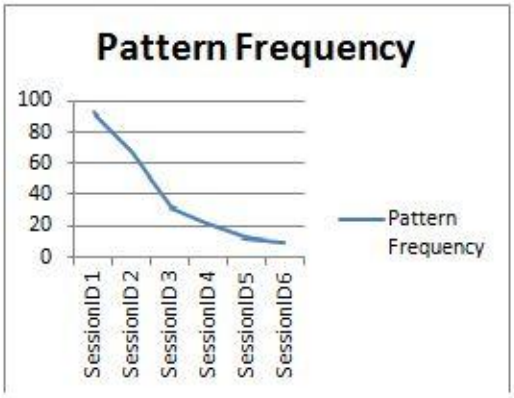

Figure 4 Pattern frequency for every session

Here this graph shows that the requests number for different country is different show the pattern which is frequent for one country can be less frequent for another country for the same request of page from user. So on the bases of the country we can predict more appropriate web page for users.

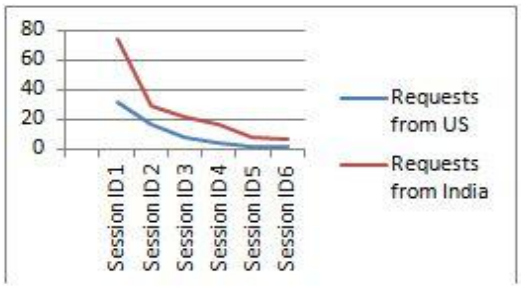

Figure 5 Session clusters and request for each session

VIII. CONCLUSION

When the data is huge in amount, we need more analysis of those data for making them more useful for users. WUM is used to understand the relationship of user and access pattern for the particular sessions. Data preprocessing will reduce the redundant data and association rule mining is used for finding frequent item set. Here two different algorithms Apriori and Eclat are compared and Eclat is better than Apriori is concluded. By Then Path traversal algorithm is useful to find frequent path by using Path Traversal graph Algorithm.Confidence Pruned Markov Model is used to evaluate the recommendation generated by using the discovered navigation pattern. This proposed approach will help to improve web server performance and to decrease the page access time.

We have used this hybrid way for small size website of less than 10 pages.The result shows that,this way is suitable for small size website.For large or medium size web site when the data is huge.For future work we can apply this to the medium and large size web site for their better performance in next page access prediction by using the users navigation pattern.

REFERENCES:

[1] Yadav, Monika, and Mr Pradeep Mittal. "Web Mining: An Introduction."International Journal of Advanced Research in Computer Science and Software Engineering 3.3 (2013)

[2] Wang, Yao-Te, and Anthony JT Lee. "Mining Web navigation patterns with a path traversal graph." Expert Systems with Applications 38.6 (2011): 7112-7122.

[3] Kewen, L. (2012, May). Analysis of preprocessing methods for web usage data. In Measurement, Information and Control (MIC), 2012 International Conference on (Vol. 1, pp. 383-386). IEEE.

[4] Nirali Madhak, Shahida Chauhan and Chintan Varnagar. "Understanding the scope of Web mining – Comprehensive study" National Conference on Emerging trends in Computer and Electrical Engineering (2014) : 51-56

[5] Panchal, Priyanka S., and Urmi D. Agravat. "Hybrid technique for user's web page access prediction based on Markov model." 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE, 2013

[6] Deshpande, Mukund, and George Karypis. "Selective Markov models for predicting Web page accesses." *ACM Transactions on Internet Technology (TOIT)* 4.2 (2004): 163-184.

[7] Kotiyal, B., Kumar, A., Pant, B., Goudar, R. H., Chauhan, S., & Junee,S. (2013, January). "User behavior analysis in web log through comparative study of Eclat and Apriori "In Intelligent Systems and Control (ISCO), 2013 7th International Conference on (pp. 421-426). IEEE

[8] Kumar, B. Santhosh, and K. V. Rukmani. "Implementation of web usage mining using Apriori and FP Growth algorithms." Int. J. of Advanced Networking and Applications 1.06 (2010): 400-404.

[9] Pal, Monti BabuLal, and Dinesh C. Jain. "An Approach for Web Pre-fetching to Enhance User Interaction of Web Application Using Markov Model."Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on. IEEE, 2014.

[10]"Apache Log Files" http://httpd.apache.org/docs/current/logs.html