Certain Issues in Web Page Prediction, Classification and Clustering in Data Mining

Srikala S[1], Geetha P[2], Sampath P[3]

[1]M.Phil Scholar, Kaamadhenu Arts and Science College, Sathyamangalam, India
[2]Asst. Professor, Department of Computer Science, Kaamadhenu Arts and Science College, Sathyamangalam, India
[3]Professor and Head, Department of CSE, Sasurie College of Engineering, Vijayamangalam. India

Abstract— *Nowadays, data mining which is a part of web mining plays a vital role in various applications such as search engines, health care centers for extracting the individual patient details among huge database, analyzing disease based on basic criteria, education system for analyzing their performance level with other system, social networking, E-Commerce and knowledge management etc., which extract the information based on the user query. The issues are time taken to mine the target content or webpage from the search engines, space complexity and predicting the frequent webpage for the next user based on users' behaviour.*

Keywords—Data Mining, Association Rule Mining, Classification, Clustering, Webpage prediction.

I. INTRODUCTION

Web mining concentrates on the analysis of World Wide Web (WWW). The overall process of discovering previously unknown, potentially valuable information or knowledge from web data is known as Web Mining. The sub areas of web mining are web content mining, web structure mining and web usage mining. Users usually interested in surfing content in the webpage. But they don't get the proper content and also they cross most of the unwanted links to reach the target link. By clicking unknowingly, those links are also stored in the weblog. Since analyzing the weblog for finding frequent webpage takes a long time to find the webpage based on users' behaviour and also would not produce an accurate prediction. There are several techniques such as support, weighted support to find the users search page within short span of time, Markov model, Weighted Markov model, association rule mining, and association rule mining with statistical features to predict the frequent webpage for the next user instead of crossing more unwanted links to reach the target link to solve problem of space complexity.

II. WEBPAGE PREDICTION

Web prediction is a classification problem in which one have to predict the next set of web pages that a user may visit based on the knowledge of the previously visited pages. Predicting users' behavior can be applied effectively in various critical applications in the internet environment. Such application has traditional tradeoffs between modeling complexity and prediction accuracy. The web usage mining techniques are used to analyze the web usage patterns for a web site. The user access log is used to fetch the user access patterns. The access patterns are used in the prediction process. Markov model and all Kth Markov model are used in web prediction. Modified Markov model enables to reduce the number of paths to reach the users' target page. This approach improves the prediction time without compromising prediction accuracy.

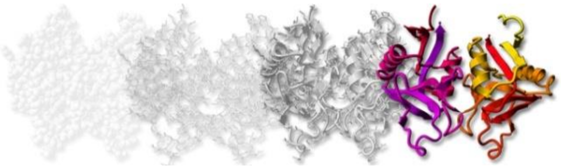

Fig.1.1: Web page prediction from different links

Prediction should be classified into two types as path-based prediction and point-based prediction. Path-based prediction is based on user's previous and historic path data, while point-based prediction is based on currently observed actions. Accuracy of point-based models is low due to the relatively small amount of information that could be extracted from each session to build the prediction model. There are different prediction models to improve accuracy namely ARM, Markov model, Modified Markov Model and Association Rule Mining with statistical features to reduce number of path level to reach target page with time and space consideration.

Web Usage Mining is concerned with the discovery of a customers access patterns when browsing the Internet.

- Clustering or Classification identifies profiles with similar characteristics.
- Association predicts correlations of items, where one set of transactions implies the existence of another.
- Sequence patterns, involves discovering patterns that indicate usage over a time period.

Frequent pattern mining algorithms are designed to find commonly occurring sets in databases. Memory and run time requirements are very high in frequent pattern mining algorithms. Systolic tree structure is a reconfigurable architecture used for frequent pattern mining operations. High throughput and faster execution are the highlights of the systolic tree based reconfigurable architecture. The systolic tree mechanism is used in the frequent pattern extraction process for the web access logs. Systolic tree based rule mining scheme is enhanced for weighted rule mining process. Automatic weight estimation scheme is used in the system. The dynamic web page weight assignment scheme uses the page request count and span time values.

The FP-growth algorithm was originally designed from a software developer's perspective and uses recursion to traverse the tree and mine patterns. It is cumbersome to implement recursive processing directly in hardware, as dynamic memory allocation typically requires some software management. The systolic tree is configured to store the support counts of the candidate patterns in a pipelined fashion as the database is scanned in, and is controlled by a simple software module that reads the support counts and makes pruning decisions.

This approach helps to reduce accessing time of target users search and also improves the memory utilization. In web prediction, the challenges faced are in both preprocessing and prediction. Preprocessing challenges include handling large amount of data that cannot fit in the computer memory, choosing optimum sliding window size, identifying sessions, and seeking/ extracting domain knowledge. Prediction challenges include long training/ prediction time, low prediction accuracy, and memory limitation.

The World Wide Web (WWW) has become a ubiquitous tool, used in day to day work, to find information and conduct business, and it is growing at an exponential rate. The activity of searching of information consists of the cycle:

i) Submitting a query to a search engine
ii) Selecting a page for browsing from the returned list of pages
iii) Navigating, i.e. link following

One can concentrate on the navigation process where the user forms trails of pages through the database graph, starting from a page chosen from the search engine's result list.

The conventional association rule problem by permitting the association of a weight with each item in a transaction that reflects the interest/intensity of the item within the transaction. Consecutively, this offers the prospect to associate weight parameters with each item of the resulting association rule. Instead of binary weights, they have assigned significant weights based on the frequency of visit and time spent by the user on each page with regards to the degree of interest.

An algorithm, Adaptive Weighted Frequent Pattern Mining is by introducing a concept of adaptive weight for each item which is capable of handling the situation where the weight (price or significance) of an item changes with time. Extensive performance analysis has demonstrated that the algorithm using adaptive weights was very efficient and scalable for WFP mining.

Link Analysis Algorithms make use of the structure present in '*hyperlinks*', sorted and displayed depending on a '*popularity index*' decided to pages linking to it and analyzed the mathematics behind these '*link analysis algorithms*' and their effective use in ecommerce applications where they could be used for displaying personalized information.

During the navigation process users often tend to "get lost in hyperspace" meaning that when following links users tend to become disoriented in terms of the goals of their original query and the relevance to the query of the information they are currently browsing; the user could refer to this problem as the navigation problem.

Accessing target webpage in the internet for the users is difficult. Web mining techniques are used to analyze the web information resources. Web content mining and structure mining methods are used to analyze the web page contents. The user access information is maintained under the web logs. The usage mining technique is used to analyze the web logs. The web logs are maintained under the web server environment. User access patterns are extracted from the web logs. The web content management and link connectivity are improved using the access patterns. The association rule mining techniques are used to extract item set relationships. Item set frequencies are used in the rule mining process. Page weights are used to denote the importance of the web pages. The weighted rule mining techniques are used to fetch the frequently accessed web pages with its weight values. An efficient weighted association rule mining approach is devised for web page recommendation. Here, the duration of a web page view and the recent access are used to discover significant page sets. For every association rule mined, time span and recent access weights are computed and combined weighted support is calculated. From the support value a webpage relevant to the user needs can be recommended. The web usage mining techniques are used to analyze the web usage patterns for a web site. The access patterns are used in the prediction process. Web prediction is a classification model attempts to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages.

Predicting user's behavior can be applied effectively in various critical applications in the internet environment. Some of the issues in user search as follows:

i) Identifying target webpage is difficult
ii) Accessing time will get increased
iii) Unwanted links occupied in weblog
iv) Frequent webpage analysis is tough
v) Setting static weighted value for newly entered link in weblog is difficult
vi) Prediction Accuracy is low
vii) Makes searching environment a tough place for searching target content

Thus the different techniques such as clustering, classification, prediction models Markov model & modified Markov model and association rule mining to predict frequently accessed page through the log file from the server. By getting the log files which contain Uniform Resource Locator (URLs) the link visited by the previous users is obtained. To find out the page prediction based on their similarity pattern can be measured with the help of links accessed by the previous users through the log files with the consideration of space and time measure.

III. CLASSIFICATION TECHNIQUES

There are several aspects to improve the prediction accuracy in accordance with different classification techniques to reduce the number of path level to improve the prediction time and also memory utilization such as

- Markov Model (MM)
- Support Vector Machine (SVM)

The Markov model approach is a powerful technique for predicting seen data; however, it cannot predict unseen data. On other hand, SVM is a powerful technique, which can predict not only for seen data but also for unseen data. One of the technique is training the SVM where the predictive power may decrease because such examples confuse the training process. To overcome these drawbacks with SVM, domain knowledge is extracted from training set and incorporated this knowledge in training set, to improve prediction accuracy of SVM by reducing number of classifier during prediction. This requires continuous attributes because they use a metric measure in their computations (dot product in case of SVM). In this implementation, bit vectors are used to represent the page IDs. The index of the page ID in vector and its numerical value is only kept, if that value is not zero. Missed attributes are assumed to have zero values. To extract more knowledge from the user sessions, the frequency matrix is used in below Figure 1.2.

The first row and column represent the enumeration of web page IDs. Each cell in matrix represents the number of times (frequency) users have visited two pages in a sequence. Freq (x, y) is the number of times user have visited page y after page x. For example, cell (1, 2) contains the number of times users have visited page 2 after Page 1.

	1	2	…	N
1	0	Freq(1,2)	..	Freq(1,N)
2	Freq(2,1)	0	..	Freq(2,N)
…	Freq(…1)	Freq(…..,2)	..	Freq(…..,N)
N	0	Freq(N,2)	..	0

Fig1.2: Frequency Matrix Table

Below example can explain the way to find out the frequent page access. Suppose for a user session A=<1,2,3,4,5,6,7> is the sequence of pages a user have visited. Suppose that a sliding window of size 5 is used. Feature extraction can be applied to A=<1,2,3,4,5,6,7> and end with the following user sessions of 5 page length: B=<1,2,3,4,5> , C=<2,3,4,5,6> and D=<3,4,5,6,7>. Note that the outcome or label of the sessions A,B,C and D are 7,5,6 and 7, respectively. This way the following four user sessions: A, B, C, and D are obtained. In general, the total number of extracted sessions using a sliding window of size 'w' and original session of size A is |A|-w+1.

Hybrid model as Markov model and support vector machine, then the output of SVM, ANN and Markov Model operates independently. Furthermore, for any session X, for which it does not appear in training set, the Markov prediction is assumed as zero. If Dempster's rule is used to combine SVM and Markov model as the body of evidence, the following equation is obtained:

$$m_{svm,markov}(c)=\frac{\sum_{A,B\subseteq\theta,A\bigcap B=C} m_{svm(A)}m_{markov(B)}}{\sum_{A,B\subseteq\theta,A\bigcap B\neq\theta} m_{svm(A)}m_{markov(B)}} \qquad (1.1)$$

In the case of WWW prediction, this formulation can be simplified because one can have only beliefs for singleton classes (i.e., the final prediction is only one web page, and it should not have more than one page) and the body of evidence itself (m(Ѳ)). This means that for any proper subset A of Ѳ for which there is no specific belief, m(A)=0.

Algorithm: Webpage prediction using hybrid model (SVM and Markov model)

Input: S User session Data

Output: Yi: Next page prediction for testing session i.

Begin

1. S Apply-Feature-Extraction(s)
2. svm-models Train svm(S)// train SVM using one *vs* all.
3. Svm-prob-model Map-SVM-model (svm-models)//map SVM output to a probability

```
4. Svm-uncertainty ComputeUncertainty(SVM)//See Eq (2)
5. Construct Markov model
6. Markov uncertainty compute Uncertainty (Markov)//See Eq (3)
7. For each testing session X in S, do
    7.1 Compute and output SVM probabilities for different pages
    7.2 Compute and output Markov probability for different pages.
    7.3 Compute using Eq.(4) and the final prediction YX
End
```

The m$_{svm}$(ϴ) and m$_{markov}$(ϴ) in Eq.(2) and Eq.(3) is computed as follows. For SVM, the margin values for SVM is used to compute the uncertainty as below

$$m_{svm}(\theta) = \frac{1}{\ln(e + svm_{margin})} \quad (1.2)$$

$$m_{markov}(\theta) = \frac{1}{\ln(e + Markov_{probability})} \quad (1.3)$$

Thus the Hybrid model is one of the approaches to improve the prediction accuracy with respect to time and space complexity.

IV. CONCLUSION

A user's targeted webpage recommendation approach must meet several aspects such as

- Reduce unwanted links
- Efficient access time (users meet target webpage directly)
- Memory utilization of weblogs

In many of the proposed schemes for users' targeted webpage directly as presented in the previous chapter, there exist one or the other disadvantage which makes the performance of the scheme reduced.

One of the important aspects is the memory utilization and efficient accessing time of users' targeted webpage. The process involved is to analyze the frequent pattern (webpage) from the weblog and then providing access to requested query of user. By applying different methods presented in the previous chapter, the access time and memory utilization of weblogs may get vary for different approach.

REFERENCES

[1] Agrawal, R & Srikant, R 1995, 'Mining Sequential Patterns', Proc. of the 11th Int'l Conference on Data Engineering.

[2] Mamoun Awad, A & Issa Khalil, 2012, 'Prediction of User's Web-Browsing Behavior: Application of Markov Model', IEEE Transactions on Systems, Man and Cybernetics-Part b:Cybernetics, vol.42, no.4.

[3] Sampath P, Amitabh Wahi & Ramya D, 'A Comparative Analysis of Markov Model With Clustering and Association Rule Mining for Better Web Page Prediction', Journal of Theoretical and Applied Information Technology, vol. 63, no.3 2014.

[4] Sampath P, Amitabh Wahi & Ramya D, 'Analysis of Web Page Prediction by Markov Model and Modified Markov Model With Association Rule Mining' International Journal of Computer Science and Technology, vol. 04, no. 1 2013.

[5] Jaiswal, M., & Patel, D. (2015). Data Mining Techniques and Knowledge Discovery Database. International Journal Of Research And Analytical Reviews, 2(1), 248-259

[6] Sampath P, Amitabh Wahi & Ramya D, 'A Comparative Analysis of Markov Model with Association Rule Mining-Statistical Features for Better Web Page Prediction' International Journal of Mathematical Sciences and Engineering (IJMSE), vol. 02, no.I, 2013.