Web Usage Pattern Detection Using Cohesive Markov Model With Apriori Algorithm

Vinukumar Luckose
Center for Advanced Electrical & Electronic System, FOEBEIT, SEGi University
Petaling Jaya, Malaysia
vinukumarluckose@segi.edu.my

Jothish Chembath
Department of CSE
Koneru Lakshmaiah Education Foundation
Vijayawada, India
jothishchembath12@gmail.com

Joe Arun Raja Ponnusamy
University Institute of Computing
Chandigarh University
Mohali, India
joearunraja@gmail.com

Sakshi Sharma
University Institute of Computing
Chandigarh University
Mohali, India
sakshi28sharma95@gmail.com

Pritpal Kaur
University Institute of Computing
Chandigarh University
Mohali, India
kaurpritpal94@gmail.com

Sajitha Smiley
Manonmaniam Sundaranar University
Tirunelveli, India
sajismiley@gmail.com

2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS) | 978-1-6654-9581-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CACIS54679.2022.9815465

***Abstract*— Web server maintains the essential user log files, recording every request to it. Web log is a record of events which includes all the user details from the time the web visitor initiated the session to the end of the session. The web usage pattern discovery to identify different states of the user access behavior on web. The design of web recommender system using a context-aware Cohesive Markov Model and Apriori clustering is proposed. The prediction rate of proposed algorithm is higher than conventional Markov model.**

Keywords*— *Web Usage mining, Weblogs, Pattern Discovery, Cohesive Markov Model, Apriori Algorithm.

I. INTRODUCTION

Web usage mining can discover and analyze web log data from WWW, using different methods. Web usage mining analyzes the details available in these log files. Web data mining techniques are used to discover interesting usage patterns from Web data [1]. Fig. 1 shows various components of web mining. The focus of the study is on web usage mining (WUM) and its associated tasks.

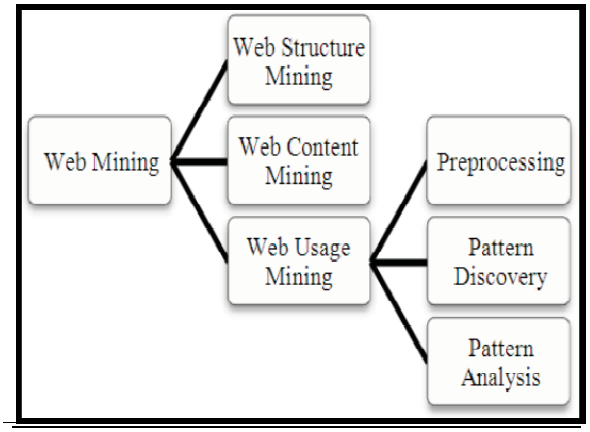

Fig 1: Components of Web Mining

A. Web Usage Mining on Web Log Files

Web data mining is applied to the World Wide Web to segregate the web data accumulated in web servers. Clustering is performed on the web log through structured mining to group similar types of data from web sessions. Web usage mining (WUM) extracts web usage pattern for finding similar user group to find applicable page group and the regularity in the use of access path[7].

II. LITERATURE REVIEW

A. Significant works

In [1], Z. Yao and X. Wang, J. Luan, (2017) found that Hidden Markov Model worked well for vertical structure website. Their technique was able to predict pages accurately by discovering hidden information. In [2], Jose Borges and Mark Levene, 2007 proposed a technique to measure the capacity of a variable-length Markov model to abbreviate user sessions' path. They were successful in proving that the accuracy of predictions increased with this ability.

In [3] Mukund Deshpande and George Karypis Markov, 2004 proposed a model-based prediction algorithm which eliminated a major portion of the states of All-Kth Order Markov model. They could show that the Markov models achieved greater precision than traditional algorithms. In [4], Xing Dongshan and Shen Junyi, 2002 discussed the serious limitations of traditional Markov model. Hybrid Markov models predicted Web access precisely, and at the same time provided high coverage and scalability.

In [5], Neetu Sahu and Pragyesh Kumar Agrawal, 2016 proposed web page prediction utilizing the weblog and web content features. Web log is another feature, which helped Markov model of third order to improve the prediction accuracy. In [6], Neetu Anand and Tapas Kumar, 2017 proposed Markov models to preprocess and analyze user Web navigation data. They indicated that Markov chain didn't work well when the input size was very large since it could be feasible only in relatively small spaces. They proposed the use of Hidden Markov Model instead.

In [7], B. Rajeswari and Dr. S. Shajun Nisha, 2018 outlined that web log file contained previous user navigation data which could very well be used to find the user access behavior. They proposed classification algorithms to predict accuracy results. They proposed a system to achieve good performance with high satisfaction using best classifiers.

In [12], Anand Charpate, Chetan Bramhankar, Prashant Gaikawad, A.D.Londhe has projected that higher order markov model is suited and found to be best for methodology to implement. Their work has mad the concept of variable length markov model when weblog are not

sufficiently created. In this case they had used the page ranking for predicting the page.

In [13], Neetu Anand, Tapas Kumar, the authors have suggested that Markov chain are used to discover web usage. This model has the ability to make predictions and can anticipate the next choice of a user.

B. *Issues and Challenges*

Web log data are diverse and huge. The data must be processed to make it have a comprehensive view, combining the varied types of data to help the agents obtain search knowledge as and when required. Information retrieval is intractable with the enormous data available in Web server. Finding out the recurrent URLs from the Web resources will be an impetus to obtain adequate results, for rapidity of Web access.

C. *Structuring web log data*

Structuring web log data for an analysis of Web log files can be of help to know the significance of Web mining. The web access history is found to contain a lot of inconsistent, incorrect and missing values incorporated in log files. Preprocessing algorithms are applied to web data to achieve efficiency in managing Web data. Clustering web pages according to the Web request pattern needs to be completed for efficient analysis, for which various clustering methods like K-means, Apriori etc. are available. Fig, 2 illustrates the process of predicting users' next web page visit and Apriori method is used for completing the process.

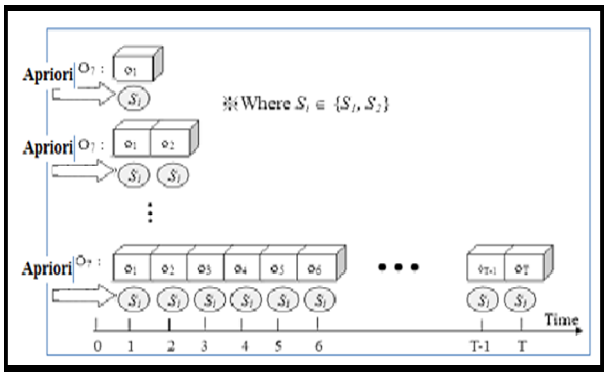

Fig 2: Choice of user preference on webpage

Extracting frequent usage patterns based on the demography aspects is difficult, since diverse Web servers have their own Web log files. A detailed analysis of all these varied log files will give significant results. Server logs can assist in recognizing the user behavior and the his choice of hyper link. Web Usage Mining (WUM) is a kind of data mining method that can be used for detecting user behavior from Web log data. This paper gives an outline of the used WUM techniques that assist in designing Web recommender systems. Finally, the researcher has proposed a solution using "Cohesive Markov Model" to address these problems.

D. *Methodology*

The system proposed aims to exhibit an ideal system for ease of user navigation. The present work proposes to use a Cohesive Markov Model that will focus on the methods that improve prediction accuracy. Fig. 3 shows the web mining process that is followed for doing a web analysis.

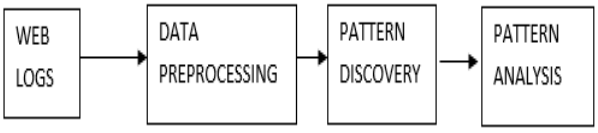

Fig 3: Web Usage Mining Process

Web usage mining applies data mining techniques on web log files to discover browser behavior knowledge. Statistics of Websites are used to improve the performance associated with improvement of web page design.

The present study is to portray the mining of weblog files in terms of behavioral patterns of web users using a Cohesive Markov Model with Apriori algorithm. This work collects the varied browsing patterns in an attempt to understand the user behavior.

III. TOOLS AND METHODS

In this study, Apriori algorithm, in combination with Cohesive Markov model, is used for clustering and prediction. Users' past browsing experience is a major source of information that assists prediction. Apriori clustering algorithm is proposed to act upon the web log information and group users according to their web page visits. This feature is used to predict the hyper link selection of web user from the appropriate cluster. Prediction can be envisioned by using Markov model in combination with the probability conditions. The probability of the web users choice of hyper link can be judged by carefully introspecting the clusters with the aid of Markov model.

A. *Prediction Accuracy*

Improvement of the prediction accuracy of statistical models [8,9,10,11] relies on the proper design of the model with appropriate training with suitable data. In this context, prediction accuracy can be achieved by combining Cohesive Markov model and Apriori clustering techniques. Clusters of webpage visits are formed according to similarities to improve the prediction process. The process is given below:

B. *The Process Steps*

- Group Web server log files using the Apriori algorithm.
- Group the Web usage sessions into clusters.
- Analyze the data sets using Markov model analysis.
- From the data set select the cluster to which each item will belong.
- Assess accuracy of Cohesive Markov model using the cluster data so formed as data to train the algorithm
- Compare the full data set with the Cohesive Markov model accuracy of the clusters.

IV. CONTEXT-AWARENESS OF COHESIVE MARKOV MODEL

Different Markov models are combined to make the resulting model efficient and less complex. This model will improve prediction accuracy and also preserve all the basic properties of All-Kth-Order Markov. Careful attention is given to eliminate the complexities involved with different

978-1-6654-9581-3/22/$31.00 ©2022 IEEE

order Markov models and simultaneously improve the performance. Our research on data sets indicates that the proposed scheme has improved prediction accuracy. The preprocessing of raw log data was carried out to create individual user sessions by using Apriori clustering technique. These sessions were used as training data. Finally, the Cohesive Markov Model (CMM) could discover the users' behavior and predict.

Cohesive Markov model is based on the assumption that the prospect of visiting a web page pi does not depend on all the web pages in total, but it takes a small set of preceding pages. Let P = {P1, P2... Pn} be the set of pages in the web site. Let W be a user session having a sequence of pages visited by a user. Let L be the number of pages in W. Page PL+1 the user will visit subsequently can be estimated as

$$p_{L+1}=\text{argmax}_{p\in P}\{P(P_{L+1}=p/p_L, p_{L-1},..,p_{L-(k-1)}\}$$

Where k is the number of former pages which recognizes the Markov models order. The prediction is performed using the probability formula as given in Equation

$$P(p_i / S_j^k) = \frac{\text{Frequency}(<S_j^k, p_i>)}{\text{Frequency}(S_j^k)}$$

All pages that satisfied by the probability condition qualify as pages which user might visit.After processing the data, construction of a suitable model is taken care of. The precision of the model depends on the transition diagrams and its parameters. Refer to Fig. 4 for the parameters of the CMM model. Apriori algorithm is used for clustering the web data to help the whole system during prediction.

Apriori uses a "bottom up" approach, where subsets that are frequent are stretched. This generated data are tested against the whole data set. The algorithm stops when no further successful stretch is possible. When a user navigates through the web, every movement made is recorded in the server. CMM can closely predict the next page visit by remembering the preferences of the user.

V. EXPERIMENTAL EVALUATION

The first step is to gather web log files from Web servers and classify them into groups of clusters. Prediction of user choice of next web page is completed by using Apriori algorithm with the Cohesive Markov model. For evaluating the performance of the conventional Markov model algorithm and Cohesive Markov model algorithm, web log dataset from two popular web sites were used (refer to Table 1).

TABLE I. DATA SET

Dataset	Meta Data			
	Code	*Period*	*Size (MB)*	*No. of Records*
NASA http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html	NASA	July-95 to August-95	205.2	34,61,612
Saskatchewan's http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html	SASK	June-95 to December-95	233.4	24,08,625

The preprocessing algorithm's performance was evaluated with the percentage of reduction acquired on the number of transactions and the amount of memory utilized.

The coverage level is less and sometimes it compromises on prediction. Higher order models had shortcomings like higher complexity; reduced coverage and sometimes bad prediction accuracy (refer to Table 2).

TABLE II. COVERAGE

Prediction Model	Dataset	
	NASA	*SASK*
LPA-EPFK means	1.6325	1.6518
Conventional Markov model with Apriori	1.4899	1.5048
Cohesive Markov Model (CMM) with Apriori	1.2184	1.1965

F1 measure (refer to Table 3) was used to measure the performance. The F1-measure is a combination of Precision P and Recall R. When precision dealt with insertion errors and substitution, Recall deals with deletion errors and substitution.

TABLE III. F1 SCORE

Prediction Model	Dataset	
	NASA	*SASK*
Random Forest	0.9013	0.9121
Conventional Markov model with Apriori	0.8241	0.8340
Cohesive Markov Model (CMM) with Apriori	0.9205	0.9127

From the results observed, the proposed CMM (Cohesive Markov Model) performs web page prediction adequately with the two datasets taken. The proposed CMM algorithm with Apriori showed an improved F1 measure over conventional Markov model with Apriori, signifying that the improvement operations used were successful in improving prediction accuracy. We can conclude from the entropy that CMM with Apriori has improved the prediction performance.

While web users are extracting information from web logs through Markov model, it will facilitate the users to see various further web sites and all related information available, the web user is searching as well as it will help to find out the expected web page to the web users quickly. Therefore, this study should expand in the near future for conducting applied research related to this the Markov model for minimizing the web browsing time and accuracy. There is a wide scope and opportunity awaiting the web mining researcher to enhance this model. Web developers and web sites can establish this Markov model in the field of data

mining for reducing the present complication and challenges, with the help of experimental research study.

VI. CONCLUSION

In this study, an analysis of navigation patterns for prediction system is analyzed. The system involves four stages. Data collection is initially completed by collecting log entries from web server. The next stage is preprocessing the data collected to remove duplicate entries. The penultimate stage is the clustering of data which group similar data. During the last stage, these clustered data are used by the proposed CMM for prediction. The researcher has observed that this model can improve the overall accuracy of prediction.

REFERENCES

[1] Z. Yao, X. Wang, J. Luan, Using Hidden Markov Model to Predict the Web Users' Linkage, Journal of Residuals Science & Technology (JRST), Volume 14 No. 3, 2017

[2] Borges and M. Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. EEE Transactions on Knowledge and Data Engineering (Volume: 19, Issue: 4, April 2007)

[3] M. Deshpande and G. Karypis. Selective markov models for predicting web page accesses. ACM Transactions on Internet Technology (TOIT) Volume 4 Issue 2, May 2004

[4] X. Dongshan and S. Junyi. A new markov model for web access prediction Computing in Science & Engineering (Volume: 4, Issue: 6, Nov/Dec 2002)

[5] Neetu Sahu, Pragyesh Kumar Agrawal, Markov Model Based Web Page Recommendations by Combining Content and Log Features, International Journal of Computing and Technology, Volume 3, Issue 11, November 2016

[6] Neetu Anand, Tapas Kumar, Prediction of user interest and behavior using markov model, International journal of scientific research in computer science and engineering, Vol.5, Issue 3, 2017

[7] B. Rajeswari, Dr. S. Shajun Nisha, Web page prediction using web mining, International Research Journal of Engineering and Technology, Volume: 05 Issue: 05, May-2018

[8] Babu, I., Balan, R. S., & Mathai, P. P. (2019). Machine Learning approaches used for prediction in diverse fields. International Journal of Recent Technology and Engineering (IJRTE), 8(2S4), 762-768.

[9] Ponraj, T. E., Balan, R. V., & Vignesh, K. (2021). Deterministic functions for measuring human protein structural variations with merit based ensemble learning scheme for native classification. International Journal of System Assurance Engineering and Management, 1-13.

[10] Ponraj, T. E., Balan, R. V., & Vignesh, K. (2021). Analysis and Prediction of Adverse Reaction of Drugs with Machine Learning Models for Tracking the Severity. Arabian Journal for Science and Engineering, 1-9.

[11] Pooja, S. B., & Balan, S. R. (2019). An Investigation Study on Clustering and Classification Techniques for Weather Forecasting. Journal of Computational and Theoretical Nanoscience, 16(2), 417-421.

[12] Anand Charpate, Chetan Bramhankar, Prashant Gaikawad, A.D.Londhe(2015). Prediction of Link and Path for User's Web Browsing Using Markov Model, IJCSMC, Vol. 4, Issue. 2, February 2015, pg.144 – 148

[13] Neetu Anand, Tapas Kumar (2017), Prediction of User Interest and Behaviour using Markov Model, Computer Science and Engineering Vol.5, Issue.3, pp.119-123, June (2017)

978-1-6654-9581-3/22/$31.00 ©2022 IEEE