ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Analyzing Prediction Accuracy for Small Scale Log Files Using Web Usage Mining Approaches

C.Thavamani [1], Dr.A.Rengarajan[2]

Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India[1]

Professor, Dept. of Computer Science Engineering, Veltech Multi Tech SRS Engineering, Chennai, Tamilnadu, India [2]

ABSTRACT: This paper emphasizes the user future request prediction using small number of previous log files without affecting the prediction accuracy level. 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 K^{th} Markov model are used in Web prediction. The framework can improve the prediction time without compromising prediction accuracy. The proposed system is to compare the prediction accuracy with the markov model, ARM, ARM-SF. The system improves the accuracy with scalability considerations. Finally the result shows which would have better prediction accuracy.

Keywords: Future Request Prediction, log files with small size, Association rule mining (ARM), Association rule mining with statistical features (ARM-SF), Markov model.

I. INTRODUCTION

Web usage mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The primary data sources used in Web usage mining are the server log files, which include Web server access logs and application server logs. Additional data sources that are also essential for both data preparation and pattern discovery include the site files and meta-data, operational databases, application templates, and domain knowledge.

WUM comprises basically three major processes namely data pre-treatment, data mining and pattern analysis. Firstly, Pre-treatment of data is done on a series on Web logs to obtain logs with minimized redundancies, user, session, transaction identification and information on path completion. Secondly, mining algorithms are applies to extract user navigation patterns which represents relationship among Web pages in a particular Web site. Lastly, pattern analyzing algorithm is applied to extract data for data mining applications.WUM automatically discovers knowledge from the data collected in Web logs.

Additionally, some models, such as association rule mining (ARM) and SVM, do not scale well with large data sets. Furthermore, some models, such as SVM and ANN, do not handle the multiclass problem efficiently because of the large number of labels/classes involved in the WPP. In this paper, we are analyzing the prediction accuracy range comparable to markov model and modified markov model with the association rule mining.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

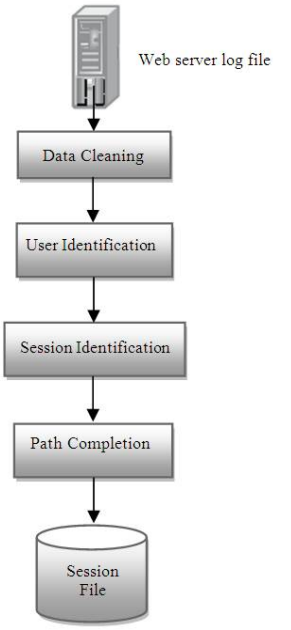


Figure 1 – Web Server Log File Pre-processing

A proposed algorithm automatically discovers pages in a website for which the location is different from where the visitors expect to find them. For this purpose "Backtrack" as a key is used for the algorithm from the point where the user will backtrack. The point where the user starts to backtrack is considered as the expected location for the page[3]. One more algorithm is devised that selects the set of navigational links to optimize the prediction time and accuracy as in [4], works on a "Real Time Management Engine" that uses historical data and on line visitation pattern of e-commerce site visitors. We present an analysis study for all K^{th} Markov Model, ARM, ARM-SF in the WPP using different N-grams and also how accuracy is increased when using even small N-grams[2].

II. LITERATURE STUDY

Researchers have used various prediction models including k-nearest neighbour (kNN), ANNs [5],fuzzy inference [6] SVMs, Bayesian model, Markov model and others. Mobasher et al. use the ARM technique in WPP and propose the frequent item set graph to match an active user session with frequent item sets and predict the next page that user is likely to visit. However, ARM suffers from well-known limitations including scalability and efficiency.

Millions of users access different Websites all around the world. When they access the network, a large amount of data is generated and is stored in Web log files which can be used efficiently as many times as user repeatedly searched the same type of Web pages recorded in the log files. These series can be considered as a web access pattern, helpful to find the user behavior. Through this personalized information, it's quite easy to predict the next set of pages user might visit based on the previously searched patterns, thereby reducing the browsing time of user. This survey too focuses on how to improve the prediction time without compromising prediction accuracy.

Avneet Saluja et al analyzed with different log files that are collected from the web servers and also tested with markov model, ARM, ARM-SF. Though prediction accuracy rate increased to a higher rate for significant N-grams log files, the *a*pproach is not feasible for the cases when the past log files are less i.e.; 'n' is small. The research especially for a few previous log files without affecting the accuracy and an in-depth analysis can be done on other features of session's log files to further improve the prediction accuracy level.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

III. WEB PAGE PREDICTION TECHNIQUES

a. Web User Behaviors Prediction System Using Trend Similarity

A trend based application system is used to analyze user's behavior and predict the future path of user based on trend similarity. It is not viable to predict the browsing behavior of current user according to the similar past behavior of other user's. Hence, a trend based prediction model is proposed to predict the future travelling path by generating ordered browsing sequence.

The system proposed which works in two phases. One is the Construction phase and another is the Prediction phase. The construction phase helps to discover useful common browsing patterns for the experts and then they use it to predict the further browsing sequences. In predicting phase, the browsing behavior of a new user is fed into the system to be compared within the prediction system so as to generate pages that can be pre-fetched to improve the browsing performance.

b. *A new classification model for online predicting user's future movement*

Web Usage Mining is usually implemented for two components on line and off line. The off line structure extracts knowledge from the historical log files and then this knowledge is used by the online component.

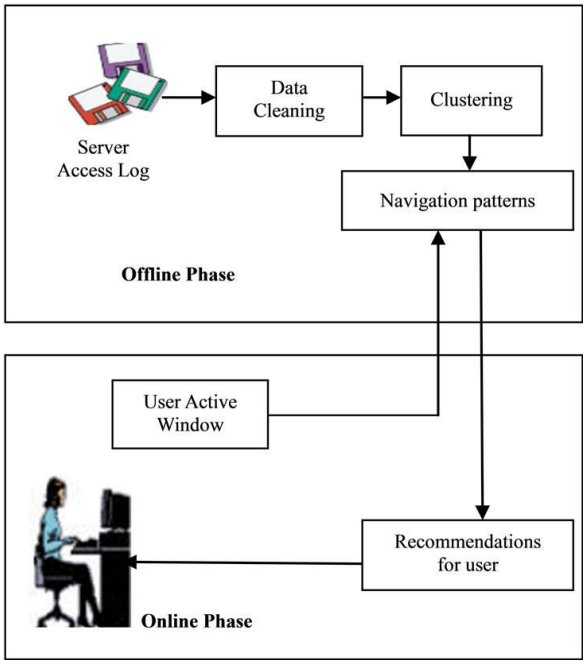


Figure 2 – On line/Off line Web page prediction

In the architecture Figure 2 classification is done using Longest Common Subsequence (LCS) algorithm. The architecture is partitioned into on line and off line phases which work simultaneously. In off line phase data pre-treatment module process the web logs and reformat it to identify all web access session. The navigational pattern mining module cluster the group sessions according to certain common properties. In on line phase, according to the URL requested and session identifier the user belongs to that session, the underlying knowledge base is updated and the list of suggestion is appended in the prediction list which finds the cluster based on LCS algorithm .Pre-treatment makes the data refined so that redundancy and anomalies can be removed in the earlier stage itself [5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

c. ***A Web-based recommendation system to predict user future movements***

Mehrdad Jalali,Narwati Mustapha, Md.Nasir Sulaiman,Ali Mamat advanced their previous work and renamed there architecture as WebPUM. In this they proposed a novel formula for assigning weights to edges of undirected graphs to classify current user activity. They used LCS algorithm to predict user near future movement and conducted two main experiments for navigation pattern mining and prediction of user's next request. In addition they found clustering patterns for user navigational behavior and quality of the used datasets CTI and MSNBC improved [8].

A Web usage mining architecture called WebPUM and proposed a novel approach to classify user navigation pattern for on line prediction of user future intentions through mining Web server logs.

d. ***A New Approach for Next Page Access Prediction***

Integration of Markov model based on sequential pattern mining with clustering showed that prediction accuracy is increased by 12% as compared to traditional Markov model. Clustering was used to identify similar access pattern from web logs using pair-wise nearest neighbour and then sequential pattern mining is performed on these identifies patterns to determine next page possible accesses. The compactness of clusters is improved by setting similarity threshold while forming clusters. When in future mining is done on these patterns, prediction accuracy will be improved as compared to the accuracy when mining is done on dissimilar access patterns. Therefore, a sequential mining technique called "Markov model" is suggested in combination with pattern discovery [10].

Markov Model provides good prediction accuracy if it is used in accordance with sequential mining [10] .Method does not consider loosely connected access sequences for mining process which can be considered as a limitation of this approach. Low order Markov Models have good accuracy however, they lack accuracy due to poor history or past web logs.

e. Prediction of User's Search Behavior -Application of Markov Model

Another variant of Markov Model for prediction of user's web browsing behavior. They focused on a new modified Markov model to alleviate the scalability issue in the number of paths. In addition to this, a new two-tier prediction framework that creates an Example Classifier (EC), based on the training examples and the generated classifiers is developed. Experiments showed that such framework can improve the prediction time without compromising the accuracy. According to their work, the next action corresponds to predicting next set of pages to be visited and the previous actions corresponds to the pages that have already been visited.

In Web prediction, the K^{th} order Markov model gives the probability that a user will visit the k^{th} page provided that the user has visited the ordered k-1 pages. For example, in second-order Markov model, prediction of the next Web pages is computed only on the basis of the two web pages previously visited.

Table I presents the All K^{th} order Markov Model which gives the prediction steps to compute the next set of pages. The function(x,m_k) is considered to predict the next page visited of session x using the k^{th} order Markov model. If m_k fails, then m_k-1 is considered using a new session x' of length k-1.This process repeats until the first-order Markov model is reached assuring that all orders fails to predict and no prediction is possible [11].

ALL K^{TH} ORDER MARKOV MODEL

Algorithm: All K^{th} Prediction
Input: user session x, of length K
Output: Next page to be visited, p

1. p ←predict(x,m_k)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

2. If p is not 0 then return p
3. x ← remove page ID from x
4. K ← K-1

5. if (K=0)return 'failure'
6. Go to step 1

7. Stop

f. ***Web page prediction using ARM***

ARM is a data mining technique that has been applied successfully to discover related transactions. Specifically, ARM focuses on associations among frequent item sets. For example, in a supermarket store, ARM helps uncover items purchased together which can be utilized for shelving and ordering processes [2]. In the following, we briefly present how we apply ARM in WPP.

In WPP, prediction is conducted according to the association rules that satisfy certain support and confidence as follows. For each rule, $R = X \rightarrow Y$,of the implication, X is the user session and Y denotes the target destination page. Prediction is resolved as follows:

Note that the cardinality of Y can be greater than one, i.e., prediction can resolve to more than one page. Moreover, setting the minimum support plays an important role in deciding a prediction. In order to mitigate the problem of no support for $X \cup Y$, we can compute prediction ($X' \rightarrow Y$), where X' is the item set of the original session after trimming the first page in the session. This process is very similar to the all-K^{th} Markov model. However, unlike in the all K^{th} Markov model, in ARM, we do not generate several models for each separate N-gram. In the following sections, we will refer to this process as all K^{th} ARM model.

g. ***Web page prediction using ARM-SF***

To finding out the better prediction of web page by including association rule mining with statistical features. One limitation of the standard approach to discovering associations is that by searching massive numbers of possible associations to look for collections of items that appear to be associated, there is a large risk of finding many spurious associations.

These are collections of items that co-occur with unexpected frequency in the data, but only do so by chance. Statistically sound association discovery controls this risk, in most cases reducing the risk of finding any spurious associations to a user-specified significance level.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

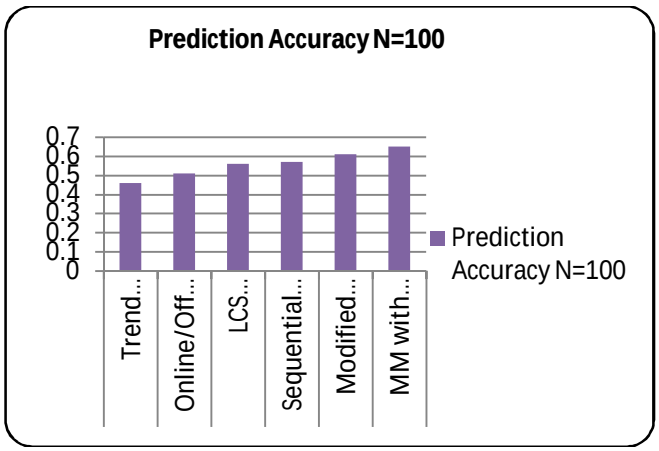
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

IV. SIMULATION AND RESULTS

S.No	Method	Prediction Accuracy for N			
		400	300	200	100
1	Trend based Prediction	0.52	0.5	0.48	0.46
2	Online based Prediction	0.57	0.55	0.53	0.51
3	LCS algorithm, Clustering	0.62	0.6	0.58	0.56
4	K^{th} order Markov Model Clustering	0.63	0.61	0.59	0.57
5	Modified Markov Model with ARM	0.72	0.67	0.64	0.61
6	MM with ARM, ARM- SF	0.8	0.75	0.7	0.65



V. CONCLUSION AND FUTURE WORK

The system is tested with specifically markov model,ARM,ARM-SF and other traditional techniques. The prediction accuracy is used as the performance metric to evaluate the quality of the system. The system is designed to successfully improve prediction accuracy using simpler probabilistic models such as Markov model,ARM, ARM-SF. Prediction accuracy rate increased to a higher rate even for the small scale log files such as N- 100.In future we extend our work with the comparison of markov model, ARM,ARM-SF and boosting and bagging model for finding better prediction accuracy for lower size web log files.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

REFERENCES

1. Avneet Saluja, Dr. bhupesh Gour, Lokesh Singh, "Web Usgae Mining Approaches for User's Request Prediction : A Survey", International Journal of Computer Science and Information Technologies, Vol 6(3), 2015, 2321-2325.
2. Sampath.P, Ramya.D, "Performance analysis of Web Page Prediction with Markov Model, ARM and ARM with Statistical Features", IOSR Journal of Computer Engineering, Vol 8, Issue 5, 2013.
3. Ramakrishnan Srikant, Yinghui Yang, "Mining Web Logs to improve website organization", in Proc. of 10th International conf. of WWW, Hong Kong, May 1-5, 2001.
4. Debra Vander Meer,Kaushik Dutta,Anindya Dutta, "Enabling scalable online personalization on the Web", in Proc. of the ACM conference on Electronic Commerce, Minneapolis, Minnesota, Oct. 17-20,2000.
5. Nasraoui.O and Krishnapuram.R,"One step evolutionary mining of context sensitive associations and Web navigation patterns," in Proc.SIAM Int.Conf.Data Mining, Arlington, VA, Apr.2002, pp.531–547.
6. Nasraoui.O and Petenes.C,"Combining Web usage mining and fuzzy inference for Website personalization," in Proc.WebKDD, 2003, pp.37–46.
7. "Enabling scalable online personalization on the Web", in Proc. of the ACM conference on Electronic Commerce, Minneapolis, Minnesota, Oct. 17-20,2000.
8. Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md.Nasir B. Sulaiman, "A new classification model for online predicting user's future movement", in International Symposium on Information Technology,pp. 1-7, Aug. 26-28,2008.
9. Mathias Gery, Hatem Haddad, "Evaluation of Web Usage Mining approaches for User's next request prediction", in WIDM ACM conf. New Orleans, Louisiana, USA, Nov. 7-8,2003.
10. A.Anitha, "A New Web Usage Mining Approach for next page access prediction", International Journal of Computer Applications, vol.8, no.11, Oct. 2010.
11. Mamoun A.Awad, Issa Khalil, "Prediciton of User's Search Behaviour :Application of Markov Model", IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, vol. 42, no. 4,Aug. 2012