Genetic Algorithm and Statistical Markov Model Fusion for Predicting User's Surfing Behavior Using Sequence Patten Mining

Mr. Ajeetkumar S. Patel[1], Prof. Anagha P. Khedkar[2]

[1]Department of Computer Engineering
Matoshri College of Engineering and Research Center
Nashik, India

[2]Department of Information Technology
Matoshri College of Engineering and Research Center
Nashik, India

Abstract

Classification technique is an important concept of data mining useful for researchers in web prediction of the web browsing page sequence's pattern analysis. The knowledge of navigated web page history of user browsing is helpful to predict the future set of page sequences that are likely to be visited by the user ahead of time. There is a wide scope for researchers to build and design prediction model based on browsing page sequences in sequence pattern mining. The main objective of prediction models is to achieve the better prediction accuracy and reduce the user latency. To achieve this objective, it is proposed to build the fusion of statistical Markov model and genetic algorithm based approach to improve the prediction accuracy. In addition, the genetic algorithm based approach is used to ease the modeling complexity of proposed system by generating optimal sequences of browsing patterns by reducing the size of search space. The proposed system is tested on the standard benchmark data sets to analyze prediction accuracy. The results outperformed by achieving 4% to 7% improvement over generalized Markov model.

Keywords: Genetic algorithm, sequence pattern mining, statistical Markov model, web prediction.

1. INTRODUCTION

In Internet era, World Wide Web (WWW) is a huge information repository available to millions of users. Such source of information can be access by the users through websites. As rapid technology development of Internet and World Wide Web, several challenges may arise like network traffic size and congestions, network latency, user latency, and websites complexity for thatthe developers and researchers has opportunity to tackle it.

Web mining applies the research of technology and application of data mining to overcome the challenges on the Internet and World Wide Web services. Web prediction technique can be established and applied many industrial applications which are essential in Internet and WWW. Some of the applications related to them are recommended system, caching system, wireless application, and search engine, prefetching and pre-catching. Therefore, it is desirable to find practical, scalable and applicable solution to overcome such challenges. Web prediction is a classification technique used to predict future set of web pages that may user visit ahead of time based on the knowledge of previously/history of visited pages. Such classification technique may improve the training and prediction process. Prediction process improvement may reduce the user's latency i.e. user's web page access times and therefore, it can ease the network traffic by avoiding unnecessary visit of web pages. Web Mining is one of the types of data mining technique that identifies sequences of page usage patterns of web data which fulfill the requirements of web applications. The process of web usage mining generally categorize in main three steps, namely, pre-processing, pattern discovery and pattern analysis. In this system the identification of web usage patterns is done based on page access sequences stored in access log at server side, thereby creating an intelligent and efficient sequence and semantic based web usage mining technique. The challenges faced in web prediction especially based on two categories namely, preprocessing and prediction. Preprocessing includes memory utilization for large amount of data sets, session identification, starting and ending of sessions, session ID assignment, N-gram assignment, domain knowledge hunt, and removal of unwanted/unnecessary data. In Prediction challenges includes low prediction accuracy, prediction time, long training/testing time, memory limitation, proper selection of prediction model, selection of hybrid approach requirement. The proposed system presents novel dictionary data structure; the fusion of genetic algorithm (GA) based approach and Hidden Markov model by applying sequence pattern mining to predict user's web-browsing behavior to improve prediction accuracy and faster prediction process. This approach extensive performance analysis shows a new empirical to give scalable solution efficiently to overcome the challenges present in web prediction process.

The contribution of proposed work is specifically summarized as follows.

- To design a technique to do preprocessing on a web server log data to map host and page accessed information using efficient data structure.
- To design a genetic algorithm for session extraction to search optimal page sequences where size of prediction model is reduce, which automatically

improves the prediction time.

- To propose a hidden Markov model using optimal page sequences to generate classifiers (features) for recommendation where it improves prediction accuracy.

The organization of this paper is as follows: In section II, presented the related work of the proposed system based on recommended system. Section III introduces the proposed system's algorithms, prediction model and architecture. Section IV is the result analysis of implemented system. Section V is the concluding section of implemented system with future perspectives and references.

2. RELATED WORK

Various prediction models used in web prediction problems are Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), K-nearest neighbor (kNN), Bayesian model, Clustering, Markov model, Fuzzy inference and others.Web prediction process of related work is mainly focused on recommended system. Joachims et al. proposed a tour guide software agent called web-watcher to assists users browsing over Internet. Such model is based on path based recommended system on kNN and reinforcement learning [1].Nasraoui and Krishnapuram Proposed fuzzy inference and clustering for web recommended system. Hierarchical Unsupervised Niche Clustering algorithm is applied on group profiles to find context-sensitive association between user session profile and different URL addresses [2] [3]. Mobasher et al. proposed Association Rule Mining (ARM) and frequent item set graph to predict user's future sets pages that are likely to be visit. To achieve this, active user's sessions are matched with available frequent item sets. The proposed ARM technique suffers limitation of scalability and efficiency [4].Su et al. proposed N-gram prediction technique to utilize path profiles of users from server log datasets to predict next page. Proposed model is based on point based prediction model [5]. Pitkow and Pirolli present predictive modeling techniques which focus on reducing the modeling complexity. To reduce modeling complexity the longest repeating subsequences technique is used for mining the surfing patterns [6].Levene and loizou presents Markov chain model to figure out the information contained in surfing trail to analyze the browsing patterns of user [7]. Hassan et al. applies the bayes rule and Markov chain to focus on navigational patterns like long sessions, page view range, page types, and categories of pages and rank problem [8]. Awad and Khan synthesize prediction models by fusionof artificial neural network and Markov model,support vector machine and Markov model. The fusion and domain knowledge exploitation focuses to improve prediction accuracy. Fusion suffers from limitation of handling multiclass and training overheads for artificial neural network. Support vector machine model suffers from handling large datasets [9][10]. Awad and Khalil proposed novel two tier prediction framework and present probabilistic model such as Markov model and association rule mining. Markov model and association rule mining delivers the result of prediction accuracy. Two tier prediction frameworksimprove the prediction time when multiple prediction models areconcerned. In addition, modified Markov model is introduced to reduce the modeling complexity. The models gives better prediction accuracy without compromising prediction time but suffers to scale on larger datasets [11].

3. PROPOSED WORK

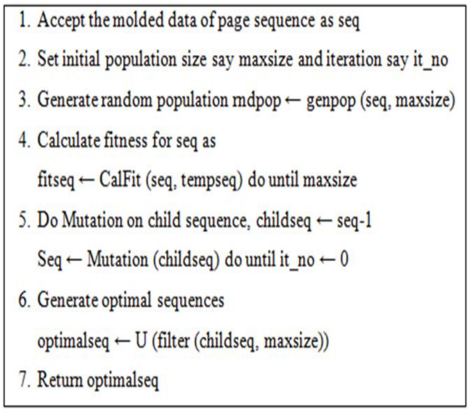
All algorithms which are defined and used to implement the proposed system specifically focused on pattern discovery of user's surfing behavior based on the probability and hidden values. This pattern discovery is represented using matrix format to apply possible value of convergence.

3.1 Pattern Discovery

Pattern discovery is the process where user surfing behavioral patterns are extracted from the formatted data of server log file. Due to this reason, preprocessing is used to convert such kind of data in semi-structured format. The output of the conversion can be used as the input to pattern discovery. Several data mining techniques did such functionality to obtain hidden patterns of users browsing behavior. Clustering, path analysis, ARM, classification algorithms are several techniques do such operations.

3.2Genetic Algorithm

In web prediction process, genetic algorithm is used due to several advantages over general prediction models. Genetic algorithm discovers high level prediction rules. The rules perform for global search and cope better with attribute interactions. It defines linkage between association and feature selection [20]. It is used as optimization tool to reset the parameters in the other classifiers. Main purpose to use genetic algorithm is that support of multiple classifiers lead to significant accuracy improvement. It also used to minimize the error rate and improve prediction accuracy. In proposed system, genetic algorithm used in pattern discovery of users browsing behavior to generate optimal page sequences where users are likely to be visit in ahead of time. The major advantage of genetic algorithm is to successfully accomplish such functionality which reduces the size of prediction model and search space. The goal of proposed system is to deploy genetic Algorithm to find the optimized solutions to investigate users surfing behavior. The major and crucial task in any mining application is that to prepare convenient target data set to which mining algorithm is applicable. Proposed system mining algorithm is Markov model prediction model used to do hunt surfing behavior. The following Figure 1 shows genetic algorithm approach of web usage mining. Genetic algorithm is local optimal search algorithm. The first step of it is to form an initial collection of sequence of pages called population which represents the possible solution to web prediction problem. Each page sequence characteristics are called a chromosome and it has an associated value called fitness value. Fitness function generates dissimilar page sequence vectors or we can say that it generates new population. But each vector has homogeneous characteristics. In given algorithm fitseq is the new population generated from tempseq dissimilar page sequence having similar characteristics. Chromosomes position is represented with predefined alphabets and it has certain values of position. This position is represented with value called alphabet is 0, 1. For new population generation the algorithm uses fitness function values to find out each page characteristics survival capacity by creating new or artificial population. This is done to improve current fitness function value from slice of old ones. After that the link and page quality of each page sequence is identified based on optimized population's average mean value. Fitness function is useful to select suitable population for the problem.



```
1. Accept the molded data of page sequence as seq
2. Set initial population size say maxsize and iteration say it_no
3. Generate random population rndpop ← genpop (seq, maxsize)
4. Calculate fitness for seq as
   fitseq ← CalFit (seq, tempseq) do until maxsize
5. Do Mutation on child sequence, childseq ← seq-1
   Seq ← Mutation (childseq) do until it_no ← 0
6. Generate optimal sequences
   optimalseq ← U (filter (childseq, maxsize))
7. Return optimalseq
```

Figure 1 Genetic Algorithm forPage Sequence Mining

Mutation phase of genetic algorithm used to randomly swap the sequences. Each individual sequences are called child sequences. Mutation changes randomly new child dubbed offspring. Those child sequences redundant individuals are removed to check for the optimization. Those optimized page sequences given to next operation which predict the best possible page that user may visit.

3.3 Hidden Markov Model

Proposed system has second major model to predict users web surfing behavior based on statistical approach called statistical Markov model dubbed hidden Markov model. Simple Markov model or Markov chain somewhat differs on functionality but results generation are almost same. State in simple Markov model is directly visible to user/observer and therefore the only parameter to it is state transition probabilities. A Hidden Markov Model (HMM) dubbed statistical Markov model is a Markov process having hidden/unobserved states. It is mathematical tool based dynamic approach used for stochastic processes (optimal nonlinear filtering problem). In hidden Markov model, output dependent on the state is visible, but states are invisible i.e. direct visible is obviate. For every possible output token, it has probability distribution for each state. Therefore sequences generated by such model give kind of information about sequence of states. In hidden Markov model the term "hidden" refers to sequence of state through which model passes. So we can say that, this state sequence is not passed to the parameters of the model and it refers as sequence based model.In general, Markov model has functionality to predict next action based on the results of previous actions [11]. In reference of proposed system for web prediction process hidden Markov model is used to predict future set of pages or next page based on history of previously visited pages. If user will visit n-th page, then it has probability P_r that must has visited in any order of n-1 pages. Such that,$P_r (P_n| \{P_{n-1}, P_{n-2},., P_{n-k} \}) = P_r (P_n | P_t)$.
Here P_t = training set example.

3.4 Architecture Model

Figure 2 shows the architecture model of proposed system where input is given as page sequences of server log data and the system provides the output inform of classifiers which shows next page that user may visit in ahead of time. This would be analyzed by query mechanism process. In between input and output stage, the proposed system is divided in to four major steps, namely preprocessing, mapping, pattern discovery using genetic algorithm and Markov model classification. These four steps are combined and named as data preprocessing and pattern discovery classification. The first step preprocessing includes identification of sessions, starting and ending of session, grouping of sessions grouping, assignment of unique ID to each session, removal/filtering of unwanted records introduced in [13]. In mapping stage, sessions are extracted to map in the tables. But before that data reading is performed from session to build the data structure. The mapped table is created for IP address and URL sequences with index to build dictionary table. In next step genetic algorithm is applied to process on IP address and page surf sequences using dictionary data structure to get optimal sequences. Then the optimal sequences are provided as input to Markov model classification. In final step of proposed system, the input is optimal page sequences given to statistical Markov model classification where optimal sequences are extracted to generate the optimal features called as classifiers for recommendation where users likely to be visit the page sequences in future.

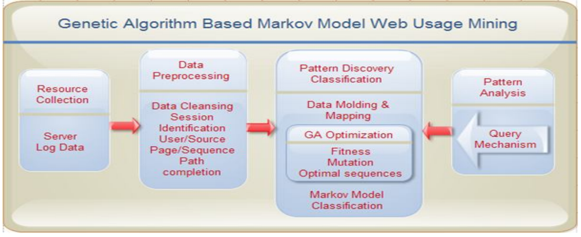

Genetic Algorithm Based Markov Model Web Usage Mining
Resource Collection
Server Log Data
Data Preprocessing
Data Cleansing Session Identification User/Source Page/Sequence Path completion
Pattern Discovery Classification
Data Molding & Mapping
GA Optimization
Fitness Mutation Optimal sequences
Markov Model Classification
Pattern Analysis
Query Mechanism

Figure 2 Architecture System of Web Usage Mining

4. RESULTS

The result is being carried out on prediction accuracy to predict user's surfing behavior using sequence pattern mining. The following sections illustrate the step by step activity performed to analyze the implemented work.

4.1 Datasets

Standard benchmark access log datasets is referred in the proposed system's implementation namely, the University of Saskatchewan's (UOFS) data set [14]. The brief statistics of such data set is shown in following Table 1[11].

Table 1UOFS Dataset Statistics[11]

Specification	UOFS
Total logrecords	2,408,625
Total sessions	172,984
Averagesessionlength	5.5
Numberofpages	5423
Dataset date (month/year)	6-12/1995

Data set is recorded at server log having contents like host name or Internet address, timestamp as date and time, request of HTTP get method, HTTP reply code, and byte transferred. This dataset is used for training and testing purposes. 60% dataset is considered for training the data set from the original set and for testing purpose 40% data set is considered from original set. The proposed system performs the data cleansing and session identification process on data set to create semi structured data set.

4.2 Evaluation

In this section, analysis of the general prediction model and implemented system is being carried out.In web prediction problem, performance is especially measured using prediction accuracy. Prediction accuracy is measured on precision value on sequence pattern prediction process. Such parameter result is specifically depends on input size of dataset.

4.3 Experiment Setup

Experimental results are presented for web prediction using fusion of hidden Markov model and genetic algorithm and compared with generalized Markov model. In the experiments, sessions are preprocessed to specific n-gram and compared to prediction accuracy. To measure the prediction accuracy, the generalization accuracy method is used by partitioning the data set into a training set and a testing set. In implemented system, sixty percent part and forty percent part of original data set is considered for training and testing purpose respectively. Each experiment is run 10 times by applying random partitioning on the data set. UOFS dataset is used for training and testing purpose of experiments to compare the proposed system and generalized Markov model of web prediction.

4.4 Prediction Setup

Consider, Testing session - t and Length - l to conduct prediction, (l-1)-gram Markov model is used and obtain prediction. Such prediction used to evaluate the accuracy of the model. Here, page sequence t is final outcome and it is longer than n-gram used in the experiment. For example, suppose page sequence t = p3, p4, p5, p6, p7. If we use second-order Markov model, then we break page sequence t into (p3, p4, p5), (p4, p5, p6) and (p5, p6, p7).

4.5 Results of N-gram Prediction Accuracy

Table 2 shows the comparison of the prediction accuracy of generalized Markov model and implemented system for various N-grams. Table results are clearly shows the improvement of improvement system over generalized Markov model. In N-gram prediction higher order of Markov model gives better outcome than lower order. It degrades continuously when N is increased.

Table 2 Prediction Accuracy for Various N-grams

	Implemented system accuracy	Generalized Markovmodel accuracy
1-Gram	0.647	0.613
2-Gram	0.412	0.377
3-Gram	0.218	0.183
4-Gram	0.197	0.162
5-Gram	0.105	0.042

Figure 3 shows the graphical representation of comparison of prediction accuracy of both models using various N-grams.

Analysis

- It is observed that prediction accuracy decreases when N increaseswhen training sessions obtained from preprocessing stage for N-gram is more than for (N+1)-gram preprocessing of sessions. Therefore, more experiences encounter for N-gram model in data set as compare to (N+1)-gram model. Hence, the prediction accuracy reduced or decreased on higher order N-gram.

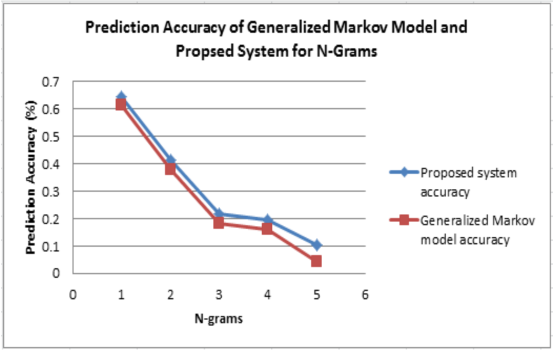


Figure 3 Prediction Accuracy for Various N-grams

- Prediction accuracy of various N-grams for the implemented model is better than the generalized Markov model.
- Figure states that when N order is increased than generalized Markov model more weakens than the implemented model.
- In figure, the N-gram analysis for prediction accuracy is performed on 5% to 10% of UOFS data set size. If the input sizes of the data set even

more,the implemented system gives even better prediction accuracy.

- If the sizes of the sessions are vary than the effect will be on prediction accuracy.

4.6 Results of Overall Prediction Accuracy

Table 3 shows the comparison of the overall prediction accuracy of generalized Markov model and implemented system using different unique IP addresses set. Each IP address set has web page sequences of user browsing and the training data set is taken as unique IP addresses set to test the model. The IP addresses set is considered as same as 5% to 10% size of whole training data set.

Table 3 Results of Overall Prediction Accuracy

Training IP Addresses	Implemented system accuracy	Generalized Markov model accuracy
100	0.394	0.346
200	0.399	0.350
300	0.407	0.356
400	0.412	0.362
500	0.415	0.376

The result of implemented model is outperformed on overall prediction accuracy. It clearly shows the difference on prediction accuracy on both models that implemented system achieves better prediction accuracy. The results of table are clearly shows that when the size of input data set is increased the prediction accuracy is also increased and it is higher than generalized Markov model. Figure 4 shows the graphical representation of comparison of overall prediction accuracy for both models.

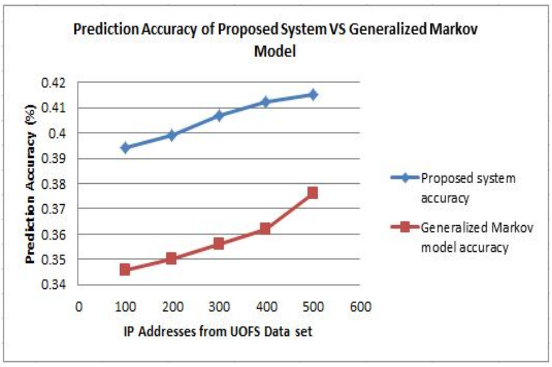


Figure 4 Overall Prediction Accuracy

Analysis

- The graph clearly shows that the efficacy of implemented model by improving prediction accuracy. For data set size of 200 IP address set, the implemented system prediction accuracy is 40% versus 35% in generalized Markov model.
- The prediction accuracy is improved in implemented system because of the generated classifiers provides appropriate selection of N-gram used for the prediction. In addition, the genetic algorithm and hidden Markov model found/search optimal hidden states to improve prediction accuracy which are not found in general Markov model.
- Graph clearly shows that prediction accuracy is increases when data set size is increases. The result is even better than current results on the data set size input is increased.
- Implemented system over generalized Markov model gives the improvement on prediction accuracy around 4% to 7%.

5. CONCLUSION

The problem of web prediction is handled in an efficient manner using the new system presented here. This new system is designed by the fusion of genetic algorithm and hidden Markov model to reduce the complexity of the general Markov model. The fusion system successfully improves the average overall prediction accuracy by 5%. To achieve the desired results, experiments were designed and performed using UOFS data set with parameters such as number of N-grams, labeling, and partition percentage. The comparative results show that smaller N-gram models performance is better than higher N-gram models in terms of prediction accuracy due to reduction in number of sessions. In future, this system can be extended by conducting in-depth analysis on the fusion of genetic algorithm and hidden Markov model for online applicability in web prediction problem. This system may be extended on analysis of prediction accuracy based on page sparsity. Besides these extensions, this system may be further investigated to check the effects of boosting/bagging.

References

[1.] T. Joachims, D. Freitag, and T. Mitchell, “WebWatcher: A tour guide for the World Wide Web,” in Proc. IJCAI, 1997, pp. 770–777.

[2.] O. Nasraoui and R. Krishnapuram, “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.

[3.] O. Nasraoui and R. Krishnapuram, An evolutionary approach to mining robust multi- resolution web profiles and context sensitive URL associations, Int. J. Comput. Intell. Appl., vol. 2, no. 3, pp. 339348, 2002.

[4.] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, “Effective personalization based on association rule discovery from Web usage data,” in Proc. ACM Workshop WIDM, Atlanta, GA, Nov. 2001.

[5.] Z. Su, Q. Yang, Y. Lu, and H. Zhang, “WhatNext: A prediction system for Web requests using n-gram sequence models,” in Proc. 1st Int. Conf. Web Inf. Syst. Eng. Conf., Hong Kong, Jun. 2000, pp. 200–207.

[6.] J. Pitkow and P. Pirolli, “Mining longest repeating subsequences to predict World Wide Web surfing,” in Proc. 2nd USITS, Boulder, CO, Oct. 1999.

[7.] M. Levene and G. Loizou, “Computing the entropy of user navigation in theWeb,” Int. J. Inf. Technol. Decision Making, vol. 2, no. 3, pp. 459–476, 2003.

[8.] M. T. Hassan, K. N. Junejo, and A. Karim, “Learning and predicting key Web navigation patterns using Bayesian models,” in Proc. Int. Conf. Comput. Sci. Appl. II, Seoul, Korea, 2009, pp. 877–887.

[9.] M. Awad and L. Khan, “Web navigation prediction using multiple evidence combination and domain knowledge,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.

[10.] M. Awad, L. Khan, and B. Thuraisingham, “Predicting WWW surfing using multiple evidence combination,” VLDB J., vol. 17, no. 3, pp. 401– 417, May 2008.

[11.] M. Awad and I. Khalil, Prediction of Users Web-Browsing Behavior: Application ofMarkov Model, IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 42, no. 4, pp.11311142, Aug. 2012.

[12.] ”Genetic Algorithm in Search Optimization” by D. E. Goldberg.

[13.] R. Cooley, B. Mobasher, and J. Srivastava, Data preparation for mining World WideWeb browsing patterns, J. Knowl. Inf. Syst., vol. 1, no. 1, pp. 532, 1999.

[14.] Internet Traffic Archieve. [Online]. Available: http://ita.ee.lbl.gov/html/traces.html