ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

A SURVEY REPORT ON PREDICTION OF NEXT ACCESSED WEB PAGE IN WEB USAGE MINING

Gourav Kumar Sharma*, Dr. R. K. Gupta
*Madhav Institute of Technology &Science, Gwalior, M.P.

DOI: 10.5281/zenodo.60586

INTRODUCTION

Data mining is the discovering insightful process, interesting, and novel patterns, as well as predictive, understandable, and descriptive models from big-scale data The objective of data mining is to recognize legal new, potentially helpful, and reasonably correlations and patterns in presenting data.

Data Mining errands can be ordered into two classifications, Descriptive Mining and Predictive Mining. The Descriptive Mining strategies, for patterns, Clustering, Sequential Pattern Discovery, Association Rule Discovery, is utilized to discover human-interpretable patterns that depict the information. In data mining, web mining is used to remove knowledge from unstructured data . In next section discuss web mining their types and concept of web usage mining, approaches, algorithms, and previously done related work.

Web Mining

Web Mining is used to remove knowledge from the raw unstructured information. The emerging web mining aims area at discovery and removing relevant knowledge that is hidden in Web related infformation, in specific in text documents published on the Web. Web mining is achive in three ways they are:
1) web usage mining
2) web content mining
3) web structure mining.

Web usage mining gives the web site design support, supplying personalization server and different business creating decision, etc. Web content mining is the procedure of mining information from the documents content or their descriptions. Web document text mining, resource discovery based on idea indexing or agent; based technology may also fall in this category. Web structure mining is the inferring knowledge process from the WWW organization and links between references and referents in the Web.

Web usage mining is used to mine knowledge based on the user log. WUM is the method of applying information mining strategy to the usage patterns discovery from Web data, targeted towards numerous applications.

The usage data collected at the various sources will represent the navigation patterns of various segments of the complete Web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns.

CONCEPT OF WEB USAGE MINING

Discovery of significant patterns from data generated through client-server transactions on one or more web servers. Classical Sources of Data:
1. Automatically generated data stored in referrer logs, server access logs, agent logs, and client-side cookies.
2. Electronic commerce and product-oriented user events (e.g. shopping cart changes, ad or product click-through, etc).
3. User ratings and/or User profiles.
4. Meta-data, page attributes page content, site structure.

APPROACHES OF WEB USAGE MINING

1. **Data collection:** collection of data is the basic level of web usage mining, the information authenticity and integrality will directly affect the following works easily carrying on and the last characteristic service's quality recommendation. Therefore it must use scientific, advanced and reasonable technology to gather numerous data. At current, towards web usage mining technology, the basic data origin has 3 kinds: middle data, server data and client data

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

2. **Data preprocessing:** Few databases are inconsistent, insufficient and including noise. The data pretreatment is to carry on a unification transformation to those databases. The outcome is that the database will to become integrate and consistent.

3. **Knowledge Discovery:** Use statistical technique to carry on the mine and analysis the pretreated information. We may determine the person or the user neighborhood's interests then construct interest model. At current the usually used machine learning approaches mostly have clustering, classifying, the relation discovery and the order model discovery. All technique has its own shortcomings and excellence, but the quite efficient technique mostly is clustering and classifying at the present.

4. **Pattern analysis:** Pattern Analysis Challenges are to filter uninteresting knowledge and to interpret and visualize the most interesting patterns to the client. First, delete the less importance models or rules from the most interested model storehouse; Another use OLAP technology and so on to carry on the comprehensive analysis and mining ; over again, let discovered data or information be visible; Finally, provide the characteristic facility to the E-commerce website.

There are wide application areas of the analysis of user web navigation behavior in web usage mining. The analysis of user web navigation behavior can help for improving the organization of the web site and improvement of web performance by pre-fetching and caching the most probable next web page in advance. Web Personalization, Adaptive web sites are some of the applications of web usage mining. Web usage mining can provide guidelines for improving ecommerce to handle business specific issues like customer attraction, customer retention, crosses sales, and customer departure.

MARKOV MODEL

Markov model is widely used for modeling the user web navigation sessions. The traditional Markov model is having its own limitation[3]. First-order Markov model is less complex but the accuracy is low because of lack of looking into the depth. As we move to the second-order Markov model it is accurate as compared to the first-order Markov model but the coverage of prediction state is less and the time complexity get increased.

Markov model is compact, easy to understand, expressive and based on a well-established theory. Markov model is widely used to model user web navigation sessions. In firstorder Markov model, each state corresponds to a single web page and each pair of viewed page corresponds to state transition. Two artificial state, start and final, are incorporated in the model. In second-order Markov model each state corresponds to sequence of two viewed web pages and so on. In Markov model, state space increases exponentially with the order of the model. Let the m is number of web pages and n is order of model then the number of states will be m^n.

Markov models have been extensively used for predicting the action a user will take next given the sequence of actions he or she has already performed. For this type of problems, 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 |S| × |A| Transition Probability Matrix (TPM), where each entry tij corresponds to the probability of performing the action j when the process is in state i. The state-space of the Markov model depends on the number of previous actions used in predicting the next action. The simplest Markov model predicts the next action by only looking at the last action performed by the user. In this model, also known as the first-order Markov model, each action that can be performed by a user corresponds to a state in the model. A somewhat more complicated model computes the predictions by looking at the last two actions performed by the user. This is called the second-order Markov model [4], and its states correspond to all possible pairs of actions that can be performed in sequence. This approach is generalized to the K^{th} -order Markov model, which computes the predictions by looking at the last K actions performed by the user, leading to a state-space that contains all possible sequences of K actions.

For example, consider the problem of predicting the next page accessed by a user on a web site. The input data for building Markov models consists of web-sessions, where each session consists of the sequence of the pages accessed by the user during his/her visit to the site. In this problem, the actions for the Markov model correspond to the different pages in the web site, and the states correspond to all consecutive pages of length K

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

that were observed in the different sessions. In the case of first-order models, the states will correspond to single pages, in the case of second-order models, the states will correspond to all pairs of consecutive pages, and so on. Once the states of the Markov model have been identified, the transition probability matrix can then be computed [5].

There are many ways in which the TPM can be built. The most commonly used approach is to use a training set of action-sequences and estimate each t ji entry based on the frequency of the event that action ai follows the state sj . For example consider the web-session W S2({P3, P5, P2, P1, P4}) shown in Figure 1. If we are using first-order Markov model then each state is made up of a single page, so the first page P3 corresponds to the state s3. Since page p5 follows the state s3 the entry t35 in the TPM will be updated. Similarly, the next state will be s5 and the entry t52 will be updated in the TPM. In the case of higher-order model each state will be made up of more than one actions, so for a second-order model the first state for the web-session W S2 consists of pages {P3, P5} and since the page P2 follows the state {P3, P5} in the web-session the TPM entry corresponding to the state {P3, P5} and page P2 will be updated. Once the transition probability matrix is built making prediction for web sessions is straight forward. For example, consider a user that has accessed pages P1, P5, P4. If we want to predict the page that will be accessed by the user next, using a first-order model, we will first identify the state s4 that is associated with page P4 and look up the TPM to find the page pi that has the highest probability and predict it. In the case of our example the prediction would be page P5.

Web Sessions:
$WS_1 : \{P_3, P_2, P_1\}$
$WS_2 : \{P_3, P_5, P_2, P_1, P_4\}$
$WS_3 : \{P_4, P_5, P_2, P_1, P_5, P_4\}$
$WS_4 : \{P_3, P_4, P_5, P_2, P_1\}$
$WS_5 : \{P_1, P_4, P_2, P_5, P_4\}$

1st Order	P_1	P_2	P_3	P_4	P_5
$s_1 = \{P_1\}$	0	0	0	2	1
$s_2 = \{P_2\}$	4	0	0	0	1
$s_3 = \{P_3\}$	0	1	0	1	1
$s_4 = \{P_4\}$	0	1	0	0	2
$s_5 = \{P_5\}$	0	3	0	2	0

2nd Order	P_1	P_2	P_3	P_4	P_5	*2nd Order*	P_1	P_2	P_3	P_4	P_5
$\{P_1, P_4\}$	0	1	0	0	0	$\{P_3, P_5\}$	0	1	0	0	0
$\{P_1, P_5\}$	0	0	0	1	0	$\{P_4, P_2\}$	0	0	0	0	1
$\{P_2, P_1\}$	0	0	0	1	1	$\{P_4, P_5\}$	0	2	0	0	0
$\{P_2, P_5\}$	0	0	0	1	0	$\{P_5, P_2\}$	3	0	0	0	0
$\{P_3, P_2\}$	1	0	0	0	0	$\{P_3, P_4\}$	0	0	0	0	1

Figure 1: Sample web sessions with the corresponding 1st & 2nd order Transition Probability Matrices.

Limitations of Markov Models:

1. In many applications, first-order Markov models are not successful in predicting the next action to be taken by the user. This is because these models do not look far into the past to correctly discriminate the different behavioral modes of the different users.
2. higher-order models have a number of limitations associated with:
    a. High state-space complexity
    b. Reduced coverage.
    c. Sometimes even worse prediction accuracy.
3. The number of states used in these models tend to rise exponentially as the order of the model increases. This dramatic increase in the number of states can significantly limit the applicability of Markov models for applications in which fast predictions are critical for real-time performance or in applications.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

LITERATURE SURVEY

J. Borges, [5], proposed a Markov model for modeling the user web navigation sessions. In this author has preprocessed the web log file then modeled it through the Markov model and finally model is used to identify the useful patterns.

F. Khalil, [3], has proposed a new framework for predicting the next web page access. Authors of [3] have used the Markov model for web prediction. If the Markov model is not able to predict the next page then the association rule are used to predict the next web page. They have also proposed the solution for ambiguity in the prediction. Ambiguity will be resolved by taking the help of association rule.

J. Borges, [4, 11, 12], proposed Higher-order Markov model for web usage mining. There are some problems associated with lower-order Markov model. Basically the low accuracy is the major limitations of lower-order Markov model. Higher-order Markov model is suffered from the state space complexity. Author has proposed the Higher-order Markov model with clustering technique to improve the effectiveness of Markov model and to reduce the state space complexity. The K-mean clustering technique has been used to reduce the state space complexity. The experimental result shows that the accuracy is improved by introducing the clustering technique in Markov model.

M. Desponde, [3], proposed the new approach for reducing the complexity of Markov model. Three approaches frequency-pruning, error-pruning and support-pruning have been used to reduce the state space complexity. They presented different techniques for intelligently selecting parts of different order Markov models so that the resulting model has a reduced state complexity and improved prediction accuracy. They have tested their models on various datasets and have found that their performance is consistently superior to that obtained by higher-order Markov models.

Bhawna Nigam et. al. [13] Dynamic Nested Markov model (DNMM) is proposed for modeling the user web navigation sessions. In this model main focus is on the coverage and time complexity of model. The DNMM is having nesting of the second-order Markov model inside the first-order Markov model or the higher-order Markov model inside the lowerorder Markov model. The DNMM uses the link list structure rather than the transition matrix of traditional Markov model (TMM).

Table I shows the example of collection of user web navigation sessions. T1 to T5 are the transaction IDs and corresponding user web navigation sessions are given in the table I.

TABLE I. COLLECTION OF USER'S NAVIGATION SESSION.

Transaction ID	Web navigation sessions
T1	P2, P3, P1, P5
T2	P2, P1, P3, P4, P5
T3	P1, P2, P5
T4	P1, P5, P4
T5	P1, P2, P4

Figure 1 show the second-order traditional Markov Model corresponds to the table I. The model is represented in the form of hypertext probabilistic matrix. In a higher-order Markov model, a state corresponds to a fixed sequence of pages, and a transition between states represents a higher order conditional probability.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

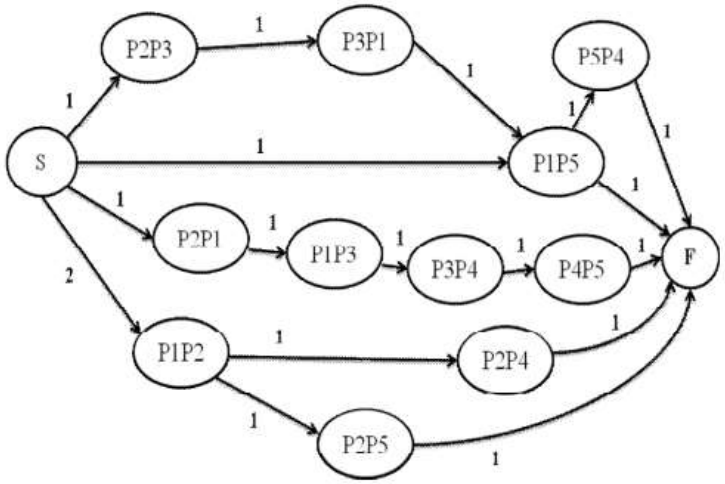

Figure 1. Second-order traditional Markov model corresponds to table I.

DISADVANTAGES

1. TMM contains information about only a particular order.
2. The exponential increment in number of states increases search space and complexity Higher-order Markov model also have low coverage problem.
3. In traditional higher-order Markov model number of states exponential increases as increase in the order of model.
4. The third-order traditional Markov model contains information about third-order only, information about first and second order cannot be extracted, and this reduces coverage and accuracy.

In DNMM this search space complexity, low coverage and low accuracy problems are targeted.

Generating of DNMM In Dynamic Nested Markov model the higher-order Markov model is nested inside the lower-order Markov model. DNMM uses the link list structure for storing the information of web page[6]. DNMM is same as the traditional Markov model with some changes so that the efficiency of model can be enhanced. This model is dynamic in nature means the addition and deletion of state can be done easily. This model uses the node structure to store the web page. All the information of a particular web page is stored in a node of that web page. In this model only one node per web page is created. In DNMM the node is a dynamic data structure rather than just name of the web page. Each node contains name of web page, count of web page and an inlink list. The inlink list is a link list in which each node contain name of a previous web page from which the current web page is traversed, count that shows number of times current web page is traversed from previous web page and an outlink list that keep track of all the corresponding to that previous web page [7]. Outlink list is a linked list whose each node contains name of next web page and its count. Now this data structure keeps track of all the previous web pages and all the next web pages corresponding to each previous page, of the current web page node. The beauty of this model is that its number of nodes is always constant which is equal to the number of web pages, no matter what the order of model is. In thirdorder DNMM every node contains data upto third-order and in fourth-order model each node contain data up to fourthorder, but number of node are always constant. In this fashion web navigation sessions are modeled in highly structured and efficient way. In DNMM each web page has unique node. All the information regarding web page is stored inside the node up to the n-order model. Figure 2 shows the node structure of second-order DNMM, W1, W2…. are the second-order inlinks to the web page Wx and the corresponding secondorder outlinks [8, 10].

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

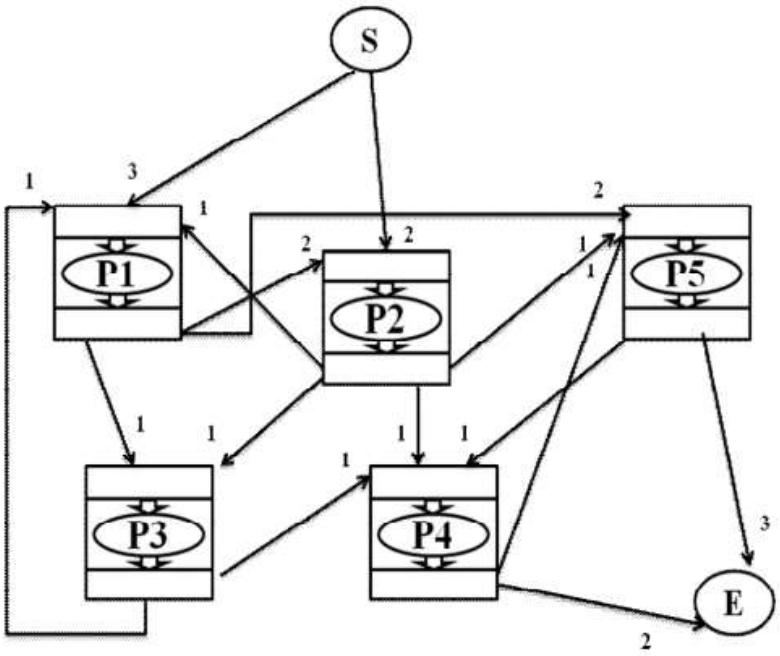

Figure 3. First-order Dynamic Nested Markov model corresponds to example of Table I.

Figure 4 shows second-order Dynamic Nested Markov model corresponds to the Table I. In the second-order DNMM history of previous web page is stored inside the node of first-order model. Similarly, it can be done for higher order models.

ADVANTAGES

1. **Time complexity:** The time is a crucial factor for the measurement of the performance of model. In DNMM the dynamic link list structure is used instead of the transition matrix of traditional Markov model. The generation time of the model is less because of the use of link list structure for storage.
2. **Coverage of the model:** The coverage of the model is defined as the ratio of number of times model is able to predict to number of request in test set. It can be said that higherorder DNMM is having the coverage of all the states of lower-order DNMM. Information about first-order cannot be extracted from TMM but it is possible in DNMM hence the coverage is increased.
3. In TMM transition matrix is used. To create the transition matrix all the individual pages must be known in advance. It is not necessary in DNMM. Once the TMM model is created, it is very tedious to update it, if a new page is introduced. On other hand to update DNMM is quite simple.
4. TMM uses transition matrix which reserve space in advance for each page to store transition information to all other pages but practically a web page has transitions to few pages so it waste lot of memory. In DNMM the required amount of memory is used.
5. Reduction of number of nodes in DNMM.

CONCLUSION

In this paper, we have surveyed about the various techniques that have been used and implemented till date. Various papers have been reviewed that can help in further work. In this paper a class of Markov model-based prediction algorithms that are obtained by selectively eliminating a large fraction of the states of the All-K th - Order Markov model have been given. Traditional and Dynamic Markov modelshave increased the efficiency and various factors that affect have been analyzed.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

REFERENCES

1. Han J and Kamber M: Data Mining: Concepts and Techniques. Second edition Morgan Kaufmann Publishers.
2. Pang Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data mining", 2009.
3. M. Desponde, and G. Karpis "Selective Markov models for predicting web page accesses" ACM Transactions on Internet Technology, vol. 4, no. 2, May 2004, pp.163–184.
4. J. Borges. "A data mining model to capture user web navigation.Ph. D. thesis", University College London, London University, 2000.
5. J. Borges, and M. Levene, "Generating dynamic higher-order Markov models in web usage mining," Proc. Ninth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), eds. A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, Oct. 2005, pp. 34–45.
6. J. Zhang, and A. A. Ghorbani, "The reconstruction of user sessions from a server log using improved timeoriented heuristics." in CNSR. IEEE Computer Society, 2004, pp. 315–322.
7. M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, "Web path recommendations based on page ranking and Markov models," Proc. Seventh Ann. ACM Int'l Workshop Web Information and Data Management (WIDM '05) , 2005, pp. 2–9.
8. R. Popa, and T. Levendovszky "Marcov models for web access prediction" 8th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Nov 2007.
9. J. Borges, and M. Levene, "Testing the predictive power of variable history web usage," J. Soft Computing, special issue on Web intelligence, 2006.
10. X. Chen, and X. Zhang, "A Popularity-Based Prediction Model for Web Prefetching," Computer, 2003, pp. 63–70.
11. J. Borges, and M. Levene, "Evaluating variable-length Markov chain models for analysis of user web navigation sessions", IEEE Trans. Knowl. Data Eng., Vol. 19, No. 4, 2007, pp. 441–452.
12. J. Borges, and M. Levene, "Data Mining of User Navigation Patterns," Web Usage Analysis and User Profiling, eds. B. Masand and M. Spiliopoulou, LNAI 1836, pp. 92–111, Springer, 2000.
13. Bhawna Nigam and Suresh Jain, "Generating a New Model for Predicting the Next Accessed Web Page in Web Usage Mining", Third International Conference on Emerging Trends in Engineering and Technology, 978-0-7695-4246-1/10, 2010.