IJCAT - International Journal of Computing and Technology, Volume 3, Issue 11, November 2016
ISSN : 2348 - 6090
www.IJCAT.org

Markov Model Based Web Page Recommendations by Combining Content and Log Features

[1] Neetu Sahu, [2] Pragyesh Kumar Agrawal

[1] Atal Bihari Vajpayee Hindi Vishwavidyalaya
Bhopal, M.P., India

[2] Institute for Excellence in Higher Education
Bhopal, M.P., India

Abstract - As the internet services are increasing day-by-day various websites are working to find new techniques for webpage recommendations. Researchers have developed new methods for increasing the accuracy of predicted web pages. This paper has utilized web content feature of the website pages for developing the Term network where terms from each page help in specifying the relations between all pages. Web log is another feature used in this work where Markov model of third order helps in improving the prediction accuracy. Experiments are performed on different dataset sizes with these feature combinations. It is observed that proposed model is better as compared to previous works. Results show that the use of Markov Model with web content feature is better for prediction.

***Keywords* - Information Extraction, Text Analysis, Feature Extraction, Text Categorization, Clustering**.

1. Introduction

Internet has become very important these days due to its wide uses in daily life. A better way of optimizing sites is to learn the user behavior pattern for presenting the next page on the other side of the server that is client/user end. This task is carried out by the ways of few activities. The first activity involved in the web usage mining procedure is preprocessing the web log files. The second activity is using the mining algorithm which is employed for finding out the pattern, whereas the final activity is analyzing the pattern which is mined by mining algorithm [1, 2].

Mining algorithm is a method to find the rules and patterns from the sequence patterns. Few examples of such rules and pattern are association rule, clustering algorithm, and sequential pattern analysis [3]. Number of researchers is working in this field of web mining which include IR, Database, predictions [4], Intelligent Agents and Topology, which gives different solutions for the web content mining and web structure mining. Web usage mining is a new area of interest and has gained lot of popularity in current time. This area includes the automated discovery and analysis of patterns in data which result in the user's interactions with one or more websites. Web access patterns are discovered to understand the users' navigation preferences and behavior by focusing on various tools and techniques.

Techniques derived from Markov model have been extensively used for predicting the actions a user can take given the sequence of actions he or she has performed earlier. The server can accordingly pre-fetch the predicted pages and cache these pages or it can pre-send this information to the user. These techniques are used effectively to help e-commerce businesses improvise their websites in better manners [5]. The focus of web usage mining is to get the model and analyze the users' behavioral patterns. It consists of three phases: Pre-processing of web data, pattern discovery and pattern analysis.

Computer applications and browsing based software majorly suffer from latency or web page prediction (user requirement) issues. Recently various efforts are made for finding a stable and efficient solution for improving web page prediction. With the help of such predictions we can

protect our customer or user and can easily reach our own requirements. We can also enhance our own business using web page prediction techniques. Furthermore, there are different tools and techniques available for web page prediction. For that purpose, the current work of developing an efficient web page prediction approach aims at studying and improving the traditional methods of web page recommendations.

1.1 Objectives

This paper takes few questions into account. For example, how can we understand a user's web access sequence to effectively recommend the next pages to the user? How can we automatically discover and represent useful knowledge for web-page recommendation given the web usage data? How to effectively learn from available historical data and discover useful knowledge of the domain and web-page navigation patterns? The authors have worked to propose a model to make effective Web-page recommendations based on the discovered knowledge.

2. Related Work

Based on the type of information used to make a particular prediction, the prediction algorithms can be broadly classified in two main groups. The first includes algorithms that predict future accesses based on the previous access patterns. To distinguish in two subgroups can be: one consists of algorithms that use of Markov models and the other one with algorithms that make use of data mining techniques. Large number of prediction algorithms based on Markov models are found in the literature and some of them provide high precision predictions but at the cost of extreme computation and lot of memory consumption [6, 7].

The second group makes use of the algorithms that analyze the web content to make certain predictions. Some authors have proposed to combine the analysis of the content with usage profiles [8], others apply neural networks to keywords extracted from HTML content and some others detect similarities in context words around links in the HTML content [9]. The proposals are based on the object popularity and the association of hyperlinks, but they do not consider the relationship among objects.

The most common strategy of presenting search results is a simple ranked list [10]. Intuitively, such a presentation strategy is reasonable for non-ambiguous, homogeneous search results; in general, it would work well when the search results are good and a user can easily sort relevant documents in the top ranked results.
In [11] web page re-ranking is performed by use of web log feature alone. Here multi-damping of user sequence is carried out on the basis of linear, page rank and generalized hyperbolic functions are used. Accuracy level is low as this paper does not include web content feature.

In [12] query optimization is done on the basis of web user behavior. In this work authors have proposed two algorithms where time bound algorithm for query classification is better as compare to other proposed algorithm which is named as QCSP (Quantified Constraint Satisfaction Problems). People attempt to infer user goals and intents by predefining some specific classes and performing query classification accordingly.

In [13] user goals as navigational, informational and categorize queries are categorized into these classes. Other works focus on tagging queries with some predefined concepts to improve feature representation of queries. However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.

3. Proposed Work

Two datasets are used in this work; first is content dataset which contains whole content of the website page-wise and the other is website weblog which contains different patterns of the users for various requirements. Block diagram of the proposed model is shown in figure 1.

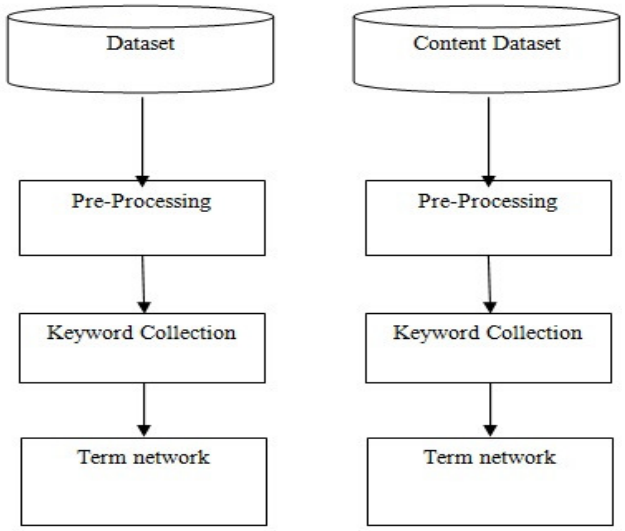

Fig. 1 Block diagram of proposed method.

3.1 Content Pre-Processing

Web pages contain keywords that need to be filtered before proceedings to next step. In order to remove stop words such as {a, an for, of,…….. }, the stop word dictionary compares all the contents of the web page. As a result, similar words are removed and rest of words is considered as the important words [16]. Now all pages generate their own important keywords while it is possible that some words may be common. These similar words help in developing the relation between pages of the website.

3.2 Keyword Collection

Important keywords obtained from above step have lots of words which may be repeated in same page. This can be understood as:
let Pi={k1, k2, k1, k3, k5, k2…….. }
here i is page number while k is important words.

So words having some repetition count, and those words which cross minimum threshold repetition are considered as the keyword of the page. Here all pages have entire keyword set. It is also possible that pages may have some similar keywords which will help in developing the relations in next step.

3.3 Term Network

Here a relation is developed between different pages of the website on the basis of terms or keywords selected from the web pages. This can be understood as:

Let Pi has keywords {k1, k2, k3}; and Pj has keywords {k1, k4, k2},

then k1, k2 are common keywords between these pages so a common link is established between these pages. Term network with two web pages is shown in figure 2.

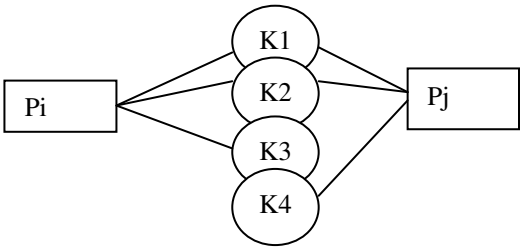

Fig. 2 Term network with two web pages.

All the website web pages terms are connected with the pages in above network in similar fashion. Now two dimension matrix is used in order to store this information where row represents the pages and column represents keywords. Hence, each cell in the matrix represents presence or absence of the keyword. Therefore, if page is having a keyword then cell contains a 1 or if it is not then a 0.

3.4 Web log Pre-Processing

Here web log dataset needs to be pre-processed. This can be understood as "https://mail.google.com/mail/page/ fold/index.html" where https, / , . , etc. are the noises in the dataset. Removal of this noise is done during this step of pre-processing [14, 15]. Two more tasks are accomplished during this step; first is to put pages in sequence then in second phase unique number is assigned to each page. This can be understood as ['mail', 'page', 'fold', 'index'].

3.5 Markov Model

In this work third order Markov Model is used where pattern from each session is found and then the proper frequency is calculated [9].

Table 1(a): Web session after preprocessing

Sessions
P1, P2, P3, P4, P5
P3, P2, P5, P1, P4
P1, P4, P2, P3, P4, P5
P2, P1, P3, P4, P5

Table 1(b): Markov Model from the dataset.

Pattern	**Target**	**Frequency**
P1, P2, P3	P4	1
P1, P2, P3	P5	2
P2, P3, P4	P5	2
P2, P1, P3	P4	1

3.6 Frequent Web Access Pattern (FWAP)

In this step frequent web patterns are filtered from the Markov Model. Some threshold value is set in this step for deciding the frequency of the webpage. As the order of

Markov Model increases threshold value needs to be reduced because it decreases the number of patterns. Therefore, all patterns having minimum frequency are considered as frequent web access patterns.

3.7 Training Module

Step 1: Collect contents of Web-pages
Step 2: Pre-process web contents
Step 3: Extract terms from Web-page
Step 4: Build the semantic network – TermNetWP

3.8 Frequent Web Access Pattern from Web log

Step 1: Collect Web logs
Step 2: Pre-Process the Web Logs
Step 3: Generates Markov Model
Step 4: Filter frequent FWAP

3.9 Testing Module

Step 1: Identifies a set of currently viewed terms in past page sequence of random user.

Step 2: Next page selection is done with the help of FWAP of Markov Model.

Step 3: Find term base similarities by selecting web page from Markov Model By using 1st-TermNet given (frequent view term pattern) FVTP. Here user FVTP F and page keywords comparison is done. So 1st-TermNet is the ratio of number of similar keywords with total number of keywords present in the F vector.

4. Experiment and Results

4.1 Experimental Setup

MATLAB 2012 version software was used in order to conduct experiment and measure evaluation results. The tests were performed on a 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM running under Windows 7 Professional operating system.

4.2 Evaluation Parameters

Accuracy: It is the ratio of number of correct page predictions to the total number of predictions.

$$Accuracy = \frac{Correct\_\Pr ediction}{Total\_Number\_of\_\Pr ediction}$$

Satisfaction: If a page is predicted and it is not opened by the concurrent sequence of user, but later it is open then it is considered as satisfaction. This is the ratio of total number of satisfaction to the total number of predictions.

$$Satisfaction = \frac{Satisfaction\_Number}{Total\_Number\_of\_\Pr ediction}$$

Execution Time: It is the execution time of the proposed or comparison method for making page predictions.

4.3 Results

Accuracy for prediction at different dataset sizes is as shown in Table 2. It is observed that the accuracy of the method proposed in this paper is better than the previous work [6].

Table 2: Comparison Results representing proposed work.

Accuracy Comparison		
Session size	**Previous[6]**	**Proposed Work**
4000	0.4808	0.5105
5000	0.4872	0.5106
6000	0.4970	0.5143

Table 3 shows that satisfaction in the proposed method is just 0.0013 percent less as compared to previous work. This is because accuracy percent of proposed work is 0.02347 percent higher and hence, satisfaction is little more in previous work.

Table 3: Comparison Results representing proposed work.

Satisfaction Comparison		
Session size	**Previous[6]**	**Proposed Work**
2000	0.6910	0.6895
4000	0.6910	0.6894
6000	0.6913	0.6905

As satisfaction and accuracy difference is about 18 times large hence, the proposed work for prediction at different dataset sizes is better as compared to the previous work [6]. Table 4 shows that execution time in the proposed work is little higher as compared to previous work but it is just 1000th part of the second and it can be considered looking at the enhanced accuracy.

Table 4: Second Comparison Results representing proposed work

Execution Time Comparison(Time in second)		
Session size	**Previous Work [6]**	**Proposed Work**
2000	0.033132	0.037885
4000	0.034837	0.03817
6000	0.034148	0.038249

5. Conclusion

Internet has become the need of the modern world by providing lots of services and tools. So to increase its efficiency is primary requirement of the researchers. This paper has contributed the page prediction work by utilizing the weblog and web content features. Here web content is used for developing the relation between the terms in form of term network. In similar fashion weblog is used to find FWAP. It can be concluded from the tables presented in the result section that the proposed model provides better results as compared to previous models on different evaluation parameters. This work will be carried forward to increase the efficiency by using other features.

References

[1] A. Harth, M. Janik, and S. Staab, "Semantic Web Architecture," in Handbook of Semantic Web Technologies, J. Domingue, D. Fensel, and J. A. Hendler, Eds.: Springer-Verlag Berlin Heidelberg, pp. 43-75, 2011.

[2] A. Balamash, M. Krunz, and P. Nain, “Performance analysis of a client-side caching/prefetching system for web traffic,” Computer Networks, Vol. 51, No. 13, pp. 3673-3692, 2007.

[3] D. Razdan, “The Next Page Access Prediction Using Makov Model,” International Journal of Electronics Communication and Computer Technology (IJECCT) , Vol. 1, No 1, 2011, ISSN: 2249-7838.

[4] G. Kollias, E. Gallopoulos and A. Grama, “Surfing The Network for Ranking by Multidamping,” IEEE Transactions On Knowledge and Data Engineering, 2014.

[5] I. Zukerman, D. W. Albrecht & A. E. Nicholson, “Predicting user’s requests on the WWW,” Proc. of the Architecture on the Limits of Reducing User's Perceived seventh international conference on User modeling, pp. 275-284, 1999.

[6] J. Domenech, J. Sahuquillo, J. A. Gil and A. Pont, “The Impact of the Web Prefetching Latency,” Proc. of the International Conference on Web Intelligence, 2006.

[7] J. M. Gascuena, A. Fernandez-Caballero and P. Gonzalez, "Domain Ontology for Personalized E-Learning in Educational Systems," In Proceedings of the Sixth IEEE International Conference On Advanced Learning Technologies, pp. 456 -458, 2006.

[8] L. Fan, P. Cao, W. Lin and Q. Jacobson, “Web Pre-Fetching Between Low-Bandwidth Clients and Proxies: Potential and Performance,” Proc. of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, pp. 178-187, 1999.

[9] M. A. Awad and I. Khalil, “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, August 2012.

[10] M. Eirinaki, D. Mavroeidis, G. Tsatsaronis, and M. Vazirgiannis, "Introducing Semantics in Web Personalization: The Role of Ontologies," M. Ackermann et al. (Eds.): Semantics, Web, and Mining, Springer-Verlag Berlin Heidelberg, pp. 147 – 162, 2006.

[11] R. Lempel and S. Moran “The Stochastic Approach for Link-Structure Analysis (SALSA) And The Tkc Effect,” Proceedings of The 9th International World Wide Web Conference, 2000.

[12] S. Grimm, A. Abecker, J. Völker, and R. Studer, “Ontologies and the Semantic Web,” in Handbook of Semantic Web Technologies, J. Domingue, D. Fensel, and J. A. Hendler, Eds.: Springer-Verlag Berlin Heidelberg, pp. 507-580, 2011.

[13] T. T. S. Nguyen, H. Y. Lu and J. Lu, “Web-page Recommendation based on Web Usage and Domain Knowledge,” 1041-4347/13/$31.00 © 2013 IEEE.

[14] T. M. Kroeger, D.E. Long and J. C. Mogul, “Exploring the Bounds of Web Latency Reduction from Caching and Pre-fetching,” Proc. of the 1st USENIX Symposium on Internet Technologies and Systems, 1997.

[15] Z. Liao, Y. Song, Y. Huang, L. W. He, and Q. He, “Task Trail: An Effective Segmentation of User Search Behavior,” IEEE transactions on knowledge and data engineering, Vol. 26, No. 12, December 2014.

[16] Y. Yalan, Z. Jinlong and Y. Mi, "Ontology Modeling for Contract: Using OWL to Express Semantic Relations," EDOC '06. 10th IEEE International Distributed Object Computing Conference, pp. 409-412, 2006.