User Next Web Page Recommendation using Weight based Prediction

Arvind Verma
PG Student
Vikrant Institute of Technology and Management, Indore, India

Balwant Prajapat
Assistant Professor
Vikrant Institute of Technology and Management, Indore, India

ABSTRACT
The World Wide Web is a source of knowledge; the knowledge is extracted from the web data. Web data is available in direct from normal web as contents to user and/or in direct forms to as the web access logs. For the web usage pattern analysis the web access logs are analysed. Web usage data used in various applications of web masters, user data recommendations, web pre-fetching and caching. In this paper using the web access log analysis, web next page recommendation system is introduced. The presented technique involves data personalization, user behavioural analysis and access patterns for recommendations.

The proposed web page recommendation system contains the K-means algorithm for finding similar access patterns of the user sessions. Additionally for classification and prediction the KNN algorithm is implemented. The model also incorporate the similar user access pattern data which is belongs from the other user therefore the proposed model also predicts the rarely accessed patterns. Thus to make the recommendations web usages data is personalized, based on URL frequencies, user navigational frequencies, session based data analysis and time based data analysis. Additionally to combine these parameters a weighted technique is used.

The proposed recommendation system is implemented using JAVA technology. And their performance in terms of accuracy, error rate, space complexity and time complexity is estimated. The experimentation with increasing amount of data provides more accurate results and also consumes less computational resources. Therefore the proposed data model is adoptable for accuracy and efficiency both.

Keywords
Web usages mining, recommendation, next web page prediction, implementation, results analysis

1. INTRODUCTION
World Wide Web contains a huge amount of knowledge and data, some of the knowledge is distributable using the contents of web pages and some of the information is not directly obtained from web pages [1]. The hidden data is available in terms of access logs or the web links organizations. The mining of such kind of web data is termed as the web mining. Therefore the web mining is an implementation of the data mining algorithms for extracting the information from web data. According to the availability of data the web mining techniques are classified in three main classes. The content of web pages are analysed using content mining, web access log are analysed under web usages mining and the connectivity and

Structure of the web pages are analysed under the structure mining techniques [2].

The paper leads to investigate about the web usage mining techniques and their applications. Additionally the key focus is made on web usages mining their algorithms. The web usage mining has a number of applications among them the web recommendation system is a popular application. Using this analysis the user navigational patterns, new trends on applications, user interest and others can be predictable. Therefore a user next web page recommendation system [3] is proposed for design in this study. The recommendation systems are frequently used with social networking, ecommerce and other applications.

The web usage data is obtained from web server's log files. The web access data can also be obtained from the proxy server access logs, ISP (internet service provider's logs) or the user cookies [4]. Among them the web server generated log files are much common in web usage mining techniques. Basically Web servers are internet infrastructure that provides hosting of web application. The web hosting servers store web documents, and that are distributed through web servers according to user request. For each user request server write the response of the system and prepare an entry on the web access server log. This log file contains a number of attributes, such as user IP address, time stamp, requested URL, responses data, browser information and other similar attributes. This data is not published publically due to security and privacy issues. But that data have expensive and private information, about the user navigations and the page visiting patterns. This information is much useful in developing various applications such as recommendation systems.

The key aim of this paper is to develop an accurate recommendation system. The proposed recommendation system is developed to enhance the accuracy of recommendation and also reduces the resource consumption during the computation. Therefore a novel methodology for enhancing the given technique is presented and developed. Further section provides the details about the proposed recommendation system.

2. PROPOSED DATA MODEL
The recommendation systems are frequently used in e-commerce web applications to enable users to find appropriate data from web sites. The use of recommendation systems increases the relevancy in results according to the user needs and behavior of navigation. That systems also able to predict most frequent pattern according to the user behaviour analysis. Thus there are two basic elements are required to predict the most desired data from available set of data. First user navigational behavior and second the predictive technique by which the model approximate the most probable data according to the user behavior.

On the other hand the user navigational behaviour can be fluctuating according to the time [5] and their navigational

environment that directly affect the user needs and their behavior of data requirements [6]. Thus the proposed technique required to involve the behavioral changes in navigational patterns [7]. In addition of that similar navigational behavior from others can also help to find the desired contents. Thus the proposed data model incorporate the time based behavioral observations and the similar navigational pattern based behavior for predicting the most appropriate data prediction. The proposed system can be understood using figure 1

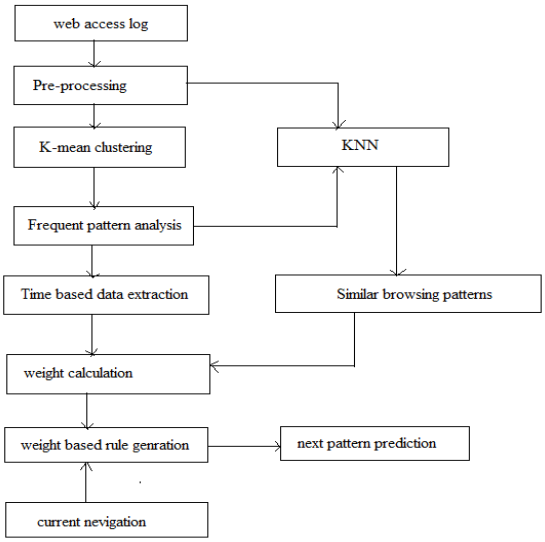

Figure 1 Proposed Model

Web access log: that is an input to the proposed system. That is a web access log file which contains different kinds of attributes i.e. IP address, time stamp, requested URL, browser information and others. Among them some of the data is required for developing the proposed recommendation system and not all the attributes are used. Therefore the pre-processing of web access log required to extract the required information from web log file.

Pre-processing: in this phase the input web log file is evaluated and the required attributes form the web access log is extracted. Remaining attributes of the web access log file is removed or leaved form the original input file.

K-means clustering: the extracted data from web access log file is used in this phase. In this phase the k-means clustering is applied on the data to prepare group of data according to the user IP address. Additionally that grouped data is according to the IP address represents the individual user's web browsing patter.

Frequent pattern analysis: the individual selected IP address based URLs are counted in this phase for finding the most frequent accessed URLs. After that their frequencies are computed for utilizing with the next phase of computation. For example the user which having the IP address 192.168.1.3 are accessed the following URLs.

Table 1 Example of Frequent Pattern

IP address	Web page	Total counts
192.168.1.3	Google	10
192.168.1.3	Facebook	12
192.168.1.3	Yahoo	8

For computing the frequency of the individual web pages accessed by the user the following formula can be used.

$$f_{URLs} = \frac{1}{N}\sum_{i=1}^{N} URLCount_i$$

Where N is the total number of pages accessed, therefore the above discussed table can be given by the table 2 with the URLs frequency.

Table 2 Example of Frequency Count

IP address	Web page	Total counts	f_{URLs}
192.168.1.3	Google	10	0.33
192.168.1.3	Facebook	12	0.4
192.168.1.3	Yahoo	8	0.26

Time based data extraction: in this similar manner the time based analysis is performed for each URLs navigated by the individual user. In order to compute the time based we can consider an example suppose the there are three URLs as demonstrated table 3 and the user works on these URLs and in all the sessions for the specified time period. Thus the total stand-up time for a URL is estimated by the following formula.

$$T_{URLs} = \frac{1}{N}\sum_{i=1}^{N} URLTime_i$$

Table 3 URL Access Time

IP address	Web page	Time in sec
192.168.1.3	Google	130
192.168.1.3	Facebook	122
192.168.1.3	Yahoo	228

Table 3 URL Access Time

IP address	Web page	Time in sec
192.168.1.3	Google	130
192.168.1.3	Facebook	122
192.168.1.3	Yahoo	228

Thus, by using the formula that can be reorganized in the following manner

Table 4 URL Time Factor

IP address	Web page	Total time	T_{URLs}
192.168.1.3	Google	130	0.27
192.168.1.3	Facebook	122	0.254
192.168.1.3	Yahoo	228	0.475

KNN: the KNN (k-nearest neighbour) algorithm is a similarity finding technique which usages the Euclidean distance for approximation of the similar patterns. Here for computing the similarity the threshold is set statically for 25% of total distance.

Similar browsing patterns: the outcome of the KNN (k-nearest neighbour) algorithm is extracted and merged with the

user's patterns data which is used in further step to approximate the most probable URLs for prediction.

Weight calculation: in order to compute the strength of next upcoming URL the computed factors are combined together. To compute the weights of the URL the following formula is used.

$$W = f_{URL} * w_1 + T_{url} * w_2$$

Where $w_1 and\ w_2$ are the weighting factors for regulating the outcomes of prediction.

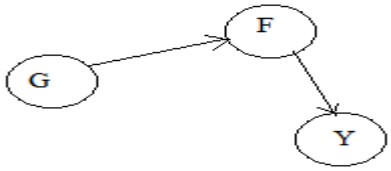

Figure 2 Weighted Graph Presentations

Weight based rule generation: after computing the weights of the extracted data the entire navigated URLs are demonstrated in the form of weighted graph for the above given example the graph is demonstrated using figure 2.

Current navigation: that is the current web page which is navigated the targeted user. Based on this web page it is required to find the prediction using the proposed data model. Therefore the system accepts the user input and predicts the next web page for the user.

Next page prediction: according to the current user input pattern system generate the prediction of next web page. According to out example the user having the current web page as Facebook then the next predicted URL is yahoo.

Functional aspects According to the figure 1 the web access log file is consumed for analysing the user behavior and the navigational pattern. Therefore the system accepts the web access log file and then the K-mean clustering is applied on data this clustering generates the subsets of data according to the user IP addresses. Now the individual IP address data is analysed for finding the user frequent access pattern from user personalized data.

Now the similar user behaviors are searched from the web log using KNN algorithm which analyze data in distance based function and most nearest patterns are listed with the help of user frequent patterns. Now the frequent patterns from the data are classified according to the time based manner and from both the similar pattern and time based patterns are used to calculate the weights for the obtained pattern. Using the weights and user navigational data a weighted graph is constructed which is used to discover the current browsing pattern and next most probable data.

3. RESULTS ANALYSIS

The given section includes the performance analysis of the implemented algorithms for the recommendation systems. Therefore the performance of algorithms are evaluated and compared in this section.

3.1 Training Time

The amount of time consumed during the training of the system is termed as the training time of the algorithm. Figure 3 shows the training time of the algorithms in terms of milliseconds. Additionally during the experimentation for extracting the performance the size of dataset is increases and according to the estimated times the best obtained results are reported. In order to represent the performance of system the X axis of diagram contains the dataset size and the Y axis shows the estimated time in milliseconds. According to the obtained results the traditional algorithm consumes less time as compared to the proposed algorithm

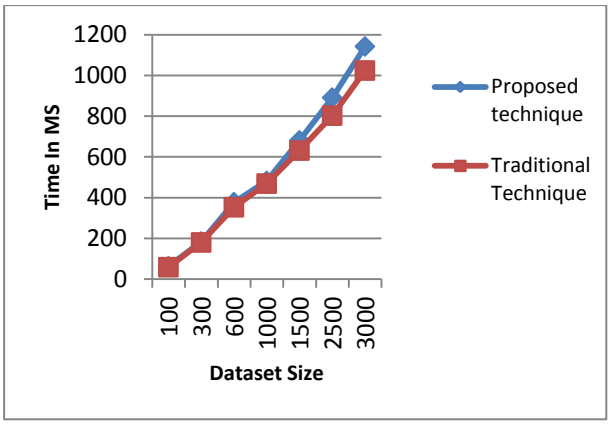

Figure 3 Training Time

The figure 6 contains the evaluated performance of the system in terms of algorithm accuracy of prediction. In this diagram the amount of accurate pattern recognition is given using Y axis and the X axis shows the amount of data to be train for the predictor. According to the comparative results the performance of the proposed algorithms remains much consistent and increasing as compared to the traditional algorithm. In order to evaluate the accuracy of algorithm a fixed amount random URLs are extracted from database and the next values are used as the actual class label during evaluation of data.

According to the observations the proposed algorithm consumes additional time as compared to the traditional algorithm. This because for improving the accuracy in prediction the system needs to calculate the additional parameters as compared too traditional method thus the time complexity of proposed algorithm is higher as compared to traditional system.

3.2 Recommendation Time

The amount of time required to evaluate the URLs form making accurate prediction is termed here as the recommendation time. That recommendation time of both the system in comparative manner is demonstrated using figure 4. In this diagram the amount of data is given in the X axis and the Y axis shows the performance obtained in terms of milliseconds. According to observations the amount of time during prediction is not much fluctuated and not also affected by the amount of data to be process. The comparative results of the systems shows the effectiveness of the proposed technique that consumes less time for computing the predicted URL as compared to traditional approach. Because during the time based data clustering reduces the amount of data to process. Thus the prediction is much frequent as compared to the traditional approach

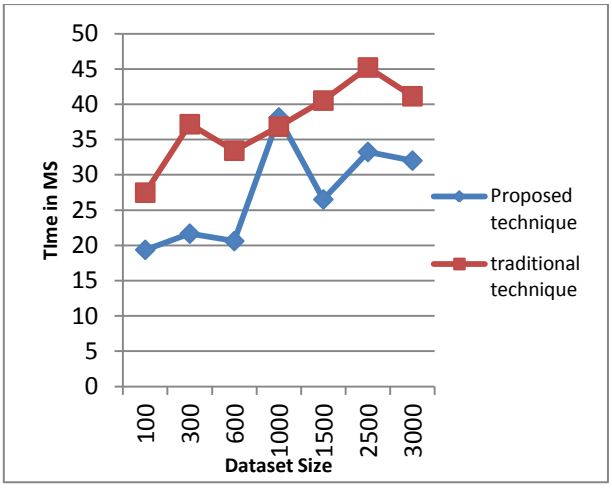

Figure 4 Recommendation Time

3.3 Memory Used

The memory consumption shows the amount of main memory required to process the algorithm task. That is also known as the space complexity in algorithm study. The figure 5 shows the memory consumption of the system with increasing size of training dataset. The amount of data is given using the X axis and the Y axis shows the amount of consumed memory during experimentation with respective amount of data in terms of kilobytes. According to the experimented results the amount of memory is similar and not more fluctuating. But the respective comparison the proposed algorithm is efficient than the traditional approach of recommendation.

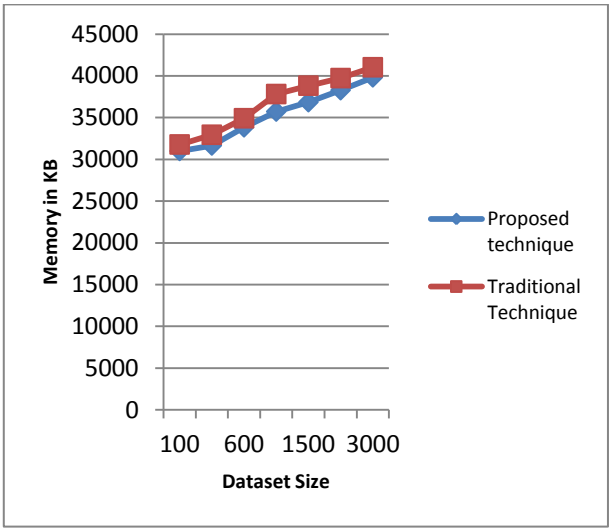

Figure 5 Memory Consumption

3.4 Accuracy

The accuracy of the predictive algorithm provides the amount of generated prediction is similar to actual outcomes. Therefore that can also be defines as the amount of correctly recognized patterns among the total samples produces to test. That can also be evaluated using the following formula:

$$\text{accuracy} = \frac{\text{total correctly identified patterns}}{\text{total input samples}} \text{X100}$$

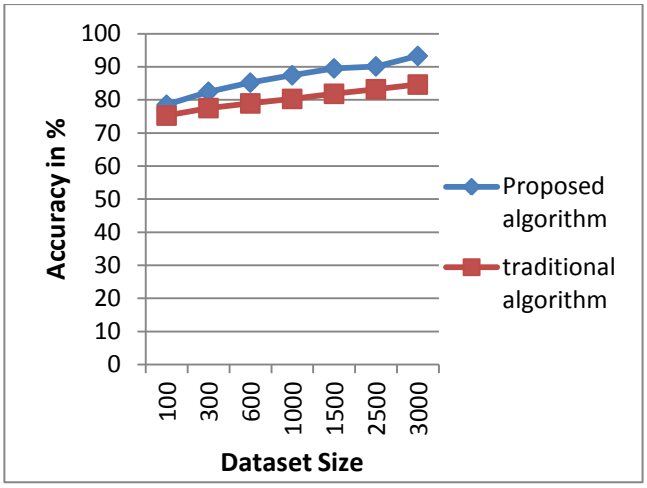

Figure 6 Accuracy

The figure 6 contains the evaluated performance of the system in terms of algorithm accuracy of prediction. In this diagram the amount of accurate pattern recognition is given using Y axis and the X axis shows the amount of data to be train for the predictor. According to the comparative results the performance of the proposed algorithms remains much consistent and increasing as compared to the traditional algorithm. In order to evaluate the accuracy of algorithm a fixed amount random URLs are extracted from database and the next values are used as the actual class label during evaluation of data.

4. CONCLUSION

Analysis and extraction of targeted data from a raw set of data is known as the data mining or data processing. In the similar ways the extraction of target patterns from the web usage data is known as the web usage mining. In this presented study first the web usages mining is explored and the different attributes relevant to web usages mining is evaluated. In order to demonstrate the web usage mining and their application in real world the web recommendation system is selected to implement.

The web recommendation system is concept by which the previous or historical user navigation data is analyzed and based on the most likely navigated technique the next web page access is predicted. In order to find an accurate and efficient data model for web recommendation a number of research articles are studied and a more promising model as given in [1] is selected for further study. The given model not only consumes the user accessed web pages that also evaluate the similar access patterns also. This concept is used to predict the web pages that rarely accessed by the user.

Therefore to design the more accurate model a weighted method is proposed. The proposed recommendation system first find the user accessed patterns form the web log and similarly the different users accessed data is also extracted using the K-mean clustering algorithm. Additionally the time based data clustering is also prepared to more refinement on patterns. After evaluation of all three parameters namely user navigational frequency, time based frequency and session wise data access pattern a combine weight for all the URLs are evaluated. These weights are further sorted and by the rank of weights the next most possible web page is predicted.

The implementation of the proposed system is performed using the JAVA framework and their performance in terms of time consumption and memory consumption is evaluated. Finally by

the cross validation process the system performance in terms of accuracy is measured. According to the obtained results the proposed system delivers the most accurate and efficient results as compared to the traditional algorithm. The performance summaries of both the algorithms are given using below given table 7.

Table 7 Performance Summary

S. No.	Parameters	Proposed system	Traditional system
1	Memory consumption	Adoptable	Adoptable
2	Training time	High	Low
3	Recommendation time	Low	High
4	Accuracy	High	Low

The proposed recommendation system is developed and their performance is evaluated. During the development and experiments the proposed model is found more effective and accurate as compared to traditional methods but the training time of the system in higher due to evaluation of additional parameters. In near future the proposed concept is explored more for reducing the training time.

5. REFERENCES

[1] Lina Yao and Quan Z. Sheng, Aviv Segev, Jian Yu, "Recommending Web Services via Combining Collaborative Filtering with Content-based Features", 2013 IEEE 20th International Conference on Web Services, 978-0-7695-5025-1/13 $26.00 © 2013 IEEE

[2] Rana Forsati, Mohammad Reza Meybodi, Afsaneh Rahbar, "An Efficient Algorithm for Web Recommendation Systems", 2009 IEEE/ACS International Conference on Computer Systems and Applications

[3] Kavita Sharma, Gulshan Shrivastava, Vikas Kumar, "Web Mining: Today and Tomorrow", 2011 3rd International Conference on Electronics Computer Technology (ICECT 2011), 978-1-4244-8679-3/$26.00 2011 IEEE

[4] Rajni Pamnani, Pramila Chawan, "Web Usage Mining: A Research Area in Web Mining", Proceedings of ISCET 2010.

[5] Quanyin Zhu, Hong Zhou, Yunyang Yan, Jin Qian and Pei Zhou, "Commodities Price Dynamic Trend Analysis Based on Web Mining", 2011 Third International Conference on Multimedia Information Networking and Security, 978-0-7695-4559-2/11 $26.00 © 2011 IEEE

[6] K. Srinivas, P. V. S. Srinivas, A. Govardhan, V. Valli Kumari, "Periodic Web Personalization for Meta Search Engine", IJCST Vol. 2, Issue 4, Oct. - Dec. 2011

[7] Bussa V. R. R. Nagarjuna, Akula Ratna babu, Miriyala Markandeyulu, A. S. K. Ratnam, "Web Mining: Methodologies, Algorithms and Applications", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[8] Sneha Prakash, "Web Personalization using web usage mining: applications, Pros and Cons, Future", International Journal of Computing Science and Information Technology, 2015, Vol.3,Iss.3, 18-26

[9] Neha Sharma & Pawan Makhija, "Web usage Mining: A Novel Approach for Web user Session Construction", Global Journal of Computer Science and Technology: E Network, Web & Security Volume 15 Issue 3 Version 1.0 Year 2015

[10] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Saudi Computer Society, King Saud University, Applied Computing and Informatics, 2015 Production and hosting by Elsevier B.V

[11] Haidong Zhong, Shaozhong Zhang, Yanling Wang, Shifeng Weng and Yonggang Shu, "Mining Users' Similarity from Moving Trajectories for Mobile Ecommerce Recommendation", International Journal of Hybrid Information Technology Vol.7, No.4 (2014), pp.309-320

[12] A. Tejeda-Lorente, C. Porcel, E. Peisc, R. Sanz, E. Herrera-Viedma, "A quality based recommender system to disseminate information in a University Digital Library", Information Sciences (2013),

[13] Renuka Mahajan, J. S. Sodhi, Vishal Mahajan, "Web Usage Mining for Building an Adaptive e-Learning Site: A Case Study", International Journal of e-Education, e-Business, e-Management and e-Learning, Manuscript submitted July 10, 2014; accepted August 29, 2014.

[14] Zahid Ansari, A. Vinaya Babu, Waseem Ahmed and Mohammad Fazle Azeem, "A Fuzzy Set Theoretic Approach to Discover User Sessions from Web Navigational Data", IEEE Recent Advances in Intelligent Computational Systems (RAICS) 2011, 978-1-4244-9478-1/11/$26.00 c 2011 IEEE

[15] Ricardo Terra, Marco Tulio Valente, Krzysztof Czarnecki and Roberto S. Bigonha, "A recommendation system for repairing violations detected by static architecture conformance checking", SOFTWARE – PRACTICE AND EXPERIENCE, Softw. Pract. Exper. 2015; 342, Published online 25 September 2013.

[16] I. Petrović, P. Perković and I. Štajduhar, "A Profile- and Community-Driven Book Recommender System", 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), MIPRO 2015/CTI.