A Survey on: Various Features and Techniques of Web Page Prediction

[1]Megha Raghu, [2]Megha Vashishtha

[1]*Mtech Scholar,* [2]*Assistant Prof.*

Abstract—Web data mining is the process of applying data mining techniques to Web data. Research in this area has the objectives of helping e-commerce businesses in their decision making, assisting in the design of good Web sites and assisting the user when navigating the Web. Keeping this goal paper has a brief survey of web page prediction with different approaches of features and methods were explained. Paper has discussed different techniques with their requirement as well. Here various comparison parameters for the algorithm comparison was done in this work as well.

Keywords:-Data mining, Social network, Product ratin, Product recommendation.

I. INTRODUCTION

Modeling the consumer web navigation behavior is now the tough task since the growth of the internet is growing rapidly. Web Usage Mining would be the field regarding web exploration which handles finding this interesting utilization pattern through the logging data. The signing information is usually stored within a file generally known as web sign file. Web sign file contains large amount of information such as IP target, date, period, web web page requested and so forth. Web sign file might be retrieved coming from web server, proxy server or even client area. This world-wide-web log contains large amount of information so it is preprocessed prior to modeling. The web log file is preprocessed and converted into the string of individual web navigation sessions. The web navigation session would be the sequence of web site navigated by a user through time window. The individual navigation program is last but not least modeled via a model. After the user navigation model is usually ready, the exploration task can be performed for seeking the interesting style. Modeling regarding web log would be the essential job in world-wide-web usage exploration. The prediction accuracy can be carried out through the modeling the web log with the accurate product. Markov product is traditionally used for modeling the Consumer web navigation sessions. The more common Markov product is having its own constraint. First-order Markov product is less complex even so the accuracy is usually low as a result of lack of looking at the level.

As specialist continue to the second order Markov it's precise in correlation with the main request Markov technique even so scoped which forecasting accuracy is less and in addition the time trouble get expanded.

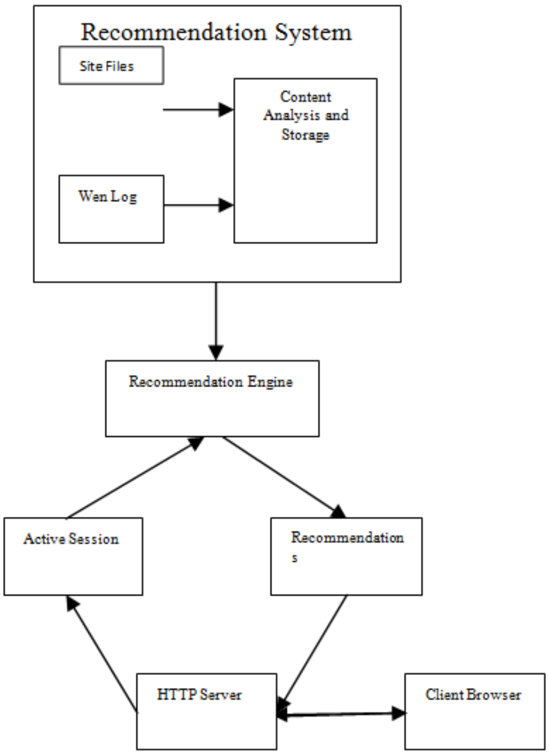

Fig. 1 Architecture of web page prediction.

There are generally wide application parts of the exploration of individual web route conduct with web use mining. The exploration of individual web route conduct may help for enhancing the organization of the site and advance of internet execution by basically pre-getting and reserving basically the most likely next site in push ahead. Web Personalization, Adaptive web destinations are a couple of the applications with respect to web use mining. Web usage digging offers rules for enhancing online business to oversee business specific issues, for example, client interest, client storage, crosses item deals, and purchaser pattern.

While browsing the Web Pages by the user, he/she leaves some valuable information in web log files. Web Log Files are used to trace user's web navigation behavior, through which researcher can easily analyze which type of information user frequently navigates from the web sites [2].

II. RELATED WORK

In [1] The paper has depicted the scientific classification of suggestion framework strategies and talked about open difficulties, issues being developed of proposed frameworks. A contextual investigation of some website page suggestion systems in view of semantic web utilization mining has additionally been performed. It is watched that the greater part of site page proposal framework structures beat the new page issue and a couple of defeat the over-specialization, adaptability and scantily issues. Demonstrate based communitarian separating and memory based shared procedures are dominative in the advancement site page suggestion frameworks.

In [2] researcher propose a hybrid recommender system where a parametric integration of co-occurrence based clusters and content based clusters provides better recommendation even for users having very less preference information, alleviating the sparsity problem. Item clustering based on co-occurrence information makes the system unique. Recommending both related and similar items are possible using the proposed hybrid approach. Observed preference list length based parameter tuning makes the recommender system learned accurately and efficiently. Researcher applies clustering based technique to resolve scalability problem.

In [3] researcher presented a bitwise frequent pattern mining (or web mining) algorithm—called BW-mine—that finds web surfer patterns for web page recommendation. To avoid these drawbacks, researcher present an alternative frequent pattern mining (or web mining) algorithm called BW-mine in this paper. Evaluation results show that our proposed algorithm is both space- and time-efficient. Furthermore, to show the practicality of BWmine in real-life applications, researcher apply BW-mine to discover popular pages on the web, which in turn gives the web surfers recommendation of web pages that might be interested to them.

In [4] proposed approach has improvised the Content Based Relevancy Algorithm which matches the keywords and the content words. A Differential Adaptive PMI algorithm that computes the semantic similarity between the query words, keywords and content words differentially with heterogeneous thresholds is implemented. The keywords and the content words are obtained from the URLs in the URL Base which contains Web Usage Information. A new strategy of Adaptive PMI is proposed for semantic similarity computation. The approach computes the semantic heterogeneity at varied thresholds to ensure that the Web Pages recommended are highly relevant to the query. Also, the strategy that is formulated is a semantic methodology for Web Page Recommendation.

In [5] The proposed system consists of three knowledge based models. To improve performance in future key information extraction algorithm is used and comparison takes place between results obtained from applying only three knowledge based models and models along with key information extraction algorithm. Recommendation is given to page from weblog records. Experimental result shows that recommendation for webpage is better by using proposed system than existing system and execution time required for the proposed system is less as compared to existing system and accuracy of proposed system is more than existing system. This can be obtain by using key information extraction algorithm.

In [10], researcher introduced a class of Markov method based forecast calculations that are gotten by specifically taking out a substantial portion of the conditions of the All-Kth - Order Markov method. Their tests on an assortment of datasets have demonstrated that the subsequent Markov models have a low state-space unpredictability and in the meantime accomplish generously preferable correctnesses over those got by the customary calculations.

In [11] researcher have effectively consolidated a few compelling forecast models alongside space learning misuse to enhance the expectation precision. Notwithstanding, the module bears costly preparing and expectation overheads as a result of the substantial number of names/classes engaged with the WPP.

In [12] exhibited Bayesian models for two things like learning and foreseeing key Web route designs. Rather than displaying the general issue of Web route they concentrate on key route designs that have functional esteem. Besides, rather than creating complex models they exhibit natural probabilistic models for learning and forecast. The examples that they consider are: short and long visit sessions, page classes went to in first N positions, scope of site hits per page classification, and rank of page classes in first N positions. They learn and anticipate these examples under four settings comparing to what is thought about the visit sessions (client ID and additionally timestamp).

In [13] enhanced the Web page get to forecast precision by incorporating each of the three expectation models: Markov model, Clustering and affiliation rules as per certain requirements. Their model, IMAC, coordinates the three models utilizing lower order Markov demonstrate. Bunching is utilized to gather homogeneous client sessions. Low request Markov models are based on grouped sessions. Affiliation rules are utilized when Markov models couldn't clarify expectations. The coordinated model has been shown to be more exact than every one of the three models executed independently, and additionally, other incorporated models. The coordinated model has less state space many-sided quality and is more exact than a higher request Markov display.

In [14], broke down and considered Markov model and all-K th Markov show in Web forecast. They proposed another changed Markov model to ease the issue of adaptability in the quantity of ways. They have utilized standard benchmark informational indexes to break down, analyze, and exhibit the adequacy of our strategies utilizing varieties of Markov models and affiliation run mining. Their examinations demonstrate the adequacy of adjusted Markov show in diminishing the quantity of ways without trading off exactness. Furthermore, the outcomes bolster their examination decisions that exactness enhances with higher requests of all Kth display.

III. TECHNIQUES

Markov Modal: In [6] web log feature is utilize to generate different orders of the web markov modal. Here as per user current user web page movement prediction of next page is done by utilizing morkov modal which give required page. Here as per the length of the user markov orders are use so storage of different size of markov modal help in different stages of the proposed work. In case of higher order markov modal if this fail then lower order markov modal will handle the situation and send the next possible page. So this step of finding the next page in lower order is continuing until possible next page is not obtained. In order to understand this consider an example, let us assume an user session s = {P1, P5, P6}, prediction of all-Kth model is performed by consulting third-order Markov model. If the prediction using third-order Markov model fails, then the second-order Markov model is consulted on the session x_ = x − P1 = <P5, P6>. This process repeats until reaching the first-order Markov model. Therefore, unlike the basic Markov model, the all-Kth-order Markov model achieves better prediction, and it only fails when all orders of the basic Markov models fail to predict.

Predict markov algorithm take session and modal number as input then find most frequent page. If it generates more than one page then, second feature will be predicted for the page selection which is keywords extracted from the web pages. There similar function take key_vector which is the collection of the keywords which is obtain from the previous page of the session, then compare the keywords of the pages in V vector. The most similar page will be the next target page of the session. This page is return to the function.

Computing HITS Algorithm [9]: In this algorithm two types of values are assigned on each page first is positive non zero weight and other is again a positive non zero hub weight. Here value of each weight are so assigned that by taking an square of the number it remain below or equal to 1. So a proper normalization of each value is done. Here page have high weight is consider as the important page or rank of the page is higher as compare to other existing page. Numerically, the mutually reinforcing relationship between hubs and authorities can be expressed as follows: if p points to many pages with large ά values, then it should receive a large h –value. In similar fashion if p is pointed to by many pages with large h -values, then it should receive a large ά-value. This motivates the definition of two operations on the weights, denoted by I and O. Given weights a p and hp, the I operation updates the ά -weights as follows, similarly the O operation updates the h-weights as follows

$$h_p \leftarrow \sum_{q:(p,q)} a_q$$

Thus, I and O operations are the basic means by which hubs and authorities reinforce one another. To find the desired ”equilibrium” values for the weights, one can apply the I and O operations in an alternating fashion, and see whether a fixed point is reached

SALSA: In [7] a random walk algorithm is proposed where bipartite hubs and authorities web graph is develop, and then proper movement is note by changing the web pages one by one. Here some of important nodes are chosen for the start of random walk, selection of those are done randomly. Now movement in walk is done by switching from one hub node to another hub node. Here selections of nodes are depend on the authority weight which are distribute as per the importance of hub. So markov modal will calculate the required weight probabilities. Let Fu be the set of pages u points to and Bu the set of pages that point to u.

$$P_a(i,j) = \sum_{k,k\in B(t)\cap B(j)} \frac{1}{|B(i)\|F(j)|}$$

Multi-Damping Method: In [14] Let Y is an adjacency matrix for the graph of nodes. Where i represent the node after which j node is choosen by the surfers with probability p’.

P’ = (Vj / V_total) = (number of logs contain j node after i node / total number of logs which contain i node)

Y(i,j) = p’

In this algorithm first Zk is calculate which is the damping coefficient & G(μ) is the Google matrix. Stochastic matrix S := P + Y. For a random web surfer about to visit the next page, the damping factor μ ∈ [0, 1] is the probability of choosing a link-accessible page. Alternately, with probability 1 − μ, the random surfer makes a transition to a node selected from among all nodes based on the conditional probabilities in vector v.

IV. EVALUATION PARAMETER

In order to evaluate this work there are different parameter present for the different techniques. The best parameter which suit this work is the precision where it give the value which is a measure of the prediction which is correctly identify by proposed model to the all the logs pass in the experiment. The other important measure is the Recall and F-score.

True Positive: When the system says page P1 and actual actual page is also P1.

True Negative: When the system says page P1 and actual page is also P2.

False Positive: When the system says no page and actual actual page is also P1.

$$\Pr ecision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\text{Re} call = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * \Pr ecision * \text{Re} call}{\Pr ecision + \text{Re} call}$$

In above true positive value is obtained by the system when the recommended web page is correct as per user choice. While in case of false positive system recommended web page is not correct as per user choice.

V. CONCLUSIONS

Mining World Wide Web has increased the dependency of the clients to make utilization of automated apparatuses for finding data assets and evaluate their usage patterns. As the next page prediction is the main motive of this work where one can generate the next page with the proper knowledge from the ontology and the web usage of the web. By the use of whole content keywords prediction of next page will be more efficient by improving the precision value, where precision is the evaluating parameter. This paper surveys various approaches with techniques of web page recommendation system. It is always desired that an algorithm will be proposed which will increase the accuracy of the page prediction for the web server, while response time be small.

REFERENCES

[1] Satyaveer Singh Mahendra Singh Aswal. ”Towards a Framework for Web Page Recommendation System based on Semantic Web Usage Mining: A Case Study. 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016)

[2] Mohammad Amir Sharif, Vijay V. Raghavan. “A Clustering Based Scalable Hybrid Approach for Web Page Recommendation”. 2014 IEEE International Conference on Big Data

[3] Fan Jiang Carson K. Leung(□) Adam G. M. Pazdor. “Web Page Recommendation Based on Bitwise Frequent Pattern Mining”. 2016 IEEE/WIC/ACM International Conference on Web Intelligence

[4] Gerard Deepak, J Sheeba Priyadarshini, M S Hareesh Babu. “A Differential Semantic Algorithm for Query Relevant Web Page Recommendation”. 2016 IEEE International Conference on Advances in Computer Applications (ICACA) 978-1-5090-3770-4/16/$31.00©2016 IEEE

[5] Ms.Priyaka Kolekar, Prof.Suchita Wakhade. “A novel approach to provide Web page recommendation using domain knowledge and web usage knowledge”.IEEE conference 2016.

[6] Mamoun A. Awad and Issa 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.

[7] R.Lempel, S.Moran,The stochastic approach for link-structure analysys (SALSA) and the TKC effect, Proceedings of the 9th International World Wide web Conference, 2000.

[8] Giorgos Kollias, Efstratios Gallopoulos, and Ananth Grama “Surfing the Network for Ranking by Multidamping”. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2014.

[9] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal www.ijsret.com, volume 3 issue 3.

[10] M. Deshpande, G. Karypis, “Selective Markov Models for Predicting Web Page Accesses,” ACM transactions on Internet Technology, volume 4, No.2, pp.163-184, May 2004

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

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

[13] F.Khalil, J. Li, H. Wang, “An Integrated Model for Next Page Access Prediction”, Inderscience Enterprises Ltd., 2009.

[14] Mamoun A. Awad and Issa Khalil, “Prediction of User’s Web-Browsing Behavior: Application of Markov Model,” IEEE Trans. Syst., Man,Cybern. A, Syst., Humans, volume 42, no. 4, pp., Aug. 2012.