International Journal of Advanced Technology in Engineering and Science
Vol. No. 09, Issue No. 04, April 2021
www.ijates.com
ijates
ISSN 2348 - 7550

A Study on User Web Surfing Recommendation Techniques and Features Extraction

Rajesh Ku.Nigam[1], Dr.Chandikaditya Kumawat[2], Dr.Manish Shrivastava[3]

[1] *Research Scholar, CSE Deptt., Mewar University, Rajasthan (India) (rajeshrewa37@gmail.com)*

[2] *Professor, CSE Deptt., Mewar University, Rajasthan (India) (chandikaditya@gmail.com)*

[3] *Professor, IT Deptt LNCT, Bhopal (India) (contct.manishshrivastav@gmail.com)*

ABSTRACT

A website success depends on visitor behavior as per published content, service, product, etc. To increase the engagement of visitor website should understand surfing pattern and present relevant pages. This analysis of page recommendation attract researcher to study web surfer and proposed methods for resolving this issue. This paper has provides a deep survey of web mining features with formulation to increase the efficiency of page recommendation. Various techniques, methodologies proposed by scholars are summarized in the paper. Techniques were chart with their benefits and drawbacks. Various evaluation parameters were also list in the paper for comparing methods of page recommendation.

Keywords - *Data Mining, Pattern Reorganization, Feature Extraction, Web mining, Clustering.*

I. INTRODUCTION

As the web users are growing day by day, the need of the networking world is becoming moderately high. So as to expand the clearness and promptness within the work great amount of labor depends on this web network [1]. This attracts several researchers for raising the performance of the network and lessens the latency time of the web, so things get less complicated and quick for the daily consumers. At this point hardware element is the means of optimizing the network however in parallel computer code additionally ought to be updated. This paper concentrates on optimizing the network power by learning the user actions for reducing the latency time of looking out the desired matter of specific interest. As websites are important supply of knowledge for pretty much all necessities, thus these necessities of individuals attract variety of individuals to produce varied services. However targeting the right client is basic demand of the service or business [2]. Analysis during this space has the purpose of serving to e-commerce businesses in their choices, helping within the style of fine websites and helping the user whereas navigating the net.

Even despite the fact that these days net users have created higher bandwidth connections, they still observe high latencies when navigating the net due to overloaded essentials, long note conversion times, and also the trip time. As a result, the reduction of the users perceived latency once browsing the net remains a vital analysis issue [3]. The reduction of the net users' perceived latency has become the topic of the many analysis efforts over the previous couple of years.

The extensively used techniques projected to lessen this latency are net caching, geographical duplication, and pre-fetching. Caching techniques are broadly put into practice currently these days since they win vital latency savings. Several transnational firms implement net replication by victimisation Content Delivery Networks [4] to lessen their websites access time however this resolution isn't possible because it is pricey and lots of small firms, organizations cannot afford it. Net pre-fetching techniques are reciprocally freelance to caching and replication techniques, so they will be applied along to attain a more robust net performance. Caching and replication techniques are widely enforced in globe; some studies have additionally investigated net pre-fetching in real environments.

Rest of paper was arrange in few section where second section has summarized different machine learning methods adopt by researcher to extract information. Third gives a brief explanation of different features used in different type of dataset, while fourth is collection of machine learning techniques. In fifth section different evaluation parameters were explained with there formulas.

II. RELATED WORK

Log Feature Based Approaches:

Awad et al. [5] proposed an approach hybrid utilizing Support Vector Machines, and the All-Kth Markov model, to determine calculation utilizing Dempster's rule. To boost the power of discrimination they apply attribute extraction of SVM. Adding together, during forecast this paper occupy domain information to decrease the number of classifiers for the development of correctness and the decrease of calculation time.

In [6] authors analyze and study Markov model and all-Kth Markov model in Web prediction. This paper proposed a new modified Markov model to alleviate the issue of scalability in the number of paths. In addition, this paper present a new two-tier prediction framework that creates an example classifier EC, based on the training examples and the generated classifiers. This paper shows that such framework can improve the prediction time without compromising prediction accuracy. This paper have used standard benchmark data sets to analyze, compare, and demonstrate the effectiveness of our techniques using variations of Markov models and association rule mining.

In [7] authors prove that polynomials of stochastic matrices can be expressed as products of Google matrices (matrices having the form used in Google's original PageRank formulation). Individual matrices in these products are parameterized by different damping factors. For this reason, this paper refer to our formulation as multidamping. This paper demonstrate that multidamping has a number of desirable characteristics: (i) for problems such as finding the highest ranked pages, multidamping admits extremely fast approximate solutions; (ii) multidamping provides an intuitive interpretation of existing functional rankings in terms of the surfing habits of model web users; (iii) multidamping provides a natural framework based on Monte Carlo type methods that have efficient parallel and distributed implementations. It also provides the basis for constructing new link-based rankings based on inhomogeneous products of Google matrices. This paper present algorithms for computing damping factors for existing functional rankings analytically and numerically.

Deepa and Raajan [11] executed the preprocessing methods to change the log file into client sessions which are appropriate for mining and decrease the range of the session file by sorting the least demanded pages utilizing the preprocessing method. Information Preprocessing is one of the significant missions before inserting mining algorithms. It changes the raw record file into client session. In this research, paper have in

brief established record file preprocessing and applied it in a CTI record file. Also, this paper create the review of the client session file. This paper have utilized filtering method to eradicate slightest demanded resources.

In [17] paper proposes a web page prediction method using a weighed support and Bhattacharya distance-based (WS-BD) two-level match. The major aim of the proposed method is to attain customer satisfaction. Initially, interesting sequential patterns are obtained using the weighed support that filters the sequential patterns obtained using a PrefixSpan algorithm based on the frequency, duration and recurrence of the web pages. Interesting sequential patterns are clustered using the proposed dice similarity-based Bayesian fuzzy clustering, and the web page is predicted using the two-level match based on Bhattacharya distance.

Content Feature Based Approaches:

Zhen Liao et. al. in [9] introduce "task trail" to understand user search behaviors. This paper define a task to be an atomic user information need, whereas a task trail represents all user activities within that particular task, such as query reformulations, URL clicks. Previously, web search logs have been studied mainly at session or query level where users may submit several queries within one task and handle several tasks within one session. Although previous studies have addressed the problem of task identification, little is known about the advantage of using task over session or query for search applications. In this paper, this paper conduct extensive analyses and comparisons to evaluate the effectiveness of task trails in several search applications: determining user satisfaction, predicting user search interests, and suggesting related queries.

In [16] authors introduces an improved product recommendation method for collaborative filtering, which is based on the triangle similarity. However, the downside of triangle similarity is that it only considers the common ratings of users. The proposed similarity measure not only focuses on common ratings but also consider the ratings of those items that are not commonly rated from pairs of users. The obtained similarity is further complemented with the user rating preference (URP) behavior in giving rating preferences.

Okura et al. [13] projected by utilizing GRUs to study further meaningful aggregation for client browsing history (browsed news), and suggest news articles with hidden feature model. In end result show a considerable development compared with the conventional word-based approach. This system has been fully deployed to online construction services and helping over ten million distinctive users everyday

Log and Content Feature Based Approaches:

In [8] paper proposes a novel method to efficiently provide better Web-page recommendation through semantic enhancement by integrating the domain and Web usage knowledge of a website. Two new models are proposed to represent the domain knowledge. The first model uses an ontology to represent the domain knowledge. The second model uses one automatically generated semantic network to represent domain terms, Web-pages and the relations between them. Another new model, the conceptual prediction model, is proposed to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge. A number of effective queries have been developed to query about these knowledge bases. Based on these queries, a set of recommendation strategies have been proposed to generate Web-page candidates.

In [15] authors proposed a recommender systems are introduced via the use of certain agents in order to provide extremely appropriate web pages for patients. The main feature of Particle Agent Swarm

Optimization (PASO) is that the creation of the algorithm is denoted by a set of Particle agents who cooperate in attaining the objective of the task under consideration. In the research method, two kinds of agents are presented: web user particle agent and semantic particle agent. PASO Based Web Page Recommendation (PASO-WPR) system is an intermediate program (or a particle agent) containing a user interface, which wisely produces a collection of info that suits an individual's requirements. PASO-WPR is carried out dependent upon incorporating semantic info with data mining techniques on the web usage data as well as clustering of pages dependent upon similarity in their semantics. As the Web pages with multimedia files are viewed as ontology individuals, the pattern of patients' navigation are like instances of ontology rather than the uniform resource locators, and with the help of semantic similarity, page clustering is carried out. For producing web page recommendations to users, the outcome is utilized. The recommender engine concentrates on the semantic info and as well exploits a particle agent to reform the outcomes of web pages recommendation. Consequently, the system response time is enhanced and as a result, creating the framework scalable. The outcomes recommend that the PASO-WPR system is improved in identifying the web page that a user is about to request while matched up other approaches.

In [18] authors proposed a hybrid generative model that can predict user behavior considering multiple factors. While work has been focused on two aspects individually: predicting repeat behavior or predicting new behavior. Model considers both aspects simultaneously during the learning process. The user-specific preference component is used to capture repeat behavior patterns, while the latent group preference component is used to discover new behavior. Besides these two components, this paper also consider the exogenous effect, which is not captured in the former two.

Ladekar A. Pawar A. et al. [10] explain a internet mining algorithm that targets at modifying the analysis of the draft's production of connection rule mining. This algorithm is being extremely utilized in internet mining. The end results achieved establish the strength of the algorithm projected in this paper.

Machine Learning Approaches:

Song et al. [12] planned a sequential DSSM model which incorporates RNNs into DSSM for suggestions. Based on conventional DSSM, TDSSM put back the network with point static characteristics, and the right network with two sub-networks to modeling client static characteristics (with MLP) and client sequential features (with RNNs).

Chen et al. [14] projected an incorporated structure with CNNs and RNNs for modified key structure (in videos) suggestion, in which CNNs are utilized to study characteristic demonstrations from key frame images and RNNs are utilized to practice the textual characteristics.

Table 1. Pros and Cos of existing works

Year	Techniques	Pros	Cons
2014 [9]	Task Trailing and Query Segmentation	Increase Page recommendation by understanding user search from browser history or task performed in last trails.	Dependency of getting the user permission for task trail sequences.
2015 [10]	Aprior Algorithm	Use web logs for understanding user behavior, hence privacy of individual not affected.	Aprior is time takes process to generate patterns from dataset
2016 [12]	Deep Neural Network Dimension Reduction Method	Utilize user and Item features for prediction. Provide data cleaning method for increasing the pattern accuracy	Prediction accuracy is low
2017 [13]	Recurrent Neural Network	User behavior (Browser History) learning in RNN	Work is only depend on user current behavior
2019 [15]	Particle Agent Swarm Optimization	Utilize web log and content features	PSO for content feature selection reduces accuracy of work
2020 [16]	Improved Triangle Similarity	User product rating for product / Page Recommendation	Other features of web was not utilizes, so new product ranking is difficult
2020 [18]	Latent Dirichlet Allocation	Read user current and repetitive for recommendation	Web log and content feature not utilize as getting information of user action is not always possible.

IV. FEATURES OF WEB MINING

Web Content Mining: Web content mining describes the automatic search of information resource available online , and involves mining web data contents. Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources.

Term Frequency: The TF is the count of category-of-words of every category in each document. So the documents term frequency for a category is the occurrence of the words in single document or article [15].

Document Term Frequency: It is the number of documents in the collection that contain a term. IDF: Inverse Document Frequency, is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word.

$$IDF(t) = log\left(\frac{N}{n}\right)$$

N represents the total number of documents in the dataset, n represents the number of documents that term t appears

TF-IDF: TF-IDF [16] (Term Frequency-inverse Document Frequency), puts weighting to a term based on its inverse document frequency. It means that if the more documents a term appears, the less important that term will be, and the weighting will be less.

$$TFIDF(t) = TF_t * log\left(\frac{N}{n_t}\right)$$

TF-IDF-CF: As per the Shortcomings of TF-IDF has, introduce a new parameter to represent the in-class characteristics, and authors have call this class frequency, which calculates the term frequency in documents within one class.

$$TFIDFCF(t) = \log(TF_t + 1) * log\left(\frac{N+1}{n_t}\right) * \frac{n_{c,t}}{N_c}$$

the number of documents where term t appears within the same class c document. Nc represents the number of documents within the same class c document.

Web Usage Mining: Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. In abstract the potential strategic aims in each domain into mining goal as: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no definite distinctions between the Web usage mining and other two categories.

Markov Model: The Kth order markov models were develop from a series of numbers. These are patterns obtain from the numeric dataset like weblog page visiting sequence [17].

Regression: As per requirement different type of regression (linear / logistic) features were extract from the numeric data [18]. Finding a feature from temporal data is done by this regression.

V. APPLICATION OF PREDICTION

Web mining extends analysis much further by combining other corporate information with Web traffic data. Practical applications of Web mining technology are abundant, and are by no means the limit to this technology. Web mining tools can be extended and programmed to answer almost any question. It can be applied in following areas:

1. Web mining algorithm provide good experience to the website user by showing relevant content as per user query.
2. Web mining algorithm provide relevant suggestion on sites bind the visitor to stay more on site.

3. The company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.
4. In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites.
5. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing
6. The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement.
7. Website internal search engine provides advanced and efficient searching capabilities[11].

VI. EVALUATION PARAMETER:

Precision of a transaction is provided as the ratio of the number of web pages appropriately predicted and the overall amount of web pages predicted.

Precision = Approximate_Correct_pages / All_predictions

Coverage is calculated as the ratio of the amount of web pages appropriately predicted and the overall amount of web pages visited by the user.

Coverage = Approximate_Correct_pages / All_Visited_Pages

M-metric is utilized with the intention of obtaining a single evaluation measure, and it is defined in this manner

M-metric = (2xPrecisonxCoverage) / (Precision + Coverage)

Execution Time: Total Time for the execution of the algorithm in the prediction of the page base on the different size of dataset.

VII. CONCLUSIONS

The Internet has become an inescapable source of information for users, enabling access to a large and increasing amount of information. As information progressively shifts from a physical medium to online content, making sense of this information is crucial to provide the best set of information resources to users. It was found that most of paper has work on weblog feature which depends on visitor previous patterns, as visitor content keywords depends on page content, so scholars involved such feature has better results. Further paper has found that latent feature extraction is an important tool for improving recommendation accuracy. Out of many techniques and methodology genetic algorithm based prediction improves result outcomes. In future scholars can develop a model that can identify user keywords for page recommendation.

REFERENCES

[1] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2, no. 1 (2000): 1-15.

[2] Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol. 1, pp. 399-403. IEEE, 2011.

[3] Srivastava, J., R. Cooley, M. Deshpande Mukund, and P. N. Tan. "Web usage mining: discovery and application of usage patterns from web data." Proceedings of SIGKDD explorations 1, no. 2 (2002).

[4] E.Tuba, R.Jovanovic, R.C.Hrosik, A. Alihodzic and M.Tuba, "Web Intelligence Data Clustering by Bare Bone Fireworks Algorithm Combined with K-Means". In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, (2018, June) (article 7). ACM.

[5] 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.

[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 1131.

[7] 13. Giorgos Kollias, Efstratios Gallopoulos, and Ananth Grama. "Surfing the Network for Ranking by Multidamping". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2013

[8] 14. Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu " Web-page Recommendation based on Web Usage and Domain Knowledge" 1041-4347/13/$31.00 © 2013 IEEE.

[9] Zhen Liao, Yang Song, Yalou Huang, Li-wei He, and Qi He. "Task Trail: An Effective Segmentation of User Search Behavior". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 12, DECEMBER 2014.

[10] A. Ladekar, P. Pawar, D. Raikar and J. Chaudhari, "Web Log Based Analysis of User's Browsing Behavior", IJCSIT - International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.

[11] A. Deepa, and P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.

[12] Yang Song, Ali Mamdouh Elkahky, and Xiaodong He. 2016. Multi-rate deep learning for temporal recommendation. In SIGIR. 909–912.

[13] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based News Recommendation for Millions of Users. In SIGKDD.

[14] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized Key Frame Recommendation. In SIGIR.

[15] R. Manikandan. "A novel approach on Particle Agent Swarm Optimization (PASO) in semantic mining for web page recommender system of multimedia data: a health care perspective". Springer Science+Business Media, LLC, part of Springer Nature 10 January 2019.

[16] Arta Iftikhar, Mustansar Ali Ghazanfar, Mubbashir Ayub, Zahid Mehmood, And Muazzam Maqsood. "An Improved Product Recommendation Method for Collaborative Filtering". IEEE Access, volume 8, July 20, 2020.

[17] P.G. Om Prakash, and A. Jaya. "WS-BD-Based Two-Level Match: Interesting Sequential Patterns and Bayesian Fuzzy Clustering for Predicting the Web Pages from Weblogs". The Computer Journal, Volume 63, Issue 1, 2020.

[18] MINH-DUC NGUYEN, AND YOON-SIK CHO. "A Hybrid Generative Model for Online User Behavior Prediction". IEEE Access, volume 8, January 7, 2020