Trivene Khede et al. 2020, 8:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Survey on Web Page Prediction Techniques and Web Mining Features

Trivene Khede, Dr. Avinash Sharma
Department of Computer Science & Engineering
Millennium Institute of Technology & Science Bhopal, India

Abstract

As websites are increasing day by day, so user behavior analysis for improving the website performance attracts many researchers. This paper given an detail survey of various techniques of web mining for next page prediction. Here feature required for the website analysis for page prediction were also discussed. This paper has provided a details survey of different researcher work as well. Paper has identified the application area of user behavior analysis for the website page prediction. Comparison of different page recommendation algorithm was done by evaluation parameter hence last section of this paper brief various set of formula.

Keywords: Information Extraction, Text Analysis, Ontology, feature extraction, text categorization, 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.

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 victimization 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.

Web pre-fetching is utilized to pre-process aim requests before the user makes a specific demand of these objects; so as to lessen the users' observes latency [4, 5]. Net pre-fetching consists of 2 steps mainly; initial, it's a necessity to create correct prediction then next user accesses. These predictions are sometimes created support to previous expertise concerning users' accesses and fondness, and also the relative hints are provided to a pre-fetching device, second, the pre-fetching device makes a conclusion that objects from the anticipated hints are reaching to be pre-fetched.

II. RELATED WORK

Hai Dong, FarookhHussain and Elizabeth Chang in [6] projected net query Classification technique that depends on net distance normalisation. During this design middle categorised queries are send to the target category by normalizing and mapping the net queries. By explaining the frequency, position and position frequency classes are stratified into 3 categories. Within the system Taxonomy-Bridging rule is employed to map target class. The Open Directory Project (ODP) is utilized to create Associate in Nursing ODP-based classifier. This taxonomy is then mapped to the target class's victimisation Taxonomy-Bridging rule. Thus, the post-retrieval question is initially classified into the ODP taxonomy, and also the classifications are then mapped into the target classes for net query. Classification of net query to the user intends and question is major duty for any data retrieval system.

MyoMyoThanNaing [7] projected "Query Classification Algorithm". To classify the net query input by the user into the user intended categories, MyoMyo Than Naing utilizes the domain ontology. Ontology is beneficial in matching of retrieve group to focus on group. User questions are extracted in Domain terms are used as input to the question classification rule. Matched terms of every domain term are extracted in more sub division. Reason the likelihood for matched classes. Then all queries are stratified by their likelihood and displays to the user's table.

In [8] web content re-ranking is prepared by the utilization of net log feature alone. Here multi-damping of the user series is prepared on the premise of linear, page rank, generalized hyperbolic features are used. As this paper has not embrace website feature thus accuracy level is low.

In [9] this analysis work, web content health recommender systems are introduced via the exercise of bound agents so as to produce very acceptable web content for patients. The essential feature of Particle Agent Swarm Optimization (PASO) is that the creation of the rule is denoted by a group of Particle agents World Health Organization join forces achieve the target of the task into account. Within the analysis technique, 2 forms of agents are presented: net user particle agent and linguistics particle agent. PASO primarily based web content Prediction (PASO-WPR) system is Associate in Nursing intermediate program (or a particle agent) containing a programmed, that sagely produces a set of information that suits Associate in Nursing individual's necessities. PASO-WPR has carried out dependent upon incorporating linguistics data with data processing techniques on the net usage information still as clump of pages dependent upon similarity in their linguistics. Because the web content with transmission files are viewed as ontology people, the pattern of patients'routing are like instances of ontology instead of the uniform supply locators, and with the assistance of linguistics similarity, page clump is carried out.

Frikhaetal.[10] concentrates on enhancing typical recommender systems through means of group action data in social media; consist of the first choice of a user still as influence from his peers who are on-line. So, ontology is based on the interest of the user is created to form predictions that are customized. Symentic social recommender system is projected for applying user interest ontology still as Tunisian Medical Tourism (TMT) ontology. Finally, social prediction rule is developed with the purpose to be used during a Tunisia tourism Website to aid users to attract in visiting Tunisia for medical reasons.

Gulzar et al. [11] projected a recommender system, which might advise as well as the user in selecting the courses based on his want. The Hybrid technique was used within the company of ontology to require out valuable data and create accurate predictions. These strategies are helpful to learners to boost their performance still as progress their satisfaction level. The given recommender systems can do higher by assuaging the constraints of basic individual recommender systems.

III. BASIC TYPES OF RECOMMENDER SYSTEMS

Recommender systems Mobasher 2007 [10] were developed to learn Web user experience in order to model the interaction between users and items described on Web-pages and to recommend the interesting items to the users. The popularity of recommender systems is increasing with the rapid growth of the Internet since the mid-1990s. In the systems, recommended items may be Web-pages (links), articles, books or products.

An intelligent recommender system will support Web users to make better decisions to rapidly reach their own target pages during a browsing session. Therefore, recommender systems become more and more important in Web-based applications, such as e-commerce, e-government, and e-services. According to Ricci, Rokach and Shapira (2011) [9], there are six types of recommender systems that vary in terms of the used knowledge, the addressed domain, and the prediction algorithm.

1. Content-based: the system recommends items that are similar to the ones that the user liked in the past. The similarity of items is calculated based on their features.
2. Collaborative filtering: in the system, an active user who is surfing the Web is suggested items that other users with similar taste liked in the past. The similarity in taste of users is calculated based on the similarity in their activity history. This technique is considered to be the most popular one in recommender systems.
3. Demographic: the system recommends items based on the demographic profile of the user.
4. Knowledge-based: the system recommends items based on explicit domain knowledge about how certain item features meet the user needs and references, and how the items are useful for the user.
5. Community-based: the system recommends items based on the references of the user's friends.
6. Hybrid recommender systems: these recommender systems combine some of the above mentioned techniques by taking the advantages of the used techniques to optimize Web prediction.

IV. FEATURES OF WEB MINING

Web Structure Mining: The goal of Web structure mining is to generate structural summary about the Web site and Web page. Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

Web Content Mining:

1. Web content mining describes the automatic search of information resource available online , and involves mining web data contents.
2. 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.
3. The Web content mining is differentiated from two different points of view: Information Retrieval View and Database View summarized the research works done for unstructured data and semi-structured data from information retrieval view.

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.

V. APPLICATION OF PREDICTION

Web page prediction system was used for the various systems like E-commerce sites, heavy web servers, etc. Here these prediction systems predict the product for the user as per his location or previous behavior of the user. So suggestion of any page for the user is based on the rating the product surrounding of the user and learn the buying behavior of the user as well. In [12] prediction of the product is done on the basis of the other user opinion, product market rating and providing community critiques.

Recommender algorithms includes processing of the user by hand or manual cross check the product lists. So collaborative filters are utilized for the study of the system, this increase the sale of the product while website usage for the user is also increased.

Recommender systems enhance Ecommerce sales in three ways: Converting Browsers into Buyers: Visitors to a Web site often look over the site without purchasing anything. Recommender systems can help consumers find products they wish to purchase. Increasing Cross-sell: Recommender systems improve cross-sell by suggesting additional products for the customer to purchase. If the predictions are good, the average order size should increase. For instance, a site might recommend additional products in the checkout process, based on those products already in the shopping cart. Recommender algorithms are a knowledge that helps business man for making one to-one marketing approach.

The recommender system analyzes a database of consumer preferences to overcome the limitations of segment-based mass marketing by presenting each customer with a personal set of predictions. Of course, recommender systems are not a complete solution.

It is still necessary to record and use other consumer data, such as preferred credit card and shipping address, to deliver complete one-to-one service to consumers.[2]
Recommender systems are technologies that can help businesses decide to whom to make an offer. Such systems allow search engines and advertising companies to suggest advertisements or offers to display based on consumer behavior.
Online predictions are preferred because they can respond immediately to the consumer’s preferences. Most of the recommender processes mentioned above can be performed entirely online.

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

Web user flexibility need optimization of web content but it has its own limits. So researcher suggest web page prediction algorithm from last few decades, which help to increase user experience and analyze behavior as well. This survey focuses on how to advance the prediction time without compromising prediction accuracy by using various algorithms of pattern discovery techniques like graph techniques of clustering and also many types of models are developed for prediction. In future, the research can be broadened for a small number of previous log files and an in-depth analysis to enhance the prediction accuracy level by using different techniques of data mining.

REFERENCES

[1]. J. Dom_enech, J. Sahuquillo, J. A. Gil & A. Pont. The Impact of the Web Prefetching Architecture on the Limits of Reducing User's Perceived Latency. Proc. of the International Conference on Web Intelligence, 2006.

[2]. I. Zukerman, D. W. Albrecht & A. E. Nicholson. “Predicting user’s requests on the WWW”. Proc. of the seventh international conference on User modeling, pages 275{284, 1999.

[3]. Balamash, M. Krunz& P. Nain. Performance analysis of a client-side caching/pre-fetching system for Web traffic. Computer Network. vol. 51, no. 13, pages 3673{3692, 2007.

[4]. T. M. Kroeger, D.E. Long & 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.

[5]. L. Fan, P. Cao, W. Lin & 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, pages 178{187, 1999.

[6]. An Ontology-based Webpage Classification Approach for the Knowledge Grid Environment by Hai Dong, FarookhHussain and Elizabeth Chang, 2009 Fifth International Conference on Semantics, Knowledge and Grid (IEEE-2009).

[7]. Myomyo Thannaing,. "Ontology-Based Web Query Classification For Research Paper Searching". International Journal Of Innovations In Engineering And Technology (Ijiet) , Vol. 2 Issue 1 February 2013.

[8]. Giorgos Kollias, Efstratios Gallopoulos, And Ananth Grama "Surfing The Network For Ranking By Multidamping". Ieee Transactions On Knowledge And Data Engineering 2014.

[9]. 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.

[10]. Nguyen TTS, Lu HY, Lu J (2014) Web-page prediction based on web usage and domain knowledge. IEEE Trans Knowl Data Eng 26(10):2574–2587.

[11]. Rani M, Nayak R, Vyas OP. "An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage". Knowl-Based (2015) Syst 90:33–48.