International Journal of Advanced Technology and Engineering Exploration, Vol 5(39)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2018.540007

Review Article

A survey and analysis of page ranking through data mining and advanced techniques

Vinamrata Singh[1*], Kailash Patidar[2] and Rajendra Prasad Sahu[3]
M. Tech Scholar, Department of Computer Science, School of Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[1]
Professor and HOD, Department of Computer Science, School of Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[2]
Assistant Professor, Department of Computer Science, School of Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[3]

Abstract
World Wide Web (WWW) makes a greater impact and dependability in today's world. In day to day life it is increases. If we consider the case of medical field, E-shopping, banking, results etc. have been affected by WWW. We cannot think a day without WWW. So the efficient usage of web data matters. In today's scenario the effectiveness of a website is depend on the users visit. This survey motivation is to meta-analysis of the related work so that new insights can be determined to find the better prediction of optimized cumulative traffic or the visit. By this analysis we also able to determine the associated optimization with respect to different domain. This study also includes the discussion on data mining and optimization techniques.

Keywords
Domain, WWW, Data mining techniques, Optimization.

1.Introduction
The World Wide Web has developed in the previous few years from a group to the greatest and most prominent method for correspondence and data scattering. Consistently, the WWW becomes by about a million electronic pages, adding to the many millions as of now on-line [1]. WWW serves as a stage for trading different sorts of data, extending from exploration papers, and instructive substance, to interactive media substance and programming [2]. The nonstop development in the size and the utilization of the WWW forces new techniques for transforming these gigantic measures of information [3]. On account of its fast and disordered development, the subsequent system of data absences of association and structure [4]. Besides, the substance is distributed in different differing arrangements. Because of this, clients are feeling here and there perplexed, lost in that data over-burden that proceeds to extend. Issues that must be managed of applicable data, including the seeking and indexing of the web content.

The formation of some meta knowledge out of the data which is accessible on the Web, and the tending to of the individual clients' requirements and diversions, by customizing the gave data and administrations [5].There are several techniques are used in the direction of page ranking. K-means clustering, fuzzy C-means and optimization techniques are used in several research works like [6-10]. This also includes the use of association rule mining as the meaningful data should be extracted [11-15].

As the mining extract the knowledge from the huge database and it is also useful in grouping based on similarity label's and class which can be a better tool for understanding and depicting the knowledge from the source of action. This phenomenon is also suggested in [15-20]. Our survey main motivation is to meta-analysis of the related work so that new insights can be determined to find the better prediction of optimized cumulative traffic. The same procedure with different way of optimization is suggested in [21]. This paper objective is to exploring the previous methodologies for finding the new insights in the direction of betterment.

*Author for correspondence

2.Literature review

In 2012, Sampath et al. [22] proposed systolic tree mechanism for the frequent pattern extraction process for the web access logs. Authors suggest that the Systolic tree based rule mining scheme is enhanced for weighted rule mining process. For the weight estimation automatic scheme is used. The dynamic web page weight assignment scheme uses the page request count and span time values. Their proposed system improves the weight estimation process with span time, request count and access sequence details.

In 2013, Nassar et al. [23] suggested clickstream data as the most important sources of information in websites usage and customers' behavior in Banks e-services. They suggested three types of web structure mining as web usage structure, mining data streams and web content. They presented integration between the web usages mining and data mining techniques for processes at different stages, including the pattern discovery phases, and introduce banks cases, that have analytical mining technique. According to the authors data mining techniques can be very helpful to the banks for better performance, acquiring new customers, fraud detection in real time, providing segment based products, and analysis of the customers purchase patterns over time.

In 2014, Azad et al. [24] proposes an algorithm for semantic-synaptic web mining and presents a method for measuring the entropy of web pages using information content. The importance of low entropy over the web mining at the combination of semantic web and synaptic web was suggested.

In 2014, Azad et al. [25] suggest the idea is to improve the accuracy and relevance of information extracting from the web. They proposed a novel model for improving the web mining and hypothesize that web mining is Semantic-Synaptic web mining. According to the authors Semantic-Synaptic web Mining interlinks the web of data to different data sources at low entropy. They combine the best ideas from the semantic web and synaptic web at low entropy and construct the architecture of Semantic-Synaptic web mining.

In 2014, Han et al. [26] suggested web log mining as the important method in web data mining. Keeping in mind the end goal to discover more esteem get to mode and decrease the information estimate from the Web, discover the information of clients and even between clients, this paper advances a technique for web log information preprocessing in view of client normal for advantages, and after that set forward a few ideas, for example, client intrigue, client intrigue similitude. The superiority of the results shows the effectiveness of the approach.

In 2014, Bhaskar et al. [27] discussed that the users habe overaloaded from the data from the web search engines like google, yahoo etc. which may affect the information extraction. They have suggested the question answering (QA) approach for the above problem. Rather than restoring a rundown of report from the ebb and flow internet searcher, QA framework gives data from an extensive arrangement of all around addressed inquiry with suitable media information. QA means to use top to bottom phonetic and media content examination and additionally area learning to return exact responses to characteristic dialect questions. They have survey and discusses the progress of multimedia question answering (MMQA) research.

In 2014, Chen [28] suggested that the by following certain web information mining methodology in electronic business administration, and as indicated by date sources by keeping to subjects of electronic trade administration, doing web date in preprocessing, change, joining, learning disclosure and mode examination, endeavors can guarantee the security of system data framework, enhance effectiveness of client relationship administration, improve electronic business site development and develop arrange activity administration framework to give choice help to electronic business and empower the ventures to be with sound intensity in the pattern of electronic business.

In 2015, Poli [29] suggested that the data mining on web is difficult for online analytic processing (OLAP) with Big data. The data mining is influenced basic by approximating the databases of Big data for learning revelation to process especially MapReducing. The surmised data is fluffy as opposed to likelihood. They have discussed fuzzy web data mining for Big data for association rules. The question preparing is examined with SQL and Xquery for fluffy information mining the fluffy Algorithms are talked about to configuration inquiries in information mining. A few cases are talked about for fluffy Web information mining.

In 2016, Mahani et al. [30] suggested that the web data mining is an emerging area in research field. Because of its different alluring and useful

administrations web is getting to be well known step by step. Web has turned into a prominent medium for data flow. Web mining is ordered into three noteworthy classifications web content mining, web structure mining and web use mining. By utilizing different sorts of procedures analysts are removing new learning and examples from the web. Be that as it may, there are different issues with respect to web information mining. A portion of the major issued are talked about in this paper. Alongside issues different impediment and difficulties are additionally talked about in this paper.

In 2016, Sinha et al. [31] suggested that the log files can grow into huge sizes in the complex systems quickly. Be that as it may, a large portion of the accessible log document pressure instruments utilize a broadly useful calculations, which don't exploit repetition particular log records. The target of this paper is to use length index preserving transformation (LIPT) method to change the log documents and lessens the noteworthy measure of repetitive characters. The connected changed document in information pressure device will successfully diminish storage room. It has been watched that LIPT is a successful method to change log records for measure decrease and the document estimate get diminished to the normal of 44% preceding the pressure. Amid the procedure of change the repetitive character gets decreased.

In 2017, Guojun et al. [32] suggested that the web data acquisition is the foundation of web data mining. Web crawler is an imperative device for Web information procurement, yet the incessant updates of Web information structures, information sources and dissemination channels, brought about high expenses of crawler program advancement and upkeep. With a specific end goal to take care of this issue, this paper composed and executed a clever dynamic crawler, which put away the information extraction standards of XPath in database, stacked the guidelines powerfully as indicated by the objective, and utilized TF-IDF strategy to compute the importance. The Web slithering tenets can be consequently procured, which made the crawler smart and dynamic, enhanced the versatility of the crawler for the mind boggling web condition, and diminished the support and refresh cost.

At long last, this paper applies the clever dynamic crawler to the danger attention to open vulnerabilities, with the strategy for information accumulation and examination of the helplessness group and the system hub internet searcher. The investigation utilized the model framework on three helplessness groups to gather and break down the information. The outcomes demonstrated that the smart dynamic crawler can understand the high-effective and adaptable information accumulation work, and established the framework for Web information mining.

In 2017, Jayamalini et al. [33] suggested WWW as the database which holding huge amount of data. It containts the data in variet that can be helpful for the users in different ways. It contains inestimable information for organizations, if mined viably. Web mining is a procedure that objectives to discover helpful data or learning from the Web page substance, hyperlink structure, and utilization or disjoin logs of sites. Web information mining is isolated into three noteworthy gatherings - Web Content mining, Web Structure mining and Web Usage mining. They have reported the basic concepts of web mining, there types, data extraction techniques and there uses.

In 2017, Singh [34] suggested that the web usage mining utilizes information digging process for the exploration of the utilization design from information brought from the web log documents. Web is the gathering of scholarly instructive foundation marry server information was break down to enable the establishment for additionally enhancing the terms and arrangements of the administration they to give. Web use is likewise helpful for enhancing or recognizing the guest of site by aditing the log records of that website. The attention is on the information gathering in web servers of scholarly instructive organization and execute.They have used Web Log Expert lite 9.3 tools for the analysis.

3.Problem statement

The observations after the study and analysis are following:

1) As the collected data of visits are more so efficint partition techniques can be used.
2) The storage structure is manage in such a way so that the data collected, inserted and updated etc. properly.
3) Association rule mining plays an important role in finding the associated ranks from the whole.
4) Tree slave structure is additionally valuable for extricating the information in broadness first pursuit way.

5) After associative ranking evolutionary alsgorithm can play an important role to show the improvement.
6) The storage structure can be taken as the dynamic slave as to rearrange properly.
7) The model with different grouping associations along with individual ranking may help.

4.Result analysis

The result analysis of the study is shown in *Table 1*.

Table 1 Analysis

Authors	Technique	Findings
Wang et al.[35]	WebSIFT	This methodology determines how to apply data mining techniques to large web data repositories in order to extract usage patterns.
Priya et al.[36]	Web Data extraction	They propose a new method for web data extraction. It has three phases. In the first phase list of web documents are selected, second phase documents are reprocessed, in the final phase results are presented to users
Lin et al.[37]	Informative Contents	They propose a new approach to discover informative contents from a set of tabular documents (or Web pages) of a Web site.
Jayalatchumy et al. [38]	Data Preprocessing	Web cleaning is the most important process as researchers say 70% of the time is spent on data pre-processing. But data cleaning becomes difficult when it comes to heterogeneous data. Maintaining accuracy in classifying the data needs to be concentrated.
Qinglan et al. [39]	SOFM	Their paper maim aim is to introduce the method, and the more SOFM parameters and internal threshold Settings and performance relationships of the algorithm.
Jamil et al. [40]	Subject identification method based on term frequency technique	They have proposed a technique for the subject identification. It may be helpful in subject's identification for the groups of text. Term frequency technique was developed by the authors. The calculation created demonstrated that joining computational phonetic technique and measurable strategy can be more viable for choosing the best subject.
Singh et al. [41]	Web data mining research: a survey	They have report the comparative study and there analysis based on different parameters.

5.Conclusion

This survey provides a direction of web mining and web traffic optimization so that future trends will be discussed and analyzed. This research emphasize the focus on the data that can be seen to enhance the optimize value in all respect. The parametric and probabilistic changes has been notices and discussed. In future proper data mining techniques with the help of bio inspired methods will change the optimization trends and the cumulative index has been increased. This will also emphasize the associated trends in the web optimization rule.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Aktas MS, Nacar MA, Menczer F. Personalizing pagerank based on domain profiles. In proceedings of WebKDD workshop: webmining and web usage analysis 2004 (pp. 83-90). ACM.
[2] Berendt B. Understanding web usage at different levels of abstraction: coarsening and visualizing sequences. In proceedings of the mining log data across all customer touch points workshop (WEBKDD '01) 2001.
[3] Berendt B. Using site semantics to analyze, visualize, and support navigation. Data Mining and Knowledge Discovery. 2002; 6(1):37-59.
[4] Baraglia R, Silvestri F. An online recommender system for large web sites. In proceedings of the IEEE/WIC/ACM international conference on web intelligence 2004 (pp. 199-205). IEEE Computer Society.
[5] Deng L, Chai X, Tan Q, Ng W, Lee DL. Spying out real user preferences for metasearch engine personalization. In proceedings of ACM WEBKDD 2004. ACM.
[6] Mareli M, Twala B. Global optimisation using Pareto cuckoo search algorithm. International Journal of Advanced Computer Research. 2017; 7(32):164-75.
[7] Li K, Cui L. A kernel fuzzy clustering algorithm with generalized entropy based on weighted sample. International Journal of Advanced Computer Research. 2014; 4(15):596-600.
[8] Li K, Li P. A selective fuzzy clustering ensemble algorithm. International Journal of Advanced Computer Research. 2013; 3(13):1-6.
[9] Liang SC, Lee YC, Lee PC. The application of ant colony optimization to the classification rule problem. In international conference on granular computing 2011 (pp. 390-2). IEEE.

[10] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. In CSI sixth international conference on software engineering 2012 (pp. 1-6). IEEE.
[11] Sadh AS, Shukla N. Association rules optimization: a survey. International Journal of Advanced Computer Research. 2013; 3(9):111-5.
[12] Dubey AK, Shandilya SK. A novel J2ME service for mining incremental patterns in mobile computing. In international conference on advances in information and communication technologies 2010 (pp. 157-64). Springer, Berlin, Heidelberg.
[13] Dubey AK, Shandilya SK. Exploiting need of data mining services in mobile computing environments. In international conference on computational intelligence and communication networks 2010 (pp. 409-14). IEEE.
[14] Sharma P. Association rule mining with enhancing list level storage for web logs: a survey. International Journal of Advanced Technology and Engineering Exploration. 2014; 1(1):15-20.
[15] Sadh AS, Shukla N. Apriori and ant colony optimization of association rules. International Journal of Advanced Computer Research. 2013; 3(10):35-42.
[16] Jamil S, Khan A, Halim Z, Baig AR. Weighted muse for frequent sub-graph pattern finding in uncertain DBLP data. In international conference on internet technology and applications 2011 (pp. 1-6). IEEE.
[17] Ashwin CS, Rishigesh M, Shankar TS. SPAAT-A modern tree based approach for sequential pattern mining with minimum support. In the international conference on applications of digital information and web technologies 2011 (pp. 177-82). IEEE.
[18] Zuhtuogullari K, Allahverdi N. An improved itemset generation approach for mining medical databases. In international symposium on innovations in intelligent systems and applications 2011 (pp. 39-43). IEEE.
[19] Gupta R, Satsangi CS. An efficient range partitioning method for finding frequent patterns from huge database. International Journal of Advanced Computer Research. 2012; 2(4):62-9.
[20] Yadav MP, Feeroz M, Yadav VK. Mining the customer behavior using web usage mining in e-commerce. In international conference on computing communication & networking technologies 2012 (pp. 1-5). IEEE.
[21] Lata S. An iterative PSO for web worth optimization through random velocity. International Journal of Advanced Technology and Engineering Exploration. 2015; 2(3):31-6.
[22] Sampath P, Ramesh C, Kalaiyarasi T, Banu SS, Selvan GA. An efficient weighted rule mining for web logs using systolic tree. In international conference on advances in engineering, science and management 2012 (pp. 432-6). IEEE.
[23] Nassar OA, Al Saiyd NA. The integrating between web usage mining and data mining techniques. In international conference on computer science and information technology 2013 (pp. 243-7). IEEE.
[24] Azad HK, Abhishek K. Entropy measurement and algorithm for semantic-synaptic web mining. In international conference on data mining and intelligent computing 2014 (pp. 1-5). IEEE.
[25] Azad HK, Abhishek K. Semantic-synaptic web mining: a novel model for improving the web mining. In international conference on communication systems and network technologies 2014 (pp. 454-7). IEEE.
[26] Han Y, Xia K. Data preprocessing method based on user characteristic of interests for web log mining. In international conference on instrumentation and measurement, computer, communication and control 2014 (pp. 867-72). IEEE.
[27] Bhaskar DB, Singh DK. Multimedia questions and answering using web data mining. In international conference on information communication and embedded systems 2014 (pp. 1-4). IEEE.
[28] Chen G. Application of web data mining technique to enterprise management of electronic commerce. In international symposium on computational intelligence and design 2014 (pp. 154-7). IEEE.
[29] Poli VS. Fuzzy data mining and web intelligence. In international conference on fuzzy theory and its applications (iFUZZY) 2015 (pp. 74-9). IEEE.
[30] Kavita, Mahani P, Ruhil N. Web data mining: a perspective of research issues and challenges. In international conference on computing for sustainable global development 2016 (pp. 3235-8). IEEE.
[31] Sinha AK, Singh V. Transformation of LOG file using LIPT technique. International Journal of Advanced Computer Research. 2016; 6(23):58-64.
[32] Guojun Z, Wenchao J, Jihui S, Fan S, Hao Z, Jiang L. Design and application of intelligent dynamic crawler for web data mining. In youth academic annual conference of Chinese association of automation 2017 (pp. 1098-105). IEEE.
[33] Jayamalini K, Ponnavaikko M. Research on web data mining concepts, techniques and applications. In international conference on algorithms, methodology, models and applications in emerging technologies 2017 (pp. 1-5). IEEE.
[34] Singh SP. Analysis of web site using web log expert tool based on web data mining. In international conference on innovations in information, embedded and communication systems 2017 (pp. 1-5). IEEE.
[35] Wang Y. Web mining and knowledge discovery of usage patterns. Cs 748T Project. 2000:1-25.
[36] Priya VS, Sakthivel S. An implementation of web personalization using web mining techniques. International Journal of Computer Science and Mobile Computing. 2013; 2(6):145-50.
[37] Lin SH, Ho JM. Discovering informative content blocks from web documents. In international conference on knowledge discovery and data mining proceedings of the eighth ACM SIGKDD 2002 (pp. 588-93). ACM.
[38] Jayalatchumy D, Thambidurai DP. Web mining research issues and future directions–a survey. IOSR Journal of Computer Engineering (IOSR-JCE). 2013; 14(3):20-7.

[39] Qinglan H, Longzhen D. Multi-level association rule mining based on clustering partition. In international conference on intelligent system design and engineering applications 2013 (pp. 982-5). IEEE.
[40] Jamil NS, Ku-Mahamud KR, Din AM, Ahmad F, ChePa N, Ishak WH, et al. A subject identification method based on term frequency technique. International Journal of Advanced Computer Research. 2017; 7(30):103-10.
[41] Singh B, Singh HK. Web data mining research: a survey. In international conference on computational intelligence and computing research 2010 (pp. 1-10). IEEE.

Vinamrata Singh had completed her B.Tech from I.E.T.E, New Delhi in 2014 in the Department of Computer Science. Currently, she is pursuing M.Tech from S.S.U.T..S University, Sehore in the department of Software Engineering.

Email: v.vaishali1990@gmail.com