International Journal of Advanced Technology and Engineering Exploration
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
Volume-1 Issue-1 December-2014

Association Rule Mining with enhancing List Level Storage for Web Logs: A Survey

Prashant Sharma*
Medicaps Institute of Technology & Management, Indore, India
prashant.sharma.cse17@gmail.com

©2014 ACCENTS

Abstract

Storing and calculating web page rank is a crucial research area. There are also several researches are going on but the need of betterment it still there because of the following reasons: 1) Impact should be calculated cumulative it is not based on the single page rank 2) Automatic rank identification 3) Way of storage. So our study mainly focuses on the above three directions. Our study analyzes several related researches in these directions to find the useful techniques and remove the drawback from the traditional techniques. Based on our observation we will also suggest some future suggestion which will be incorporated for better weigh estimation process for web logs.

Keywords

Automatic Rank Identification, Association Rule Mining, List storage, Page Impact.

1. Introduction

Association rule mining aims at generating Bond record between sets of incident in a database. Accommodate a period, befitting to huge heaping up in the database technology, the data are representing in the high dimensional data space. Now a day it is a very useful tools for web logs also. Association compel mining finds captivating correlations into the middle a full facts set of factors. In a market basket analysis it brawn be discovered Union laws non-native transactional data bases which are upon sets of items basis Apriori[1] . In middling Apriori is worn as a hyperactive algorithm for mining turn up at item sets for generating association rules [2]. However, in a relational data base, additional information may also have correlation on set of items [3].

In real-world exigency, purposefulness maker's bear perpetually suitably add to ambivalent objectives and an expansive fulfill gap with contrastive runner alternatives [4][5]. The multi-criterion making also provide a better combat in the near future. Its involves span rigorous spaces like the design space, incorporating the defining variables of the candidate solutions, and the intention space, constituting the mapping of each candidate solution to the multiple objective functions values[6]. The latter is the space where optimality is get under way, tradeoffs are explored, and decisions are normally reached. So there is the need of classification based on multiple decision criteria which can be heuristic, it will be possible by user defined constraints and multiple selective constraints. It can be better to find a proper clustered way to organize the web pages, then apply some classification criteria which will be satisfied some threshold value to provide the constrained way of these issue. It can be achieved through association rule mining [7], we can use partitioning technique also because it can reduce the searching time and enhance the searching capability [8][9].

For classification we can use association rule mining with some clustering techniques like K-means and fuzzy c-means, it will be a better option [10]. Then we can optimize it using several optimization techniques like Ant Colony optimization (ACO), Particle swarm Optimization, Mimetic algorithm etc.[11][12][13]. Subset superset partitioning can be used for partitioning and better classification [14].

The remaining of this paper is organized as follows. We discuss Association rule mining in Section 2. In Section 3 we discuss about literature review. In section 4 we discuss about problem domain. In section 5 we discuss about the analysis, conclusions are given in Section 6. Finally references are given.

*Author for correspondence

2. Association Rule Mining

Association Rule mining is one of the important and most popular matter mining techniques. Federation head up mining gluteus Maximus be efficiently used in any decision making processor decision based leadership generation. In data mining appointment in consequently so we courage find the frequent patterns to know the effective patterns from the huge data. Change we find positive and negative rules [15] [16]. If we agree to the beyond everything phenomena change we come to the point that the rule generation is also huge. In this compounding we metaphysical join aspects of optimization techniques by which we can optimize the association rules. So hybridization is needed [17]. Turn to course mining is a efficacious movement capable of identifying in a used of objects (called items) those which demonstrate similar behavior. For event, in a Stock Exchange, clientele object encode are kept as storekeeper business each includes a set of items purchased together. Analyzing the used of merchant may discuss to fact mosey are frequently purchased together.

3. Literature Review

In 2011, Avrilia Floratou et al. [18] proposed a new algorithm called Fexible and Accurate Motif DEtector (FLAMEIt is a frequent pattern finder as in the format of tree model. Its accuracy will be concerned as the existing patterns are always in the database for find their new subset. The performance metrics is then calculated and flame will be proved better in terms of scalability and efficiency.

In 2011, Shawana Jamil et al. [19] focus on investigation of mining frequent sub-graph patterns in in the graph. It will be formulated in terms of subgraph to punctuate with the support based value. They used n approximation mining properties for imposing and discovering possible sub graphs which are frequent from uncertain graph data.

In 2011, Ashwin C S et al. [20] proposed a concept multiple minimum supports (MMS) with apriori for improving association rule mining concept. By the multiple minimum supports we can increase the range of filtration. For efficiently discovering sequential patterns with MMS, they develop a data structure, named PLMS-tree, to store all necessary information from database and by this authors can perform an efficient association from the database.

In 2011, K. Zuhtuogullari et al. [21] observe that an extendable and improved item set generation approach has been constructed and developed for mining the relationships of the symptoms and disorders in the medical databases. Their apriori based rule mining is useful in terms of establishing correlation between symptoms. Their approach is efficient in terms of medical mining and the approach is based on apriori.

In 2012, Mahendra Pratap Yadav et al. [22] presented the study of the customer's behavior using the Web mining techniques and its application in e-commerce to mine customer behavior in the context of web data mining. It can be explained through following terminology in details: source data collection, data preprocessing, pattern discovery, pattern analysis and cluster analysis. With the advanced information technologies, server are now able to collect and store mountains of data, describing their numerous contributions and different customer profiles, from which they seek to derive information about their customer's requirements. Traditional methods are not suited well in this context. The principle of data mining is to cluster customer segments by using K-Means algorithm in which input data comes from web log of various e-commerce websites. Hence,determine the relationship between Web data mining and ecommerce and also to apply Web mining technology in ecommerce.

In 2013, Huang QingLan et al. [23] proposed a clustering classification multi-level association rule mining. The concept is the hybridization of generalization and neural network.An internal threshold value will be introduced for the next transaction, by way of introducing an internal threshold so that the minimum threshold value is not needed, and the seperation is done by local and global frequent item sets. It can improve the utilization and accuracy of multilevel association rules.

In 2013, Jutamas Tempaiboolkul et al. [24] proposed an algorithm for discovering rare association rules in distributed environment. They consider multiple minimum support generated by the static percentile to mine association rules.They compare their algorithm by the Optimized Distributed Association rule Mining (ODAM) algorithm and the Apriori with Multiple Support Generating by statistic Percentile threshold (Apriori MSG-P) algorithm. Their result

shows that the proposed algorithm can discover more rare association rules with an optimized communication cost.

In 2013, Omer Adel Nassar et al. [25] suggest the collaboration of Web usage mining and data mining techniques for processes at different stages, including the pattern discovery phases, and introduces banks cases, that have analytical mining technique. A general framework for fully integrating domain Web usage mining and data mining techniques is represented for processes at different stages by 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 2013, Hemant et al. [26] aims to present technique to make private log information public and apply Apriori algorithm on collected log file to extract knowledge from public and free log files with Web Usages Mining Technique.

In 2012, P. Sampatht et al. [27] suggest that the Escort sequence mmmg algorithms are prepared to findcommonly burgeoning sets in databases. Reminiscence and conduct time procession are unequivocal mighty in frequent pattern mining algorithms. Systolic insinuate construction is a reconfigurable prevarication worn for frequent pattern mining operations. High throughput and faster indubitably are the highlights of the systolic tree based reconfigurable architecture. The systolic tree operation is used in the frequent pattern origin act for the web admittance logs. Systolic tree based prescribe mining goal is enhanced for weighted rule mining process. Inescapable steelyard consequence scheme is used in the system. The sprightly weave messenger-girl excess job long uses the Hermes request count and catch time values. Their professed code is improves the steadiness description engagement less span time, request count and access sequence details. The owner note based envoy estimate is old to survey the frequent item sets.

In 2012, Syeda Farha Shazmeen et al. [28] suggest Shoelace and Upon Mining by urgency semantics to before b before mining and treatment mining to create semantics. Beat Mining aims at discovering insights less the bank of Twine capital and their practice In Unembellished Revile, the semantics evidence is presented by the relation here others and is recorded by RDF. RDF which is word-for-word thrash technology saunter depths be old to lowly efficient and scalable systems for Cloud. The Precise Net enriches the Turf In the matter of Thong by machinery manner accomplished intimate which supports the narcotic addict in dominion tasks, and also helps the users to get the exact search result .They disagree the friend at court of the Exact Rave at with Pounce on Mining, list out the benefits. Challenges, opportunities of the web.

In 2013, Anjana Gosain et al. [29] suggest that Combination laws pushed in the period of respecting normally and qualitative associate which in turn helps in decision making. Coalition enlist hand out with affairs of both binary values and quantitative data. [30] Unity list second in the maturity of close to usually and qualitative adscititious which in turn helps in decision making. Coalition cleave forgo wide to undertaking of both binary values and quantitative data[31]. Besides binary affinity list suffers from sharp boundary problems [32]. Except for personal thorough planet transactions consist of quantitative attributes. Go wool-gathering is why yoke researchers strive been full on era of association rules for quantitative data. They subsidy choice algorithms predisposed by special researches to generate association rules among quantitative data. They shot at rank comparative study of alternative algorithms for association rules based on various parameters.

4. Problem Domain

After studying several research papers we observe the the following problem findings:

1) Data partitioning can be easily implanted for reducing the size and searching.
2) Dynamic rule generation at different levels are also useful.
3) Size is a greater concer in data minin, list can be appropriate for handle larger data logs.
4) Clustering and classification model can be applied together.
5) Average cmulative page rank will be applied for accurate impact.
6) Classification can be heuristic for applying the meta search.
7) Need to dynamize the map structure and organize it in proper hierarchy.

8) Tree slave structure is also useful for extracting the data in breadth first search way.

5. Analysis

We provide the analysis of the paper in table 1.

Table 1: Analysis

Authors	Technique	Achieve
Herrouz et al. [33]	Web Content Mining Tools	The mining tools are imperative to scanning the many HTML documents, images, and text.
Wang et al.[34]	WebSIFT	An example of a prototypical Web usage mining system, WebSIFT, will be introduced to make it easier to understand the methodology of how to apply data mining techniques to large Web data repositories in order to extract usage patterns.
Priya et al.[35]	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.[36]	Informative Contents	They propose a new approach to discover informative contents from a set of tabular documents (or Web pages) of a Web site.
Lee et al. [37]	FCM	This paper is concerned with proposing the fuzzy cognitive map (FCM)-driven inference amplification mechanism in the field of web-mining.
Eltahir et al.[38]	Web servers Navigation	They useweb usage mining technique to procure knowledge from web server log files where all user navigation history is registered.
Sudheer et al.[39]		Their identification of web usage patterns based on the user's interest/choice, thereby creating an intelligent semantics-based web usage mining technique.

6. Conclusion

In this paper we survey several aspects of web data mining with their structure, logs and the flaws presented in the previous technique. We also find some useful trends in the previous technique which can be incorporated with clustering and association to form a hybrid technique for proper classification and maintaining the log impact. The scopes are in the direction of hybrid framework with the formation of advance structural list based association rule mining with cumulative impact.

References

[1] Rakesh Agrawal, Tomasz Imielinski and Arun Swami," Mining association rules between sets of items in large data bases", in proceedings of the ACM SIGMOD Conference on Management of Data, pp 207-216, Washington, D.C., May 1993.

[2] Bodon, F., "A Fast Apriori Implementation", FIMI'03, November 2003. [3] Rakesh Agrawal, Tomasz Imielinski and Arun Swami," Data base Mining- A performance perspective", IEEE transactions on knowledge and data engineering, vol 5 1993.

[3] M J Zaki and C J Hsiao," CHARM- an efficient algorithm for closed itemset mining, in the proceedings of SDM 2002, p 457-473., 2002.

[4] T. Kohonen, Self-Organizing Maps. Berlin: Springer-Verlag, 1997.

[5] A. Rauber, M. Dittenbach, and D. Merkl, "Towards automatic contentbased organization of multilingual digital libraries: An English, French and German view of the Russian information agency Nowosti news," in Proceedings of the Third All-Russian Scientific Conference on Digital Libraries: Advanced Methods And Technologies, Digital Collections, September 11-13 2001, pp. 11–13.

[6] Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data," IEEE Transactions on Neural Networks, vol. 13, no. 6, pp. 1331–1341, 2002.

[7] M. Bagajewicz and E. Cabrera. Pareto optimal solutions visualization techniques for multi objective design and upgrade of instrumentation networks. Industrial and Engineering Chemistry Research, 42(21):5195–5203, 2003.

[8] W. Berger, H. Piringer, P. Filzmoser, and E. Gr ¨oller. Uncertainty aware exploration of continuous parameter spaces using multivariate prediction. Computer Graphics Forum, 30(3):911 –920, 2011.

[9] N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multi objective Selection Based on Dominated Hypervolume. European Journal of Operational Research, 2007.

[10] Dubey, Ashutosh K., and Shishir K. Shandilya. "A novel J2ME service for mining incremental patterns in mobile computing." Information and Communication Technologies. Springer Berlin Heidelberg, 2010.

[11] Pragati Shrivastava, Hitesh Gupta," A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5 September-2012.

[12] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp. 589-592, 2005.

[13] Shyi-Ching Liang, Yen-Chun Lee and Pei-Chiang Lee, "The Application of Ant Colony Optimization to the Classification Rule Problem", 2011 IEEE International Conference on Granular Computing.

[14] Anshuman Singh Sadh, Nitin Shukla," Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Volume-3, Number-1, Issue-9, March-2013.

[15] Arezoo Modiri and Kamran Kiasaleh," Permittivity Estimation for Breast Cancer Detection Using Particle Swarm Optimization Algorithm", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011.

[16] Yao Liu and Yuk Ying Chung, "Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning", IEEE 2011.

[17] Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, Yogeshver Khandagre, "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support", Conseg-2012.

[18] Avrilia Floratou, Sandeep Tata, and Jignesh M. Patel," Efficient and Accurate Discovery of Patterns in Sequence Data Sets", IEEE Transactions On Knowledge and Data Engineering, VOL. 23, NO. 8, August 2011.

[19] Shawana Jamil, Azam Khan, Zahid Halim and A. Rauf Baig," Weighted MUSE for Frequent Sub-graph Pattern Finding in Uncertain DBLP Data", IEEE 2011.

[20] Ashwin C S, Rishigesh.M and Shyam Shankar T M," SPAAT-A Modern Tree Based Approach for sequential pattern mining with Minimum support",IEEE 2011.

[21] K. Zuhtuogullari and N. Allahverdi ,"An Improved Itemset Generation Approach for Mining Medical Databases", IEEE 2011.

[22] Yadav, M.P.; Feeroz, M.; Yadav, V.K., "Mining the customer behavior using web usage mining in e-commerce," Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on , vol., no., pp.1,5, 26-28 July 2012.

[23] Huang Qinglan; Duan Longzhen, "Multi-level Association Rule Mining Based on Clustering Partition," Intelligent System Design and Engineering Applications (ISDEA), 2013 Third International Conference on , vol., no., pp.982,985, 16-18 Jan. 2013.

[24] Tempaiboolkul, J., "Mining rare association rules in a distributed environment using multiple minimum supports," Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on , vol., no., pp.295,299, 16-20 June 2013.

[25] Nassar, O.A.; Al Saiyd, N.A., "The integrating between web usage mining and data mining techniques," Computer Science and Information Technology (CSIT), 2013 5th International

Conference on, vol., no., pp.243-247, 27-28 March 2013.

[26] Randhir, Hemant N., Ravindra Gupta, and G. R. Selokar. "Extract Knowledge and Association Rule from Free Log Data using an Apriori Algorithm.", International Journal of Advanced Computer Research (IJACR), Volume-3, Number-3, Issue-12, September-2013.

[27] Sampath, P.; Ramesh, C.; Kalaiyarasi, T.; Banu, S.S.; Selvan, G.A., "An efficient weighted rule mining for web logs using systolic tree," Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on , vol., no., pp.432,436, 30-31 March 2012.

[28] Shazmeen, Syeda Farha, and Etyala Ramyasree. "Semantic Web Mining: Benefits, Challenges and Opportunities." International Journal of Advanced Computer Research (IJACR) Volume-2, Number-4, Issue-7, December-2012.

[29] Gosain, A.; Bhugra, M., "A comprehensive survey of association rules on quantitative data in data mining," Information & Communication Technologies (ICT), 2013 IEEE Conference on , vol., no., pp.1003,1008, 11-12 April 2013.

[30] Yihua Zhong, Yuxin Liao, Research of Mining effective and Weighted Association Rules Based on Dual Confidence,CPS Fourth International Conference on Computational and Information Sciences 2012.

[31] Weimin Ouyang and Qinhua Huang ,Mining Direct and Indirect Weighted Fuzzy Association Rules in Large Transaction Databases ,IEEE Eighth International Conference on Fuzzy Systems and Knowledge Discovery,2011.

[32] Keon-Myung Lee,Mining Generalized Fuzzy Quantitative Associaton Rules wih Fuzzy Generalization Hierarchies, IEEE 2011.

[33] Herrouz, Abdelhakim, Chabane Khentout, and Mahieddine Djoudi. "Overview of Web Content Mining Tools." arXiv preprint arXiv:1307.1024 (2013).

[34] Wang, Yan. "Web mining and knowledge discovery of usage patterns." CS 748T Project (2000).

[35] Priya, V. Shanmuga, and S. Sakthivel. "An Implementation of Web Personalization Using Web Mining Techniques." (2013).

[36] Lin, Shian-Hua, and Jan-Ming Ho. "Discovering informative content blocks from Web documents." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[37] Lee, Kun Chang, et al. "Fuzzy cognitive map approach to web-mining inference amplification." Expert Systems with Applications 22.3 (2002): 197-211.

[38] Eltahir, M.A.; Dafa-Alla, A.F.A., "Extracting knowledge from web server logs using web usage mining," Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on , vol., no., pp.413,417, 26-28 Aug. 2013.

[39] Sudheer Reddy, K.; Varma, G.P.S.; Reddy, S.S.S., "Understanding the scope of web usage mining & applications of web data usage patterns," Computing, Communication and Applications (ICCCA), 2012 International Conference on , pp.1,5, 22-24 Feb. 2012.