International Journal of Engineering & Technology, 7 (2.7) (2018) 948-951

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A literature review: big data and association rule mining

Meenakshi [1] *, Rainu Nandal [2]

[1] Ph. D. Scholar U.I.E.T (M.D.U), India.
[2] Assistant Professor U.I.E.T (M.D.U), India.
**Corresponding author E-mail: meenakshimalik16@gmail.com*

Abstract

In Today's modern and advanced era, huge amounts of data have become available on hand to developers and choice makers. Big data successfully handles datasets that are not only large, but also very high in velocity and variety, which difficult to handle using conventional techniques, methods and tools. Multilevel association rule mining plays a very vital role in distributed environment in analysis of big data for preparing different Marketing strategies. As compared to Single Level rule, more precise and prominent information is provided by multilevel association rule and additionally from the hierarchical dataset it generates the conceptual hierarchy of knowledge. This paper aims to analyze Data Mining Technique named Multilevel Association rule, which provides additional important information in comparison to single level rule, and it also invents conceptual hierarchy of information/data from the hierarchical dataset. Tools and techniques of Big Data have also been reviewed in detail.

Keywords: *Big Data; Data Mining; Hierarchical; Multilevel Association Rule; Single Level Rule*

1. Introduction

Data Mining is nothing but extraction of information from already existing huge sets of data. In other words, it can be stated that data mining is the method of pulling out knowledge and specific information from data [1]. The data mined or extracted by this process can be used in many ways by industries these days.

As on date we have plenty of Data Mining methods and techniques. One of those techniques is 'Association rule mining'. It's the task in which relationships of data are explored. This is used for the market basket analysis for specific company who is concerned in identifying stuff or products that are being purchased together as well as frequently. It supports in establishing patterns, inter-relations, associations or informal structures among sets of various items in different transaction databases. Hence it supports in knowing customer purchasing habits by establishing associations and correlations between the different items that customers collect in their "shopping basket"[2].

Association rules are used to establish and analyze the relationship among frequently purchased item sets in relational and transactional databases.

Let us suppose that I = {I1, I2…, Im} is a set of different items. Let us suppose that 'D' is a set of database connections where each of transaction T is a set of items in such a way that T ⊆ I. Let 'S' be a set of items. A transaction T will be said to contain S if and only if S ⊆ T. Here an association rule is being shown in the form A→C, where A ⊆ I, C⊆ I and A ∩ C = ∅ [3] [4].

This rule shows the "IF THEN" relationship between the item set A with the item set C. The association rules are created on the basis of support and confidence.

Support: Frequency of the occurring event. The computing of occurrence frequency can be taken as the probability that the 2 events (A and C) occur in the same transaction. The equation of support is as under.

Support (A→C) = P (A∩ C)

Confidence: Proportion of frequency of co-occurring events (A and C) to the frequency of ancestor event (A). The equation of confidence is as under

Confidence (A→C) = P (C|A) = P (A∩ C)/P (A)

Concept hierarchy: It organizes concepts or available data in the form of hierarchy. Then this concept is used to express knowledge in useful, brief and facilitative knowledge of mining at multiple levels of abstraction. For instance, consider following rules, in which "Stallion Chassis" is an antecedent of "Ashok Leyland" based on the concept hierarchy. Support(A →C) = Support(A ∪ C) = P(A ∪ C) Confidence(A →C) = P(C |A) = Support(A ∪ C) Support(A) Stallion Chassis→ Super Structure, Ashok Leyland→ Super Structure. The concept hierarchy complexity can be explained in expressions of internal nodes numbers and height of each internal node. It efficiently measures the interestingness of the knowledge rules discovered. "Multilevel association rules" are obtained from data mining at multiple levels of concept hierarchy [5].

2. Preliminaries

2.1. Big data

In Today's Technology and science age, Big data is hottest topic of research and it has a great scope in each segment of our surrounding world; such as Market, weather, physical conditions, SS, and much more. Big data is presently considered as data sets of sizes more than capability of normally utilized tools in usual software for managing and capturing data. Big data is yet in its immature stage and it is clearly evident from its indistinct explanation, limited application and privacy barriers for enveloping implementation [6].As it is

an application-oriented area, so definitely it is required to incorporate field acquaintance and data into information systems, which are just like usual database systems and which have a thorough mathematic base, execution mechanism and a set of designed rules.

2.2. History and evolution of big data

Although the word "BIG DATA" is new but the base concept is very old. Even in the 1940s-1950s, decades prior to anyone spoken or heard of the term "big data," businesses used to utilize basic analytics (essentially statistics in a business worksheet that were examined by hand) to find out trends and calculations of business. Certainly Big data enables the business to run speedily and efficiently. In the past gathered information could be used for taking futuristic decisions whereas currently big data provides immediate references and results for an industry to work faster and having a competitive edge.

2.3. Big data analytics

Since ages, multi-structured data is getting generated by humans and machineries. The diagnostic value, scope, and power of stored data were supposed to bloom for sure one day. In nut shell, data is a strategic asset which can supports govt. as well as private organizations in enhancing their capabilities and competencies in a big way. So there was a special and urgent requirement of a tool or software that can handle vast and mammoth volumes of data to analyze and represent.. The well-timed entry, performance and the awesome recognition of the Hadoop [7] frameworks has proven as a boon to derive actionable insights by reducing difficulty levels and allowing making speedy decisions. Hadoop-based analytical products are competent enough to analyze and process data type and quantity through thousands of products server clusters.

2.4. Uses of big data

Government
Big data usage in government organization both in central and local allows saving and resource utilizations in field of improvement, modernization, cost and productivity.
International development
At international level, it supports in exploring opportunities in cost-optimization enabling improvement in making fast decisions in field of significant improvement fields like service sector, health segment, crime control, public safety and natural calamity. Manufacturing Big data utilizes sensory data of various vital inputs (like; power, pressure, airflow, voltage, controller data etc) of manufacturing sector to predict upcoming uncertainties. It supports in achieving Predictive manufacturing which continually proceeds towards near-zero downtime and minimum rejection levels.
Healthcare
Area of Healthcare has been strengthened by Big data by introducing predictive analysis, assisting tailored medication and specific analytics, waste reduction, automatic reporting of patient's external and internal data. This includes electronically maintained data of health records, sensor data, patient generated data and various other forms of complicated to process data [7] [8].

2.5. Characteristics of big data

Big Data has a lot of characteristics mentioned as V's characteristics. These characteristics of the Big Data are outcome of many researches, data analysis and study. So far 9V's characteristics have been listed down: (Velocity, Volume, Validity, Veracity, Visualization, Variety, Variability, Volatility, and Value) [9].
Three major Vs are elaborated below in fig 1:

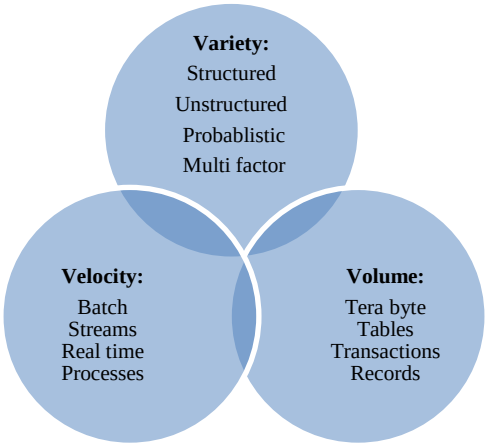

Fig. 1: Characteristics of Big Data.

3. Literature review

R.A. Angryk, F.E. Petry [10] investigates function of fuzzy concept hierarchies for pulling out multilevel information from huge datasets. The unique process of induction of fuzzy hierarchical has been analyzed and further it has been extended with new distinctiveness which further improves applicability of the innovative approach to technical and systematic data mining. We get introduced a consistent model of fuzzy induction and further it is applied to mine generalized association rules. It states that utilization of fuzzy concept hierarchies provides additional flexibility which in turn results in improved modeling of dependencies of real-life and increased satisfaction levels of induction process.
Authors N.E. Oweis, M.M. Fouad, S.R. Oweis, S.S. Owais, V. Snasel [11] developed a "parallel association rule mining algorithm" based on Map Reduce paradigm by using "Lift interestingness" measured (MRLAR). MRLAR can successfully hold huge datasets with a very large quantity of nodes. This straightly eliminates the requirement of additional calculations.
Big table is nothing but a spread storage system used for handling organized data which is planned to range to a mega sizes: Thousands of servers having data in petabytes.
F. Chang, J. Dean, W.C. Hsieh, D.A. Wallach, M. Burrows, R.E. Gruber [12] describes the design, implementation of big table and also details the simple data model which is provided by Big table, that in return provides dynamic control on format and data layout to client's support.
Authors J.H.C. Yeung, C.C. Tsang, K.H. Tsoi, B. Kwan, C. Cheung, A.P.C. Chan, P.H.W. Leong, [13] presents the application of a MR(map-reduce) library holding parallel field programmable gate arrays and "graphics processing units" called as GPUs. This paper further shows that, supported by the MR methodology and library, large systems can be developed with elevated output/efficiency and using brief issue details on mixed GPU/FPGA-based customized computing machines.
P.N. Tan, V. Kumar, J. Srivastav [14], describes numerous important properties that should be studied so that correct measure for a provided application can be selected. These properties have been studied in comparison using 21 measures which have initially been developed in various areas such as SS, Science, Machine learning, data mining and statistics.
Finally, this paper presents an algorithm for choosing a small set of patterns in a way that field specialists can easily find a way that suits their necessities in best way by grading or positioning this small set of patterns.
T. Brijs, K. Van hoof, G. Wets [15] proposed a summary of few prevailing "measures of interestingness" and also comments on properties of same. Overall, interestingness measures are divided into objective and subjective measures. This paper emphasis on objective measures of interestingness only.
G. Shaw, Y. Xu, S. Geva [16] proposes the elimination of rules of hierarchically redundant from multi-leveled datasets. It additionally

states; As Association rules can easily be derived from the representation, so the subsequent conciseness presentation of 'non-redundant association rules' is totally loss free. Experimentations proved that multilevel non-redundant rules can be effectively generated by extension.

Ultimate aim of this paper is to reduce the rule set size in order to make the usefulness and quality better with no triggering to any type of information loss.

Y. Xu, G. Shaw, Y. Li [17] initially presents meaning for 'redundancy' and a brief representation called Reliable basis for presenting 'non-redundant association rules'. Then the paper shows that hierarchically redundant rules can be removed from multi-level datasets. It proposes an approach which removes 'hierarchical redundancy' by using frequent closed generators and item sets.

Y. Xu, G. Shaw, Y. Li [18] proposes a novel method for improving data mining situation by studying the rules mined from original concept level to attain multilevel rules. Dynamic concept hierarchies are also supported by this proposed method. It also discusses about the calculation problems of multilevel association rules at specific level, rule support and the adjustment among suggested possible solutions, commonsense and specific patterns.

T. Fadi, S. Hammoud [19] overcomes from the issue of very huge data mining by suggesting a new rule called Map-Reduce association rule mining (MR-ARM) technique .A format of hybrid data transformation is used for finding recurrent items and rules generation, speedily. MR-ARM implementation done on Weka and Hadoop. The results show that MR-ARM is very valuable instrument for mining of association rules from big datasets from distributed environment.

From the view point of big data, in this paper authors Kalyan P. Subbu, Athanasios V. Vasilakos [20] presents context aware computing. Firstly classification of accessible work on the base of sense platforms is proposed and afterwards it discusses the latest advancement and progress in this area of Big data context aware systems centered on how such systems deal with a variety of challenges of big data. Secondly it explains the "Context aware computing" called as CAC that facilitates applications for awareness of the context by creating inferences by supplying elegant intelligent services to the user.

The projected scheme by Navroop Kaur, Sandeep K. Sood [21] provides approximation of the data characters of big data stream in expressions of variability, velocity, variety and volume.

The projected system intends allotment of proper assets to big data stream based on its FOUR Vs.

Daniele Apiletti, lenaBaralis,Tania Cerquitelli, PaoloGarza, Fabio Pulvirenti and LucaVenturini [22] presents analysis of Spark-based scalable algorithms and Hadoop attending to the recurrent item set mining issue in the domain of Big Data via comparative analysis of experiments and theory.

Drew Schmidt, Wei-ChenChen, ike Matheson and George Ostrouchov [23]explains that without addition of overhead of memory and excess computational overheads, 'R' exploits extremely scalable parallel libraries like ScaLAPACK and PBLAS through pbdR packages and this is done by non inclusion of big extent of overhead of memory and computational overhead.

For management of data produced by networks of wireless multimedia sensor, Cihan Küçükkeçeci, Adnan Yazıcı[24] offers a big database model based on graph database. It generates synthetic data using simulator.

4. Technology used

4.1. Hadoop

Hadoop is a JAVA programming language and an open source framework, which allows processing of huge data sets in a distributed or parallel computing environment. Two main components of Apache Hadoop are Map Reduce (MR) and HDFS [25].

4.2. Hadoop distributed file system (HDFS)

HDFS is system that is planned to stock up very huge sets of data dependably and then to stream these sets at very elevated bandwidth to user applications. 'HDFS' is a file system which is having block-structure and it breaks a file into fixed size blocks for storing it into numerous machines; default block size of block is 64MB. Hadoop has two types of machines that work in a master-slave style, a Name Node as master machine and a number of Data Nodes as slave machines. [26][25] The Name Node allocates block ids to the blocks of a file and stores metadata (name of file, authorization, copy, each block's location) of file system in main memory which provides rapid access to this information. Data Nodes are nothing but the individual machines in the clusters that store and recover the replicated blocks of several files.

4.3. Map reduce

Map Reduce is backbone of Hadoop. It is the hub of Hadoop and a programming paradigm enabling mass scalability across plentiful servers in cluster of Hadoop. Hadoop Map Reduce is a software construction for simply scripting applications which route massive volumes of data in-parallel on big hardware clusters in a mistake-tolerant and dependable way. It is used as functional programming in artificial intelligence. It has also been reintroduced by GOOGLE to resolve the issue of big data analyzation. [27]

MAP and REDUCE are two key functions to be applied. The Map Reduce framework operates on (key, value) pairs [28] [29]. Ecosystem of hadoop as shown in fig.

Table.1: Hadoop Ecosystem

CRUNCH	HIVE	SOLAR INDEXER	PIG	SEARCH	IMPALA
Spark of Map Reduce					
Resource Management YARN Storage HDFS, Hbase					

4.5. Apache spark

Apache Spark is an innovative structure for spread computing which has been planned and made functional to enhance tasks of low-latency and to stock midway data and outcomes in memory. It is logically appropriate for machine learning. [30].It is a super-fast cluster computing technology, which has been designed for rapid computation. Based on Hadoop Map Reduce, it spreads the Map Reduce model to proficiently use it for additional types of computations, which comprises stream processing and interactive queries. In-memory cluster computing is the key feature of Spark that enhances an application's processing speed. Spark covers a varied series of workloads like interactive queries, iterative algorithms, batch applications and streaming. Other than supporting this workload in a defined system, it decreases the management load of upholding separate tools [29].

5. Conclusion

In this paper the big data's content, usage, characteristics, tools and techniques have been reviewed with the help of various Tools and techniques of Big Data. Multilevel association rule has also been reviewed which imparts additional and important information with respect to single level rule. With supportive LIVE example it has been shown that MAR helps in knowing customers buying behavior through establishing links and correlation among various items that are purchased by defined set of customers in defined time period.

References

[1] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2004.

[2] E.R. Omiecinski, Alternative interest measures for mining associations in data-bases, IEEE Trans. Knowl. Data Eng. 15(1) (2003) 57–69.

[3] E. H¨ullermeier, R. Kruse, and F. Hoffmann (Eds.): Interestingness Measures for Association Rules within Groups IPMU 2010, Part I, CCIS 80, 2010. Springer-Verlag Berlin Heidelberg 2010, 298–307.

[4] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proc. Int. Conf. of ACM–SIGMOD on Management of Data, 1993, pp.207–216

[5] H.J. Hamilton, D.R. Fudger, Estimating DBLearn's potential for knowledge dis-covery in databases, Comput. Intell. 11(2) (1995) 280–296.

[6] Shui Yu • Song Guo, Big Data Concepts, Theories, and Applications, Springer, ISBN 978-3-319-27763-9,

[7] P.Raj and Sathish A. P. Kumar, Big Data Analytics Processes and Platforms Facilitating SmartCities, 2017 JohnWiley & Sons,.

[8] www.sas.com/en_us/insights/analytics/big-data-analytics.html

[9] Ishwarappa,Anuradha, Introduction on Big data 5vs characterstics and Hadoop teachnology, Science Direct, procedia computer science 48(2015) 319-324.

[10] R.A. Angryk, F.E. Petry, Mining multi-level associations with fuzzy hierarchies, in: 14th IEEE Int. Conf. on Fuzzy System, 2005, pp.785–790.

[11] N.E. Oweis, M.M. Fouad, S.R. Oweis, S.S. Owais, V. Snasel, A novel MapReduce Lift Association Rule Mining Algorithm (MRLAR) for big data, Int. J. Adv. Com-put. Sci. Appl. 7(3) (2016).

[12] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, R.E. Gruber, Bigtable: a distributed storage system for structured data, ACM Trans. Comput. Syst. 26(2) (2008) 1–14.

[13] J.H.C. Yeung, C.C. Tsang, K.H. Tsoi, B. Kwan, C. Cheung, A.P.C. Chan, P.H.W. Leong, MapReduce as a programming model for custom computing machines, in: Proc. 16th IEEE Symposium on Field-Programmable Custom Computing Machines, FCCM'08, 2008.

[14] P.N. Tan, V. Kumar, J. Srivastava, Selecting the right objective measure for asso-ciation analysis, Inf. Sci. 29(4) (2004) 293–313.

[15] T. Brijs, K. Vanhoof, G. Wets, Defining interestingness for association rules, Int. J. Inf. Theories Appl. 10(4) (2003) 370–375.

[16] G. Shaw, Y. Xu, S. Geva, Eliminating redundant association rules in multilevel datasets, in: 4th Int. Conf. on Data Mining, Las Vegas, USA, 2008, pp.14–17.

[17] Y. Xu, G. Shaw, Y. Li, Concise representations for association rules in multilevel datasets, J. Syst. Sci. Syst. Eng. (2009) 53–70.

[18] Y. Xu, G. Shaw, Y. Li, Concise representations for association rules in multilevel datasets, J. Syst. Sci. Syst. Eng. (2009) 53–70.

[19] T. Fadi, S. Hammoud, Mr-arm: a map-reduce association rule mining frame-work, Parallel Process. Lett. 23(03) (2013) 1350012.

[20] Kalyan P. Subbu, Athanasios V. Vasilakos, Big Data for Context Aware Computing – Perspectives and Challenges, 7 October 2017, Big Data Research,

[21] Navroop Kaur, Sandeep K. Sood, Efficient resource management system based on 4Vs of big data streams, 2017, Big Data Research.

[22] DanieleApiletti, ElenaBaralis, TaniaCerquitelli, PaoloGarza, FabioPulvirenti∗, LucaVenturini, Frequent Itemsets Mining for Big Data: AComparative Analysis, 2017, Elsevier.

[23] DrewSchmidt, Wei-ChenChen, MikeMatheson, GeorgeOstrouchov, Programming with BIG Data in R: Scaling Analytics from One to Thousands of Nodes, 2016, Elsevier.

[24] Cihan Küçükkeçeci, Adnan Yazıcı, Big Data Model Simulation on a Graph Database for Surveillance in Wireless Multimedia Sensor Networks, 2017, Big Data Research.

[25] Apache Hadoop, http://hadoop.apache.org/, 2015.

[26] S. Singh, R. Garg and P. K. Mishra, "Review of Apriori Based Algorithms on MapReduce Framework," 2014 International Conference on Communication and Computing (ICC - 2014, 2014, pp. 593–604.

[27] J. Woo, Apriori-map/reduce algorithm, in: Int. Conf. on Parallel and Distributed Processing Techniques and Applications, PDPTA 2012, Las Vegas, 2012.

[28] M. Bakratsas, P. Basaras, D. Katsaros, L. Tassiulas, Hadoop MapReduce performance on SSDs for analyzing social networks, Big Data Research, July 11, 2017.

[29] Tom White, Hadoop: The Definitive Guide, April 2015: Printed in the United States of America.

[30] Jian Fu, Junwei Sun, Kaiyuan Wang, SPARK—A Big Data Processing Platform for Machine Learning. 978-1-5090-3575-5/16 $31.00 © 2016 IEEE.