Review of Algorithm Based on Semantic Similarity to Mine New Association Rules

Daljeet Kaur, Gagan Kumar

Computer Science and Engineering Department, MIET College, Mohri Kurukshetra, Haryana, India

Abstract— The problems of mining association rules in a database are introduced. Most of association rule mining approaches aim to mine association rules considering exact matches between items in transactions. A new algorithm called "Improved Data Mining Based on Semantic Similarity to mine new Association Rules" which considers not only exact matches between items, but also the semantic similarity between them. Improved Data Mining (IDM) Based on Semantic Similarity to mine new Association Rules uses the concepts of an expert to represent the similarity degree between items, and proposes a new way of obtaining support and confidence for the association rules containing these items. Rapid advances in both data storage and data capture technologies have resulted in a marked increase in the amount of data being stored in the database of businesses, government agencies and scientific sectors. Because of these advances, millions of records are being generated and stored, each of them containing tens or hundreds of fields. Many of these datasets are expanding on a daily basis. Traditional analyses of these types of datasets, involving human experts who manually analyze the data, are clearly no longer adequate. Association rule mining [5] finds interesting associations and/or correlation relationships among large set of data items. It provides information in the form of "if-then" statements. Association rules shows attribute value conditions that occur frequently together in a given dataset. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. A typical and widely used example of association rule mining is Market Basket Analysis. This paper Surveys some of the algorithms that have been introduced so far.

Keywords: Association Rule Mining, item set mining, transactional item set, candidate generation.

I. INTRODUCTION

DATA mining (DM), also known as knowledge discovery in databases (KDD), has been recognized as a new area for database research. This positive and evolutionary cycle is now occurring in area named data mining or knowledge discovery in database for efficiently discovering interesting rules from large collections of data. Informative knowledge discovering and new valuable data finding in database are very attractive in various business scenes.[1]

Data mining (DM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns such as association rules. Data mining has been defined as "the nontrivial extraction of implicit, previously Unknown and potentially useful information from data "and "the science of extracting useful information from large data sets or databases"[1]. It involves sorting through large amounts of data and picking out relevant information. It is usually used by businesses and other organizations, but is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimentation. Although data mining is a relatively new term, the technology is not. Companies for a long time have used powerful computers to sift through volumes of data such as supermarket scanner data, and produce market research reports. Continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy and usefulness of analysis.

It is an analytic method aimed to survey data (basically huge number of data connected to marketplace or company, shouted large data) in order to pursue reliable outlines and/or logical connections amid variables and next to validate the aftermath by requesting the outlines to new data. The aim of data excavating is to mine data from data and change it into construction suitable for more use or prediction. Data excavating is a critical pace in Vision Creation from Data (KDD) [1] procedure or supplementary models like Cross Industry Average Procedure for Data Excavating (CRISP-DM) [2].

Emerging as an affecting earth, Data excavating can be believed as convergence of several streams such as statistics, databases, contraption discovering, data science, and countless others. In supplement, according to the way utilized in the discovery of data, supplementary technical disciplines can be applied,

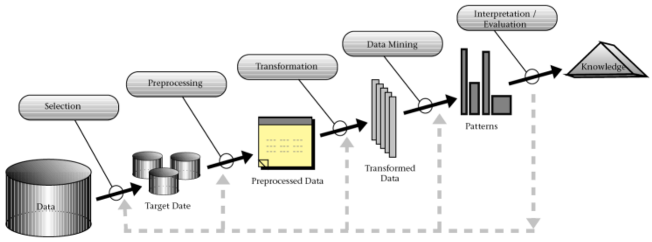

Figure 1 Data Mining as a step in KDD process

like furry logic, ANN [3], inductive software design, high-performance computing or vision presentation techniques. Due to the contribution of a expansive scope of disciplines to data excavating, we anticipate excavating scutiny to produce a expansive collection of data excavating systems. Therefore, critical task is to furnish an unambiguous categorization of data excavating arrangements, so that users can facilely differentiate amid these arrangements and choose on those that are best suited for their requirement. The data excavating tasks can be believed of generally two types: illustrative and predictive.

(a) Descriptive mining task: Clustering
Clustering is unsupervised method of partitioning task points by usual clusters, shouted clusters, such as points of a cluster are extremely comparable, as the points across hierarchical clusters are additionally dissimilar, created density established graphs and spectral clustering. It begins alongside the clustering methods, that contain K-means [5][6] and Expectation-Maximization (EM) algorithms [7]. K-means being a voracious algorithm minimizes the squared error of points from their corresponding cluster way, and implements hard clustering whereas every single point is allocated to just one cluster. A little supplementary methods of illustrative excavating are association law excavating or sequential outline invention.

(b) Predictive mining task: Classification
The association task is supervised discovering method to allocate individual unlabeled points to corresponding clusters[8]. Formally, a classifier is a purpose M that predicts the class label y for a given input instance x, that is, y = M (x), whereas y {C1, C2, , Ck} is the forecasted class label (a categorical attribute value) approximated as O(d2). The naive Bayes classifier [9] assumes all qualities to be autonomous therefore making the estimation of merely O(d). To onset alongside, it needs a training set that specifies a little points alongside correct clusters to that these points belong. The decision tree classifier comes alongside the strengths of producing models moderately simpler to comprehend in analogy to others. The prop vector contraption (SVM) [10] is the most accepted classifier for several setback domains. Its aim is to maximize the margin amid clusters by discovering the optimal hyper plane. Moreover, SVM kernel traps are helpful in discovering the non-linear borders, that though correspond to precise linear hyper plane of a little "non-linear" high-dimensional space. An vital task in the association is to assess how well the models. Separately from association deviation detection or regression can additionally be utilized as predictive methods.

(c) Association rule mining

Association law excavating [11], one of the most vital and well analyzed methods of data excavating, was early gave in. It aims to remove interesting correlations, recurrent outlines, associations or casual constructions amid sets of items in the deal databases or supplementary data repositories. Association laws are extensively utilized in assorted spans such as telecommunication webs, marketplace and chance association, catalog manipulation etc. Assorted association excavating methods and algorithms will be briefly gave and contrasted later.

Association law excavating is to find out association laws that gratify the predefined minimum prop and assurance from a given database. The setback is normally decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are shouted recurrent or colossal item sets. The subsequent setback is to produce association laws from those colossal item sets alongside the constraints of negligible confidence.

Formal Definition
Let $I=I_1, I_2, \dots , I_m$ be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called itemsets, and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent, the rule means X implies Y.

There are two vital frank measures for association laws, support(s) and confidence(c). As the database is colossal and users concern concerning merely those oftentimes bought items, normally thresholds of prop and assurance are predefined by users to drop those laws that are not so interesting or useful. The two thresholds are shouted negligible prop and negligible assurance respectively. Support(s) of an association law is described as the percentage/fraction of records that encompass $X \cup Y$ to the finished number of records in the database. Presume the prop of an item is 0.1%, it way merely 0.1 percent of the deal encompass buying of this item.

Confidence of an association law is described as the percentage/fraction of the number of deals that encompass $X \cup Y$ to the finished number of records that encompass X. Assurance is a compute of strength of the association laws, presume the assurance of the association law $X \Rightarrow Y$ is 80%, it way that 80% of the deals that encompass X additionally encompass Y together.

(d) Transactional Data Mining

Association law excavating (ARM) identifies recurrent item-sets from databases and generates association laws by pondering every single item in equal value. But in real dataset the items are extremely disparate in countless features for number of real globe requests, for example retail marketing, web log, etc. The difference amid items makes a forceful encounter on the decision making in these applications. Therefore, established ARM is not a good choice for such setback domains. ARM is the early choice for marketplace hamper analysis. It is additionally functional in discovering correlations, recurrent outlines or co-occurrences, and associations amid items in transactional sets. ARM uses measures like prop, assurance, lift and insufficient others to find association laws for patterns. ARM is a two pace procedure such that early pace attempts to find out all recurrent item-sets alongside deals by employing minimum prop worth whereas in subsequent pace minimum assurance is utilized to produce laws for items.

II. Related Work

Longbing Cao et al, in "Combined Mining: Discovering Informative Knowledge in Complex Data" 2011 [12], the authors describe Enterprise data mining applications often involve complex data such as multiple large heterogeneous data sources, user preferences, and business impact. In such situations, a single method or one-step mining is often limited in discovering informative knowledge. It would also be very time and space consuming, if not impossible, to join relevant large data sources for mining patterns consisting of multiple aspects of information. It is crucial to develop effective approaches for mining patterns combining necessary information from multiple relevant business lines, catering for real business settings and decision-making actions rather than just providing a single line of patterns. The recent years have seen increasing efforts on mining more informative patterns, e.g., integrating frequent pattern mining with classifications to generate frequent pattern-based classifiers. Rather than presenting a specific algorithm, this paper builds on their existing works and proposes combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods on demand. They summarize general frameworks, paradigms, and basic processes for multifeature combined mining, multisource combined mining, and multimethod combined mining. Novel types of combined patterns, such as incremental cluster patterns, can result from such frameworks, which cannot be directly produced by the existing methods. A set of real-world case studies has been conducted to test the frameworks, with some of them briefed in this paper. They identify combined patterns for informing government debt prevention and improving government service objectives, which show the flexibility and instantiation capability of combined mining in discovering informative knowledge in complex data.

Romero, C. et al, in "Educational Data Mining: A Review of the State of the Art" 2010 [13], the authors describe Educational data mining (EDM) is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyze educational data in order to study educational questions. This paper surveys the most relevant studies carried out in this field to date. First, it introduces EDM and describes the different groups of user, types of educational environments, and the data they provide. It then goes on to list the most typical/common tasks in the educational environment that have been resolved through data-mining techniques, and finally, some of the most promising future lines of research are discussed.

Schluter, T. et al, in "About the analysis of time series with temporal association rule mining" 2011 [14], the authors describe This paper addresses the issue of analyzing time series with temporal association rule mining techniques. Since originally association rule mining was developed for the analysis of transactional data, as it occurs for instance in market basket analysis, algorithms and time series have to be adapted in order to apply these techniques gainfully to the analysis of time series in general. Continuous time series of different origins can be discretized in order to mine several temporal association rules, what reveals interesting coherences in one and between pairs of time series. Depending on the domain, the knowledge about these coherences can be used for several purposes, e.g. for the prediction of future values of time series. They present a short review on different standard and temporal association rule mining approaches and on approaches that apply association rule mining to time series analysis. In addition to that, they explain in detail how some of the most interesting kinds of temporal association rules can be mined from continuous time series and present an prototype implementation. They demonstrate and evaluate their implementation on two large datasets containing river level measurement and stock data.

Chun-Wei Tsai et al, in "Data Mining for Internet of Things: A Survey" 2014 [15], the authors describe It sounds like mission impossible to connect everything on the Earth together via Internet, but Internet of Things (IoT) will dramatically change their life in the foreseeable future, by making many "impossibles" possible. To many, the massive data generated or captured by IoT are considered having highly useful and valuable information. Data mining will no doubt play a critical role in making this kind of system smart enough to provide more convenient services and environments. This paper begins with a discussion of the IoT. Then, a brief review of the features of "data from IoT" and "data mining for IoT' is given. Finally, changes, potentials, open issues, and future trends of this field are addressed.

Yang, Z. et al, in "Association Rule Mining-Based Dissolved Gas Analysis for Fault Diagnosis of Power Transformers" 2009 [16], the authors describe This paper presents a novel association rule mining (ARM)-based dissolved gas analysis (DGA) approach to fault diagnosis (FD) of power transformers. In the development of the ARM-based DGA approach, an attribute selection method and a continuous datum attribute discretization method are used for choosing user-interested ARM attributes from a DGA data set, i.e. the items that are employed to extract association rules. The given DGA data set is composed of two parts, i.e. training and test DGA data sets. An ARM algorithm namely Apriori-Total From Partial is proposed for generating an association rule set (ARS) from the training DGA data set. Afterwards, an ARS simplification method and a rule fitness evaluation method are utilized to select useful rules from the ARS and assign a fitness value to each of the useful rules, respectively. Based upon the useful association rules, a transformer FD classifier is developed, in which an optimal rule selection method is employed for selecting the most accurate rule from the classifier for diagnosing a test DGA record. For comparison purposes, five widely used FD methods are also tested with the same training and test data sets in experiments. Results show that the proposed ARM-based DGA approach is capable of generating a number of meaningful association rules, which can also cover the empirical rules defined in industry standards. Moreover, a higher FD accuracy can be achieved with the association rule-based FD classifier, compared with that derived by the other methods.

Mukhopadhyay, A. et al, in "Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II" 2014 [17], the authors describe This paper is the second part of a two-part paper, which is a survey of multiobjective evolutionary algorithms for data mining problems. In Part I , multiobjective evolutionary algorithms used for feature selection and classification have been reviewed. In this part, different multiobjective evolutionary algorithms used for clustering, association rule mining, and other data mining tasks are surveyed. Moreover, a general discussion is provided along with scopes for future research in the domain of multiobjective evolutionary algorithms for data mining.

Kantardzic, M. in "Advances in Data Mining" 2011 [18], the authors describe This chapter contains sections titled: Graph Mining Temporal Data Mining Spatial Data Mining (SDM) Distributed Data Mining (DDM) Correlation Does Not Imply Causality Privacy, Security, and Legal Aspects of Data Mining Review Questions and Problems References for Further Study

Maqbool, O. et al, in "Metarule-guided association rule mining for program understanding" 2005 [19], the authors describe Software systems are expected to change over their lifetime in order to remain useful. Understanding a software system that has undergone changes is often difficult owing to the unavailability of up-to-date documentation. Under these circumstances, source code is the only reliable means of information regarding the system. In the paper, association rule mining is applied to the problem of software understanding i.e. given the source files of a software system, association rule mining is used to gain an insight into the software. To make association rule mining more effective, constraints are placed on the mining process in the form of metarules. Metarule-guided mining is carried out to find associations which can be used to identify recurring problems within software systems. Metarules are related to re-engineering patterns which present solutions to these problems. Association rule mining is applied to five legacy systems and results presented show how extracted association rules can be helpful in analysing the structure of a software system and modifications to improve the structure are suggested. A comparison of the results obtained for the five systems also reveals legacy system characteristics, which can lead to understanding the nature of open source legacy software and its evolution.

Lihua Zhang et al, in "CSRule: Efficient mining of closed strong association rules in the resource effectiveness matrix" 2013 [20], the authors describe Association rules which are used to mine resource failure, can reduce the number of wrong alarm resources to be replaced. In this paper, they proposed an efficient algorithm: CSRule, for mining closed strong association rules based on merging strategies. Firstly, it generates all pairs of association rules which only have two resources; secondly, all the closed strong association rules are mind using merging strategies. In order to improve the mining efficiency of the algorithm, CSRule algorithm adopts several pruning strategies to mine closed strong association rules without candidate maintenance. For the escape of low efficiency on secondary mining based on definition, CSRule algorithm can mine closed strong association rules at one time. The experimental results show their algorithm is more efficient than traditional algorithm.

III. Conclusion

Association law excavating has a expansive scope of requests encompassing marketplace health diagnosis, scutiny, exploration, protection and countless more. In this paper, we surveyed the trials of continuing association law excavating approaches. The established algorithms of association laws excavating encounter countless well recognized setbacks in exercise for example, the algorithms do not always revisit the aftermath in a reasonable period, as the set of association laws and generated Candidates can quickly produce rapidly. For Transactional data excavating the procedure becomes even tough compounded by the fact that discovering transactional database is costly. The larger the set of recurrent item sets the extra the number of laws gave to the user most of them are redundant. A extra generalized way for excavating item set above synthetic transactional data is required; our upcoming work will complement this task.

IV. References

[1] Pazzani, Michael J. "Knowledge discovery from data?." Intelligent systems and their applications, IEEE 15, no. 2 (2000): 10-12.

[2] Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining." In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp. 29-39. 2000.

[3] Craven, Mark W., and Jude W. Shavlik. "Using neural networks for data mining." Future generation computer systems 13, no. 2 (1997): 211-229.

[4] Zhai, Zhongwu, Bing Liu, Hua Xu, and Peifa Jia. "Clustering product features for opinion mining." In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 347-354. ACM, 2011.

[5] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." In KDD workshop on text mining, vol. 400, no. 1, pp. 525-526. 2000.

[6] Wagstaff, Kiri, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k-means clustering with background knowledge." In ICML, vol. 1, pp. 577-584. 2001.

[7] Moon, Todd K. "The expectation-maximization algorithm." Signal processing magazine, IEEE 13, no. 6 (1996): 47-60.

[8] Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005).

[9] Murphy, Kevin P. "Naive Bayes classifiers." University of British Columbia (2006).

[10] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9, no. 3 (1999): 293-300.

[11] Hipp, Jochen, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. "Algorithms for association rule mining—a general survey and comparison." ACM sigkdd explorations newsletter 2, no. 1 (2000): 58-64.

[12] Longbing Cao; Huaifeng Zhang; Yanchang Zhao; Dan Luo; Zhang, C.,"Combined Mining: Discovering Informative Knowledge in Complex Data",IEEE,Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,2011

[13] Romero, C.; Ventura, S.,"Educational Data Mining: A Review of the State of the Art",IEEE,Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,2010

[14] Schluter, T.; Conrad, S.,"About the analysis of time series with temporal association rule mining",IEEE,Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on,2011

[15] Chun-Wei Tsai; Chin-Feng Lai; Ming-Chao Chiang; Yang, L.T.,"Data Mining for Internet of Things: A Survey",IEEE,Communications Surveys & Tutorials, IEEE,2014

[16] Yang, Z.; Tang, W.H.; Shintemirov, A.; Wu, Q.H.,"Association Rule Mining-Based Dissolved Gas Analysis for Fault Diagnosis of Power Transformers",IEEE,Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,2009

[17] Mukhopadhyay, A.; Maulik, U.; Bandyopadhyay, S.; Coello, C.A.C.,"Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II",IEEE,Evolutionary Computation, IEEE Transactions on,2014

[18] Kantardzic, M.,"Advances in Data Mining",Wiley-IEEE Press,Data Mining:Concepts, Models, Methods, and Algorithms,2011

[19] Maqbool, O.; Babri, H.A.; Karim, A.; Sarwar, M.,"Metarule-guided association rule mining for program understanding",IET,Software, IEEE Proceedings -,2005

[20] Lihua Zhang; Zhengjun Zhai; Miao Wang; Guoqing Wang,"CSRule: Efficient mining of closed strong association rules in the resource effectiveness matrix",IEEE,Biomedical Engineering and Informatics (BMEI), 2013 6th International Conference on,2013