ISSN:0975-9646

Review on Mining Association Rule from Semantic Data

Kalyani A.Kale
Computer Science and Engg. Department
Amravati University.
P.R.M.C.E.A.M Badnera,
Amravati [MH] India

Prof **R.P.Sonar**
Computer Science and Engg. Department
Amravati University.
P.R.M.C.E.A.M Badnera,
Amravati [MH] India

Abstract: **The amount of ontology and semantic annotations for various data of broad applications is constantly growing. This type of complex and heterogeneous semantic data has created new challenges in the area of data mining research. Association Rule Mining is one of the most common data mining techniques which can be defined as extracting the interesting relation among large amount of transactions. Since this technique is more concerned about data representation, it is the most challenging data mining technique to be applied on semantic data. Moreover, the Semantic technologies offer solutions to capture and efficiently use the domain knowledge. This paper presents review on researches done in association rule mining.**

Keywords: Data mining, Association rules, semantic data

I. INTRODUCTION:

The main Purpose of data mining is to disclose the hidden information from the database. Due to the growth of data volume in an organization sectors like banking, marketing, telecommunication, manufacturing, and transportation etc, a different technique for deletion of repetitive data and conversion of data to more usable forms has been proposed under data mining. Data mining also known as knowledge discovery is used to discover useful patterns from the database. Many techniques have been developed in data mining amongst which association rule mining is very important. Apriori is one of the best algorithms for the association rule mining. The Apriori algorithm discover the frequent patterns from database whose support and confidence must satisfy the minimum support and confidence.

II. DATA MINING:

Data mining is an inter-disciplinary subfield of computer science, also called as "Knowledge Discovery in Databases" process. It is a process to extracting useful and interesting knowledge from large datasets. Data mining is the computational process involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The knowledge modes data mining discovered have a variety of different types. Including models like association model, classification model, class model, sequence pattern, etc. The overall goal of the data mining process is to extract information from a database and transform it into an understandable structure so that anyone can use it in future. The actual data mining task extract previously unknown interesting patterns such as groups of data records by cluster analysis, unusual records called anomaly detection and dependencies by association rule mining. Mining association rules is one of the most important aspects in data mining.

III. SEMANTIC DATA:

The Semantic data is organized in such a way that it can be interpreted meaningfully without human intervention. Since, 1970 to current semantic data issued in a wide variety of data management systems and applications. Based on relationships between stored symbols and the real world it is a software engineering model. The designed Goals of Semantic Data system is to represent the real world as accurately as possible within some data set. There is linear and hierarchical organization of data to give certain meanings like in below example. Semantic data allow the real world within data sets by representing, machines to interact with worldly information without human interpretation. This semantic data is organized on binary models of objects, mostly in groups of three parts consisting of two objects and their relationship. Consider example, if one wanted to represent a pen is on a letter book, the organization of data might look like: PEN LETTER BOOK. The objects (pen and letter book) are interpreted with regard to their relationship. The data is organized linearly, telling the software that as PEN comes first in the line, it is the object that acts. i.e., the position of the word makes the software to understand that the pen is on the letter book and not that the letter book is sitting on the pen. Databases designed in this concept have greater applicability and are easily integrated into other databases. Since, this semantic data is developing from 1970; its uses are growing on increasing and reach too many important applications. It has very important applications for the enterprise world. Database Management Systems can be integrated with one another and compared. It is helpful model for streamlining the relationship between company and vendors, making database sharing and integration much simpler.

IV. ASSOCIATION RULES:

Association Rule Mining (ARM) is the most important and researched techniques of data mining. ARM was first introduced by Agrawal et al. 1993.It is association tools for

analyzing customer purchasing habit, such as market-basket analysis. ARM aims to extract interesting frequent patterns, association among set of items or database. Association Rules are if/then statements that help to discover relationships among unrelated data in a data repository. Many algorithms are proposed for finding frequent item sets for large datasets Association rule uses two criteria support and confidence to identify the relationships and rules are generated by analyzing data for frequent if/then pattern. Association rules are generally needs to satisfy a minimum support and a minimum confidence at the same time.

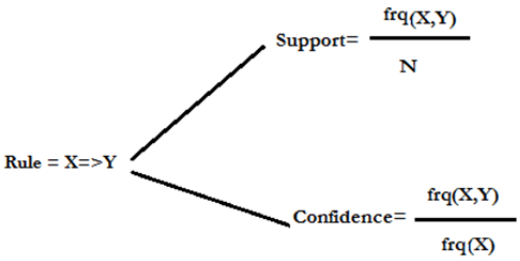

Fig.1. Association Rule

V. ASSOCIATION RULE MINING:

In Data Mining Association rule learning is a method for discovering interesting relations between variables in large database. Association rule discovers interesting association/correlation among a large set of data items. The sales of Super market would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. This information will help business to know the behavior of the customers. Shopping centers also uses the association rule mining to place the items next to each other so that user buy more items .Another application of Association mining is the goggle auto complete, where we type a word it searches frequently associated words that user type after that particular word.

VI. CONCEPT OF ASSOCIATION RULE MINING:

Support(S) of an association rule is defined as the percentage/fraction of records that contain X∪Y to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1percent of the transaction contain purchasing of this item. (If x and y are two items in database then both comes together).

Support (XY) = Support count of (XY) / Total number of transaction

Confidence(C) of an association rule is defined as the percentage/fraction of the number of transactions that contain X∪Y to the total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X⇒Y is 80%, it means that 80% of the transactions that contain X also contain Y together.

Confidence (X|Y) = Support (XY) / Support (X)

Item*:* It is a field of transactional database.

Consider the following Transactional database Table-I:

Table-I: Transactional Database

Transaction ID	Milk	Bread	Butter
1.	1	1	1
2.	0	0	1
3.	0	0	0
4.	1	1	1
5.	0	1	0

In Table I, 1 represent the presence of item and 0 represent the absence of items. Now let's count the support and confidence.

Consider X= milk and Bread, Y = Butter.

Support {milk, Bread} → {Butter} = Support (X→Y)
= 1/5
= 0.2(20%)

Confidence {Milk, Bread}→{Butter} = Confidence (X→Y)
=0.2/0.4
=0.5(50%)

Support says that milk butter and bread all purchased together while confidence says that whenever milk and bread purchased there is also possibility of butter.

Association rule mining usually split into two separate steps:

1. Apply minimum support value to find all frequent item sets in a database. This steps required more attention
2. Form rules by using frequent item sets and the minimum confidence value.

VII. GENERATION MODEL OF ASSOCIATION RULE:

Association Rules Generation contains many process, they can easily understand by the following related model. The model of Association Rule Generation is in Fig. 2.

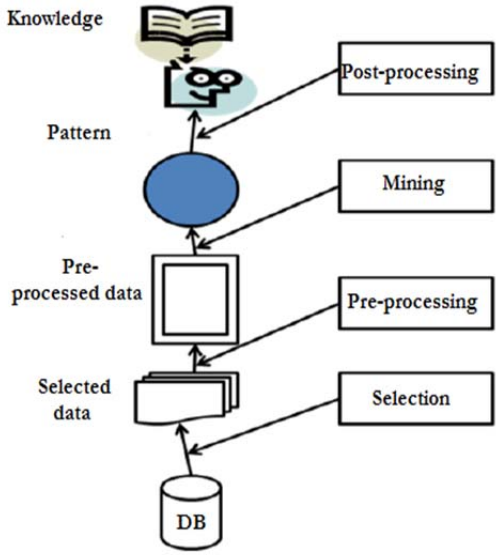

Fig.2. Model of Association Rule Generation

Association Rule Generation contain of processes as selection of database from the large repository, then preprocessing on selected data, after that mine the candidate frequent item sets from the preprocessed data, then prune the frequent item sets according to a given threshold. Such a way rules are generated then association rules are mined according to given support and confidence.
ARM is mainly used for mining the frequent and infrequent item sets from the large databases. It is based on two principles - support and confidence. Association rules are the if/then sentences to show the relationship among the item sets. One of the most important ARM algorithms is Apriori algorithm.

- **Apriori Algorithm:**

The Apriori algorithm is one of the most popular algorithm in the mining of association rules in a centralized database. Farah Hanna AL-Zawaidah , Yosef Hasan Jbara in [3] proposed the Apriori Algorithm for finding the frequent itemsets .The Apriori Algorithm proposed to finds frequent items in a given data set. The name of Apriori is based on the fact that the algorithm uses a prior knowledge of frequent item set properties. Apriori employs an iterative approach known as a level wise search, where k item sets are used to explore (k+1) itemsets . This algorithm contains a number of passes over the database. During pass k, the algorithm finds the set of frequent item sets Lk of length k that satisfy the minimum support requirement. Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data. To illustrate the concepts, there is a small example from the supermarket domain. The set of items is I = {milk, bread, butter, beer} and a small database containing the items. An example association rule for the supermarket could be {milk, bread} => {butter} meaning that if milk and bread is bought, customers also bought butter. Apriori uses bottom up strategy. It is the most famous and classical algorithm for mining frequent patterns. Apriori algorithm works on categorical attributes. Apriori uses breadth first search

VIII. LITERATURE REVIEW:

Mining Association Rules is one of the most important in data mining. Association rules are of interested in database researchers and data mining users. Since 90s, different approaches of data mining have been proposed for discovering useful knowledge from very large semantic datasets. A survey of previous research in the area is provided below:

- "Ashraf Sadat Heydari Yazdi,and Mohsen Kahani" in their paper titled "A Novel Model for Mining Association Rules from Semantic Web Data" has stated, two general phases in semantic association rule mining system:1)semantic transaction production and 2)running semantic association rule mining algorithm on them. The algorithm is rewritten to deal with semantic transactions and semantic rules, with their predefined format in the ontology will be resulted [1].
- "Rakesh Agrawal, Tomasz Imielinski and Arun Swami" in their paper titled "Mining Association Rules between sets of Items in Large Database" has stated that if there is a large database of customer transactions, the Memory reclamation algorithm defined in the paper incorporates buffer management, novel estimation and pruning techniques. They also present results of applying this algorithm to sales data customer commercial obtained from a large retailing company, which intriguing shows the effectiveness of the algorithm [2].
- "Farah Hanna AL-Zawaidah and Yosef Hasan Jbara" in their paper titled "An Improved Algorithm for Mining Association Rules in Large Databases" stated that, Mining association rules in large databases is a topic of data mining. The approach proposed in this paper is derived from the conventional Apriori approach with features added to improve data mining performance. The approach to attained desired improvement is to create efficient new algorithm out of the conventional extensive one by adding new features to the Apriori approach. The proposed mining transaction and algorithm can efficiently discover the association rules between the large items in large database. They have performed extensive experiments and compared the performance of their algorithm with existing discovering algorithms found in the literature [3].
- "S.C.Punitha, P. Ranjith Jeba Thangaiah and M. Punithavalli" in their paper titled "Performance Analysis of Clustering using Partitioning and Hierarchical Clustering Techniques" stated the HAC method which gives various algorithmic aspects, including well-definiteness and computational properties. The basic idea of the algorithm in HAC method is to merge documents based on their similarity into clusters. This method starts with each example in its own cluster and iteratively combines them to form larger and larger clusters. The effectiveness of this technique is improving the search efficiency over sequential scans method [4].
- "Peter Fule and John F. Roddick" in their paper titled "Experiences in Building a Tool for Navigating Association Rule Result Set" stated the model IRSetNav which has capabilities in redundant rule reduction, subjective interestingness evaluation, item and item set pruning, related information searching, text-based item set and rule visualization, hierarchy based searching and tracking changes between data sets using a knowledge base. It also incorporates several techniques that have found to be useful for speeding up the knowledge discovery process. And also reduce iterations in the knowledge discovery process by reducing its iterative nature [5].
- "Mohd Helmy Abd, Mohd Norzali Haji Mohd and Mohamad Mohsin" in their paper titled "Discovering Web Server Logs Patterns Using Generalized Association Rules Algorithm" focused on the aspect

of web usage mining. They stated that as commercial companies as well as academic researchers developed an array of tools that perform several data mining algorithms on log Files coming from web servers in order to identify user behavior on a particular web site. Performing this kind of investigation on the web site can provide information that can be used to better accommodate the user's needs [6].

- "Ming-Cheng Tseng ·Wen-Yang Lin and Rong Jeng" in their paper titled "Updating generalized association rules with evolving taxonomies" stated the problem of updating the discovered generalized association rules under evolving taxonomies. For this purpose they proposed two algorithms Diff _ET and Diff_ET2 are used for updating generalized frequent item sets. And evaluation showed that both algorithms are effective and have good linear scale-up characteristics [7].
- Zahir Tari and Wensheng Wu in their paper titled "ARM: A HYBRID ASSOCIATION RULE MINING ALGORITHM" stated that Most of the approaches for association rule mining focus on the performance of the discovery of the frequent item sets. They are based on the algorithms that require the transformation of from one representation to another, and therefore excessively use resources and incur heavy CPU overhead. They Proposes a hybrid algorithm that is resource efficient and provides better performance. In addition, they propose a comparison algorithm (CmpApr) that compares candidate item sets with a transaction, a filtering algorithm (FilterApr) that reduces the number of comparison operations required to find frequent item sets. ARM has better performance and scales linearly [8].

IX. CONCLUSION:

Association rule mining is an interesting topic of research in the field of data mining. The paper gives a basic idea about the terms related to association rule mining. Association rules are basic data mining tools for initial data exploration usually applied to large data sets, seeking to identify the most common groups of items occurring together. There are various association rule mining algorithms. This paper studied the most frequently used association rule mining algorithms i.e. Apriori algorithm which is used for discovering all significant association rules between items in a large database of transactions. This paper also studied and represents the reviews on the recent researches done in the field of association rule mining. However, association rule mining is still in a stage of exploration and development.

REFERENCES:

[1] Ashraf Sadat Heydari Yazdi, Mohsen Kahani, "A Novel Model for Mining Association Rules from Semantic Web Data" in Engineering Faculty Ferdowsi University of Mashhad, 978-1-4799-3351-8/14/$31.00 ©2014 IEEE

[2] Rakesh Agrawal, Tomasz Imielinski and Arun Swami, "Mining Association Rules between Sets of Items in Large Databases" in *IBM Almaden Research Center650 Harry Road, San Jose, CA 95120 2012*

[3] Farah Hanna AL-Zawaidah , Yosef Hasan Jbara," An Improved Algorithm for Mining Association Rules in Large Databases" in *World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 311-316, 2011.*

[4] S.C. Punitha, P. Ranjith Jeba Thangaiah and M. Punithavalli, "Performance Analysis of Clustering using Partitioning and Hierarchical Clustering Techniques" in International Journal of Database Theory and Application Vol.7, No.6 (2014), pp.233-240

[5] Peter Fule and John F. Roddick, "Experiences in Building a Tool for Navigating Association Rule Result Set" Copyright c 2004, Australian Computer Society, Australasian Workshop on Data Mining andWeb Intelligence (DMWI04), Dunedin, New Zealand.Conferences in Research and Practice in Information Technology, Vol. 32.

[6] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Mohamad Farhan Mohamad Mohsin, "Discovering Web Server Logs Patterns Using Generalized Association Rules Algorithm" in Intech ISBN: 978-953-307-067-4, 2010

[7] Ming-Cheng Tseng ·Wen-Yang Lin · Rong Jeng, "Updating generalized association rules with evolving taxonomies" in Appl Intell (2008) 29: 306–320 DOI 10.1007/s10489-007-0096-5

[8] Zahir Tari and Wensheng Wu, "ARM: A HYBRID ASSOCIATION RULE MINING ALGORITHM" in Springer journal, 2006

[9] V. Nebot, R. Berlanga, "Finding association rules in semantic web data, Knowledge-Based Systems", Vol. 25, 2012

[10] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data – The Story So Far. In: IJSWIS, Vol. 5, Issue 3, Pages 1-22, 2009.

[11] Mahendra Thakur, Geetika S. Pandey, "Performance Based Novel Techniques for Semantic Web Mining" in IJCSI International Journal of Computer Science Issues, Vol. 9, January 2012

[12] Jaideep Srivastava_y, Robert Cooleyz, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" in ACM SIGKDD, Volume 1, Jan 2000.

[13] Mingzhu Zhang and Changzheng He, "Survey on Association Rules Mining Algorithms" in Advancing Computing, Communi., Control and Management, LNEE 56, pp. 111–118 © Springer-Verlag Berlin Heidelberg 2010

[14] Karin Becker a,*, Mariângela Vanzin, "O3R: Ontology-based mechanism for a human-centered environment targeted at the analysis of navigation patterns" in K. Becker, M. Vanzin / Knowledge-Based Systems 23 (2010) 455–470 0950-7051/$ 2010

[15] ANDREA TAGARELLI and SERGIO GRECO, "Semantic Clustering of XML Documents" in ACM Transactions on Information Systems C_ 2010 ACM 1046-8188/2010/01-ART3 $10.00

[16] Ansaf Salleb-Aouissi, Christel Vrain Cyril Nortet,Xiangrong Kong, Vivek Rathod, and Daniel Cassard "QuantMiner for Mining Quantitative Association Rules" in Journal of Machine Learning Research 14 (2013) 3153-3157

[17] Janki M. Padaliya, Arjun V. Bala, " Mining Data Using Various Sequential Patterns Mining Algorithm in Semantic Web Environment" in International Journal of Engineering Development and Research © 2014 IJEDR | Volume 2,| ISSN: 2321-9939

[18] Ali Harb and Kafil Hajlaoui , "Enhanced Semantic Automatic Ontology Enrichment" in *10th International Conference on Intelligent Systems Design and Applications* 978-1-4244-8136-1/10/$26.00_c 2010 IEEE

[19] Tasawar Hussain, Muhammad Abdul Qadir and Sohail Asghar, "Fuzzification of Web Objects: A Semantic Web Mining Approach" in IJCSI International Journal of Computer Science Issues, Vol. 9, March 2012

[20] Ms. Arti Rathod, Mr. Ajaysingh Dhabariya and Mr. Chintan Thacker, "A Review on Association Rule Mining and Improved Apriori Algorithms" in International Journal of Scientific Research in Computer Science (IJSRCS) Vol. 1, Issue. 1, Sep. 2013