Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 4, March-2019, ISSN 2230-7540

Customer Relationship Management using Apriori Algorithm and Frequent Item Sets in Data Mining

Manoj Semwal[1]* Rahul Joshi[2]

[1] M.Tech Scholar, Department of Computer Science and Engineering, Doon Institute of Engineering & Technology, Dehradun, India

[2] Assistant Professor, Department of Computer Science and Engineering, Doon Institute of Engineering & Technology, Dehradun, India

Abstract – Data mining (DM) is method of discovering exciting pattern & information via huge quantity of data and also known as knowledge discovery from data. In this we start via investigate information in pattern extraction via Customer Relationship Management (CRM) datasets. Association rule mining (ARM) has been considered broadly in Knowledge Discovery in Databases (KDD) field for pattern extraction (PE) & there exist several known algorithms to execute. The support & confidence thresholds are normally utilized to direct search for exciting patterns. KDD from databases includes usage of different methods and algorithms such as Association rules, Predictions, Decision trees, Artificial intelligence. Advanced data mining techniques are used to discover relationships and other hidden patterns. ARM is among essential DM applications. From our literature survey, I observed that nearly all of pattern mining system is extensive; few practical issues may happen while amount of things in every record are very huge. DM is termed as "Nontrivial pulling out of implicit, formerly unknown & potentially helpful information from data" & "science of extracting useful information from wide data sets or databases". It is core rule of KD method that has data selection, pre-processing & cleaning, transformation & reduction, evaluation, & visualization. In this paper we will discuss about Customer Relationship Management using Apriori Algorithm

----------X----------

INTRODUCTION

DM has been extended ahead of the limits to relate to several form of data analysis. The frequent definitions of DM or KDD are nontrivial extraction of implicit, formerly unidentified, unknown, & potentially functional information from data. This encompass numerous various technical demands like clustering, data summarization, knowledge classification rules, finding dependency networks, analyzing change, & to detect anomaly

DATA MINING:

Discovering knowledge from databases includes usage of different techniques and algorithms such as Association rules, Classification, Predictions, Decision trees, Artificial intelligence. These techniques also include neural networks, clustering and neural networks. Advanced DM methods are utilized to determine relationships & other hidden patterns. ARM is among significant DM applications. Steps present in knowledge discovery are:

1. **Data selection**: recovery of relevant data from databases.
2. **Pre-processing & cleaning**: Elimination of noise & inconsistent data, finding & dealing with missing values.
3. **Transformation & reduction**: data sets are condensed to lesser size may through sampling or outline statistics. As ex. tables of data may be replaced via expressive statistics like mean & standard deviation.
4. **Data mining**: Intelligent methods are selected for pattern extraction.
5. **Evaluation**: patterns recognized via DM process are interpreted, for instance, determining clinical significance of findings.
6. **Visualization**: knowledge representation techniques like pie charts & graphs are utilized to present mined knowledge to user.

7. *Descriptive mining* repeatedly extracts fresh or useful information from huge databases & present discovered information in naturally understandable terms for human analysis. ARM is well-studied *descriptive mining* process in KDD area [21-24]. Their primary strength lies in their major easy-to-read power & their being relatively easy to comprehend, thus making them appropriate for incorporation into decision-making processes.

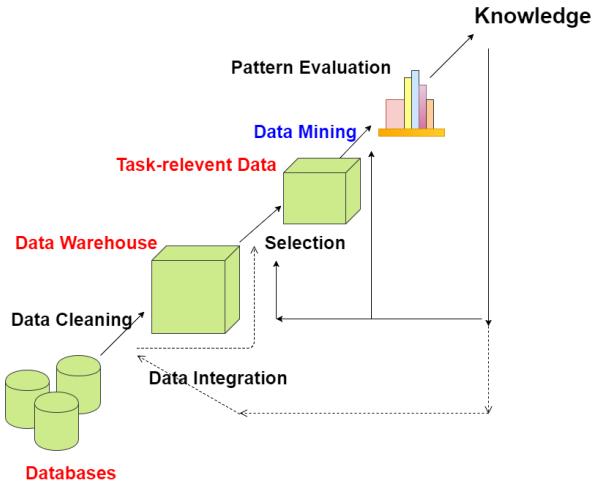

Figure 1 Knowledge discovery process

Descriptive mining repeatedly extracts fresh or useful information from huge databases & present discovered information in naturally understandable terms for human analysis. ARM is well-studied *descriptive mining* process in KDD area [21-24]. Their primary strength lies in their major easy-to-read power & their being relatively easy to comprehend, thus making them appropriate for incorporation into decision-making processes.

Data warehouse

A DW is Relational Database Management System intended particularly to gather requirements of On Line Transaction Processing systems. It can be freely distinct as any centralized data repository or a data warehouse can be queried for business benefits. DW and mining potential can be improved if suitable data has been composed & preserved in DW of huge data storage. Data warehousing is an influential method may extract archived operational data & conquer inconsistencies among unusual legacy data formats. In addition to integrating data throughout an organization, despite of location, format, technology or communication necessities, it is probable to fit in supplementary or expert and professional information.

DATA MINING TECHNIQUES

Associations rules where connecting of event is there that is one connected with other example the events of purchasing bread and butter. Association rules are generated. Support, confidence is calculated. Threshold for minimum support as well as of minimum confidence is assumed. Only rules satisfying both the thresholds are considered to be true otherwise they are considered to be false. Accuracy is acknowledged as the confidence. Coverage is regarded as support. Accuracy implies probability of precedent to be true in case the antecedent is true. Existence of high accuracy implies that this is the rule which is extremely dependable. The number of records on which rule applies to is indicated by coverage. A rule that has high coverage is the rule that is used very regularly.

Path or sequence analysis pattern where on event is leaded by another event, such as child brother and diapers which are purchased that time.

Classification –In which new patterns are identified such as the coincidence made by purchasing plastic sheet and also cello tapes.

Clustering -Where various clusters are made of unknown puts to know them better such as geographical locations and perfumes.

Forecasting-Which is discovering of patterns form a king reasonable predictions

REVIEW OF LITERATURE

Agrawal and Agrawal (2017 Agrawal, R., & Agrawal, J. (2017): Explained information of the WEKA Tools Clustering Analysis Algorithm. Paper-defined clustering is a technique used in a number of fields, including image analysis, pattern recognition & statistical data analysis. Clustering is a information partition into comparable item sets. Each cluster includes different items that are similar to them and unlike other sets objects. Some clustering algorithms are used for cluster production. Method of data clustering to discover clusters in databases of spatial cancer. Computer Applications International Journal, 10(6), 9–14. DOI:[Google Scholar] 10.5120/1487-2004). WEKA instrument used to compare various algorithms for clustering. It has been used because it offers the user with a better interface than other information mining instruments.

Amira, Pareek, and Araar (2015 Amira, A., Vikas, P., & Abdelaziz, A. (2015): Offered ARM algorithms are frequently used to discover all database laws to satisfy certain minimum assistance and minimum confidence limitations. The amount of rules produced decreased the adaptation of ARM algorithm to mine only to specific subset of association rules where in past studies classification class attribute is allocated to right side. A traffic accident data set was collected

from Dubai Traffic Department, UAE in this research.

FP-growth: **Novitasari, Hermawan, Abdullah, Sembiring, and Herawan (2015 Novitasari, W., Hermawan, A., Abdullah, Z., Sembiring, R. W., & Herawan, T. (2015):** Two significant drawbacks in Apriori-like algorithms are the candidate set generation and tests submitted. A new data structure called frequent pattern tree (FPTree) was launched to cope with this issue. Subsequently, FP-Growth was created on the basis of this data structure & is presently benchmarked & rapid algorithm for Lee, Kim, Cai.

Han (2003 Lee, Y. K, Kim, W. Y, Cai, Y. D, & Han, J. (2003): FP-Growth's advantages are that it needs to scan the transaction database twice. First, to calculate a list of different items sorted by descending order, it scans the database and removes unusual items. It then scans to compress database into a FP Tree framework & recurrently mines FP-Tree to build its conditional FP-Tree.

Shrivastava and Panda (2014 Shrivastava, A. K., & Panda, R. N. (2014): Several algorithms have been created to mine association rules from enormous databases. The Apriori algorithm provided by authors is best prevalent algorithm to mine dataset association rules. There are several instruments available for executing the Apriori algorithm. WEKA is an open source machine-learning algorithm software instrument. A research defined by WEKA is the compilation or collection of information mining instruments using the association rules. Rules of association created by analyzing information for different samples and using normal assistance and reliability to define the most significant relationships. In data mining, they are divided into distinct classes and used to conduct the activities in the WEKA.

Tanna and Ghodasara (2014 Tanna, P., & Ghodasara, Y. (2014): Discussed the use of Apriori for repeated pattern mining through WEKA. Apriori algorithm outlined in the paper is very efficient in extracting repeated groups for the laws of Boolean association. This article concludes that Apriori is the simple algorithm that applied from the transaction database for the mining of repeated patterns. Paper provided the WEKA tools used to apply the Apriori algorithm for association rule. Authors have practiced the Apriori algorithm to use the WEKA GUI to obtain association laws that have minSupport= 50% and min trust= 50%. They attempted to introduce the Apriori algorithm for adequate study job and used W EKA to refer to the association rule mining method.

Slimani and Lazzez (2014 Slimani, T, & Lazzez, A. (2014): The additional property of this algorithm is that database is not utilized at all to count candidate item assistance set after first pass. Instead, an encoding is used for this purpose of the applicant item sets used in the past pass. Objectset mining, linear pattern mining, linear rule mining, and association rule mining are the most critical tasks of frequent pattern mining methods. Apriori algorithm is one of the originally suggested framework that addresses issues of association rule. The AprioriTid and AprioriHybrid algorithms were provided in synchronicity with Apriori. The AprioriTid algorithm is performed equivalently as well as Apriori for lower problem sizes, but the efficiency degraded by applying it to huge issues twice as slow.

Bansal and Bhambhu (2013 Bansal, D., & Bhambhu, L. (2013): It has been noted that the association rule transacts with frequent itemsets as a result of many association algorithms such as the Apriori algorithm used in broadly actual apps for vitality. In this document, writers involve the use of ARM to extract patterns that often occur within dataset & explain application of WEKA's Apriori algorithm method from dataset that collects demeaning offences against females in session court. This article uses WEKA tool to study 2 association rule algorithms Apriori algorithm & Predictive Apriori algorithm & match results of both algorithms.

OBJECTIVE

In the present study the following are the objectives:

1. To find out best rules of association for DM.

2. Improve the computational time for processing

PROPOSED METHODOLOGY

With the rapid development of e-commerce apps, vast amounts of information accumulate in months not years. DM also termed as KDD identifies anomalies, correlations, patterns & trends for predicting results. Apriori algorithm (AA) is classic data mining algorithm. It is utilized for frequent itemsets (FI) & association rules to be used for mining. It is designed to work on a database that contains many transactions, such as products carried to a shop by clients. It is very crucial for efficient market basket analysis and it enables clients to more easily purchase their products, which improves market revenues. It was also used for the identification of adverse drug reactions in the healthcare sector. It generates rules of association indicating what all drug combinations and patient features lead to ADRs.

Association rules (AR)

AR learning is important & well-explored system for influential relations between variables in huge databases. As formal definition of issue of AR

prsented via Rakesh Agrawal, President & Founder of Data Insights Laboratories.

Let $I=\{i1,i2,i3,\ldots,in\}$ be set of n attributes termed as items and $D=\{t1,t2,\ldots,tn\}$ be set of transactions. It is known as database. Every transaction, *ti* in *D* has unique transaction ID, & it made up of subset of itemsets in *I*.

A rule can be defined as an implication, $X \longrightarrow Y$ where *X* & *Y* are subsets of $I(X,Y \subseteq I)$, & they have no element in common, i.e., $X \cap Y$. *X* & *Y* are antecedent & consequent of rule, correspondingly.

Let's take simple example from supermarket sphere. The example that we are considering is quite minute & in practical conditions, datasets have millions or billions of transactions. The set of itemsets, *I* ={Onion, Burger, Potato, Milk, Beer} & database having of 6 transactions. Each transaction is tuple of 0's & 1's where 0 shows absence of an item & 1 presence.

Conviction

The conviction of rule can be distinct as:

$conv(X \longrightarrow Y)=1-supp(Y)I1-conf(X \longrightarrow Y)$

For the rule {onion, potato}=>{burger}

Undefined control sequence \implies

The conviction value of 1.32 means that rule {onion,potato}=>{burger} would be incorrect 32% more often if the association among X & Y was an accidental chance.

AA has property that any subset of frequent itemsets must be frequently utilized & this property helps in dropping finding space. Iterative approach is employed via algorithm that k-itemsets are utilized for exploring (k+1)-itemsets.

Thus, it is separated into 2 step functions that utilized to find frequent item sets: join & prune actions.

Ck which is superset of Lk that may have members which is frequent or not. But overall of FI are members of Ck. Apriori property is utilized here for dropping size of Ck. A database scanleds to purpose of Lk via determining count of every Ck candidate. Thus all candidates that have support count no less than support threshold are in Lk.

Table 1 Transactional Data

TID	ITEMS
T1	A,B,E
T2	B,D
T3	B,C
T4	A,B,D
T5	A,C
T6	B,C
T7	A,C
T8	A,B,C,E
T9	A,B,C

Let us consider a transaction database as shown in following table: Min_Sup=2, Min_conf=70%

CONCLUSION

Data mining is the cumbersome job in every research related activity. One of biggest challenges in DM is to reduce processing time with appropriate accuracy. Research with high accuracy and less processing time could be achieved through rule induction and rule mining. It leads handling large number of data with minimum number of rules. Apriori lgorithm could process large datasets in very much less time with minimum set of induction rules comparatively to traditional methods. In the present study some analytical tools i.e. Weka Tangera etc were used to show the graphical representation of analysis.

REFERENCES

1. Raval Kalyani M. (2012). "Data Mining Techniques." *International Journal of Advanced Research in Computer Science and Software Engineering*2.10: pp. 439-442.

2. Moharana, U. C., and S. P. Sarmah (2014). "Determination of optimal kit for spare parts using association rule mining." *International Journal of System Assurance Engineering and Management*Springer, Sweden: pp. 1-10.

3. Liu Yanxi (2010). "Study on application of apriori algorithm in data mining." *Computer Modeling and Simulation, 2010.ICCMS'10.Second International Conference on*.Vol. 3.IEEE,2010.

4. Yang, Jun, et. al. (2013). "An Improved Apriori Algorithm Based on Features."*Computational Intelligence and Security (CIS), 2013 9th International Conference on*. IEEE, 2013.

5. Al-Maolegi, Mohammed, and Bassam Arkok (2014). "An Improved Apriori Algorithm for Association Rules." *International Journal on Natural Language Computing (IJNLC),* 3.1.

6. AL-Zawaidah, Farah Hanna, YosefHasanJbara, and A. L. Marwan (2011). "An Improved Algorithm for Mining Association Rules in Large Databases." *World of Computer Science and Information Technology* 1.7: pp. 311-316.

7. Lee, Dong Gyu, et. al. (2013). "Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction." *Journal of medical systems* 37.2: pp. 1-10.

8. Verma, Srivastava, Chack, Diswar, and Gupta Verma, M, Srivastava, M, Chack, N, Diswar, A. K, & Gupta, N. (2012). A comparative study of various clustering algorithms in data mining. International Journal of Engineering Research and Applications (IJERA), 2(3), pp. 1379–1384.

9. Amira, Pareek, and Araar Amira, A., Vikas, P., & Abdelaziz, A. (2015). Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai. *International Journal of Soft Computing and Engineering (IJSCE)*, *5*(4), pp. 1–12.

10. Han Lee, Y. K, Kim, W. Y, Cai, Y. D, & Han, J. (2003). CoMine: Efficient Mining of Correlated Patterns. *In ICDM*, *3*, 581–584. November.

11. Shrivastava and Panda Shrivastava, A. K., & Panda, R. N. (2014). Implementation of Apriori algorithm using WEKA. *KIET International Journal of Intelligent Computing and Informatics*, *1*(1), p. 4.

12. Slimani and Lazzez Slimani, T, & Lazzez, A. (2014). Efficient analysis of pattern and association rule mining approaches. *Journal of Information Technology and Computer Science (IJITCS)*, *6*(3), pp. 70–81. DOI: 10.5815/ijitcs.2014.03.09

Corresponding Author

Manoj Semwal*

M.Tech Scholar, Department of Computer Science and Engineering, Doon Institute of Engineering & Technology, Dehradun, India

manojsemwal1@gmail.com