2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)
March 17–18, 2021, Amity University Dubai, UAE

2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) | 978-1-6654-2921-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIKE51210.2021.9410737

Comparative Analysis of Market Basket Analysis through Data Mining Techniques

Shish Kumar Dubey
Research Scholar
Department of Computer Science and Engineering
Jaipur National University,
India shishdubey@gmail.com

Sonu Mittal
Department of Computer Science and Engineering
Jaipur National University
Jaipur, India
sonumittal.research@gmail.com

Seema Chattani
Visiting Faculty
Manipal Academy of Higher Education, Dubai
United Arab Emirates
seemaaswani03@gmail.com

Vinod Kumar Shukla
Department of Engineering and Architecture
Amity University Dubai
United Arab Emirates
vinodkumarshukla@gmail.com

Abstract- Market basket analysis is a technique for evaluating buyer's preferences in order to find the connection between various items in the cart. The exploration of these relationships help the vendor to propound the sales strategy by considering the frequent purchased of items and with this kind of approach data-mining techniques best fits in analyzing and implementing the logic. The points of comparisons, which include the concept of buying patterns from the consumer end and the production pattern from the company, end which alternatively helps in procuring or buying the product. Evaluating the activities of business consumers is very important and this can be achieved by various data mining techniques available. This paper provides a comparative study of two widely used data mining techniques in understanding the frequent activities of buyer i.e. Association Rule Mining (ARM) and Collaborative filtering (CF) technique used in product recommendation.

Keyword(s): *Market Basket Analysis (MBA), Data Mining, Association Rule Mining (ARM), Product Recommendation system.*

I. INTRODUCTION

Nowadays, in various fields such as retail industries, real estate, banking sectors etc., massive quantities of data are stored in the databases. Using that whole information stored in database is not necessarily useful for the user. Therefore, it is very important to derive from the broad chunk of data the valuable piece of data. This method of extracting valuable data from data is referred to as data mining. Data mining is the process of analyzing large data volumes to uncover the useful information from the data that enables businesses to solve challenges, reduce risks and seize new opportunities.

Data mining is basically a practical based topic and not a theoretical topic as in data mining, we as a data miner is mainly interested in those techniques that are used for finding patterns in data. These patterns may provide some idea or insights which helps in better decision making capabilities. One such example is a scenario where a customer who has switched loyalties, as in situations where certain contact lenses are prescribed. The data generated as output is in the form of predictions of whether a customer will switch for a prediction or not or the kind of lens he or she may opt for or may be prescribed under certain circumstances.

There are various learning approaches that look for descriptions based on structures of what is learnt, for example the descriptions that may become complex when expressed as rules and these are the ones that may be described earlier or may even be decision trees. Hence such descriptions serves to explain what it means to people as the can be easily understood by them and they explain the basis for new predictions. Based on several examples relating to machine learning and data mining, the knowledge acquired explicitly are the structural descriptions and are at least important in performing well on new examples. People often use data mining to gain knowledge and not to just acquire predictions [1].

Data mining helps in business to take correct business decisions and to understand the consumer behavior [2]. The process of discovering the associations between buyer's buying patterns is known as Market Basket Analysis and the methodology used to discover the associations between different products that the buyer has purchased is known as Association Rule Mining [3].

There is a contrast between understanding the relationship between various items dependent on the acquisition of the purchaser and prescribing the items to the customer dependent on the past buys. This paper presents the near investigation of the information digging procedures utilized for Market Basket Analysis to discover the Association and Product recommendation. Association rule mining (ARM) is utilized for identification of association between a large set of data items. Because of the expanded volume of information being put away in data sets, few enterprises are utilizing mining affiliation rules from their data sets. This will assist the ventures with executing cross displaying and taking different business choices [4].

Product Recommendation frameworks are frameworks that look to give exact suggestions to a client. A recommendation framework can be constructed utilizing a wide range of methods and ideas inside Data Mining. Collaborative Filtering is the strategy utilized in information mining to construct Product recommendation. The collaborative filtering algorithm is divided into two groups: User-based and Item based.

978-1-6654-2921-4/21/$31.00 ©2021 IEEE

Challenges in Data Mining

Challenges in data mining can be broadly explained and categorized under variety (or heterogeneity), volume (or scale) and velocity (or speed) which are the top three challenges. Data safety and security is also a very important aspect of this, while dealing with sensitive data. Data security algorithm like water marketing [5] and steganalysis[6] or similar to this can be adopted for more data safety and security.

Heterogeneity and Variety: Data mining techniques are often used to understand or learn the unknown patterns or relationship of significance from small datasets that are consistent and organized. Data in various sources poses different types and representation forms. Hence, these data might be linked or connected, somewhat relating to one another and represented in a manner that may not be consistent.

Thus, mining a dataset is a big problem and we cannot even imagine the level of difficulty in doing it even before we reach such a similar situation. Heterogeneity in data mining simply means that it is not an option but a requirement to deal with structured, semi-structured and unstructured data instantaneously. In modern day databases, structured data well fits while semi-structured data may fit partially. On the other hand, unstructured data will never fit. Those data that are unstructured or may be semi-structured are usually kept in files. Hence, it is so in data-intensive and those sectors that is scientific computation zones. Thus, earlier unmanageable hidden patterns or knowledge inhabited at the nodes within dissimilar big data are now revealed with the help of the heterogeneity which is the feature of big data.

***Scalability*:** High scalability and mining tools that are the part of data management are the essential requirements of huge volumes of data or scale of big data. Great scale of big data endures further potential perceptions that we have no chance to learn from conventional data. Since we are confident about the methods that, if misused well, may result in extraordinary scalability as necessary by upcoming data and its related mining systems to cope and mine data as cloud computing has already proved better elasticity by using the power of parallel computing designs which shows the need for scalability in interacting with capacity challenges of big data. Advanced graphical user interface or language centred facilities also swift active system-user communication.

Speed/velocity: Speed really matters for big data as ability of fast accessing and mining data isn't a simple job as it is a requirement for data streams often termed as a public format of big data, that needs processing/mining task to be done within a specific time frame or else mining/processing task turn out to be less valuable or useless. Certain examples are the need of real-time demand from tremor forecast, stock market estimation and agent-based self-directed exchange (buying/selling) systems. Hence speed is applicable to scalability as successful or moderately solving anyone benefits the other one [7].

Market demand for Data Mining

Amidst the crisis caused by the Corona virus, the worldwide marketplace for data mining tools was projected at around US$634.7 Million in the year 2020. By 2027, it is estimated to influence a market of US$1.3 Billion which is growing at a CAGR of 11.1% in the years 2020-2027. One of the sectors analysed in the report, that is Services, is likely to record 10.9% CAGR and reach US%871.5 Million by the completion of the analysis period. Data Mining Tools marketplace in the US is valued at US$187.2 Million in the year 2020. Market scope of US$233.1 Million by the year of 2027 is expected in China at a trailing CAGR of 10.7% in the years 2020-2027. Japan and Canada are some of the noteworthy markets which may grow at 9.7% and 9.4% respectively over the years 2020-2027 [8].

Market Basket Analysis

Market Basket Analysis as well known by the term association rule learning or affinity analysis which is basically a data mining procedure that is being used widely in the field of education, nuclear science, bioinformatics and marketing. In this MBA method, the purchase behavior of the buyer is analyzed and this information is handed over to the retailer so that it can help retailers in better decision-making [9]. In a world of competitive markets, to sustain a good position in the market is always a challenge for organizations as it always depends on the organizations capabilities of decision making and understanding the behavior of customers [10]. Hence examining customer-buying pattern is crucial for an organization [11].

Data mining Techniques

There are many data mining techniques and following are some of the techniques used:

Association: Association is one of the best methods used. Here, a pattern is revealed based on the connection of an item and other items in a transaction. For instance, in MBA, it is used to detect the products that consumers often purchase together [12].

Classification: New objects presented in the market are studied and its features are examined and then assigned to a predefined class. For example, to classify credit applicants as low, medium or high risk.

Prediction: In this feature, missing or unknown attribute values are predicted. For example, forecasting the next week sale with the help of currently available data.

Clustering: Here, data is organized into sub-groups or clusters by data mining such that the points or groups are similar to each other and different as possible from points in another group. This is also termed as an unsupervised classification.

Outlier Analysis: Explaining exceptions and identifying them are done by data mining. For example in MBA, some transaction that usually happens is termed as an outlier [9].

II. RESEARCH OBJECTIVE

This study aims to examine how Market Basket Analysis (MBA) is useful in finding the association in consumer's buying behavior and giving the Product recommendation to consumers based on consumer's past purchases. There are different data mining algorithms available to find out the frequent trends in consumer's buying pattern and also to give the product recommendation on the basis of past purchases made by the consumer. In this paper we have focused on mainly two algorithms: Apriori Algorithm to find out the frequent buying pattern of the consumer and Collaborative Filtering algorithm to give the Product Recommendations. We have conducted the comparative study of these two algorithms to find out the differences and similarities between the product association and product recommendation.

III. LITERATURE SURVEY

We have focused on presenting various areas in this section where data mining algorithms are used. The current algorithms developed by researchers in the sense of association rule mining in the MBA and collaborative filtering for product recommendations are outlined in this section.

Market Basket Analysis: Analysis of the Market basket is a modeling methodology often referred to as affinity analysis, it helps in recognizing those things which are probably going to be bought together by noticing regular itemset. The market basket problem expects we have some huge number of items. Client purchases the subset of things according to his/her necessities and advertiser gets the data that which things client has taken together. The advertiser utilizes this data to put the things on various position [13].

In order to understand consumer behavior, the data collection of past transactional knowledge is analyzed. Only the information about the transactions over a certain period such as a day or a week etc. was available on personal computers back in time. Change in product scanning and bar code technology, however, enables data on transactions to be stored on a regular basis. As a result, vast volumes of data are obtained and processed. Due to restricted database functionality, such data sets are typically stored in higher-level storage. For example, to improve the functionality of the database and to process the data, if anyone buys a bottle of milk, they also prefer to buy packet of oats at the same time.

So, Milk => Oats.

MBA is used in determining the location of goods within a store. If a client purchases a bottle of milk, he is much more likely to purchase a packet of oats. It would make consumers tempted to purchase one item with another to hold the milk and oats next to each other in a supermarket.

Association Rule Mining: Association Rule Mining (ARM) is a tool widely used to find a link in a set of different objects, to find frequent trends in a transactional database, relational databases or any other repository of knowledge. In retailing, clustering and classification, the two applications of Association Rule Mining are widely used in Market Basket Analysis (MBA). A standardized protocol is accompanied by the ARM method to discover standard objects in a client's cart. Three traditional ways of calculating association are available [14].

Measure 1: Support. This states how common an itemset is, as calculated by the proportion of transactions in which an itemset appears.

Measure 2: Confidence. This states how likely item B is to be bought when item A is purchased. It is represented as {A -> B}. This is determined by the proportion of Item A transactions that are also calculated by Item B.

Measure 3: Lift. This states how likely item B is to be purchased when item A is purchased, thereby tracking how popular item B is.

Apriori algorithm for finding frequent itemset: The data set of the Apriori algorithm is evaluated to decide which combinations of items frequently take place together. The mostly used algorithm to find the association rules is the Apriori algorithm, being used in a basket-item relationship.

Basic idea behind this algorithm is [15]:

- Only a large item set can be an item set if all its subsets are large item sets.
- It is possible to accept collections of products that have minimal support.
- From frequent item sets, association rules can be created.

Collaborative Filtering Algorithm for Product Recommendation: Collaborative filtering is a way of predicting (filtering) the tastes of a consumer automatically by collecting preferences or taste data from many users (collaborating). It suggests items by identifying other users with a similar preference; it uses their opinion to suggest items to the active user [16]. CF works on finding the similar categories of products using the historical dataset and based on that it gives the product recommendation. With some added scalability Collaborative Filtering will also work for large datasets with no repetitions as it is used for finding the similarities instead of finding the association between products purchased. Collaborative filtering classes comprise two classes which are User-based and Item-based:

User-based: It is the calculation of the similarity of target users to other users.

Item-based: It is the calculation of the similarity between objects to determine the association between items that target user's rate or communicate with and the other items.

The CF recommendation process is split into the following four stages.

- ***User-item rating matrix construction:*** After browsing the buying activity, user rating data is collected and then cleaned, transformed. Finally, the data is entered to obtain a user-item rating matrix.
- ***Similarity computation:*** Cosine similarity (COS), ACOS (Adjusted Cosine Similarity) and so on are among the common methods for CF recommendation techniques to calculate similarity. After calculating the similarity between the users, sorting is performed to analyze each user's similarity with other users.
- ***Neighborhood selection***: The optimal k-nearest neighbors are selected to enforce the expected set, or set the similarity threshold, according to the result of the similarity ranking between users, and users who surpass the threshold are selected as the target user's neighbors.
- ***Rating prediction and item recommendation:*** After obtaining the target users nearest neighbor set, we use the similarity as a weight to get the prediction of the unrated item by the target user and shape a Top-N list to suggest to the user.

IV. RELATED WORK

Apriori Algorithm: Consider the following information on which apriori calculation can be applied to discover the affiliation rule for frequent item set [17].

Table -1: Sample Transactional Dataset

Transaction ID	Items
1	Milk, Banana, Oats
2	Milk, Oats
3	Milk, Biscuits
4	Cheese, Bread, Banana

Step 1: Check data for the number of individual items in the dataset.

Item set	Support
{Milk}	3
{Oats}	2
{Banana}	2
{Biscuits}	1
{Cheese}	1
{Bread}	1

Step 2: Compare the count of item support with minimal support (50 percent)

Item set	Support
{Milk}	3
{Oats}	2
{Banana}	2

Step 3: Generate the set of Items and its support from the above table.

Item set	Support
{Milk, Banana}	1
{Milk, Oats}	2
{Oats, Banana}	1

Step 4: Discard the itemset with lowest support.

Item set	Support
{Milk, Oats}	2

Step 5: Hence the frequent itemset is: {Milk, Oats}.
Therefore, the association rule can be set as

Milk->Oats or Milk->Oats

It can then be determined that these are the frequent items purchased in a combination by the clients, and the marketing plan can be decided for the store accordingly [18].

Collaborative Filtering Algorithm: Instead of similar users, the Collaborative Filtering algorithm seeks similar objects, i.e., items that clients prefer to purchase together. The steps for measuring the similarities between all products are illustrated in the following steps [19]:

- *For each Product in Transactional dataset, Product1*
- *For each client C who purchased Product1*
- *For each product Product2 purchased by client C*
- *Record that a client purchased Product1 and Product2*
- *For each Product Product2*
- *Compute the similarity between Product1 and Product2*

IV. Conclusion

Association rule mining is a technique for computing a principle of similarity between objects - when things go together exceedingly often in a basket, session, and purchase. The underlying math overlaps a lot and is down to the co-occurrence counting. Association rule mining aims to search for groups of items, but not just pairs, and focuses on doing so efficiently.

The basis of one classic form of recommendation is often similar item pairs, where the best recommendations are the most similar - often occur with - other items in your history. Thus, the mining association rule is almost a subset of one type of collaborative filtering. The primary difference between ARM and CF is, in association rule mining it is usually the "session" (which products appear together in the same session) and computed across all users. While in user- or item-based CF the unit is the user (which products have been consumed by the same user), computed across all user sessions (i.e., regardless of whether they were consumed together or not).

References

[1]. H Witten, I., 2016, "Data mining" [online] http://etonline-digitallibrary.com/bitstream/123456789/2358/1/1306.pdf, Accessed on FEB 2021

[2]. M. J. A. B. & G. S. Linoff, Data Mining Techniques, 2nd ed. Wiley, 2004.

[3]. D. Hand, H. Mannila, and P. Smyth, Principles of Data Mining Cambridge, vol. 2001. 2001.

[4]. M. Dhanabhakyam and M. Punithavalli, "A survey on Data mining algorithm for market basket analysis," Glob. J. Comput. …, vol. 11, no. 11, pp. 1–7, 2011.

[5]. Anil, V. K. Shukla and V. P. Mishra, "Enhancing Data Security Using Digital Watermarking," 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 2020, pp. 364-369, doi: 10.1109/ICIEM48762.2020.9160090.

[6]. Shankar, D.D., & Shukla, V.K. (2018). Result Analysis of Cross-Validation on low embedding Feature-based Blind Steganalysis of 25 percent on JPEG images using SVM. 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 1-5.

[7]. Che, D., Safran, M. and Peng, Z., 2013, April. From big data to big data mining: challenges, issues, and opportunities. In International conference on database systems for advanced applications (pp. 1-15). Springer, Berlin, Heidelberg.

[8]. "Global Data Mining Tools Industry"[online] https://www.reportlinker.com/p05798317/Global-Data-Mining-Tools-Industry.html?utm_source=GNW

[9]. Kaur, M. and Kang, S., 2016. Market Basket Analysis: Identify the changing trends of market data using association rule mining. Procedia computer science, 85, pp.78-85.

[10]. Raorane, A.A., Kulkarni, R.V. and Jitkar, B.D., 2012. Association rule–extracting knowledge using market basket analysis. Research Journal of Recent Sciences ISSN, 2277, p.2502.

[11]. Gupta, S. and Mamtora, R., 2014. A survey on association rule mining in market basket analysis. International Journal of Information and Computation Technology, 4(4), pp.409-414.

[12]. Bharati, M. and Ramageri, M., 2010. "Data mining techniques and applications"[online] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.898.6657&rep=rep1&type=pdf

[13]. S. Gurudath and M. B. Analysis, "Market Basket Analysis & Recommendation System Using Association Rules Market Basket Analysis & Recommendation System Using Association Rules Shruthi Gurudath Submitted in partial fulfillment for the degree of Master of Science in Big data management and ," no. August, 2020.

[14]. M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," Procedia Comput. Sci., vol. 85, no. Cms, pp. 78–85, 2016, doi: 10.1016/j.procs.2016.05.180.

[15]. S. Gupta and R. Mamtora, "A Survey on Association Rule Mining in Market Basket Analysis," Int. J. Inf. Comput. Technol., vol. 4, no. 4, pp. 409–414, 2014.

[16]. S. Gong, "A colloborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering," J. Softw., vol. 5, 2010.

[17]. V. Pathan, Palande, Shende, Patil, "A study on Market Basket Analysis and Association Mining," Proceeding Natl. Conf. Mach. Learn., 2019.

[18]. R. Agrawal, "Fast Algorithms For Mining Association Rules In Datamining," Int. J. Sci. Technol. Res., vol. 2, no. 12, pp. 13–24, 2013.

[19]. B. N. Miller, J. A. Konstan, and J. Riedl, "PocketLens: Toward a personal recommender system," ACM Trans. Inf. Syst., vol. 22, no. 3, pp. 437–476, 2004, doi: 10.1145/1010614.1010618.