North Asian International Research Journal of Sciences, Engineering & I.T.

ISSN: 2454-7514 Vol. 7, Issue-2 February -2021

IRJIF. I.F. : 3.821

Index Copernicus Value: 52.88 Fulfill MCI Criteria Citation Index

A Peer Reviewed Refereed Journal

DATA MINING IN THE RETAIL INDUSTRY

MOHMMAD ILYAS MALIK

ABSTRACT

In the fast pace and technologically advanced age that we live in, the volume of data now available to businesses is phenomenal. With the use of point of sale technology, both online and in stores, combined with the widespread use of customer loyalty cards, it has never been easier for retailers to collect vast amounts of data on customers and their purchases. The important task they face is deriving value from this data. Data mining involves applying algorithms to large datasets in order to find interesting or useful patterns. This discovery of knowledge can then be utilized by retailers to plan product placements in order to maximize cross-selling of products, target specific customers based on previous purchases or to suggest favorite or similar items to customers. This area of knowledge discovery is known as Market Basket Analysis and involves analyzing a customer's transaction, or basket, and attempting to identify trends or patterns of items regularly purchased together. A simple concept, however, the implementation can become very complex when attempting to build a generalized model to suit the needs of retailers with differing needs and constraints. This paper will look at several methods in the area of Market Basket Analysis.

***KEY-WORDS**: Data, Mining, Retail, Industry*

INTRODUCTION

In the fast pace and technologically advanced age that we live in, the volume of data now available to businesses is phenomenal. With massive amounts of data being gathered every second, from transactional data to personal data, to social media and telecommunications, the ability of companies to effectively and cleverly identify useful patterns in order to gain valuable insights has become essential. Data mining, the practice of examining large databases to extract useful patterns and knowledge, has become crucial for understanding customer habits and behaviors as well as allowing businesses to gain an advantage over their competitors. According to the authors of Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, the definition of data mining is "a business process for exploring large amounts of data to discover meaningful patterns and rules" (Linoff and Berry, 2011, p.2). Although referring to a specific industry which is primarily focused on customer data, this is in line with the view that data mining, in simple terms, is a process of examining large amounts of data in order to identify useful patterns and similarities. Knowledge discovery in databases, or KDD, incorporates these data mining techniques, utilised across many industries including manufacturing, retail, telecommunications, pharmaceuticals

and fraud detection to name a few. The process includes the preparation, cleaning and tidying of data in order to then apply appropriate algorithms to the data. The results are then interpreted and evaluated to determine if any useful patterns have been discovered.

In this paper, we will briefly discuss data mining and knowledge discovery and move on to investigate data mining techniques, specifically in a retail space where companies have massive amounts of customer data at their fingertips. We will explore some examples of how clever algorithms are applied to identify consumer trends and in turn, target customers with specific products through an area known as Market Basket Analysis.

DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Today, the need for advanced computational and statistical theories and techniques is quite evident given the volume of data being collected at phenomenal rates across every industry. How to make use of, and more importantly, derive value from these massive amounts of data is a rapidly growing field called knowledge discovery in databases (KDD). As the name would suggest, seeking out useful patterns, information and knowledge from large databases is the goal, but the trick is to do so in a clever and efficient manner. Fayyed et al., (1996) described KDD as a process that can identify understandable patterns in databases that have potential for being useful. They map out the process in Figure 1 below.

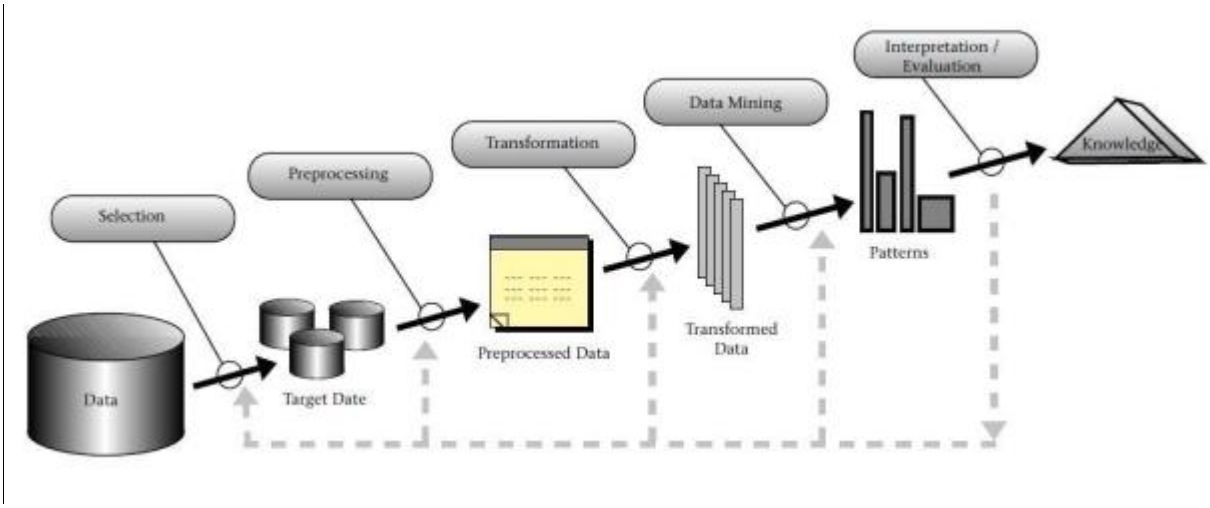

Figure 1. The steps of the KDD process. Fayyed et al. (1996)

From this diagram, we can see that the process involves first selecting data appropriately and preparing or tidying data to remove inconsistencies in the pre-processing stage. Data can then be transformed to reduce dimensions and to have an appropriate structure for fitting a model. In the data mining step, an appropriate machine learning algorithm such as a clustering or classification technique is applied to the

data. Following this, the results are analyzed and interpreted, and, in this stage, we hope that knowledge is discovered from the data through useful patterns or trends.

This data mining stage of this process is arguably the most important stage and the one on which we will focus. Identifying the correct algorithm, applying appropriate criteria and even coping with the size and diversity of a database are just some of the factors that can impact the results of the KDD process.

MARKET BASKET ANALYSIS

With the massive amounts of transactional data now available through point of sale systems, a colossal amount of product and customer information is now at the fingertips of retailers and it is crucial for them to derive value from this. Since the implementation of point of sale technology, data mining and knowledge discovery techniques have come a long way to identifying patterns in customer purchases. Further to this, with the introduction of customer rewards cards, it has been possible to link customers with their purchases, providing an even greater potential for insights into the habits of certain demographics of shoppers, geographical trends and the ability to predict customer behaviour (Linoff and Berry, 2011). This area of knowledge discovery is called Market Basket Analysis and will be the main data mining technique I will look at in this paper.

ASSOCIATION RULE MINING

Market Basket Analysis attempts to identify associations between products. It uses datasets of customer transactions to look for frequent combinations of the same products to identify relationships between items in order to optimise cross-selling of products. This is invaluable to retailers for several reasons; to inform strategic layouts in stores, to suggest favourites or similar products, to target customers with marketing campaigns or even the design and layout of store catalogues. The method behind market basket analysis is the implementation of association rule mining.

Association rule mining is a technique used to find rules in large amounts of data based on relations between unrelated items. The two metrics involved in these rules are support and confidence which are used to obtain the interestingness of association rules. In simple terms, support is the percentage of transactions that the rule can be applied to and confidence is a ratio of the number of transactions in which a rule is satisfied against the number of transactions in which it is applicable. There is a minimum value assigned for each, support and confidence, and these thresholds when exceeded define when a set of items is deemed to be frequent (Dongre, et al., 2014).

Mining these rules happens in two steps. Firstly, a clever sorting algorithm is applied to identify frequently occurring sets of items in the database, known as item sets, and following this, the rules are extracted based on these frequent item sets. The threshold values vary depending on the type of items involved or the intention of the retailer. Indeed, choosing these thresholds is a difficult process in itself, as a restrictive threshold could produce too few rules and a lenient one can produce too many. As the use of this technique can vary greatly across datasets depending on types of products, retailers and indeed the intention of the model, there is a large range of models and algorithms out there intending to optimise specific situations. Below we will look at a few of the algorithms and examples of their implementation. We will look at the widely used Apriori model, the Generalized PROSET model and the application of Multi Attribute Value Theory.

THE APRIORI ALGORITHM

The Apriori algorithm was first proposed in 1994 by Rakesh Agrawal and Ramakrishnan Srikant in their paper 'Fast Algorithms for Mining Association Rules' (Agrawal and Srikant, 1994) to deal with the task of identifying association rules between items in large datasets of sales transactions. This paper had a significant impact on the research to follow as noted by "Since the introduction of the Apriori algorithm, it has been considered the most useful and fast algorithm for finding frequent item sets" (Venkatachari and Chandrasekaran, 2016, p.58). They note that improvements have since been made to increase the effectiveness and efficiency of the original algorithm, but most algorithms developed since have been based on this original concept. In their case study on Mumbai Retail Store, they compare the implementation of the Apriori algorithm with another called the FP Growth algorithm. Here they found, on their sample of 300 observations, that the Apriori algorithm performed much faster than that of the FP Growth algorithm (Venkatachari and Chandrasekaran, 2016).

So how does the Apriori Algorithm work? The word Apriori comes from the Latin 'a priori' meaning 'from the earlier' or 'deductive', and so this gives us an idea of how the algorithm will work. The method is described well in the paper, 'The role of Apriori algorithm for finding the association rules in Data mining' by Dongre et al., (2014). As mentioned above, a minimum threshold is set for a support variable in order to gauge interestingness. The algorithm searches the database for items that satisfy this minimum support value. This step then repeats each iteration adding items to item sets one at a time, provided they satisfy the minimum support value, and this is how the frequent sets are created. An important aspect that lends itself to the speed of the algorithm is that the order of the items in the item sets is not important. However, when the association rules are created the order of these is important.

Rules are of the structure A -> B in that if a customer purchases A, that implies the purchase of B also. So now, the confidence of the association rule is calculated by the support value of the whole item set divided by the support value of item A. Any rule that has a confidence value of less than the assigned threshold is removed, and so the set of rules is built. The same paper concludes that varying minimum values of confidence generate different sets of rules but a high level of confidence filtered rules more accurately (Dongre, et al., 2014).

In the paper 'An Improved Apriori Algorithm for Mining Association Rules', Yuan (2017) highlights some deficiencies and suggested improvements to the algorithm. Firstly, the algorithm requires repeated scanning of the database which consumes time and secondly, many item sets are created requiring large amounts of memory. The proposed alternative, called T_Apriori was found to improve the speed of the rule mining significantly.

With the growing volumes and complexity of data being collected, the issue of storage and its cost is a concern when using complex algorithms that may not be using memory efficiently. Verma, N et al., (2017) attempted to address this need for resources by developing a MapReduce Apriori algorithm to better deal with big data. A similar approach is taken by Pandagale, et al., (2016), in which they use Hadoop-HBase for mining association rules using the MapReduce Apriori algorithm.

Over two decades after its inception, the continued research into variations of the original Apriori Algorithm solidify it as an extremely dependable algorithm for association rule mining, the problems now are in dealing with the 3 V's of big data, Volume, Velocity and Variety.

THE GENERALIZED PROFSET MODEL

The Generalized PROFSET Model, named from Profitability per Set, is a model put forward by Brijs, et al., (2001). The rationale behind this model is based on calculating the profitability of each frequent item set in order to determine the cross-selling potential between items. This model was one of the first attempts at addressing the issue that when using association rule mining there is no consideration for the business value of the associations found. An excellent example used to illustrate this is the purchase of an expensive bottle of wine combined with oysters having the same inherent association as that of milk purchased with cereal. Clearly one of these is more valuable to the retailer than the other and this should be considered when striving for an optimal model.

As mentioned above, measures such as support and confidence are used to determine the interestingness of association rules. This PROFSET model goes a step further and suggests that a rule is only considered interesting in the situation where it can be used in the decision-making processes of the business. "The key idea of the model is that products should not be selected based on their individual profitability, but rather on the total profitability that they generate, including profits from cross-selling" (Brijs, et al., 2014, p. 3). Following this, in a basket where there are overlapping frequent sets, the model must be able to appropriately allocate the profit margin. This highlights the importance of knowing what the authors call the purchase intention of the customer, in order to allocate profit margin to the most profitable frequent set in the basket. As the analysis is retrospective, it is impossible to simply ask the customer their purchase intention, so the support and confidence measures come in to play in quantifying the probability of the occurrences of the frequent sets.

In the paper 'A new approach of inventory classification based on loss profit', Xiao, et al., (2011) highlight some weaknesses of the PROFSET model for their purposes. Mainly, as the strength of relations between items is not considered, it is not possible to assign items a relative ranking which is an important aspect when it comes to the classification of items. Where a basket contains more than one frequent set, it is not simple to determine which frequent set represents the purchase intention of the customer. This conveys how complex the theory around identifying frequent sets can be given that the purpose of the outcome varies depending on the retailer's needs.

CONCLUSION

The area of research around association rule mining is driving improvements in Market Basket Analysis. Improvements in efficiency and effectiveness of existing algorithms, as well as the development of new techniques, are powering the retail industry forward as a formidable area for data mining and knowledge discovery.

Data mining has been described as a "Competitive Weapon" in retail industries (Hormozi and Giles, 2004). This is a fitting description given the opportunity for retailers to gain an advantage over competitors by investing in KDD. It would be naive of retailers not to utilise the abundance of customer and transactional data available.

This paper has looked at several sources highlighting methods through which data mining and KDD are used in a retail environment. Focusing on Market Basket Analysis and association rule mining, it's clear that a broad range of techniques can be applicable depending on the environment and items involved for individual retailers. KDD in retail can become very complex when considering the classification of items, assumptions around customer intentions and the prediction of customer behaviours in very large datasets. Through the basic Apriori algorithm, we

saw the method behind one of the most dependable algorithms and the basis for many other association rule mining algorithms to follow. We also saw how research is being carried out to improve on this existing algorithm and adapt it to suit the growing volume and variety of data that is available today.

Moving on to the generalized PROFSET model, we saw how the aspect of profitability could be incorporated into a model for cross-selling items, bringing another level of complexity with this additional dimension to impact on retailer's ability to make knowledge-driven decisions. The ability for a model to be able to incorporate profitability into its outcome is an invaluable source of knowledge with huge potential for real impact on knowledge-driven decisions.

As I see it, the toughest challenge for the area of knowledge discovery in databases will be the three V's of big data, Volume, Velocity and Variety. The ability for algorithms to cope with the rapidly growing size of databases and the complexity of data will be crucial in order to be able to identify useful patterns from large databases and derive value from them.

REFERENCES

1. Agrawal R, Srikant R (1994) 'Fast algorithms for mining association rules', 20th VLDB conference, pp 4S7-499
2. Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K. and Wets, G., 2001. A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model. arXiv preprint cs/0112013.
3. Dongre, J, Prajapati, G.L, and Tokekar, S.V, (2014) 'The role of Apriori algorithm for finding the association rules in Data mining'. 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), p. 657.
4. Fayyad, U., Piatsky-Shapiro, G. and Smyth, P. (1996), 'From Data Mining to Knowledge Discovery in Databases', AI Magazine, 17(3), pp. 37-54. Available at: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131 [Accessed 14 June 2019]
5. Hormozi, A. M. and Giles, S. (2004) 'Data Mining: A Competitive Weapon for Banking and Retail Industries', Information Systems Management, 21(2), pp. 62–71.
6. Linoff, G.S. and Berry, M.J.A (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 3rd ed. Indiana: Wiley Publishing Inc. p.2
7. Pandagale, A.A and Surve, A.R, (2016), 'Hadoop-HBase for finding association rules using Apriori MapReduce algorithm', 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), p. 795-798.
8. Venkatachari, K. and Chandrasekaran, I.D., (2016), Market Basket Analysis Using FP Growth and Apriori Algorithm: A Case Study Of Mumbai Retail Store. BVIMSR's Journal of Management Research, 8(1), pp. 56-63.
9. Verma, N. and Singh, J. (2017) 'An intelligent approach to Big Data analytics for sustainable retail environment using Apriori-MapReduce framework', Industrial Management & Data Systems, 117(7), p. 1503-1521.
10. Yuan, Xiuli. (2017). An Improved Apriori Algorithm for Mining Association Rules. Available: https://doi.org/10.1063/1.4977361. Last accessed 01/06/2019