A new approach for association rules mining using computational and artificial intelligence

Yoseph, Fahed; Heikkilä, Markku

Published in:
Journal of Intelligent and Fuzzy Systems

DOI:
10.3233/JIFS-200707

Published: 01/01/2020

Document Version
(Peer reviewed version when applicable)

Document License
Publisher rights policy

A New Approach for Association Rules Mining Using Computational and Artificial Intelligence

Fahed Yoseph[a],*, Markku Heikkilä[b]
[a] *Faculty of Social Sciences, Business and Economics, Åbo Akademi University, Turku, Finland, fyoseph@abo.fi*
[b] *Faculty of Social Sciences, Business and Economics, Åbo Akademi University, Turku, Finland, maheikki@abo.fi*

Abstract. Market Intelligence is knowledge extracted from numerous data sources, both internal and external, to provide a holistic view of the market and to support decision-making. Association Rules Mining provide powerful data mining techniques for identifying associations and co-occurrences in large databases. Market Basket Analysis (MBA) uses ARM to gain insights from heterogenous consumer shopping patterns and examines the effects of marketing initiatives. As Artificial Intelligence (AI) more and more finds its way to marketing, it entails fundamental changes in the skills-set required by marketers. For MBA, AI provides important ways to improve both the outcomes of the market basket analysis and the performance of the analysis process. In this study we demonstrate the effects of AI on MBA by our proposed new MBA model where results of computational intelligence are used in data preprocessing, in market segmentation and in finding market trends. We show with point-of-sales (POS) data of a small, local retailer that our "Åbo" model increases mining performance and extract important marketing insights to assess both demand dynamics and product popularity trends. Additionally, the results show how, as related to the 80/20 percent rule, 78% of revenue is derived 16% of the product assortment.

Keywords: Association Rules Mining, Artificial intelligence, Market Intelligence, Small and medium-sized retailer

1. Introduction

Small and medium-sized enterprises (SME) form the backbone of the economy employing more than 50% of the private workforce [1]. In many nations the SME are the engine of the social and economic development [2]. Small and medium-sized retailers (SMR) used to be the preferred shopping distribution channels of the middle and lower classes [3]. However, many SMR have been driven out of business, due to financial difficulties caused by shifts in economies, demographics, and technologies that all seem to be against traditional retail shopping models and paradigms [4], [5]. Consumers' shopping habits and rituals are changing fast, and the traditional shopping in retail stores is now on the fast decline [6]. This is due to a larger, cultural change that has dramatically altered consumers' expectations, forcing them to increasingly reevaluate their experience over vendors, and led to preferences of sustainability over low prices, convenience over wide choice, and personalization over standardization. But this time is also a unique opportunity for the SMR to reinvent their competitive strategies with new methods of gaining market intelligence, and to better understand their consumers and competitors in their business planning [7], [8]. Modern consumers are well informed, making the timing of product lifecycles and their profitability even greater a concern for the retailers [9].

The aim of sifting through often very large transactional databases is to recognize and analyze complex patterns to identify product trends and precisely reach consumers' specific shopping demands. This was once an insurmountable process due to human limitations, but now is an every-day reality thanks to modern data processing power [10].

*Corresponding author. E-mail: fyoseph@abo.fi.

Retailers are today moving to new kinds of modern marketing methods to reach consumers, who increasingly are online, and to generate more revenue for their brand rather than spending marketing efforts and budget on conventional ways of marketing, such as mass marketing [8]. Therefore, it is imperative to implement marketing intelligence where brands and retailers acclimate to the relevant and, at times, drastic shifts or business environments where they risk losing competitive positions [12], [13]. Many businesses have adopted practical, heuristic rules to guide their marketing strategies, such as the 80/20 percent rule of Pareto. Known also as the rule of the Vital Few it is supposed to boost the efficiency of the organizations' performance [14]. According to Marshall [15], understanding the 80/20 principle is essential to learning how the business prioritizes marketing tasks and many examples show how about 20% of products and consumers usually generate 80 % of the organization's profits [16]. Thus, the question that arises and is interesting for the marketers is: With marketing intelligence, how much of the 80 percent can the SMR afford to eliminate without taking the risk of losing stature in the industry?

In section 2 we present a review of literature relevant to our topic. The overall process that we test in our analysis is presented in section 3. In section 4 we present the results of the comparative analysis.

2. Market Intelligence for the SMR

Market intelligence (MI) is systematic way of gathering and utilizing market information. As such, it is considered as an ongoing effort to increase the competitive ability of the strategic marketing processes. MI systematically collects large amounts of high-velocity consumer data from all relevant sources, processes it to get an accurate view of the prospective markets to enter and ascertains the changing trends in the marketing environment [17]. Artificial Intelligence (AI) is an increasingly integral part of MI strategy, such strategy being nowadays considered as vital for retailers interested in increasing their market share [18]. While MI used to improve the business model and assist several different marketing goals at a high level, it's more and more used to inform decisions related to competitors, products, and consumer trends or behavior [19],[20]. As a branch of computer science, AI provides computing methods of human-like behavior, such as the ability to think, learn by example, doubt, act, see, and speak [21], to business analytics. AI research has traditionally been influenced and inspired by nature, by the way humans accomplish various tasks [22].

Data mining (DM) is the AI based engine in the process known as Knowledge Discovery in Databases (KDD) involving inference algorithms that explore and analyze large quantities of data, develop mathematical models and discover significant patterns [23]. DM was one of the most important technologies to hit the retailing industry during 2010's, as a powerful analytical tool for finding overlooked and useful information [23]. When it comes to explaining how most successful retailers in the world got to where they are today, and how they will continue to dominate the market in the future, many experts and scholars point to their proficiency in AI data mining [24]. According to Hall [22], incorporating AI techniques to anticipate consumer's next purchase and improve the consumer journey into marketing will offer a systematic process to bridge the gap between data science and execution.

One of the most central AI methods used in discovering knowledge is Association Rules Mining (ARM) used for Market Basket Analysis (MBA) that aims at finding regularities in the shopping behavior of consumers and helping uncover relationships between seemingly unrelated data in relational databases [25]. The first and arguably most influential algorithm for efficient association rule discovery is *Apriori*. Proposed by Agrawal and Srikant [26] in 1994, it is commonly used in ARM and is useful especially in discovery of frequent itemsets in large transactional databases [25]. As a categorical (non-numeric) algorithm, Apriori uses prior knowledge of frequent itemset patterns to help businesses enhance sales performance. It has been deployed to support operations in several industries like healthcare, electricity supply, manufacturing, railway safety management, education, retail, finance, and monitoring the quality for web services [27]. It is performed as a two-step process: frequent itemsets determination and association rules generation, the first step requiring more processing time [28].

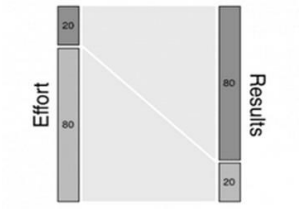

Fig. 1. The 80/20 percent rule of Pareto [14]

According to Hegland [29], Apriori suffers from two limitations. Firstly, as being an NP-hard algorithm, its performance is related to time and space complexity as it takes many potentially irrelevant databases scans when mining candidate item support in a large database, and the rules generated by these association mining techniques are large and often difficult to understand. Secondly, Apriori algorithm treats all items in the database equally by considering only the presence and absence of an item within the transaction. Said et.al. [30] stated that Apriori is not able to answer complex queries like identifying the important preoperative predictors. It lacks intelligence to accurately assess consumer's buying habits and product profitability embedded in product popularity trends and has difficulties in finding rarely occurring events [30]. Yet another major limitation is that it has difficulties in finding rarely occurring, yet meaningful events, because they don't meet the support threshold set to Apriori [31].

Point-of-sale (POS) transactions data can potentially have rare but interesting patterns. However, current methods for finding rare (infrequent) patterns are computationally expensive due to the heaviness of candidates generation of the Apriori [31]. Improved versions, such as Apriori Tid and Apriori All, as well as Direct Hashing and Pruning algorithm have been developed [32], [33]. Recently, Srivastava et. al. [34] proposed the Bi-Phase method that includes item utilities to complement Apriori algorithm to find utilities in item sets, in addition to their frequency. High confidence is attained, but the limitation of the approach lies in low support. Zhao [35], proposed frequent itemsets mining algorithm BITXOR based on the efficient bit table, which uses the bit table and represents itemsets with a binary sequence. Doshi and Joshi [27] proposed a comparative analysis of Apriori and Apriori with a hashing algorithm. The study aims to compare Apriori and Apriori with hashing algorithm to find which algorithm is better in providing an accurate result in less time. Venkatachari and Chandrasekaran [36] proposed MBA to see how different products in a grocery store assortment interrelate and how to exploit these relations by marketing activities. The model uses FP-growth and Apriori algorithm to provide mining association techniques about co-occurrences and co-purchases of products in the study that collected data from a retail store with a sample of 300 purchases. Singh et. al [37] report that Apriori algorithm and the enhanced traditional parallel and distributed algorithms have only been experimented on smaller datasets and are not efficient and scalable to manage big data.

As the literature illustrates limitations of the Apriori algorithm, as well as some of the notably enhanced versions of Apriori, it has been a good motivation to investigate the possibilities of AI and advanced database methods. In this research we propose new and efficient techniques embedded into MBA model. The ultimate objective would be to automate consumer shopping behavior forecasting and construct a comprehensive outlook on the consumers to make data-driven conclusions for marketing decisions. We also investigate the 80/20 rule at the product level. We expect our new model to tackle some of the issues existing in previous ARM algorithms.

For a retailer to succeed in this vibrant state of change industry. Brands should move towards a consumer-centric approach by bringing consumer insights into the process of decision making. A systematic process of extracting information and practical consumer's insight from POS data to enhance the decision-making process in key areas such as consumer relationship management, is needed [38]. MI plays a vital role in the formulation of marketing initiatives to achieve the goal [39]. AI has been widely studied in the literature. However, there is a lack of scientific studies referring to the AI in modern marketing, and especially to the transaction data. Most of these studies examine more generic aspects of marketing, such as consumer behavior and conversion optimization [40].

AI technologies, like modern data mining, play a vital role in the field of marketing. It involves identifying hidden patterns with complex statistical methods to gain insights into consumer shopping behavior [38]. Today, the SMR compete in rapidly changing digital markets [39]. Massive waves of change have been surfacing in many businesses, quite often originating from improvements in MI. Businesses are confronted with different social and economic changes. Smith [41] writes that their ability to survive is at risk if MI is not adapted. MI helps businesses to position correctly in today's market and minimize the retailer's exposure to risk. For the SMR, relationship with consumers is the key factor to brand loyalty, repeated store visits, and ultimately, sales conversions. However, this relationship has been affected by recent social and economic shifts. This change has prompted the retail industry to be more strategic in their marketing, seeking new channels of reaching customers, and to develop a deep

understanding of their customers. This has required a shift from traditional (mass) marketing to modern (focused content) marketing [42].

To large extent, the SMR still rely on traditional (non-digital) marketing. However, the need to accommodate to existing changes is obvious but lack of innovation and the traditional way of looking only in the rear-view mirror of known consumer shopping behavior, with the methodological assumption that consumers are like one another, are the main problems. Often, traditional marketing fails to find critical facts about individual consumers, which may lead the SMR to consume limited and valuable resources on serving unprofitable customers, while big retailers, with larger resources, have already spent on modern marketing methods to be up to date with all the current trends. Marketing based on MI leverages a scalable, measurable approach to understand broader audience and opens the flood of opportunities to build close relationship between brands and consumers. Yoseph and Heikkilä [43] developed a market segmentation model to help SMRs move from traditional marketing to target marketing. The MS model helped identifying VIP customers who are on the verge of shifting their purchases to competitors. The model helps in identifying products that are highly profitable and purchased by the most profitable consumers' segments. Davenport [44] writes about how marketing and consumer behavior will change in the future and proposes a multidimensional AI framework to understand the impact of involving AI in business operations. The objective would then be to augment human managers rather than replacing them.

Theodoridis [40], presents machine learning model that fit to many digital marketing situations. In contrast to general nature of machine learning, MBA is most often applied by large retailers who with their large POS databases and larger resources can run extensive number of database scans and utilize generated rules more effectively. As they do not have resource constraints like their smaller competitors, the large retailers can can include their mining results into their customer relationship strategies. However, for the SMR with less resources to compete in analytical resources, the association rules generation should have a different strategic point of view.

The objective of our proposed approach is essentially to achieve the maximum benefits of the MBA model. To achieve these objectives, we need to apply several steps algorithmically. Fig. 3 shows the flow chart to illustrate the (life cycle) phases of our algorithmic model. Each phase is described in detail in the next section, along with the related tasks required to complete it.

3. The proposed Market Basket Analysis model

In this section, we present the improvements we have made to the classical Apriori algorithm.

3.1 Apriori

The Apriori algorithm is a stepwise process for finding association rules [25] where data is mined to find frequent itemsets whose properties are then used as rules for creating customer offerings. The process of knowledge generation starts from finding all products, in other words, itemsets of $k = 1$, and counting the number of shopping baskets they occur in to get the candidate - C_1. Then the minimum support threshold $supp_{min}$ is used to set the minimum limit for finding the products frequently enough in the shopping baskets, for instance $supp_{min} = 0{,}001$ means that the product should occur in at least once in every thousand shopping baskets. After finding the set of all these 1-product frequent itemsets L_1 is defined. This is then further analyzed on k-levels $k = \{2, 4, \ldots\}$ with two rules:

1. All subsets of a frequent itemset must be frequent (Apriori property), and
2. All subsets of an infrequent itemset will be infrequent.

The strength of an association rule can be measured in terms of its support, confidence and lift. The support is the rate of occurrences of an itemset in a transactional database. The confidence refers to the likelihood that an item Y is also bought if item X is bought. The lift of a rule is the probability or ratio of all the items in a rule occurring together (Support) divided by the product of the probabilities of the items on the left and right-hand side occurring as if there was no association between them [12].

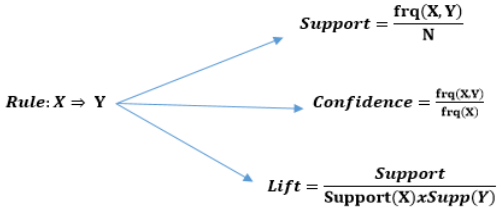

Fig. 2. Analytical parameters of Apriori

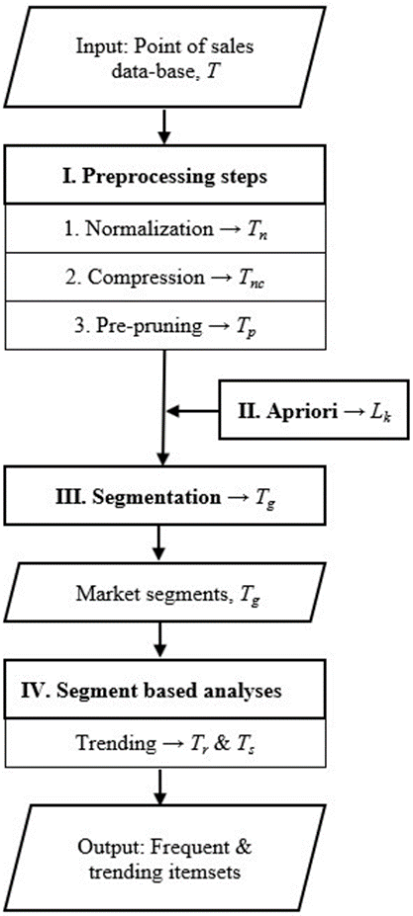

Fig. 3. Flow chart of the proposed MBA model

3.2 Approach to Enhance Apriori

The advantages of standard Apriori algorithm are inherited in our model. Better efficiency and enhanced intelligence complement the Apriori algorithm.

Our proposed approach focuses on providing viable fixes to achieve the following

1. Avoid rescanning the database.
2. Reduce the size of candidate itemsets.
3. Accelerate both joining and pruning processes
4. Introduce additional intelligent features as variables into the algorithm

As stated before, the approach consists of 6 stages. In the first three stages, the performance of Apriori is in the pinpoint. The fourth, fifth, and sixth stages contribute to the improving of the computational intelligence (CI) generated by the algorithm. Here we summarize how this model is applied to improve performance and calculation accuracy. The efficiency of the algorithm is measured by reducing the number of database scans (i.e. k-itemsets) with new approaches to Normalization, Compression, and Pre-pruning, and hence reducing the size of the dataset.

In our proposed MBA model (Fig. 3) the data is first preprocessed. It is retrieved from as a POS dataset, transformed into a separate data warehouse schema T and unified into T_n by converting identification codes into a new unified 6-digit long codes.

The second step compresses T_n into T_{tc}. Each transaction, with each purchased item in it, is converted into one single horizontal line that enables a horizontal scan. Using a horizontal scan instead of vertical improves scan performance. Here, the algorithm scans one single record instead of many. Considering that the average size of a retail POS database is about 100 Gigabytes, this will significantly decrease the number of database scans.

The third step prunes T_{tc} into T_p. This is critical for performance as itemsets not passing the support and confidence thresholds required by the Apriori algorithm are eliminated. With this step, performance decline in the classic algorithm is avoided as the many re-scans of the database and generation of sub-items are no more needed. In classical Apriori, the algorithm scans the entire database in every iteration. In our approach, the algorithm first scans the database to generate all frequent itemsets (large itemsets), by doing a single pass over the POS database to count the support of the candidates, which results in the set of 1-itemsets.

Unification

In the unification phase transaction item codes are converted from the long, mixed characters and digits layout to a new unified fixed-length format with 6 digits (Table 1). As Singelis [26] points out, in a typical relational database management process a horizontal storage manager is used. The reduction in digits improves the performance of computation. Technically, the smaller the data types are, the faster the data retrieval is, because they occupy less physical space on the hard disk, memory, and generally requires fewer CPU cycles to process [27].

Table 1. Unification of items codes

Original item ID	Unified ID
100000225823	007701
AG4039234	007902
0741-MS-20002147	008740

Compression

A typical POS database is stored in a fashion where each record is on a (horizontal) row. For example, one transaction with the purchase of 1 shirt, 2 pairs of pants, and 5 pairs of socks are stored in the database on three rows. Apriori first creates a table that contains the support count of each item in the POS dataset by scanning the entire database. This is then the candidate set containing each product in the database, C_1. Then the algorithm compares support counts of the items in this level-1 candidate set with the minimum support threshold, $supp_{min}$. Items with the support count less than the $supp_{min}$ are rejected and itemset L_1 is generated from items that pass the threshold test.

Next, the classical algorithm would run another complete database scan (iteration) to generate a candidate set C_2 by joining all the level-1 items C_1 to the next level (the join step), comparing support, counts in C_2, rejecting infrequent 2-item candidates and their subsets in L_1 (rule 2) and resulting in the new level-2 frequent itemset L_2. This process would then be iterated until no candidate itemset in the set L_k in the reached item level-*k* qualifies to C_k. Thus, the set of highest frequent itemset(s) L_{k-1} would have itemsets with 1 to *k-1* products.

The disadvantage of classical Apriori is that it generates a lot of meaningless candidate itemsets. Here, our proposed model differs entirely from the classical Apriori. Now we introduce a new technique to accelerate the joining and pruning process that we see as a fix to the classical Apriori algorithm.

The compression (one-time process) consists of two steps

1. Convert the original itemsets (dataset T_n) from vertical (product) structure to horizontal (transaction) structure, which generates the dataset T_t with *t* transactions.
2. Compress the horizontal itemsets (dataset T_t) to an even shorter horizontal dataset T_{tc}, by combining similar transaction itemsets into one row and then count the number of occurrences, *c* for the transaction *t*.

Pre-pruning

The next improvement is achieved by the pre-pruning that executes a database mapping to avoid repeated re-scanning of the POS database and eliminates the itemsets that are having infrequent subsets.

The pre-pruning process creates a dataset T_p from the level-1 itemsets, L_1. Next, as difference form the classical Apriori, candidate sets $\{C_2, C_3, ..., C_k\}$ and frequent sets $\{L_2, L_3 \ldots, L_{k-1}\}$ are generated by using only the newly generated T_p rather than rescanning the entire original POS database. This step can significantly enhance the performance and reduce the size of the often very large transactional POS database. Now, C_2 is generated form L_1 with the join step using the same process as in the generation of C_1 and L_1. As the approach only uses the T_p rather than the original POS database the efficiency difference to the classical Apriori is increased in each join step.

3.3 Implementation of the Market Basket Analysis with Market Intelligence features

In this section we present the MBA model implementation for the purpose of developing Market Intelligence for the SMR. We use PL/SQL language in our implementation. As an extension of SQL, PL/SQL is an effective and scalable programming language for processing transaction databases. We utilize features of PL/SQL to send entire blocks of statements to the database server simultaneously. Consequently, network traffic is considerably reduced, and the code can be divided into smaller modules. This enables easy conversion between other applications and other programming languages, as well.

Once we have performed our three first steps with and selected our pruned dataset T_P, we start the fourth phase of the model that, along with the steps of classical Apriori, generates new candidate level-k itemsets using the frequent (k−1) itemsets found in the previous iteration. The algorithm eliminates all candidate whose support count is less than the support threshold.

3.4 Improvement on Marketing Intelligence

The retail industry relies heavily on trends and up-to-date factual data to make key strategic business decisions. Most recent transactions can be used to determine existing and expected trends, and together with older transactions, they can be used as a source of knowledge. However, the older the data used in finding trends, the less it tells about the current product trending. Therefore, our objective with this research is to highlight the importance of the most recent purchases and use them to bring up-to-date trend information to complement MBA. For retail managers, such information together with association rules gives additional intelligence to make better-informed marketing decisions, such as special offerings to certain customers and market segments.

The significance of tracking sales trends reflecting the popularity of the product is vital to any business trying to maintain growth and increase the bottom line.

Product trends are highlighted by the change in the direction of profit from one period to another or, in the most positive case, by experiencing a surge in the product's popularity. Obviously, the upward trending leads to the increase in product profitability over time, while the downward trending has the opposite effect. The retailer may be products that have not yielded profit quarterly or semi-annually but may in the future have increased profitability due to a recent change in popularity trends. In the opposite case, an established product that recently has generated profits may have a recent declining trend that questions its future profitability. When looking for a product's trending patterns in the POS data, the retailer can determine both opportunities and potential risks.

If the product trend is declining, the retailer can make timely decisions such as to spend more on the product sales efforts, reduce prices or discontinue, but if a certain product is selling off the shelves, the retailer should stock inventory accurately across channels. Retailers often fail because of their inability to find trending products to sell. Finding one product that sells the best or the characteristics of trending products has become an arduous task, especially when every retailer is trying to follow the same marketing strategy. There are many obstacles preventing retailers from getting trending insights. Our approach offers a dynamic insight based on empirical data into the products' and itemsets' trending profitability, identifies the history of sentiments attached to the product and provides an early warning indicator of potential popularity issues, as well as deeper insights into customer's shopping, likes and dislikes. With these tools the retailer can then pinpoint areas of success and failure, then marketers can properly forecast products' future perceptions and maneuver for the future.

To consider how the customer perceives the product, and how satisfied the customer is with the retailer prices and services, is a central concern of any SMR. If the average sales of the product decline continuously, the marketing manager can conclude that the product is falling from a beneficial category to a non-beneficial category, or that customers are purchasing from other sources. In a similar manner, but contrary to the previous case, the marketing manager makes positive conclusions about the product if the sales go up and the product is very profitable and ranks it accordingly. To find trends early is especially critical for the SMR, who need to allocate their resources efficiently in order to compete against their larger rivals. With all this in mind, we introduce time series analysis into MBA to explain and anticipate itemsets lifetime value.

In our model, we determine the trending factor that emphasizes the most recent purchase for each itemset. First, we compute a slope, T_S for the sales trend from sequential sales data C_g, where g denotes the time sequence of six months for the itemset C. Then, we introduce the attenuation factor att_g that is used to discount the importance of earlier itemsets and highlight the importance of the most recently realized sales in the trend. The attenuation factor, att_g is calculated with an exponential function of the path length through the medium, also known as the arithmetical natural value of signal transmission over long distances. In other words, the strength or intensity of a signal or digits is reduced with *att*. This method has long been used along with other numerical methods to generate and validate data, and to de-noise signal data, and it is widely used in telecommunications, engineering, optic fibers, and ultrasound applications [28; 29]. Attenuation is a common method to handle weakening signals, due to for example distance. In signal processing, this factor is often considered as an attenuation coefficient, α that multiplies signal strength based on the transmission conditions (transmitted distance, transmission method, transmitting material etc.). We use the attenuation coefficient as a factor to account for recency of an itemset. The longer the time to the itemset in the analysis, the higher the attenuation coefficient is set. In other words, the value of the itemset is discounted based on its historical occurrence to the point of analysis, or – respectively – accentuated based on its recency.

Each attenuation factor for time segment g, att_g is calculated with the chosen attenuation coefficient α for each itemset based on the occurrence of the itemset from current time ($g = 0$), where time segments g can have values $g = \{0, -1, -2, -3\}$.

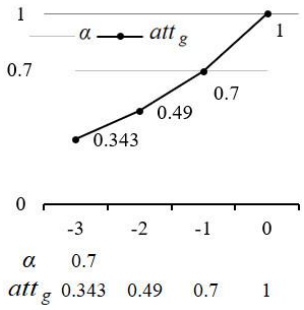

Fig. 4. Time segment slopes with attenuation factor α

The trend of the itemset sales T_S is calculated as a slope from sales figures in all selected time segments. For instance, for an itemset with four-time segments g = {0, -1, -2, -3}, if we set the calculated itemset attenuation coefficient, α to 0.7, we get the current slope (g = 0) by multiplying with $att_0 = 1$, the previous one (g = -1) by multiplying with att_{-1} = 0.7, the one before that (g = -2) by multiplying with att_{-2} = 0.49 and the first slope (g = -3) by multiplying with att_{-3} = 0.343. By this way, we decrease the importance of the signal of older itemsets in T_S.

In MBA, the transactions rule could be: {*Jeans*, *Jacket*} => {*Socks*}. This means that if a customer has a transaction that contains *Jeans* and *Jacket*, then they are likely to be interested in also buying *Socks*. Before unlocking commercially relevant insights from our POS database, we want to understand and make a conclusion about how viable the client's business proposition is.

Finding new trending products can be challenging. By the time the retailer manages to source them, the popularity may be over, and the demand has vanished. By knowing what products will be popular before they peak, i.e. having caught the wave before surfing on it, the retailer will be able to profit from growing sales. Our approach starts with computing the itemsets purchase amount slopes over the predefined historical time segments and then determines a present value for past itemsets purchase slopes. We calculate the total purchase amount slope with a smaller effect of old, already visited slopes. As retailers, this information enables us to source and spot popular high demand itemsets. It will also reveal information about infrequent profitable itemsets often missed by the classical Apriori algorithm. These could potentially contribute to the 20% most profitable items and offer the retailer a broader vision of the viability of itemsets.

Our algorithm uses a segmentation technique to subdivide the extracted POS dataset into six-time segments C_g, g = {1,…,6}, where each segment constitutes of a time period of four months.

The aim is to capture two critical factors in marketing, the time based-segmentation, and occasion-based segmentation. Segmentation with time-based dimensions can be highly effective. Some stores work longer hours than others, some products and services are sold only at certain times and dates of the year (e.g., Eid Holiday, Month of Ramadan, Christmas, the New year, etc). On the other hand, segmentation with occasion-based dimensions is based on the observation that customers tend to behave and think differently on different occasions, and such differences can be beneficial for product segmentation.

To conclude, the fifth and sixth steps of our model aim at improving the level of intelligence provided to the MBA, as compared to the classical Apriori. With the fifth step we count the trending popularity T_r and purchase power slope T_s for all itemsets, and then continue by comparing the current time segment C_g over a specific period. The sixth set step, the time series attenuation technique, is used to capture the itemset purchase slope T_s, to strengthen the effect of the most recent purchased itemsets, while justifying the importance of the previously purchased itemsets by inserting a dynamic calculated slope factor T_s.

3.5 Measuring the Pareto Rule and Effect

As mentioned previously, one of the objectives of this research is to find if there is empirical support to the 80/20 rule at the product level, and therefore, we examine the total sales driven by the top 20% of products. To identify the proportion of total sales from the top 20% of products, we use the following formula (2) to measure the Pareto ratio at the product level.

If the Pareto Ratio reaches the level of 0,8 we have found evidence to support the Pareto rule for product sales, and even with levels close to 0,8 the Pareto effect is considerable.

4. Comparative Analysis of Computational Intelligence

To verify our model, we have used point-of-sales data of a retailer in the Middle East. The retailer sells diverse products like clothing, footwear, cosmetics, electronics and accessories, and purchases a variety of

products, goods and brand products from various manufacturers, vendors and wholesalers. Some products are purchased from local vendors and wholesalers, and some goods are shipped directly from manufacturers from China and Bangladesh, for example.

This section contains analytical comparisons between the classical Apriori and our novel, improved model based on analysis with POS data of this small retailer. It should be noted that due to the novelty of our approach our point of view has not been to give rigorous proofs to our model or any step within the model. Instead, as the model is developed in cooperation with business practitioners, our analysis is motivated by the experienced benefits in every-day use of the model.

4.1 Database Size Performance

The first two methods to increase algorithmic performance are based on the unification and compression of the data. These methods aim at reducing the time consumed in the generation of candidate itemsets C_k. In other words, we want to organize data in such a way that ensures effective insertion, retrieval, pruning optimization, and transaction reduction of data.

To verify performance improvement achieved by unification, compression, and pre-pruning we ran the algorithm several times with our experimental dataset. The two compression steps reduce the database size and reorganize the dataset content (Fig. 5.) The first step reduced the database size to the level of 48% from the original database. The second compression step reduced the size of the database further to the level of 41% from the original and 84% from the previous step. The last pre-pruning step is a key step to avoid redundant association rules, and the analysis shows database size reduction to the level of 11% from the original POS dataset and 27% from the previous compression. Analysis results in show how various stages in our approach effect on the physical database size and the total database rows count. Both the database size and row count have dramatically decreased already after the first compression (48% and 11% of the previous state, respectively). The second considerable reduction takes place after pre-pruning that decreases the twice compressed database (27% and 55% of the previous stage, respectively). We can see that compressing the database has a considerable effect on the row count while the pre-pruning has the largest effect on the size of the data file.

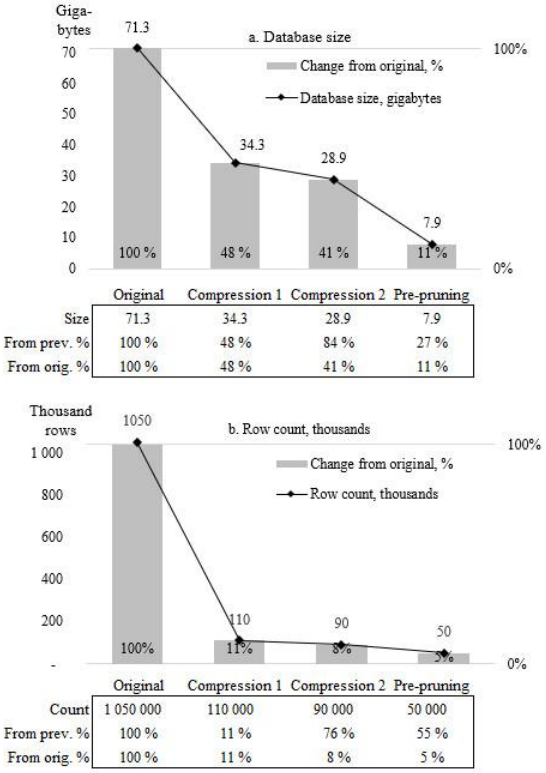

Fig. 5 Effect of the approach on the database size (panel a. top) and row count (panel b. bottom)

4.2 Operational Performance

Precise operational metrics in this research measure database speed, maximum memory consumption and database storage capability. The analysis takes six operation metrics (see Table 2).

Table 2. Operational performance metrics

Operational metric	Description
Database scans, 1000s	Nr. of database scans in thousands
Records count, 1000s	Number of records in itemsets in thousands
Size, GB	Size of the database
Total reads, 10e+10	Time each algorithm I/O takes (Database size * Database scans)
Mining process	Repeated instructions until all conditions have been met
Exec. time (min)	Total time in minutes consumed

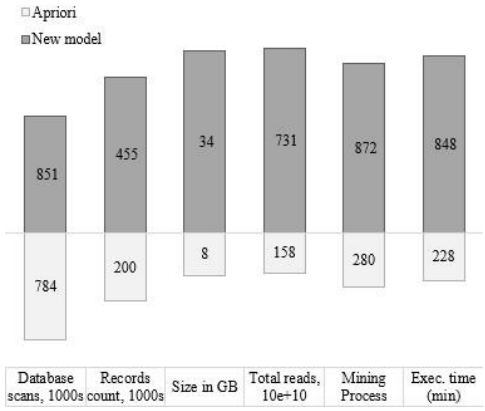

Fig. 6. Operational performance, test results

Results show how our approach uses less time (about a fourth of the time Apriori uses) and less memory, needs less physical space than the Apriori. The results of the comparative experiments demonstrate that the convergence speed and accuracy of the approach significantly enhances the performance in all performance operational metrics, and in that way provides a more efficient association rules discovery than the classical Apriori algorithm.

4.3 Finding Meaningful Association Rules (Artificial Intelligence)

Here, we present the analysis of association rules generated with our approach. To obtain the most appropriate market baskets and trend results, our general recommendation to the retailer is to experiment with different support and confidence thresholds to obtain the most appropriate values. While there are too many rules to be able to look at individually, for our analysis, we will focus on the top 5 rules with the largest lift. *The lift ratio* is the factor by which the co-occurrence of A and B exceeds the expected probability of A and B co-occurrence, had they been independent. Therefore, the higher the lift ration, the higher the chances of A and B occurring together.

shows 5 frequent itemsets rules with the largest lift, and a closer look at these rules seems to make a strong intuitive sense. For example, the first rule might represent the sort of itemsets purchased for getting ready for the autumn and winter season. The second rule is most likely a set of gifts purchased for a young woman used for day to day outing. The probability of the characteristic of the third shoppers strongly suggests a family shopper buying Women Sketchers with girls matching Socks. The fourth rule is most likely for a single woman bought for working out. The probability of the fifth rule itemset is more of a purchase preceding the 'Eid Muslim holidays where this type of perfume is very popular for exchanging gifts.

use three different two-product itemsets[1]. Figures 7 and 8 show the behavior of Itemsets 1-4, over six-time segments with both the original itemset trend and the attenuated itemset trend. The blue line depicts the original itemset trending and the blue line the trend after the time series attenuation.

In this section, we display the power of the approach by showing a sample of itemset trends, both downwards and upwards. For such an analysis we use visual tools to highlight the effect trending capabilities of attenuation. To show the effect of attenuation we

Table 3. Five rules with the largest lift

	Rule	Supp.	Conf.	Lift
1	{Unisex Flat Front Pant, Unisex Casual Pant (Grey}, {Girls' Short Sleeve white shirt, Boys' Short Sleeve white shirt} => {School bag}	0.0097	0.236	38.01
2	{Women Oud Saffron Ro by Perry Ellis}, {Musk Oud-6ml Roll-on Perfume Oil Box 6} =>{Bakhoor Hajar Al Aswad - Exotic Arabic Incense}	0.0082	0.191	13.44
3	{Sketchers Women's D'Lites - Bright Blossoms}, {Sketchers Girls' Dyna-Air} => {Girls' 3 Pack Heart Crew Socks }	0.0073	0.149	17.03
4	{Women Sketcher GOrun Mojo – Radar}, {Boys' S Lights: Sketcher Energy Lights – Merrox} => {Women's 6 Pack Half Terry Low Cut Socks Box}	0.0061	0.117	14.82
5	{Maybelline New York Lash Sensational Mascara-9.5 ml}, {Maybelline New York Master Camo Concealer Palette - 6 g, Light 1} => {Maybelline New York Superstay Nail Polish}	0.0057	0.087	14.00

[1] Itemset 1: Boys' S Lights: E-II Sandal Beach Glower; Girls 6 Pack Low Cut Cutie Critter Socks

Itemset 2: Local made Unisex Flat Front Pant – Girls' Short Sleeve white shirt – Boy's Short Sleeve white shirt

Itemset 3: Women Oud Saffron Ro by Perry Ellis; Musk Oud Roll-on Perfume Oil Box of 6

Itemset 4: Unisex Black Lights Sketchers; Unisex Retro Polarized Sunglasses; With Aris Aris Perfume & Deo Spray For Women Gift Set

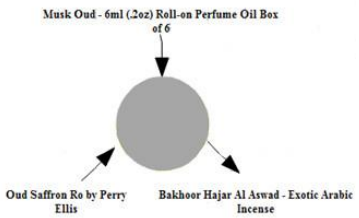

Fig. 7. Rule nr 2 visually.

4.4 Finding Trends and Popularity in The Itemsets (Artificial Intelligence)

Fig. 8, upper panel, shows how itemset 1 was trending upwards in the first period of the analysis, mainly during summertime. After that the demand decreased, the popularity goes down in later time segments. The grey line drawn after attenuation for the time series shows that the itemset is still trending downward (the latest segment 6 in the grey line). However, the degree of downwards trending is smaller. This does not necessarily mean that customers are not actively buying the products of itemset 1. Instead, there might be a marketing disconnect between the product and the customer, or lack of perceived value from the customer perspective. The retailer should evaluate the viability of this itemset and maybe find a way to reverse or slow downward trending for these products.

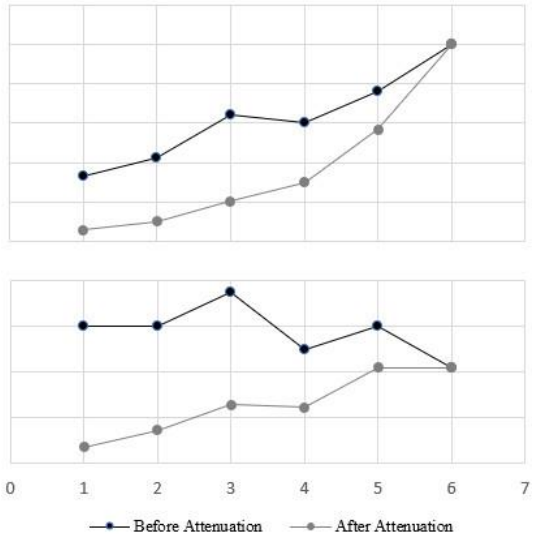

Fig. 8 Trend attenuation, itemsets 1 (top) and 2 (bottom)

Similarly, the bottom panel of Fig. 8 depicts itemset 2 over the six-time segments. The itemset has gained popularity in the middle periods of the analysis (segments 3 and 5), but the overall result of this itemset is trending downwards (segments 4 and 6). After attenuation, the itemset trend is mostly upwards, and we know that the increase occurred in the winter season (segments 3 and 5). The conclusion could be that there are customers who are not guided or driven by fashion trends and are only willing to wait to pay less for the itemset out of the season.

The upper panel of Fig. 9 depicts itemset 3. This itemset has also been gaining popularity in five-time segments. However, there is a slight downward trend in the fourth segment suggest that this itemset is less popular during this period of time? No, because after attenuation also showing steady popularity, where the sales of itemset 3 gradually start building in August all the way to December. When we dig further into the features of Itemset 3, we see that it is a unisex black Lights Sketchers, with unisex retro polarized sunglasses that are always purchased together with Aris Aris Perfume & Deo Spray for Women gift set. It is surprising that such uncorrelated products, and belong to different categories are the best up-warding trending all throughout the year.

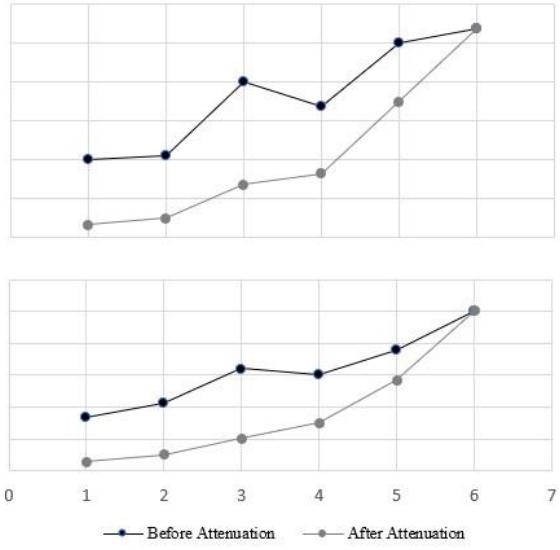

Fig. 9. Trend attenuation, itemsets 3 (top) and 4 (bottom)

Similar to itemset 3, itemset 4 has been gaining popularity in five-time segments. However, there is a slight downward trend in the fourth segment suggest that this itemset is less popular during this period of time? No, because after attenuation also showing steady popularity, where the sales of itemset 4

gradually start building in August all the way to December. When we dig further into the features of itemset 4, we see that it is a unisex black Lights Sketchers, with unisex retro polarized sunglasses that are always purchased together with Aris Aris Perfume & Deo Spray for Women gift set. It's astonishing that such uncorrelated products and belong to different categories are the best up-warding trending all throughout the year.

4.5 Testing the Pareto Rule

In the last section of our analysis, we run a simple summation to test the viability of the Pareto rule in our data. Table 4 shows how ranking the highest performing products of our analysis and then splitting the data at 16.2% of the ranked products implies 77.8% of total sales. Thus, there is rather strong a compliance in the data to the Pareto rule. This imbalance implies consequences on marketing. By avoiding to look at the pattern that the rule provides, and failing to gain appropriate marketing intelligence to find the best selling products, there is a danger that the retailer mistargets marketing efforts. Furthermore, for our data, with the majority of customers coming from an Arab background, our market baskets imply that there is a strong attachment to certain brands and products that deliver the premium sales of 77.8%.

Table 4. Summary of profit-generating products

Category	80/20 products	Other Products	Total
Nr of products	1 805	9 310	11 150
% of products	16.2%	83.5%	100%
Sales	61 097 000	17 455 218	78 552 218
% of sales	77.8%	22.2%	100%

5. Conclusion

In this research, we applied a structured customer basket analysis, and multidimensional analysis was carried out to answer the research questions. Two unsupervised algorithms implemented on Oracle POS data warehouse to forecast consumer shopping behavior. The extracted analysis results were extensively analyzed and cross-validated by judgmental experts (The Client). The analysis sheds light on who is the ideal customer. Digging deeper into the analysis, and looking at the characteristic of the purchased products (see section Results show how our approach uses less time (about a fourth of the time Apriori uses) and less memory, needs less physical space than the Apriori. The results of the comparative experiments demonstrate that the convergence speed and accuracy of the approach significantly enhances the performance in all performance operational metrics, and in that way provides a more efficient association rules discovery than the classical Apriori algorithm.

4.3 Finding Meaningful Association Rules), we find that products with an Arab background, especially those targeted to women, are the ideal products with the highest associations. In addition, most of the extracted association rules show that products targeted to female customers are associated with other specific products. Also, looking at time periods, back to school season is the ideal shopping timing for the retailer. Arabian Women Perfumes, female Sketchers Shoes, and female T-shirts purchased with pants are the best selling products. While Women / Girls Sketchers Shoes purchased with Socks accessories are the best trending products.

The proposed MBA model has proven to be scalable and robust, the, results enable retailers to catch old and new trends in a continuously changing market. Now, the client is systematically involved in an iterative process of knowledge discovery, and the type of extracted information is valuable if the retailer is interested in marketing activities such as cross-selling or targeted campaigns. Our "Åbo" model has shown far more superiority in terms of computational performance, and artificial intelligence over the classic Apriori algorithm. The model was able to formulate product popularity trend over time with the new conception of calculating frequent Itemsets (Support, Confident, Lift). Finally, the main research contribution can be summarized into the following. This research is a significate step to demonstrate the importance of designing and developing synergies when incorporating three different and vital areas of research, namely market research, consumer shopping behavior, and Computational and Artificial Intelligence, to forecasting consumer's shopping behavior and develop market intelligence for the SMRs industry. Next, we summarize our paper with our answers to the research questions.

Does the approach improve the computational performance and the intelligence of the classical Apriori association rules mining algorithm?

The approach has shown computational efficiency as compared to the classical Apriori. Firstly, from the vantage point of performance, the comparative analysis demonstrated that the new methods in our approach enhance the computational efficiency and provides efficient association rules discovery. The time taken to scan database transactions is reduced, and the time used in generation of candidate itemsets has also been considerably decreased. Moreover, the size of the database is decreased. Secondly, from the vantage point of computational intelligence, the time aspect is more profoundly addressed in the calculation of product performance measures. In our approach, the probability is calculated by logically examining the dates, where most recent transactions are spotted in our model. Giving the retailer the ability to understand the holidays shopping seasons across key products categories better. Also, the approach picks interesting infrequent itemsets that miss the minimum support threshold and use these to detect shopping patterns.

Are there any empirical pieces of evidence of Pareto's 80/20 rule at the product level?

Analysis of POS data suggests that the Pareto rule is valid and implies that more focused marketing efforts are needed for improved efficiency of sales. The analysis results extracted by our methods can help the SMR to optimize their marketing processes, increase the visibility of their products and increase sales on product promotion and advertising. In addition, with also the ability to review different scenarios they can select optimal marketing strategies. Finally, this approach helped a retailer in Kuwait to adjust their product assortment for the year 2019 to better meet customer preferences and they want to repeat the analysis for the year 2020.

For future work, we are investigating the incorporation of Product Lifetime Value measure before implementing ARM.

References

[1]. Messina, P. (2019). Finance for SMEs: European Regulation and Capital Markets Union: Focus on Securitization and Alternative Finance Tools. Kluwer Law International BV.

[2]. Turner, R., & Ledwith, A. (2018). Project Management in Small to Medium-Sized Enterprises: Fitting the Practices to the Needs of the Firm to Deliver Benefit. Journal of Small Business Management, 56(3), 475-493.

[3]. Kashani, K., Jeannet, J. P., Horovitz, J., Meehan, S., Ryans, A., Turpin, D., & Walsh, J. (2005). Beyond traditional marketing: innovations in marketing practice. John Wiley & Sons.

[4]. Omar, O. E. & P. Fraser (2010). The role of small and medium enterprise retailing in Britain. UH Business School Working Paper, University of Hertfordshire.

[5]. Bakhtiari, S., Breunig, R. V., Magnani, L., & Zhang, J. (2020). Financial Constraints and Small and Medium Enterprises: A Review. IZA Discussion Paper

[6]. Kashani, K., Jeannet, J. P., Horovitz, J., Meehan, S., Ryans, A., Turpin, D., & Walsh, J. (2005). Beyond traditional marketing: innovations in marketing practice. John Wiley & Sons.

[7]. Goyat, S. (2011). The basis of market segmentation: a critical review of the literature. European Journal of Business and Management, 3(9), 45-54.

[8]. Chiu, W., Kim, T., & Won, D. (2018). Predicting consumers' intention to purchase sporting goods online: an application of the model of goal-directed behavior. Asia Pacific Journal of Marketing and Logistics, 30(2), 333-351.

[9]. Jones, M. A., & Jones, N. A. (2018). Systems and methods of controlling the distribution of products in retail shopping facilities. U.S. Patent Application No. 15/803,343.15/803,343.

[10]. Zhan, Y., Tan, K. H., & Huo, B. (2019). Bridging customer knowledge to innovative product development: a data mining approach. International Journal of Production Research, 57(20), 6335-6350.

[12]. Berry, M. J. A., & Linoff, G. S. (2004). Data mining techniques second edition - for marketing, sales, and customer relationship management. Wiley.

[13]. Stubseid, S., & Arandjelovic, O. (2018). Machine learning based prediction of consumer purchasing decisions: the evidence and its significance. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

[14]. Koch, R. (2011). The 80/20 Principle: The Secret of Achieving More with Less: Updated 20th anniversary edition of the productivity and business classic. Hachette UK.

[15]. Marshall, P. (2013). 80/20 Sales and Marketing: The Definitive Guide to Working Less and Making More. entrepreneur Press.

[16]. McCarthy, D. M., & Winer, R. S. (2019). The Pareto rule in marketing revisited: is it 80/20 or 70/20? Marketing Letters, 30(2), 139-150.

[17] Sterne, J. (2017). Artificial intelligence for marketing: practical applications. John Wiley & Sons.

[18].Igbaekemen, G. O. (2014). Marketing intelligence as a strategic tool for competitive edge.British Journal of Marketing Studies, 2(5), 17-34.

[19].Unemyr, M., & Wass, M. (2018). Data-Driven Marketing with Artificial Intelligence: Harness the Power of Predictive Marketing and Machine Learning. Magnus Unemyr AB

[20]. Franco, M, Magrinho, A and Silva, JR (2011). Competitive intelligence: a research model tested on Portuguese firms, Business Process Management Journal, Vol. 17, No. 2, pp.332 – 356

[21]. J. McCarthy, (1998). What Is Artificial Intelligence? Working paper, Stanford University.

[22]. Hall, J. (2020). How Artificial Intelligence Is Transforming Digital Marketing. [online] Forbes. Available at: https://www.forbes.com/sites/forbesagencycouncil/2019/08/21/how-artificial-intelligence-is-transforming-digital-marketing/#3ddc7d0121e1 [Accessed 8 Mar. 2020].

[23]. Palace, B. (2016). Data Mining, a technology note prepared for Management 274A, Anderson Graduate School of Management at UCLA, Spring 1996, Available at: https://web.archive.org/web/20161119003144/http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm [Accessed, 8. Mar. 2020]

[24]. Omari, A. (2008). Data Mining for Retail Website Design and Enhanced Marketing, Doctoral dissertation, University of Düsseldorf, Germany.

[25]. Szymkowiak, M., Klimanek, T., & Józefowski, T. (2018). Applying Market Basket Analysis to Official Statistical Data. Econometrics, 22(1), 39-57.

[26]. Agrawal R, Srikant R (1994). Fast algorithms for mining association rules. In: Proceeding of the 20th international conference on very large databases, VLDB, pp 487–499.

[27]. Doshi, A. J., & Joshi, B. (2018). Comparative analysis of Apriori and Apriori with hashing algorithm. International Research Journal of Engineering and Technology (IRJET), 5(1), 976-979.

[28]. Ng, A. & Soo, K. (2017) Numsense! Data Science for the Layman: No Math Added. Annalyn Ng & Kenneth Soo

[29]. Hegland, M. (2003). Algorithms for association rules. In Advanced lectures on machine learning (pp. 226-234). Springer, Berlin, Heidelberg.

[30]. Said, I. U., Muhammad, J. M., & Gupta, M. K. (2015). Intelligent Heart Disease Prediction System by Applying Apriori Algorithm. International Journal, 5(9), 887-891.

[31]. Arunkumar, M. S., Suresh, P., & Gunavathi, C. (2018). High Utility Infrequent Itemset Mining Using a Customized Ant Colony Algorithm. International Journal of Parallel Programming, 1-17.

[32]. Sun, L. N. (2020). An improved apriori algorithm based on support weight matrix for data mining in transaction database. Journal of Ambient Intelligence and Humanized Computing, 11(2), 495-501.

[33]. Rathod, S., & Sharma, A. (2016). Survey Paper for Enhancement of Apriori Algorithm. International Journal of Engineering and Management Research (IJEMR), 6(2), 808-811.

[34]. Srivastava, N., Gupta, K., & Baliyan, N. (2018). Improved Market Basket Analysis with Utility Mining. In Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) (pp. 26-27).

[35]. Zhao B.G., Liu Y., (2015). An efficient Bittable Based frequent itemsets mining algorithm. Journal of Shandong University (Natural Science), (5), 5.

[36]. Venkatachari, K., & 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), 56.

[37]. Singh, S., Garg, R., & Mishra, P. K. (2015). Performance analysis of apriori algorithm with different data structures on hadoop cluster. arXiv preprint arXiv:1511.07017.

[38]. Akinkunmi, M. (2018). Data Mining and Market Intelligence: Implications for Decision Making. Synthesis Lectures on Engineering, 13(1), 1-181.

[39]. Sharp, S. (2012). Competitive Intelligence Advantage: How to Minimize Risk, Avoid Surprises, and Grow Your Business in a Changing World (Wiley). Revista Inteçigência Competitiva, 2(2).

[40]. Gkikas, D. C., & Theodoridis, P. K. (2019). Artificial Intelligence (AI) Impact on Digital Marketing Research. In Strategic Innovative Marketing and Tourism (pp. 1251-1259). Springer, Cham.

[41]. Smith, D. (2019). Book review: Rohit Bhargava, Non-obvious 2019: How to predict trends and win the future.

[42]. Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A (2020). The impact of big data market segmentation using data mining and clustering techniques. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-15.

[43]. Yoseph, F., & Heikkila, M. (2018). Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. In 2018 International Conference on Machine Learning and Data Engineering (iCMLDE) (pp. 108-116). IEEE.

[44]. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(1), 24-42.