

A Strategy for Mining Association Rules Continuously in POS Scanner Data

Egídio Loch Terra
Faculdade de Informática – PUCRS
Porto Alegre – Brazil
egidio@inf.pucrs.br

Adilson Adão Borges Júnior
IGR - Université de Rennes 1
Rennes - France
adilson.borges@fnac.net

Abstract: In Retail Organizations the operational procedures generate a huge volume of data. To analyze it manually is practically impossible, then alternatives such as data mining are used. One important kind of Data Mining application is Association Rules, which is fully applicable to retail data, commonly called market basket, generated in Point-of-Sale(POS) scanners data. In Association Rules, products with linked behavior are revealed, being very useful for decision making. The managerial implications of mastering the formation of baskets of products are important for retailers in a general way. At the same time, practical aspects in Association Rules discovering in retail organizations are difficult to overcome, given the volume involved. In this paper, we describe a strategy for mining retail data continuously. The strategy used is divide-and-conquer, where we combine small pieces solution in a rules base, thus providing both historical and up-to-date knowledge. We present a successful case study based on our strategy altogether with extracted results and analysis.

I. INTRODUCTION

Knowledge Discovery in Databases (KDD) or *Data Mining* has been an area of intense research since its beginning in early 90s and many applications have been developed in distinct domains. The goal of such applications is to retrieve knowledge buried in large volumes of data stores in database[4][5]. Extracted knowledge is of a wealthy value for organizations, providing them precise and up-to-date information required for decision making.

In retail organizations, a high volume of data is created through transactions made daily. To analyze these data it is necessary some kind of appropriate framework, environment or tool, and this is where KDD play its role. One classical KDD application is Association Rules [1], where elements are verified to check if their behavior is linked. In Market Basket, this is exactly what should be done, products within a market basket are checked to verify if they are associated. In these organizations, data is generated in Point-of-Sales (POS), equipped with scanning system, enabling transactions to be fully automated.

The managerial implications of mastering the formation of baskets of products are important for retailers in a general way, opening many alternatives in their decision making. For example, a profitability analysis for each product can be made by taking into consideration the basket to which it belongs and considering different margins based on the time

of turnover, as already practiced in organizations with advanced management techniques. At the same time, it is common, especially when the strategy is leadership of costs, that the retailer promotes a given product with zero, or even negative margin, in order to generate turnover and more sales in the store in a general way.

Although easily stated, Association Rules practical application are rich in details that can turn into execution problems. It must be decided what kind of rules are useful, which are repetitive and novel, and consequently interesting. Also, it is not an easy task to know for how long time they must be extracted and stored, and, if stored, how to retrieve the most interesting ones. In this paper we present a strategy for mining POS scanner data continuously, addressing these questions, thus allowing decision makers to access both historical and up-to-date knowledge extracted from transactions. A case study is also presented, describing strategy application, choices made and results extracted.

This paper is organized as follows. First, Section 2 presents KDD and Data Mining aspects with particular focus on Association Rules. Section 3 describes POS common operations and information. In Section 4 our strategy is presented, based on previously described POS characteristics. The case study is presented in Section 5. In Section 6 conclusion and future work are presented.

II. DATA MINING AND ASSOCIATION RULES

In [5] Data Mining is regarded as one of the existing steps in Knowledge Discovery in Databases (KDD) Process, which is the "non-trivial process of extracting patterns from data that are useful, novel and comprehensive". This process treats all parts in whole analysis task, from understanding goals and collecting data. to results verification and validation. Basically, the KDD process can be viewed in three parts: 1) goal definition and data preparing, where existing data, related to the goal is collected, cleaned and enriched as much as possible; 2) Analysis, where data is transformed and applied to one or more Data Mining algorithms; 3) Results Interpretation, validation and deployment, where useful, novelty process output is used as a benefit to the organization.

Although Data Mining requires high computational efforts, the preceding steps of preparation for data mining accounts for between 75 and 85% of the overall process

effort [3][4]. On the other hand, the Data Mining step is the part of the process where knowledge is extracted. Since this task is not human-performable, due to the quantity and complexity involved, though special attention should be taken in the mining algorithm choice and parametrization.

There exist many distinct Data Mining techniques, each one with many algorithms. [1] gives a first taxonomy to these techniques depending on the author's point of view, calling them *class problems*: a) Association Rules, where events are verified in order to determine if their behaviour is linked (in business point of view); b) Classification, where events in business are labeled within some classes and then a description is searched for them, so new events can be pre-classified; c) Sequences, where events are studied to check if their occurrences are subsequents, or if there exists sequence patterns in business events. These class problems incorporate most of the interesting techniques from the business point of view, except for the lack of clustering techniques. In Clustering, elements or events are analyzed to check if they behave similarly, or if unlabeled group patterns exist. In [5] techniques are further investigated and are called *Data Mining Tasks* and include new techniques, such as clustering, summarization and deviation detection.

The Association Rules' most commonly used example is the relationship between products in Market Baskets [1]. It is stated in an antecedent and consequence set : $A \Rightarrow B$, where A is an antecedent and B a consequent; or $A, B \Rightarrow C$, where there are two antecedents and one consequence. [7] states association rules as "an interesting combinatorial problem", but it is very hard to predetermine if combinations occur and to what level, two products, three products, etc... These peculiar features allied to its potential results application pushed attention to association rules.

Although easily stated, Association Rules demand some choices to be made. The first is the parameters threshold. In these rules there are originally two parameters to be adjusted: Support and Confidence. The first one determines how often rules must occur within all existing events, which gives an overall rule significance. Confidence is used to verify the internal strength of the rule. This means that an element's occurrence must not be much higher from others that appear in same the rule, e.g. if one element appears in all transactions then it will be associated to all products and will appear in all rules, and this might not be interesting in the business point of view. Adjusting parameters is not an easy task, since their values work as a pruning measure : the higher the support and confidence the fewer the rules generated, but there are no perfect/predetermined values. Choosing parameters is an empirical procedure, and based on first rules extracted they can be readjusted.

Another choice to be made is the use of quantitative continuous values in contrast with qualitative discrete/boolean ones. It is different to say that elements A

and B are associated (e.g, $A \Rightarrow B$) than to say that A with amount of 30 and B with amount of 50 are associated ($A(30) \Rightarrow B(50)$). In the second case, an event with A and 20 altogether with B and 10 would not fit the rule and, therefore, would be discarded. Treating quantitative values has been addressed in [8].

The last choice to be made is the number of consequences and antecedents. This alters the mining procedure because having found a pair of elements (antecedent and consequent), it must be decided if the process continues to find more antecedents or consequences still satisfying the parameters' threshold. As the events are not necessarily linked, rules are not transitive, e.g $A \Rightarrow B$ and $B \Rightarrow C$ does not necessarily mean $A \Rightarrow C$.

III. POINT-OF-SALE SCANNER DATA

In retail organizations, data from business transactions are generated in a high frequency and, as result, a high volume is created. Each time a transaction is made in a Point-of-Sale (hereafter called POS), a transaction record is created and, in some cases, immediately sent to a local server.

Due to the sheer volume of the data, the organizations can not store all transactions for a long period. By the time new data is generated, operational procedures are made, such as accumulating the total amount of money spent and taxes calculation. In some cases, stock is also altered by this data, particularly when the EDI (Electronic Data Interchange) and ECR's (Efficient Consumer Responses) technologies are used. POS scanners also generate action records, such as POS opening and closing time, payment method and data, etc...

In each transaction record there is, at least, information about product purchased, quantity and price, and there could be one or more products in one transaction. Every product must be uniquely identified in a transaction and, often, products have their barcode stamped. This identification is attributed by the product's manufacturer. Unfortunately or not, it is not true that all products are industrialized, e.g. natural produce such as fruit and vegetable have no barcode associated to them. As all of these products are sold, they must coexist inside the system. Consequently, products without manufacturer identification are attributed by a system which creates one.

Another important point is the aggregate level of analysis. If we consider stock-keeping units (SKU) in a supermarket, the number of products, and consequently the association possibilities, is too big to consider. One approach to deal with this problem is to work with the products' categories. We can see the relationship between coffee and milk, and, if this association is interesting, we can explore the association between the coffee and milk's brands.

Organizations often has internal categories for products they sell, and sometimes, these form a Concept Hierarchy [6]. This hierarquical opportunity can be explored in order to reduce the combinatorial problem existing in discovering association rules.

IV. A STRATEGY FOR DATA MINING IN POS SCANNER DATA

The nature of POS scanner data introduces many problems whilst working with a Data Mining application. If the data quality problem is not so serious as to affect the extracted results, then the quantity becomes the focus of existing difficulties in Data Mining.

A. POS Scanner facts

The number of daily transactions can easily reach ten thousand even in small stores. In each transaction the number of products can reach 200, but on average this number would be close to 10 [2]. The number of transactions have a great variance among different types of stores. Take those numbers as fact, then 100.000 (10.000 transaction x 10 products average) products sold should be counted every day. Moreover, if choice is a low level aggregation (~100.000 distinct products) and association rules with one antecedent and one consequent, then an astronomic 4 trillion possible rules should be checked on 10.000 transactions. If transactions are accumulated in a period longer than one day, this situation worsens, e.g in a monthly basis 260.000 transaction would occur and be analyzed.

B. Divide-and-Conquer Strategy

The divide-and-conquer strategy is very simple though powerful. It is based on problem decomposition, solution and combination. In Market Basket, the whole problem is to find Association Rules in a set of all transaction within a period. This problem can be divided in time intervals, for example in days or weeks. Association Rules are discovered in all resulting sets and then combined.

From business point of view, Data Mining process is executed when decision makers want information about their task. If knowledge is continuously extracted, then the decision making task is better supported, and mainly, executed faster because rules are readily available.

A daily frequency to generate association rules can also be used to break down the combinatorial problem involved and supply decision makers with knowledge, this corresponds to divide the problem and solving each fragment. The output is a set of solutions that must be combined to solve the entire problem. To do so, a rules base is created to store daily results and to enable later retrieving

of interesting rules. From the adoption of a rules base a new parameter is established : *stability*. A rule is interesting if, within a specific period, its *stability* satisfies a minimum threshold, or if the rule occurs at least as many times as defined in the *stability* parameter (in percentual), e.g. if the rule happens every day in the period analyzed, *stability* will be 100%.

With good Support and Confidence choice, this strategy allows a continuous data acquisition, mining and storing of interesting rules, which makes raw scanner data unnecessary after this daily process (in association rules point of view). Rules base size can grow faster depending on support and confidence threshold but will be far smaller than storing transaction data. Stability and period can be defined by the time of analysis, enabling a flexible use of knowledge extracted, for example, it can be compared to rules in one day against rules generated in same day one year before.

V. CASE STUDY : DIVIDE-AND-CONQUER STRATEGY IN PRACTICE

A. Strategy Parameters and Choices

This strategy was applied to two supermarket stores in Porto Alegre, Brazil. Data from all the purchases made by consumers in the period of one hundred and twenty days were used. These two supermarkets are equipped with scanning systems for reading barcodes and will be identified as store A and store B.

The associations were generated on a daily basis, being accumulated in a general Rules base that has the following configuration: Group (the products in the association), Support, Confidence and date (transaction date).It was decided to use discrete product rules, where products were analyzed to check if they are or not present in transaction, quantities involved in product purchase were not treated.

The aggregation level was defined to treat products closer to their functions and far from brands and major physical aspects. This choice reduced the number of possible products in transactions to 4500, reducing the total of possible rules to 10 million. The side effect of this choice is that products in transaction were to be modified because they are in a lower level, exactly on the granularity level, and as a result, all products in all transactions must be transformed prior to Mining. This additional transformation considerably reduced the time spent in analysis, although including extra-time to get data ready. All data was converted to meet analysis aggregation requirements. Additionally, all recorded operational transactions were removed by the time of this conversion, leaving only the useful information for data mining.

To limit daily rules, the extraction support parameter was to set a minimum limit of 1%. Therefore, products that were not within 1% of the buying tickets in a given day were

excluded and were not considered for the rule formation. This reduced the number of rules to be tested within one day and consequently, the groups that didn't represent 1% in the total number of transactions that day were also excluded. That does not harm the global analysis, just excluding the products that are not significant in the total sales of the stores.

An analysis of all the special offers in the period was carried out, and the days in which the associated products were used in special offers were excluded from the base of rules. To do so, a database was created for all of the products used in special offers in the period analyzed, so that they do not affect the final results of associations. It should be pointed out that the period in which the products were out of the associations base was exactly the same as the one in which they were on special offer, and late effects of these offers were not considered, though they can still exist.

B. Extracted Results

The resulting rules base stored more than 6000 extracted rules. Some of them were frequent, others happened just few days. After having rules base ready to query, it was decided to compare results from one store against the second one. The following results are one of possible alternatives to use our strategy, many others can be executed.

In our study, we choose to use store A as a reference. In other words, we applied the criteria for selecting the associations from store A and later checked their presence in store B. Our methodology allowed us to find many associations. For space constraints, we'll show only six associations. These associations are classified in usual and non-usual. In the first one, the products involved have shared uses or applications, which justifies consumers' choice (e.g. coffee and milk). The non-usual associations are defined as the one in which there is no direct link between the associated products in terms of use or application (e.g. milk and detergent) [2]. The results from store A are presented in Table 1. Store B results are shown in Table 2.

TABLE 1 - Store A Results

<i>Association Rule</i>	<i>Stability</i>	<i>Conf.μ</i>	<i>σ Conf.</i>
Usual:			
<i>Ham & bread</i>	100%	52.27%	4.65
<i>Carrots & tomatoes</i>	100%	55.17%	6.76
<i>Sugar & long-life package milk</i>	100%	48.96%	5.40
Non-Usual:			
<i>Clothes conditioner & special rice</i>	96%	34.10%	5.60
<i>Detergent & long-life package milk</i>	96%	52.57%	6.81
<i>Onion & cheese</i>	97%	25.39%	4.16

TABLE 2 - Store B

<i>Association Rule</i>	<i>Stability</i>	<i>Conf.μ</i>	<i>σ Conf.</i>
Usual:			
<i>Ham & bread</i>	100%	47.04%	5.72
<i>Carrots & tomatoes</i>	96%	52.31%	6.06
<i>Sugar & long-life package milk</i>	100%	40.79%	5.77
Non-Usual:			
<i>Clothes conditioner & special rice</i>	27%	28.68%	3.49
<i>Detergent & long-life package milk</i>	37%	49.18%	6.40
<i>Onion & cheese</i>	71%	22.71%	3.83

Source: Raw Data

The first block of products presented in Table 1 is made up of usual associations. It should be pointed out that these associations, starting from the culture of the area where the research was developed, presents products with common uses and applications. In store A, the three associations were present in 100% of the valid days. Meanwhile, in store B, the association between carrots and tomatoes was present in 96% of the valid days, which is also a high index, and the other associations were present everyday. According to the concept of stability of the association defined by the author, presence above 95% of the observed period and standard deviation below 10, the usual associations are stable in the two stores.

The confidence expresses the relationship between the two products that make up the association. Therefore, we can

see that on average, in store A, 52.27% of the consumers that bought ham also bought bread, 55.17% of those that bought carrots also bought tomatoes and 48.96% of those who bought sugar also bought long-life package milk. The averages of the confidences were significantly higher for the associations found in store A than to those for store B ($p < 0.01$). This may indicate that even the usual associations, though stable in both stores, have different intensities.

The second block presents non-usual associations. Products classified do not have a common use or application. In this group, the first association to be shown is between clothes conditioner and special rice, which was observed in 96% of valid cases in store A and in 27% in store B. In store A, as we can observe in Table 1, the mean confidence of the association was 34.10%, that is to say, on average, approximately 34% of the consumers that bought clothes conditioner also bought special rice.

The non-usual associations in store A were stable over the analyzed period (all of them presented percentages of occurrence higher than 95% of valid days and standard-deviation below 10). However, in store B, this occurrence was lower, implying non-stable associations. Clothes conditioner and special rice were present in 27% of valid days in store B, while detergent and long-life milk in 37%. The association between onion and cheese was the one that presented the largest percentage of occurrence in store B, 71%, nevertheless well below the 95% suggested for a stable association.

This element leads the researchers to suppose that the non-usual associations must be very peculiar to the store in which they were found, and may not be found at any other store. Now, the usual associations were already observed in both stores and may repeat themselves for the same culture (stores of the same area, for example).

The confidences of associations are also higher in store A than in store B ($p < 0.01$). Even with a smaller percentage of occurrence in store B, the confidence of associations was high. Therefore, there are indications that the usual associations tend to repeat within the same group of consumers, or consumers with a similar profile, while the non-usual associations should be relatively specific to the store, following the same degree of consumers specificity.

As for the confidence of usual and non-usual associations, no significant differences were observed between the associations, which presented a wide range of values. Therefore, on average, 52.27% of consumers that buy ham also buy bread in store B, and 52.57% of consumers that buy 1 kg powdered detergent also buy long-life milk in store A. The ranges of confidence as a classificatory criterion for the associations is probably not the best way to visualize them.

VI. CONCLUSION

In the retail systems, the data base analysis is a very important marketing strategy instrument. Today, the retailer has all this incredibly useful information but he doesn't know what to do with it. Our methodology to extract information from the enormous data base is the principal contribution of this paper. It makes it possible to extract the information from the data base and help the decider to take the better decision. Strategy adopted enable decision maker to access historical and up-to-date association rules extracted from their, usually discarded, transaction data. Extracted rules are kept in a rules base, enabling flexible queries that meets users needs. We have shown one possible analysis when our strategy along rules base creation were used.

The managerial implications of mastering the formation of baskets of products are important for retailers in a general way. For example, from our extracted results would be possible to establish a smaller margin on rice considering the percentage of consumers that also buy clothes conditioners, with a higher margin.

However, it is currently extremely difficult to know if consumers attracted by special offers are profitable, that is, if they really buy products with higher margins, or if they simply buy only the products that are on special offer, pressing the margins down. Knowing these baskets would allow the retailer to plan these types of promotions more carefully, ensuring safer the commercial returns for the retailer.

It was not addressed the quantity involved in Association Rules, what would lead to a more precise information, though harder to extract and to choose parameters threshold. This work can be further enriched by linking customer profiling to rules extracted, thus allowing better targeted promotions with even higher profits.

ACKNOWLEDGMENT

We wish to thank Mark Fielding for his help during english review stage.

REFERENCES

- [1] Agrawal, R.; Imielinski, T., Swami, A. Database Mining : A performance perspective. IEEE transactions on Knowledge and Data Engineering, vol. 5 n° 6 December 1993.
- [2] Borges, A.A Jr. A formação de cestas de produtos em situações de compras repetidas: um estudo de caso em uma rede de supermercados de Porto Alegre. Master Thesis. UFRGS-RS. Porto Alegre, 1998.
- [3] Brachman, R. J., Anand, T. The Process of Knowledge Discovery in Databases: A Human-Centered Approach,

Em : Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

- [4] Cabena, P. et al. *Discovering data mining: from concept to implementation*. Prentice Hall, 1998.
- [5] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. *From Data Mining to Knowledge Discovery : An Overview*. Em : Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.
- [6] Han, J.; Cai, Y.; Cercone, M. *Data-Driven Discovery on Quantitative Rules in Relational Databases*. *IEEE transactions on Knowledge and Data Engineering*, vol. 5, No. 1, February 1993.
- [7] Mannila, H.; Toivonen, H. ; Verkamo, I. *Improved Methods for Finding Association Rules*. Internal Paper. Helsinki University. Finland. 1994
- [8] Srikant, R; Agrawal, R. *Mining Quantitative Association Rules in Large Relational Tables*. Proc. of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996.