Temporal Visual Profiling of Market Basket Analysis

Francisco Javier Moreno Arboleda, Giovanni Pérez Ortega, Jaime Alberto Guzmán Luna

***Abstract*—Market basket analysis allows analysts to understand the behaviour of customers. In this paper, we propose a novel technique to generate the profile of a customer with regard to his/her product purchase history. After obtaining the profile of a customer, we present a visual technique, to compare in a friendly and interactive way the profiles of customers over time. To show the expediency of our proposal, we performed experiments with a dataset of purchases of a retail market. The results showed that our proposal can be useful for stock planning and identifying customers with similar buying tendencies.**

***Index Terms*— Market basket analysis, customer profiling, product classification, association rules, visualization.**

I. Introduction

MARKET basket analysis (MBA) is a key data mining technique for discovering associations between products purchased by customers [1]. These associations are usually expressed through association rules [2]. Informally, an association rule, specifies the probability that the occurrence of one set of items (the antecedent) implies the occurrence of another set of items (the consequent). For example, if a customer purchases bread and milk, what is the probability that his/her purchase includes jam. To identify the most relevant association rules, measures such as confidence, support, and lift [2], [3] are used.

However, MBA goes beyond obtaining association rules. Through MBA, it is possible to analyze the behavior of the customers, e.g., what products or categories of products they prefer, what payment methods they use, how often they purchase, and how these aspects evolve over time. This brings us to *customer profiling*. Customer profiling allows businesses to understand their customers. Customer profiles describe customers based on a set of attributes [4]. For example, customer profiling might consider the following aspects: geographic, demographic, behavioral (e.g, attitude toward a product), psychographic (e.g., interests, and opinions of customers), economic (profit, e.g, using the recency, frequency, and monetary (RFM) model [5]), or a combination of these ones, among others.

In this paper, we present a method to obtain the profile of a customer regarding his/her *product purchase history*, i.e., characterize a customer according to the products or categories of products he/she purchases. For this, we consider for each product that the customer purchased: the total number of units purchased of the product and the total number of his/her purchases in which the product was included. Consider the following example. Suppose that the total number of purchases of a customer during a specific period of time was 20 and that the total number of apples he/she purchased was 100. Now, suppose that the customer purchased the 100 apples: i) in one single purchase and ii) in the 20 purchases, i.e., on average 5 apples in each purchase. Case i) suggests that the customer *is not* a frequent buyer of apples while in case ii) he/she *is*. We consider these aspects when analyzing how a product characterizes a customer according to their purchases, see Section III. After obtaining the profile of a customer, we present a visual technique to compare the profiles of customers *over time* (e.g., compare the profiles of two customers month by month, during a year). This allows analysts, e.g., detecting customers with similar profiles and can also help manage inventory, fine-tune marketing campaigns, design promotions, and forecast sales for certain periods of time.

Considering that most of the works: i) in MBA focus on association rules, see Section II, and ii) in customer profiling (based on product transaction history) focus on classifying customer on specific segments ("binning"), see works at the end of Section II; we believe that our proposal is a contribution to the identification, visualization, and comparison of customers profiles *over time*.

The paper is organized as follows. In section II, we review related works. In section III, we present our proposal for profiling a customer. In section IV, we present our experiments and discuss the results. Finally, in section V, we conclude and consider further work.

II. Related Works

In [6] a system that recommends a set of complementary products is proposed to complete the purchase of a customer. The system applies case-based reasoning. Thus, if a customer is making a purchase and has chosen, e.g., a pair of shoes, the system searches for past purchases (i.e., similar cases) where this product has been included and retrieves the additional products that have been included in such purchases, e.g., polish and laces, to generate recommendations.

In [7] a method is proposed to determine the customers' purchasing behavior and thus satisfy the unknown

Manuscript received May 3, 2021; revised December 14, 2021.

Francisco Javier Moreno Arboleda is an associate professor at the Departamento de Ciencias de la Computación y de la Decisión, Universidad Nacional de Colombia, Sede Medellín, Colombia (phone: 604-425-5376; e-mail: fjmoreno@unal.edu.co).

Giovanni Pérez Ortega is an associate professor at the Departamento de la Ingeniería de la Organización, Universidad Nacional de Colombia, Sede Medellín, Colombia (e-mail: gperezo@unal.edu.co).

Jaime Alberto Guzmán Luna is an associate professor at the Departamento de Ciencias de la Computación y de la Decisión, Universidad Nacional de Colombia, Sede Medellín, Colombia (e-mail: jaguzman@unal.edu.co).

expectations of the customers. For this, association rules are generated with the Apriori algorithm. The authors state that the results can help design stores, e.g., placing in the same section, products that are usually purchased together.

In [8] the authors state that most of the algorithms for MBA *do not* consider the changes that association rules may undergo due to future purchases. Therefore, they propose an algorithm called ARM (Association Rule Miner) Predictor, based on periodic mining, i.e., performing data mining after certain periods of time. Association rules are classified as atypical and suggestions. As the data increases, a rule could change of classification.

In [9] a desktop-application to forecast the purchases of customers is proposed. The Apriori algorithm is applied and the association rules are found. The authors state that the association rules can help design promotions and manage inventory.

In [10] a recommendation system is proposed that uses data mining techniques, collaborative filtering, and content-based filtering. This system characterizes customers by a value called *lifetime* that is obtained from the RFM model (recency, frequency, and monetary), where recency represents the time of the last purchase, frequency the number of purchases in a period of time, and monetary the amount of money that the customer spent in that time. Clusters with customers with similar lifetimes are formed using the *k*-means algorithm. These clusters are then separated into time-ranges to get more precise recommendations.

In [11] a technique is proposed to know how much inventory to stock in a commercial establishment according to the historical demands of customers. The Apriori algorithm is used to generate the association rules. Additionally, an artificial neural network is applied to predict the demand of the products.

In [12] the authors consider a database with m transactions (purchases) and n products. A transaction t is represented by a vector of n positions (products), where each position has a 1 or a 0 depending on whether the product was purchased or not in t. The authors calculate a matrix of similarity *sim* of order $n \times n$, where $sim(i, j)$, represents the probability that the products i and j are in the same transaction. Association rules are obtained with the Apriori algorithm. To generate a recommendation to a transaction t (that is in process): i) each of its products is selected and the most similar products are searched in *sim* and ii) the association rules related to the similar products of t are selected.

In [13] the authors consider p products. A transaction t is represented by a vector of length $p + 1$ where the first position is the Id of t and the other positions will have a 1 or 0 depending on whether the product was purchased or not in t. A network is created, its nodes represent transactions and products. A product p is connected to a transaction t if p was purchased in t. Subsequently, this network is transformed into a *co-purchased product networ*k, where the nodes are the products and the weight of the connections between them represents in how many transactions they were purchased together. Finally, an algorithm is applied to detect *overlapping communitie*s and thus find groups of products strongly connected.

With regard to MBA visualization, in [14] a method to analyze and visualize business data is proposed. A data set of products is taken (it includes data such as product name, unit price, units in stock, and supplier) and visualization techniques are applied, e.g., bar graphs that allow comparing the units in stock and the unit prices.

A feature that stands out in MBA is the visualization of association rules. For this, three techniques are usually applied: i) scatter plots, ii) matrix visualization, and iii) graphs. An analysis of these techniques is presented in [15] and [16].

For example, in [17] an application called GraphMiner is proposed. The antecedent and the consequent of each rule are represented by circles. The two circles are connected by an arrow. The higher the support, the thicker the arrow, and the higher the confidence, the darker the arrow. Furthermore, unlike other similar proposals [15], [16], the author states that his proposal allows selecting (filter) the association rules by keywords, by support, and by confidence.

In general, when working with huge volumes of data, analysts try to find the most relevant association rules, which is a time-consuming task. In [15] an R-extension package called *arules* is presented. It offers visualization techniques, such as scatter plots for association rules and frequent itemsets that help identify in an interactive way, the most relevant rules.

In [18] the authors present a visualization framework, called Scalable Association Rule Visualization (SARV). SARV is based on matrices and graphs. The matrix view is a $n \times n$, where n is the number of items. Each cell, corresponding to row r and column c, represents the support of the rules that have item r in their antecedent and item c in their consequent. Users can select items from the matrix view and the relationships among their corresponding rules are shown in a graph view.

In [19] a similar visualization technique is presented. They use an R-extension package called *arulesViz* [16]. A matrix of $n \times n$ is created along with a balloon plot with groups of items as columns and groups of items as rows. The size of the balloon in a cell represents the aggregated support of the both corresponding groups of items.

In [20] association rules are shown using a plot called SGI MineSets Rule Visualizer. A set of products is placed on each axis (x and y) and the relationship between the products of each axis generates association rules. They are represented with bars perpendicular to the plane, each with a ring around it. The height of the bars represents the confidence and the height of the ring represents the support.

With regard to customer profiling based *specifically on product transaction history*, we identified the following works.

In [21] the authors propose an approach for classifying customers in bins (i.e., "binning" or "bucketing", [22]) considering their total number of transactions and items purchased over a period of time (e.g., 30 months). An algorithm compute the number of bins which are labeled considering the number of items, e.g., “1-5”, “6-13”,”14-28”, and so on.

A similar work is [23] where customers of a mobile phone company are profiled through a sample survey and cluster analysis. Four customer groups were identified: “Functional people”, “Practical people”, “Techno-funs”, and “Matures

silent".

In [24] individual profiles from customers' transactional histories are created. The profile contains facts about a customer (demographic information) and rules describing his/her behavior, e.g., "*When purchasing bread, John Doe usually buys milk*" and "*On weekends, John Doe usually spends more than $80 on groceries*." Thus, they apply rule association discovery methods individually to each customer.

In [25] is presented a mining approach for finding rules between the customers (demographic characteristics) and the clusters of products they purchase in existing locations. These rules are intended to help in predicting sales for a new location. The customers with similar preferences are grouped into the same category, e.g., "*If Age = 35 and Gender = M then beer and cigarettes is common*", "*If Age = 15 and Gender = F then Ice-cream and chocolates is common*".

In [26] is proposed an extension of the RFM model called LRFMP (length, recency, frequency, monetary, and periodicity) for classifying customers in the grocery retail industry, where length is the time between the customer's first and last purchases and periodicity the standard deviation of the customer's inter-purchases times. They use data from a grocery chain in Turkey. Five customer groups were identified: "*high-contribution loyal customers*", "*low-contribution loyal customers*", "*uncertain customers*", "*high-spending lost customers*", and "*low-spending lost customers*". In a later work [27], a data mining approach for analyzing and profiling customers in video on demand services is proposed. For customer segmentation, the LRFMP model is employed alongside the *k*-means and Apriori algorithms to generate association rules between the identified customer groups and content genres. They use data from a digital broadcasting platform in Turkey. Four customer groups were identified: "*high consuming-valuable subscribers*", "*less consuming subscribers*", "*less consuming-loyal subscribers*", and "*disloyal subscribers*".

In Tables Ia, Ib, and Ic we summarize these works.

TABLE Ia.
SUMMARY OF RELATED WORKS FOR MBA

Ref	Technique/ algorithm	Software	Usefulness
6	Case-based reasoning		Generate purchases recommendations
7	Apriori	Mining software Tangara	Determine patterns of purchasing behavior
8	Periodic mining, Apriori	ARM-Predictor	Generate association rules considering purchases development
9	Apriori		Forecast purchases
10	Data mining, collaborative filtering, content-based filtering, Apriori, *k*-means		Generate purchases recommendations
11	Back-propagation artificial neural network, Apriori		Manage stock
12	Apriori	Integration Scheme	Generate purchases recommendations
13	Graph mining techniques, algorithm for overlapping community's detection		Identify groups of products strongly connected.

TABLE Ib.
SUMMARY OF RELATED WORKS FOR ASSOCIATION RULES VISUALIZATION

Ref.	Software	Visualization technique
14	Classification, clustering, and Apriori	Bar graphs
15	arules (an R extension package)	Scatter plot, two-key plot
17	GraphMiner	Graph
18	SARV	Matrix, graph
19	arules, arulesViz	Matrix, balloon plot
20	SGI MineSets Rule Visualizer	Three-dimensional plot

TABLE Ic.
SUMMARY OF RELATED WORKS FOR CUSTOMER PROFILING BASED ON PRODUCT TRANSACTION HISTORY

Ref.	Technique/algorithm	Profile technique
21	An ad-hoc algorithm	Binning
23	Sample survey, cluster analysis	Binning
24	Apriori	Demographic facts and individual association rules
25	Apriori	Binning
26	LRFMP and Apriori	Binning
27	LRFMP, *k*-means, and Apriori	Binning

III. PROFILING A CUSTOMER

Next, we present a method to get the profile of a customer with regard to their product purchase history.

A tree [28] is a set of nodes that represents a hierarchical parent-child relationship with the following properties:

1) Each node in the tree has a single parent node, except for one node, called *root*, which has no parent.
2) When a node has no children, it is called a *leaf* and when it has at least a child it is called a *branch*.
3) Due to the hierarchical relationship, the parent of a node cannot be any of its descendants (children, grandchildren, etc.).

In our case, we define a *Product Classification Tree* (PCT) for the classification of products.

Let $C = \{c_1, c_2, ..., c_n\}$ be a set of branches of the PCT, where c_i represents a category of products. $c_i = (c_{id}, c_{name}, c_{percentage}, c_{parent})$, where $c_{id} \in I^+$ is the unique identifier of each category, c_{name} (a string) is the name (also unique) of the category, $c_{percentage} \in R^+$, $0 \leq c_{percentage} \leq 100$, represents the importance of the category (its calculation is explained below) to the customer (according to his/her purchases), and c_{parent} is the identifier (c_{id}) of a node $c_j \in C$ ($c_i \neq c_j$), i.e., c_j is the parent category of c_i (if c_i is the root then c_i has no parent). Example: $c_3 = (3, \text{Fruits}, 20, 1)$.

Let $P = \{p_1, p_2, ..., p_m\}$ be a set of leaves of the PCT, where p_i represents a product. $p_i = \{p_{id}, p_{name}, p_{weight}, p_{percentage}, p_{parent}\}$, where $p_{id} \in I^+$ is the unique identifier of each product,

p_{name} (a string) is the name of the product (also unique), $p_{weight} \in I^+$ and $p_{percentage} \in R^+$, $0 \leq p_{percentage} \leq 100$, (their calculations are explained below) represent the importance of the product to the customer (according to his/her purchases), and p_{parent} is the unique identifier (c_{id}) of a node $c_i \in C$, i.e., the category (parent) to which p_i belongs. Example: p_8 = (8, Apple, 100, 15, 3).

In a PCT, the intersection of the unique identifiers of the categories with the unique identifiers of the products is empty.

We show an example of a PCT in Fig. 1.

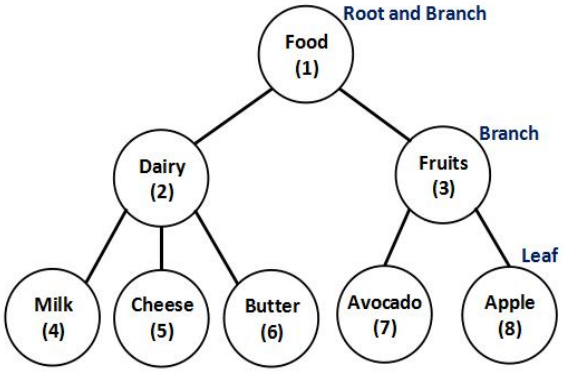

Fig. 1. Example of a PCT. Only the names and the unique identifiers (in parentheses) of each node are shown

Let $T = \{t_1, t_2, ..., t_k\}$ be a set of transactions (purchases) of a customer. $T_i = (t_{id}, D)$, where $t_{id} \in I^+$ is the unique identifier of the transaction and $D = \{d_1, d_2, ..., d_q\}$ is a set of details. $d_i = (d_{product}, d_{num\_uni})$, where $d_{product}$ is the unique identifier (p_{id}) of a product $p_i \in P$, and $d_{num\_uni} \in I^+$ is the number of units purchased of the product p_i in d_i. There can not be in D two details d_i, d_j, $i \neq j$, with the same $d_{product}$, i.e., there can not be two details in the same purchase with the same product. Example: $T = \{t_1, t_2, t_3, t_4\}$, where:

t_1 = (1, {(4, 2), (5, 5), (6, 2)}),
t_2 = (2, {(4, 3), (5, 5), (7, 3)}),
t_3 = (3, {(4, 2)}), and
t_4 = (4, {(6, 1), (8, 7)}).

p_{weight} and $p_{percentage}$ calculation

Attributes p_{weight} and $p_{percentage}$ indicate the importance of a product p to a customer based on his/her purchase history. For computing these two attributes, we propose two options.

a) **Regularity**: it considers the frequency with which the customer purchases the product p. Thus, the greater the number of purchases in which the customer includes *at least* one unit of p, the greater the value of p_{weight} and $p_{percentage}$. This option also considers the total number of purchased units of p, see (1).

b) **Amount**: it considers only the total number of units purchased of the product p, regardless of in how many purchases it is included. Thus, the greater the number of units purchased of p, the greater the value of p_{weight} and $p_{percentage}$, see (1).

The p_{weight} of each product (leaf), $p_i \in P$, is calculated as follows. Let be:

- T a set of transactions of a customer.
- *num_trans* the total number of transactions of the customer, i.e., $|T|$.
- *trans_by_product* the total number of transactions of T in which the product p_i is included (p_{id} of p_i).
- *total_num_uni* the total number of units purchased from the product p_i in T; for this, we consider all the transactions in T which include the p_{id} of p_i and we get the sum of d_{num_uni}.

Then, we apply (1):

a) Regularity:

$$p_{weight} = total\_num\_uni * trans\_by\_product/num\_trans \quad (1)$$

b) Amount:

$$p_{weight} = total\_num\_uni$$

For option a) (regularity), the rationality of the equation is as follows. Suppose a customer purchased 100 apples and has 20 purchases in total.

- If he/she purchased the 100 apples in one single purchase, then he/she is *not* a frequent buyer of apples and p_{weight} = 100 * 1/20 = 5.
- If he/she purchased the 100 apples in two purchases, then p_{weight} = 100 * 2/20 = 10.
- If he/she purchased the 100 apples in the 20 purchases (e.g., 5 in each purchase), then p_{weight} = 100 * 20/20 = 100.

Thus, the equation penalizes: the lower the number of purchases in which the customer includes the product, the lower the value of p_{weight}. Note that option b) is equivalent to option a) when *num_trans* = *trans_by_product*.

The results of applying (1) considering the set of transactons T of the previous example with both options, are shown in Table II.

TABLE II
p_{weight} BY REGULARITY AND BY AMOUNT

	a) **Regularity**	b) **Amount**
p_{id}	p_{weight}	p_{weight}
4	5	7
5	5	10
6	2	3
7	1	3
8	2	7

After calculating the p_{weight} of each product (leaf), its $p_{percentage}$ is calculated as follows. Let *sum_weights* be the sum of the weights (p_{weight}) of the leaves, i.e., the set P. Then, we apply (2) to each leaf:

$$p_{percentage} = p_{weight} * 100/sum\_weights \quad (2)$$

Thus, $p_{percentage}$ represents the percentage of the node's weight with regard to the sum of the weights of all the leaves.

In Table III we show the results with both options, a) and b), for our current example.

Note that regardless of the option (regularity or amount), the sum of the percentages ($p_{percentage}$) of the leaves (products) is 100.

TABLE III
$p_{percentage}$ BY REGULARITY AND AMOUNT

	a) Regularity	b) Amount
p_{id}	$p_{percentage}$	$p_{percentage}$
4 (Milk)	33,33	23,33
5 (Cheese)	33,33	33,33
6 (Butter)	13,33	10
7 (Avocado)	6,67	10
8 (Apple)	13,33	23,33

The results, Table II and Table III, show the effect of profiling the customer by regularity and by amount. For example, for the products with p_{id} = 4 and p_{id} = 8, the customer purchased the same number of units (7) but in a different number of transactions (3 for product 4 and 1 for product 8). Thus, in option a) (regularity), the product with p_{id} = 4 gets a p_{weight} = 5 and the product with p_{id} = 8 gets a p_{weight} = 2. The product with p_{id} = 8 got a lower weight because this product is included *only* in one purchase. On the other hand, in option b) (amount), where the amount is considered (but not the number of transactions in which the product is included) the two products get the same weight (p_{weight} = 7).

Now, to calculate the percentage of a branch c, i.e., $c_{percentage}$, we proceed as follows. Let $Prchildrenc$ = {pr_1, pr_2, ..., pr_k} be the set of the percentages ($p_{percentage}$) of all the k children of c. Then, we apply (3):

$$c_{percentage} = \sum pr_j,\ 1 \le j \le k,\ pr_j \in Prchildrenc \qquad (3)$$

As a result of these calculations, we get the PCT of Fig. 2 (where the customer is profiled by amount).

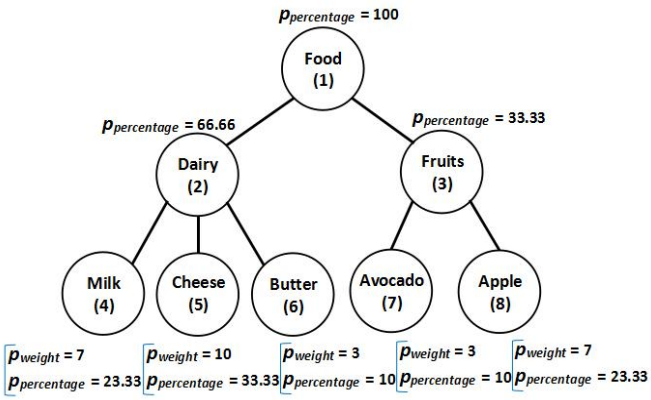

Fig. 2. Example of a PCT

For example, for Fruits (c_{id} = 3), $c_{percentage}$ = 10 + 23.33 = 33.33. Next, we present an algorithm, *customerProfiling*, which calculates p_{weight} and $p_{percentage}$ for each leaf and $c_{percentage}$ for each branch of a PCT based on the transactions (T) of a customer.

```
Algorithm customerProfiling(PCT, T, profileOption)
Input:
PCT: Product Classification Tree.
T: set of transactions (purchases) of a customer.
profileOption: 0 profiling by regularity, else profiling by amount.
Output:
PCT with pweight and Ppercentage calculated for each leaf and cpercentage for each branch.
BEGIN
1. num_trans = |T|; //Total number of transactions of a customer.
2. sum_weights = 0; //Sum of the weights of the leaves.
3. P = PCT.getLeaves(); //Get the set of leaves (products) of the PCT.
   //Calculate pweight of each leaf:
4. FOREACH p ∈ P
5.    pid = p.getId(); //Get the unique identifier of p.
6.    trans_by_product = T.getTransbyProduct(pid); //Get the total number of transactions in which p is included in T.
7.    total_num_uni = T.getTotalNumUni(pid); //Get the total number of units purchased of p in T.
8.    IF (profileOption = 0) //Profiling by regularity.
9.       pweight = total_num_uni * (trans_by_product/num_trans);
10.   ELSE //Profiling by amount.
11.      pweight = total_num_uni;
12.   END IF
13.   PCT.updatePweight(pid, pweight); //Assign pweight to the leaf (p).
14.   sum_weights = sum_weights + pweight;
15. END FOREACH
   //Calculate Ppercentage of each leaf, call function percentageLeaves, see below.
16. percentageLeaves(PCT, P, sum_weights);
17. //Calculate cpercentage of each branch, call function percentageBranches, see below.
18. percentageBranches(PCT, P);
END customerProfiling
Function percentageLeaves(PCT, P, sum_weights)
BEGIN
1. FOREACH p ∈ P
2.    pid = p.getId(); //Get the unique identifier of p.
3.    ppercentage = p.getPweight(pid) * 100/sum_weights;
4.    PCT.updatePpercentage(pid, ppercentage); //Assign ppercentage to the leaf (p).
5. END FOREACH
END percentageLeaves
Function percentageBranches(PCT, P)
BEGIN
1. C = PCT.getBranches(); //Get the set of branches (categories) of the PCT.
2. FOREACH c ∈ C //Branches are taken in a botton-up way, i.e., from the penultimate level (the level of the branches immediately above the leaves) to the root level.
3. cid = c.getId(); //Get the unique identifier of c.
4. sumPercentagesChidren = PCT.getSumPpercentagesChildren(cid); //Sum of the percentages of the children of c.
5. PCT.updateCpercentage(cid, sumPercentagesChidren); //Assign sumPercentagesChidren, i.e., Cpercentage, to the branch c.
5. END FOREACH
END percentageBranches
```

IV. Experiments

For the experiments, we used a dataset of purchases of a retail market [29]. The dataset includes 387875 purchases

from 312 customers during one year (2011) and covers 3665 different products. The dataset was filtered as follows. We remove purchases of: i) customers with less than 10 purchases during the year and ii) products that had less than 20 purchases during the year. After this filtering, the dataset was left with 113099 purchases, 281 customers, and 1550 different products. The 1550 products were manually classified into three categories: Decoration (D), Toys (T), and Utensils (U); and into five subcategories:

- In category D: Home (HO), Personal (PE), and Party (PA).
- In category U: Kitchenware (K) and Hardware (HA).
- Category T has no subcategories.

Each customer was profiled in each month of the year in the two options (by regularity and by amount). Thus, e.g., for a customer, all his/her January 2011 purchases were considered and the *customerProfiling* algorithm (with the two profiling options) was applied. Then, all his/her February 2011 purchases were considered and the algorithm was again applied. The same process was performed for the other ten months of 2011. Fig. 3 shows the January 2011 purchases for the customer with id = 15311. The colors represent the days the customer made purchases. The size of each bar is related to the number of units he/she bought of each product.

Fig. 4 shows the PCTs for the customer with id = 15311 during 2011, profiled by amount. All experiments were conducted on a Intel(R) Core (TM) i3/2.4 GHz processor with 8GB of RAM. The average time for computing a PCT (for a month) was 2.5 minutes. Data were stored on a Oracle database and processed with PL/SQL

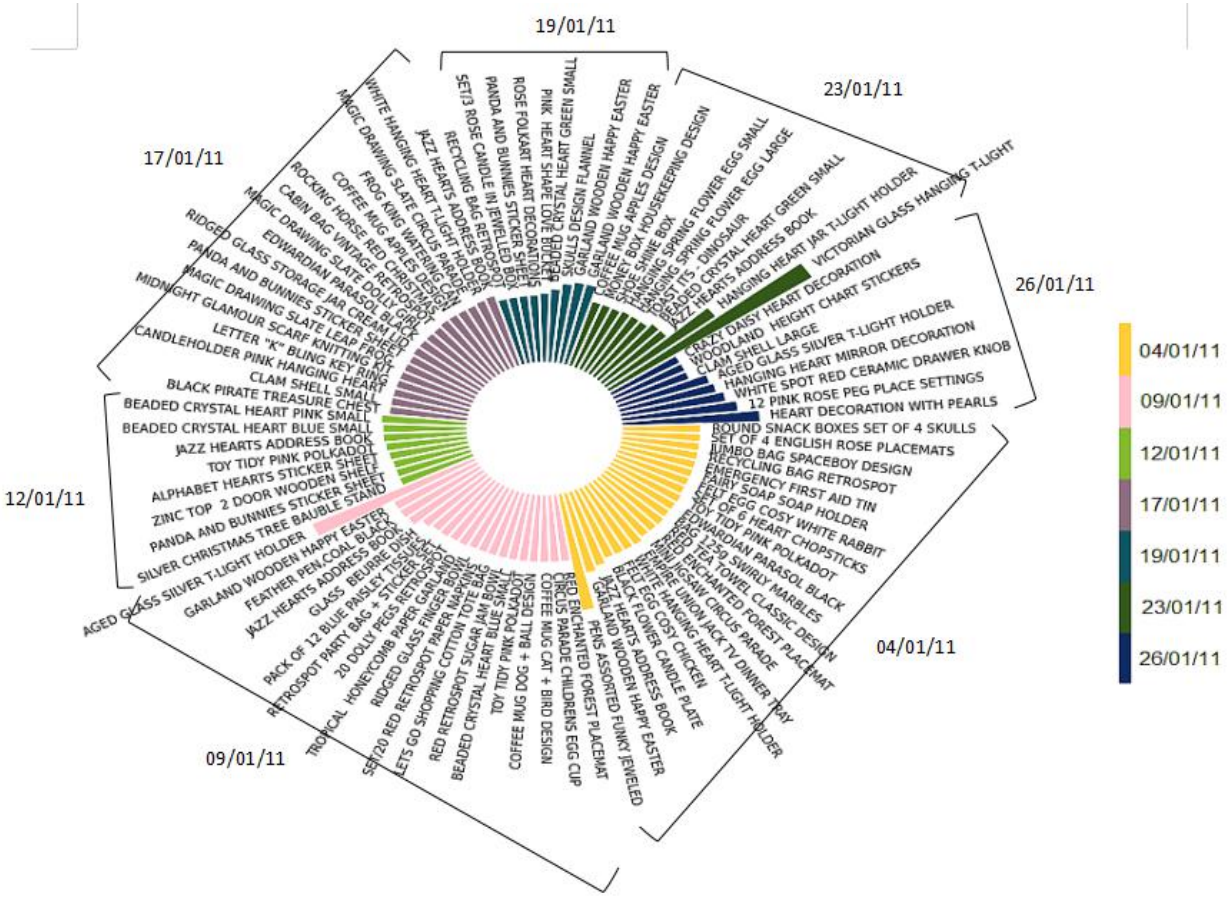

Fig. 3. January 2011 purchases for the customer with id = 15311.

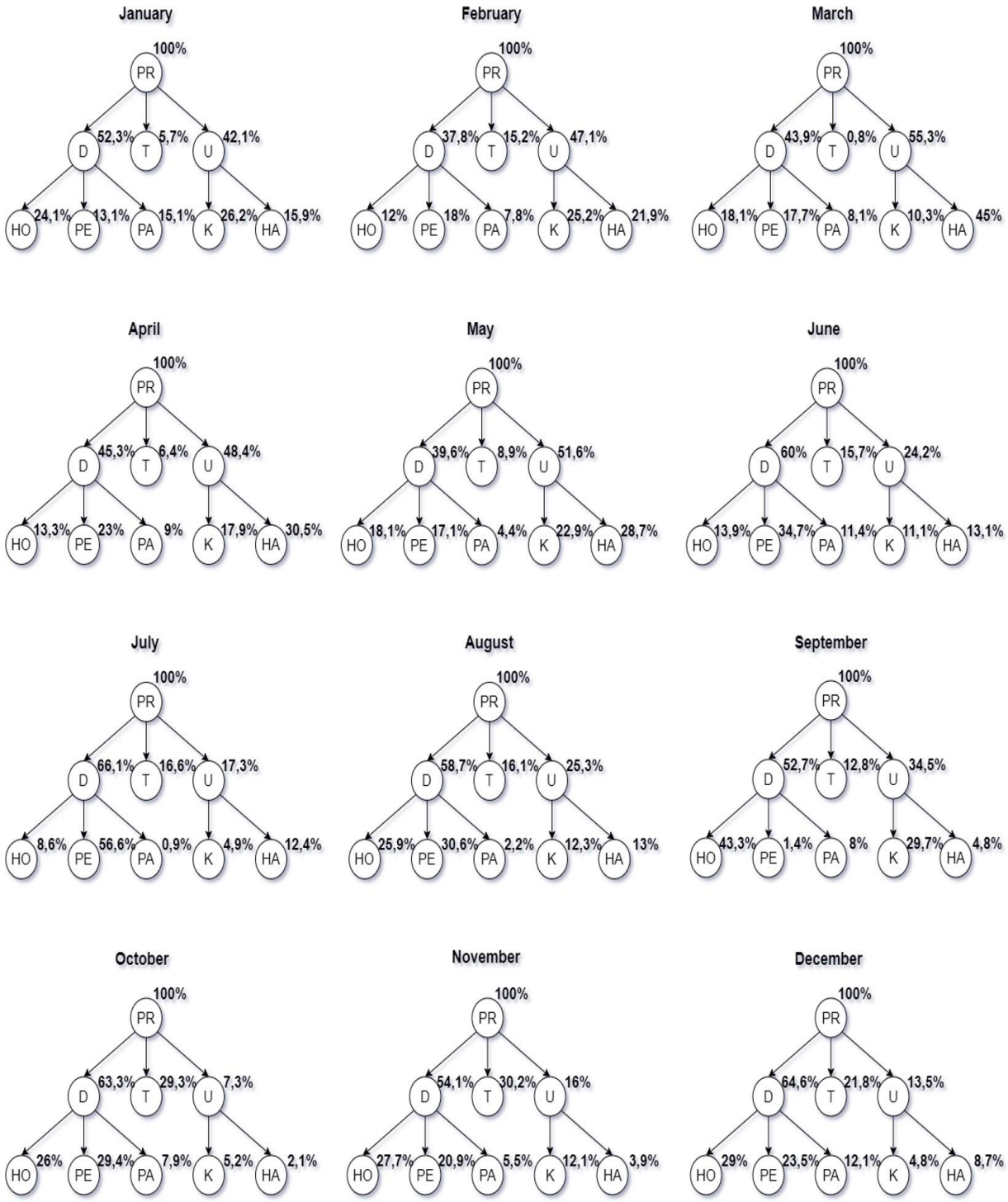

Fig. 4. PCTs for customer 15311, profiled by amount.
Conventions: PR = Products, D = Decoration, T = Toys, U = Utensils, HO = Home, PE = Personal, PA = Party, K = Kitchenware, and HA = Hardware

Note that in the PCTs of customer 15311: i) there is an increase in purchases of category T in the last quarter of the year when festivities such as Halloween and Christmas take place and ii) category D is the one that he/she purchases the most throughout the year (it was the highest in 8 of the 12 months).

We also obtained the PCTs for the same customer, but profiled by regularity. The results are shown in Fig. 5.

In the PCTs of Fig. 5, the two trends mentioned for the PCTs of Fig. 4 persisted (here, category D was the highest in 9 of the 12 months).

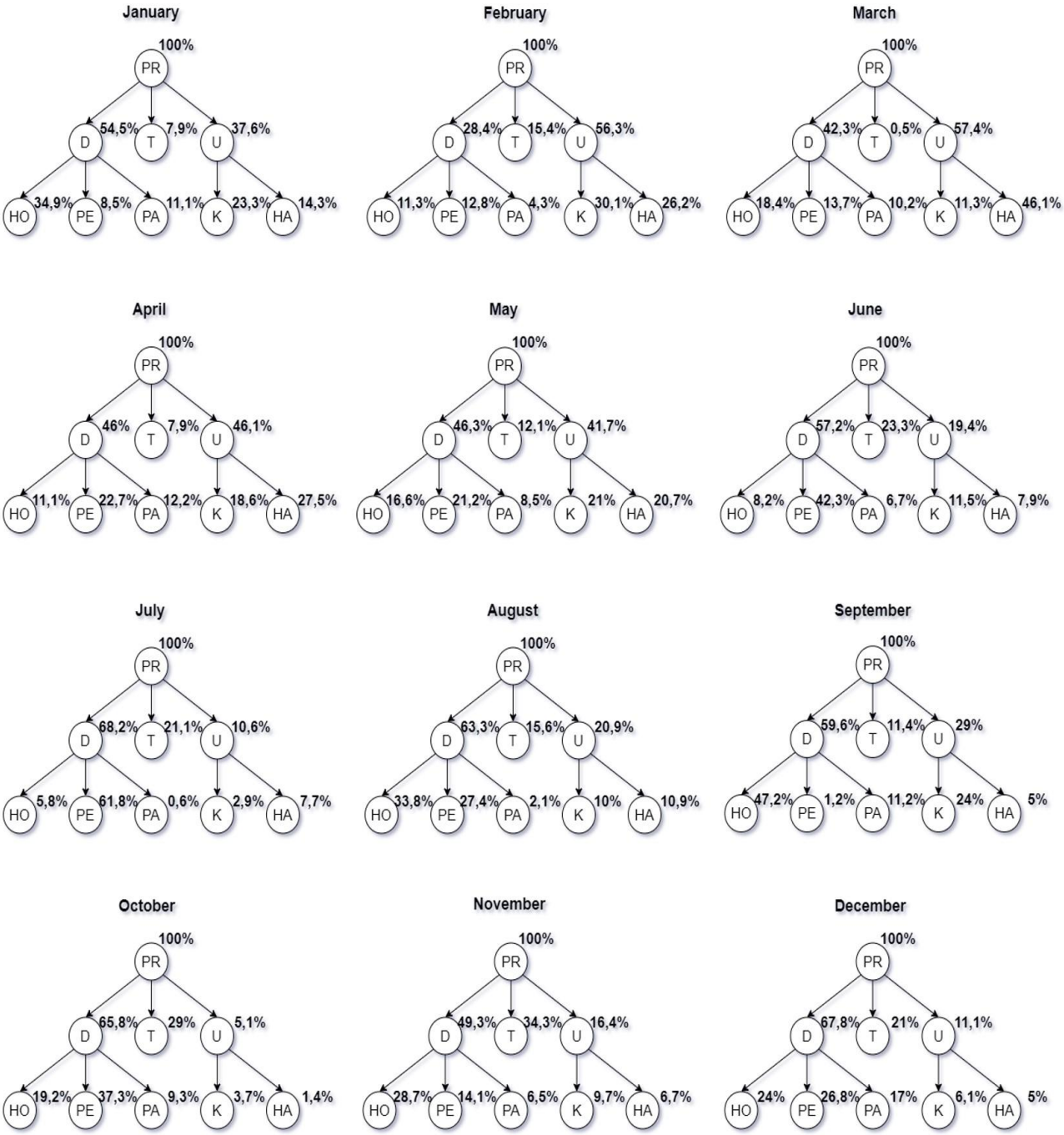

Fig. 5. PCTs for customer 15311, profiled by regularity.
Conventions: PR = Products, D = Decoration, T = Toys, U = Utensils, HO = Home, PE = Personal, PA = Party, K = Kitchenware, and HA = Hardware

On the other hand, when comparing the PCTs of the two profiling options for this customer, we note that: if we consider any node of a PCT in Fig. 4 and consider the corresponding node in Fig. 5, the maximum variation was 10.8%: in January in the profiling by amount, the HO subcategory scored 24.1% compared to 34.9% in the profiling by regularity. This suggests that this customer prefers to make purchases from this subcategory *throughout* the month.

We also got the PCTs (for the two profiling options) for the customer with id = 14606 and compared with the corresponding PCTs of customer 15311. Thus, our proposal allows analysts to compare customer profiles. For the comparison, we define a color system for the nodes:

- A green node means that there was no variation (less than or equal to a given threshold, e.g., our threshold was 10%) between the corresponding nodes of the customers' PCTs, i.e., they are considered "similar".
- A red node means that there was a variation (greater than the threshold 10%) between the corresponding nodes of the customers' PCTs, i.e., they are considered "not similar".

Thus, *the greener nodes there are, the more similar are the profiles (PCTs) of the two customers*. In Fig. 6 we show the results of the January PCTs comparison for the two customers (by amount). In Fig. 7 we show the total number of units purchased by each customer on each day of January 2011. In figures 8 and 9 we show the results for the 12 months

for the two profiling options.

The following is highlighted:

- In May, in profiling by regularity, the two customers had a very similar behavior (7 green nodes out of 9). This was the month when their PCTs were most similar. A similar behavior ocurred in November in the profiling by amount (7 green nodes out of 9).
- April and September, in profiling by regularity, were the two months in which the two customers were least similar (6 red nodes out of 9). A similar behavior occurred in February and March in the profiling by amount (6 red nodes out of 9).
- In November, the leaves were mostly green (in both profiling options). Furthermore, note that the similarity of the parent nodes has no relation to the similarity of the leaves.
- In September, the leaves were mostly red (in both profiling options), but most of the branches were green. This shows that although customers 15311 and 14606 do not have similar purchases of specific products, they *do* purchase products of the same subcategories.

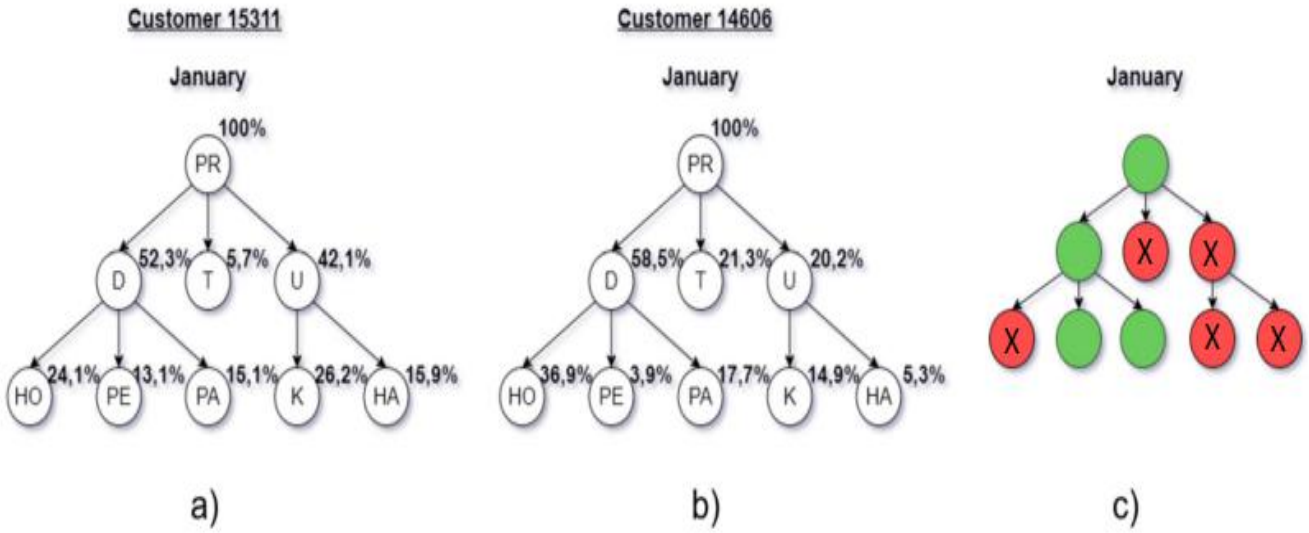

Fig. 6. Comparison of January PCTs of customers 15311 and 14606 (profiled by amount): a) PCT of customer 15311, b) PCT of customer 14606, and c) Resulting PCT (comparison): green represent similar nodes (variation ≤ 10%) and red (X circles) not similar ones.

Fig. 7. Total number of units purchased by customers 15311 and 14606 on each day of January 2011.

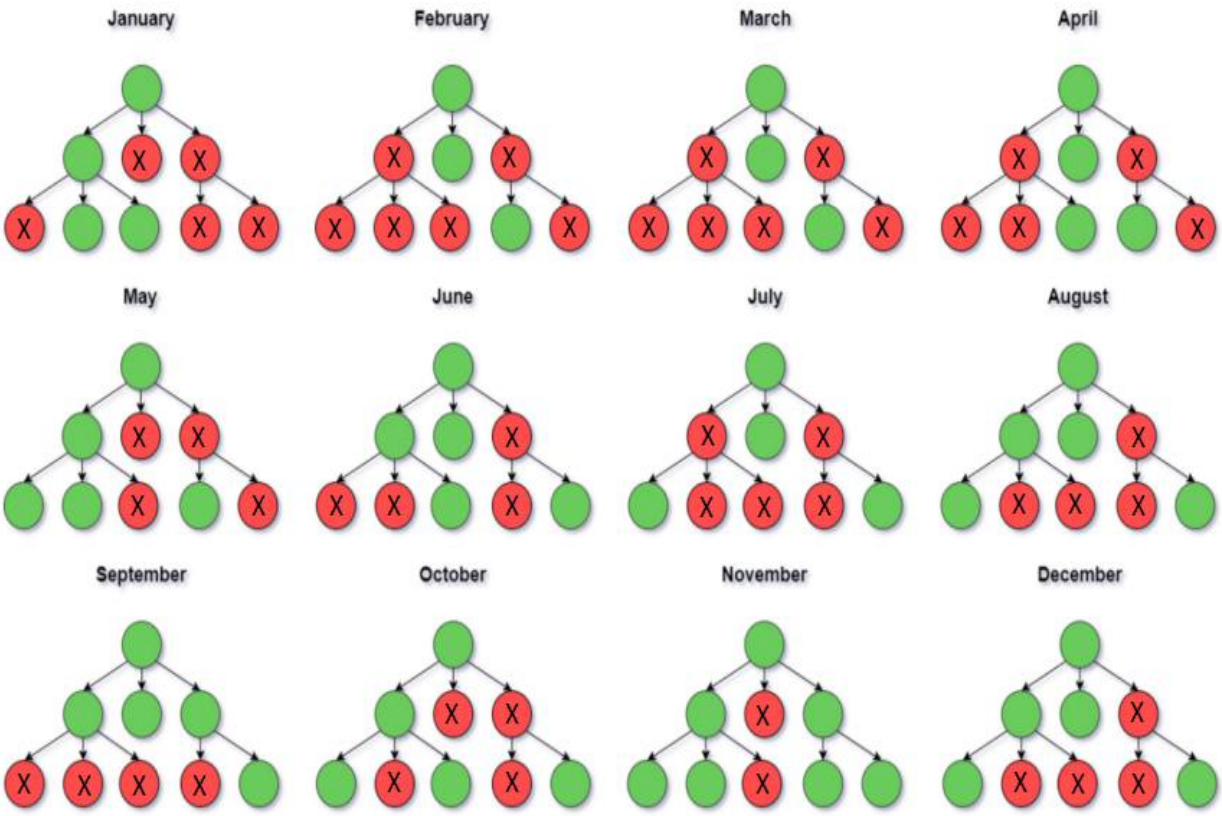

Fig. 8. Comparison of the PCTs of customers 15311 and 14606 (profiled by amount)

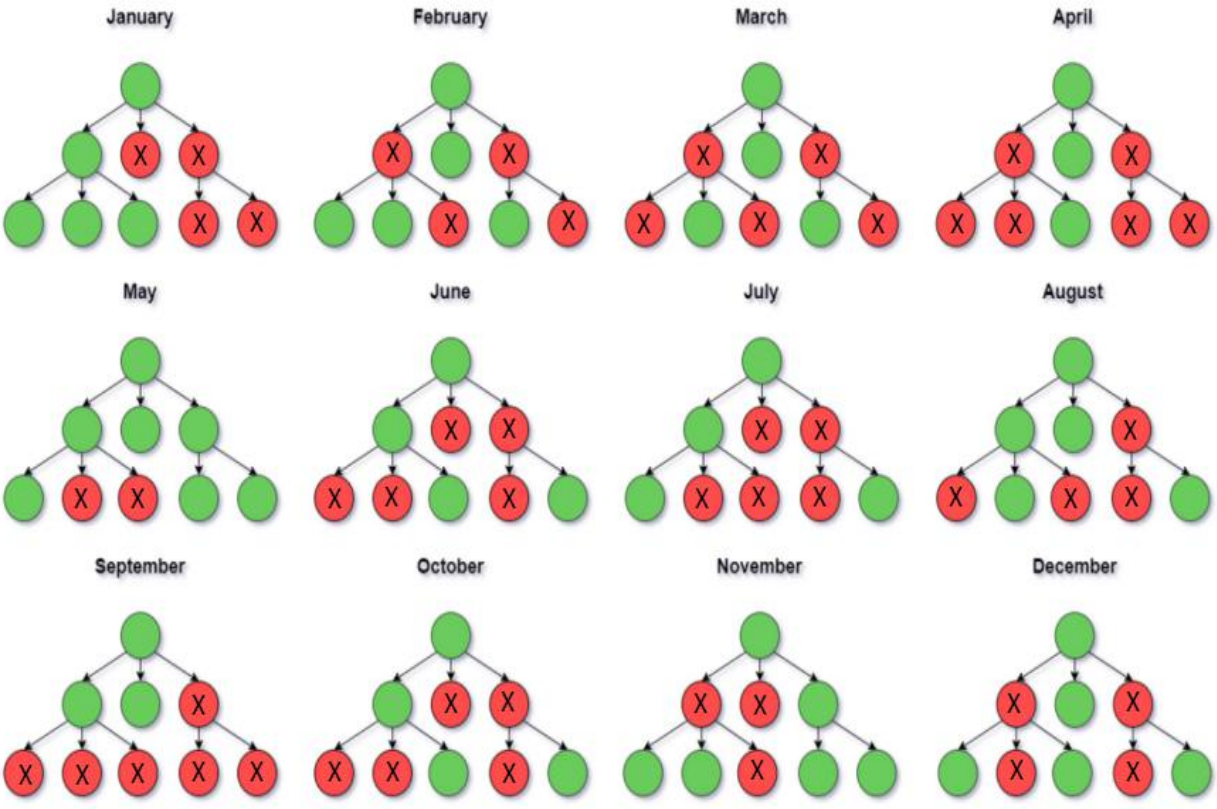

Fig. 9. Comparison of the PCTs of customers 15311 and 14606 (profiled by regularity)

In addition, to get more comprehensive results, 30 pairs of customers were randomly chosen to get indicators when comparing them. The results are summarized in Table IV.

TABLE IV.
RESULTS FOR THE PAIRS OF CUSTOMERS COMPARISON (30 PAIRS)

Pairs of customers who had more than three months in which all their PCTs had:	Number of pairs (by amount/by regularity)
All their nodes green.	4/2
At least two green nodes in the categories.	24/21
At least three green nodes in the subcategories	18/16
All their nodes green red.	2/1
At least two red nodes in the categories.	20/19
At least three red nodes in the subcategories	17/15

Next, we considered the 80 customers with more purchases during the year. Each of these customers had more than 30 purchases. Each customer was profiled in each month of the year in the two profiling options. Thus, we got 1920 PCTs because we have 80 customers, 12 months, and 2 profiling options (50 * 12 * 2 = 1920). Then, we found for each profiling option *the average of each node in each month* as follows: we got 960 PCTs (profiling by amount), i.e., 80 PCTs for each month (because we analyzed 80 customers). Thus, we averaged the 80 values (percentages) of each node considering the 80 PCTs. The same procedure was applied for the 960 PCTs (profiling by regularity). In this way, we obtained an "*average PCT*" *for each month in each profiling option*. That is, we got a "*representative PCT"* that summarizes the behaviour of the 80 customers with more purchases during the year in each month in each profiling option. The results are shown in tables V and VI.

The following is highlighted:

- In both profiling options, there was an increase in purchases of the categories D and T and the subcategory T in the last quarter of the year when festivities such as Halloween and Christmas take place.
- In both profiling options, categories D and U were the ones with more purchases throughout the year.
- In both profiling options, subcategories HO and PE were the most stable throughout the year. They were around 15% throughout the year in both options.
- In general, the results between the two profiling options showed a similar behaviour in all categories and subcategories. This suggests that the differences are especially noticeable with individual customers and small groups of customers (e.g., less than 20). With larger groups, 80 customers in the experiment, the differences between the two profiling options decrease.

Thus our proposal can be useful to identify, e.g., for individual customers or small groups of customers (e.g., less than 20) which products are sold: i) more by amount than by regularity, and thus have more units in stock of these products in certain months and ii) more by regularity than by amount, and thus have units of stock of them throughout the year or throughout several consecutive months.

In addition, we can also identify products that are usually sold together (they have a similar percentage and greater than a given threshold) in some consecutive months to design promotions that increase their purchases and store sufficient units of these products in certain months.

On the other hand, consider a new customer who begins his/her purchase history in the second semester of 2020. To forecast his/her purchases for the first semester of 2021, we could find similar customers to him/her in the second semester of 2019 and analyze how these customers behaved in the first semester of 2020. This could give us an idea of how the new customer will behave in the first semester of 2021. In a more general way, given a group of customers with a purchasing behavior in a specific period (e.g., the first quarter of a year), we can forecast the behavior of the group for the same period of the next year and try to apply marketing strategies to boost sales of certain products. Similarly, we can:

- Compare the corresponding periods of two years to determine the variation in purchases. Thus, we could determine if it is advisable to increase or decrease the units of stock of certain products in these periods.
- Identify groups of customers who change their preferences, e.g., from one year to the next, to adjust the inventory and meet customer expectations [30].

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we presented an algorithm to get the profile of a customer from their purchase history. The profile corresponds to a tree of products and product categories (a PCT) that characterize the customer *over time*. The PCT can be defined by the analyst (products and categories). In addition, we propose two profiling options: by regularity and by amount.

We also propose a visual technique to compare customer profiles. This allows analysts to observe in a friendly and interactive way the similarity between customers in a specific period of time (e.g., one day, one month, one semester) and over time according to these periods (i.e., in our proposal *it is possible to observe the evolution of their similarity*). We believe this is a contribution in the area of MBA since most works focus on the generation, analysis, and visualization of association rules (even in this approach, few works consider the evolution of rules). In MBA, time analysis is essential as it is unrealistic to think that customer behavior is static. Thus, our method considers this aspect, both in the generation of the profile and in the comparison of customers profiles.

On the other hand, from the results of our basic experiments, we can perceive the expediency of our proposal in fields such as inventory management, marketing, and sales forecasts, among others.

We plan to extend our profile comparison method for groups of customers. For example, given a group of customers, each with their monthly PCTs, get the monthly PCTs that characterize the group. This could generate indicators that represent the group's cohesion with regard to a given set of products and categories.

We also plan to involve in the PCTs generation other attributes related to customer purchases. For example, include the payment method, shipping methods, enviromental preferences (e.g., if the customer purchases remanufactured or new products [31]) and other ways to categorize the products. According to this last aspect, for a customer *different PCTs could be obtained for the same period of time*.

TABLE V.
AVERAGE PCT OF THE 80 CUSTOMERS WITH MORE PURCHASES (PROFILED BY AMOUNT)

Node	January	February	March	April	May	June	July	August	September	October	November	December
Products (PR)	100	100	100	100	100	100	100	100	100	100	100	100
⌊Decoration (D)	38.9	40.3	37.5	35	33.2	41	42.8	37.3	40.9	59.9	61	62.3
⌊Home (HO)	11.4	16.3	11.9	9.4	18	16.6	13.1	9.7	8.1	13.7	20.4	22.1
⌊Personal (PE)	17.6	15.1	18.1	16.6	9.9	13.3	19.9	14.5	18.3	7.9	10.8	16.9
⌊Party (PA)	9.9	8.9	7.5	9	5.3	11.1	9.8	13.1	14.5	37.3	29.8	31.6
⌊Toys (T)	15.4	11.1	4.5	3.8	5.5	8.2	8	7.9	17.8	28.8	30.3	23.3
⌊Utensils (U)	45.7	48.6	58	61.2	61.3	49	49.2	54.8	41.3	11.3	8.7	14.4
⌊Kitchenware (K)	18.5	17.7	29.8	29.2	31.2	27.7	33.1	32.2	21.6	8.5	6.1	7.1
⌊Hardware (HA)	27.2	30.9	28.2	32	30.1	21.3	16.1	22.6	19.7	2.8	2.6	7.3

TABLE VII.
AVERAGE PCT OF THE 80 CUSTOMERS WITH MORE PURCHASES (PROFILED BY REGULARITY)

Node	January	February	March	April	May	June	July	August	September	October	November	December
Products (PR)	100	100	100	100	100	100	100	100	100	100	100	100
⌊Decoration (D)	33.2	29.5	33.3	41.1	39.9	41.9	34.4	30.3	39.7	55.3	52.5	59.9
⌊Home (HO)	13.3	7.7	9.8	20.5	20.1	18.7	8.1	9.1	11.6	12.7	10.7	14.5
⌊Personal (PE)	9.7	11.3	14.5	13.4	11.2	12.1	14.2	10.7	10	13.2	15.4	15.5
⌊Party (PA)	10.2	10.5	9	7.2	8.6	11.1	12.1	10.5	18.1	29.4	26.4	29.9
⌊Toys (T)	10.1	9.5	5	4.3	3.8	5	6.5	6.3	10.2	25.1	28.3	19.5
⌊Utensils (U)	56.7	61	61.7	54.6	56.3	60.1	59.1	63.4	50.1	19.6	19.2	20.6
⌊Kitchenware (K)	23.7	27.8	30	26.6	25.5	29.7	32.8	34.5	29.8	11.5	8	7.6
⌊Hardware (HA)	33	33.2	31.7	28	30.8	30.4	26.3	28.9	20.3	8.1	11.2	13

REFERENCES

[1] S. Li, (2017, Sep 24). A gentle introduction on market basket analysis — association rules (Online). Available: *https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce*

[2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington D.C., 1993, pp. 207–216.

[3] S. Tufféry, *Data Mining and Statistics for Decision Making*. New York: John Wiley & Sons, 2011.

[4] F. Kushnazarov, "Consumer Life Cycle and Profiling: A Data Mining Perspective," in *Consumer Behavior and Marketing,* M. Reyes, Ed. London: IntechOpen, 2020, pp. 1–16.

[5] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: using iso-value curves for customer base analysis," *Journal of Marketing Research*, vol. 42, no. 4, pp. 415–430, 2005.

[6] A. Gatzioura and M. Sànchez-Marrè, "A case-based recommendation approach for market basket data," *IEEE Intelligent Systems*, vol. 30, no. 1, pp. 20–27, 2014.

[7] M. A. Alan and A. R. Ince, "Use of association rule mining within the framework of a customer-oriented approach," *European Scientific Journal*, vol. 12, no. 9, 2016.

[8] M. Kaur and S. Kang, "Market basket analysis: identify the changing trends of market data using association rule mining," *Procedia computer science*, vol. 85, pp. 78–85, 2016.

[9] F. Kurniawan, B. Umayah, J. Hammad, S. M. S. Nugroho, and M. Hariadi, "Market basket analysis to identify customer behaviours by way of transaction data," *Knowledge Engineering and Data Science*, vol. 1, no. 1, pp. 20–25, 2018.

[10] F. Rodrigues and B. Ferreira, "Product recommendation based on shared customer's behaviour," *Procedia Computer Science*, vol. 100, pp. 136–146, 2016.

[11] A. Mansur and T. Kuncoro, "Product inventory predictions at small medium enterprise using market basket analysis approach-neural networks," *Procedia Economics and Finance*, vol. 4, pp. 312–320, 2012.

[12] K. Tatiana and M. Mikhail, "Market basket analysis of heterogeneous data sources for recommendation system improvement," *Procedia Computer Science*, vol. 136, pp. 246–254, 2018.

[13] I. F. Videla-Cavieres and S. A. Ríos, "Extending market basket analysis with graph mining techniques: a real case," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 1928–1936, 2014.

[13] G. Ramesh, T. Rajinikanth, and D. Vasumathi, "Explorative data visualization using business intelligence and data mining techniques," *Int. J. Appl. Eng. Res.*, vol. 12, pp. 14008–14013, 2017.

[15] M. Hahsler, "arulesViz: interactive visualization of association rules with R," *The R Journal*, vol. 9, no. 2, pp. 163–175, 2017.

[16] M. Hahsler and S. Chelluboina, "Visualizing association rules: Introduction to the R-extension package arulesViz," Technical report, pp. 223–238, 2011.

[17] P. Weichbroth, "The visualisation of association rules in market basket analysis as a supporting method in customer relationship management systems," in *Knowledge Acquisition and Management*, M. Nycz and M. Owoc, Eds. Breslavia: Wrocław University of Economics, 2011, pp. 136–145.

[18] Y. A. Sekhavat and O. Hoeber, "Visualizing association rules using linked matrix, graph, and detail views," *International Journal of Intelligence Science*, vol. 3, no. 1A, pp. 34–49, 2013.

[19] M. Hahsler and R. Karpienko, "Visualizing association rules in hierarchical groups," *Journal of Business Economics*, vol. 87, pp. 317–335, 2017.

[20] D. A. Keim, M. Ankerst, and M. Sips, "Visual data-mining techniques," in *Visualization Handbook*, C. H. Hansen and C. R. Johnson, Eds. London: Elsevier, 2004, pp. 831–843.

[21] E. T. Apeh, B. Gabrys, and A. Schierz, "Customer profile classification using transactional data," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, Salamanca, 2011, pp. 37–43.

[22] N. Wang, J. Wang, Y. Zhang, and T. Li, "Two-dimensional Bin-packing Problem with Rectangular and Circular Regions Solved by Genetic Algorithm," *IAENG International Journal of Applied Mathematics*, vol. 51, no. 2, pp. 268–278, 2021.

[23] C. Liberati, F. Camillo, and K. Athappilly, "Profiling a customer base through a sample survey: a data mining approach for business solution," in *Proc. of Annual Midwest Business Administration Association International Conference*, Chicago, 2007, pp.55–65.

[24] G. Adomavicius and A. Tuzhilin, "Using data mining methods to build customer profiles," *Computer*, vol. 34, no. 2, pp. 74–82,

2001.
[25] D. Bhanu and P. M. Soundararajan, "Retail marketing segmentation and customer profiling for forecasting sales," in *Proc. of the 12th international conference on Networking, VLSI and signal processing*, Cambridge, 2010, pp. 230–239.
[26] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: a case study," *Marketing Intelligence & Planning*, vol. 35, no. 4, pp. 544–559, 2017.
[27] S. Guney, S. Peker, and C. Turhan, "A Combined Approach for Customer Profiling in Video on Demand Services Using Clustering and Association Rule Mining," *IEEE Access*, vol. 8, pp. 84326–84335, 2020.
[28] L. A. Székely and H. Wang, "On subtrees of trees," *Advances in Applied Mathematics*, vol. 34, no. 1, pp. 138–155, 2005.
[29] D. Chen, (2012, Aug 27) Online retail data set, (Online). Available: *https://archive.ics.uci.edu/ml/datasets/Online+Retail.*
[30] Okolie P. Chukwulozie, Obika E. Nnaemeka, and Nwuzor I. Chigoziri, "Quality and Productivity Management," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2018, WCECS 2018, 23-25 October, 2018, San Francisco, USA, pp758-761.
[31] Swee Kuik, Toshiya Kaihara, Nobutada Fujii, and Daisuke Kokuryo, "The Optimal Product Disposition Decision for Product Returns towards Sustainable Manufacturing," Engineering Letters, vol. 24, no.3, pp301-306, 2016.