Expert Systems With Applications 100 (2018) 1–16

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Retail business analytics: Customer visit segmentation using market basket data

Anastasia Griva[a,*], Cleopatra Bardaki[a], Katerina Pramatari[a], Dimitris Papakiriakopoulos[b]

[a] ELTRUN: The e-Business Research Center, Department of Management Science and Technology, Athens University of Economics and Business, 47A Evelpidon Str., Room 801, Athens 113 62, Greece

[b] Department of Business Administration, Technological Educational Institute of Athens, 28 Agiou Spyridonos Str., Egaleo, Athens 122 43, Greece

ARTICLE INFO

Article history:
Received 7 September 2017
Revised 15 December 2017
Accepted 20 January 2018
Available online 2 February 2018

Keywords:
Customer visit segmentation
Retail business analytics
Shopper behavior
Clustering
Data mining

ABSTRACT

Basket analytics is a powerful tool in the retail context for acquiring knowledge about consumer shopping habits and preferences. In this paper, we propose a business analytics approach that mines customer visit segments from basket sales data. We characterize a customer visit by the purchased product categories in the basket and identify the shopping intention or mission behind the visit e.g. a 'breakfast' visit to purchase cereal, milk, bread, cheese etc. We also suggest a semi-supervised feature selection approach that uses the product taxonomy as input and suggests customized categories as output. This approach is utilized to balance the product taxonomy tree that has a significant effect on the data mining results. We demonstrate the utility of our approach by applying it to a real case of a major European fast-moving consumer goods (FMCG) retailer. Apart from its theoretical contribution, the proposed approach extracts knowledge that may support several decisions ranging from marketing campaigns per customer segment, redesign of a store's layout to product recommendations.

1. Introduction

The increasing capabilities of business analytics tools and techniques and data-driven decision-making have risen in the agenda of many businesses. Retailers have long recognized that data-driven decision-making could improve decision quality (Kowalczyk & Buxmann, 2015). Since customer satisfaction affects profitability, i.e. the key to business success, retailers want to embrace a more customer-centric approach and find out innovative ways to understand their customers and satisfy them (Anderson, Jolly, & Fairhurst, 2007; Linoff & Berry, 2011). Consequently, they seek means to exploit the collected customer data (e.g. what they purchase, how they move in the stores etc.) and extract knowledge that facilitates effective decisions and offer extra value to their demanding customers. Taking advantage of business analytics, they look for patterns in customer shopping behavior with the purpose to provide tailor-made services that suit the specific shopping needs and preferences of the different customers.

There are studies that utilize various kinds of data, ranging from demographics to sales data and Radio Frequency Identification (RFID) data, to identify patterns in customer purchases. Many of these studies focus on customer segmentation, with the objective to support the customization of the retail service offering to the different customer segments. Current research on *customer segmentation* utilizes the *complete purchase history (all shopping visits) of a customer* in order to identify customer groups (e.g. Aeron, Kumar, & Moorthy, 2012; Boone & Roehm, 2002; Chen, Kuo, Wu, & Tang, 2009; Khajvand, Zolfaghar, Ashoori, & Alizadeh, 2011; Liao & Chen, 2004; Miguéis, Camanho, Falcão, and Cunha, 2012; Park, Park, & Schweidel, 2014). These studies examine shoppers' behavior via looking at the entirety of the products a shopper has purchased, regardless of whether this took place in one or more visits and try to segment shoppers based on this behavior. Other researchers utilize market basket analysis to examine the *associations between the products/ items* purchased during a shopper's single visit (e.g. bread→ milk), i.e. they look for answers to questions such as 'which products are bought together' (e.g. Agrawal, Imieliński, & Swami, 1993; Cil, 2012; Srikant & Agrawal, 1995; Tang, Chen, & Hu, 2008).

The aforementioned studies overlook the shopping purpose of a single customer visit, because they either examine the entirety of a customer's shopping history or focus on the association between specific products that customers purchase during a single visit. However, marketing researchers who talk about different shopping trip types, e.g. fast refilling trip or major monthly trip (Bell, Corsten, & Knox, 2011; Walters & Jamil, 2003), have stressed

* Corresponding author.
E-mail addresses: an.griva@aueb.gr (A. Griva), cleobar@aueb.gr (C. Bardaki), k.pramatari@aueb.gr (K. Pramatari), dpapak@teiath.gr (D. Papakiriakopoulos).

https://doi.org/10.1016/j.eswa.2018.01.029
0957-4174/

the need to understand a single customer visit. Alike practitioners have coined the term "*shopping mission*" to refer to the intention behind a shopper's visit (ECR Europe, 2011). The ultimate goal of this research was to delve deeper into and comprehend the customers' shopping behavior and intentions per visit and, thus, enable retailers to provide customers' satisfactory services tailored to their requirements per visit. We propose a business analytics approach that utilizes clustering techniques to identify segments of customer visits. We analyze retail data in basket level and produce groups of customer visits based on the product categories the customers have purchased during each visit to a physical retail or web store. We suggest that the resulting mix of product categories that prevails each visit segment reflects the shopping intentions of the respective customers that visited the stores. In brief, we generate segments of customer visits and, then, we attribute to each segment the shopping intention behind the visits. Let's assume that the prevailing product categories purchased during the shopping visits of a mined segment are biscuits, chocolates, beverages, ice creams, beers, soft drinks and chips. Then, we conclude that the shopping intention of the respective customers was to buy "snacks and beverages".

In addition, we do not overlook the significant effect of the product taxonomy in the effectiveness and validity of our clustering results, since data analytics studies have highlighted that this may seriously affect the knowledge discovery process and the data mining results (Albadvi & Shahbazi, 2009; Cho, Kim, & Kimb, 2002). More specifically, product taxonomies are often unbalanced and have characteristics hindering the performance of data mining algorithms. Thus, it matters for example whether we should refer to a can of sparkling orange juice of brand XYZ as sparkling beverage, as beverage, or as orange juice. For that reason, this research also suggests a semi-supervised feature selection approach that uses the product taxonomy as input and extracts the features (product categories) as output. This approach is used to adjust the original product taxonomy tree, and it takes into account both the frequency of product purchases and the product semantics to tackle with data skewness problems.

We demonstrate our customer visit segmentation approach by putting it in practice. We have performed and report the results of this new segmentation approach for a major European fast-moving consumer goods (FMCG) retailer that has provided Point-of-Sales (POS) data for over a year.

The remainder of the paper is organized as follows. Next, we summarize the relevant literature and pinpoint how this study differs from the extant ones. The proposed approach and its application are described in the next two sections. Finally, we conclude with the main outcomes of the paper, the theoretical contribution and the practical implications of our approach; and some highlights of further research.

2. Background

Business analytics techniques have been developed that can connect and manage very large datasets to enable broader and deeper analysis than previously possible (Phan & Vogel, 2010; Provost & Fawcett, 2013). The relevant research investigates methods to analyze data of different domains, in order to acquire deeper, unprecedented knowledge that can improve business decision-making. Hence, businesses that apply business analytics methods to their data can facilitate decisions and improve their performance.

Retailers collect and store voluminous and several types of data about their customers daily, ranging from customer demographics, to data that indicate how customers move into the physical or web stores, what products they put in their baskets or try on in the fitting rooms, what products they purchase etc. Since the data volume, variety and velocity have far outstripped the capacity of manual analysis (Chang, Kauffman, & Kwon, 2014), one of the greatest aspiration of retailers is to find innovative ways to exploit the collected datasets that remain unutilized. Up until now, retail analytics efforts have mainly dealt with identifying shopping behavior patterns hidden in the sales data. We can classify the pertinent studies found in the literature in two broad categories: those focusing on customer segmentation based on a customer's purchase history and demographics data, and those utilizing basket data to find associations between the products a shopper purchases during a visit. These two streams of research are further discussed below.

2.1. Customer segmentation

Retailers like Tesco, Metro and Wal-Mart have recognized the need of data-driven decision-making. Mainly, they utilize business analytics tools to gain a competitive advantage in areas such as marketing e.g. cross-selling, in-store behavior analysis, customer segmentation and multi-channel experience (European Commission, 2014). Among their greatest endeavors is to identify the different customer groups visiting their stores, understand the specific needs and preference of each segment, and offer suitable services with a view to satisfy them e.g. by tailoring their marketing mixes (Boone & Roehm, 2002).

Researchers have responded to the retailers' interest for effective customer segmentation and many studies have appeared that utilize various kinds of data. Customer segmentation is the process of dividing heterogeneous customers into homogeneous groups on the basis of common attributes and is essential for handling a variety of customers with rich sets of diverse customer preferences more efficiently (Hong & Kim, 2012). Customer segmentation studies have utilized one or more of the following types of data (Cui, Wong, & Lui, 2006; Hong & Kim, 2012; Miguéis et al., 2012): (a) Demographic data, e.g. gender, age, marital status, household size etc.; (b) Geographic data, e.g. area of residence or work etc.; (c) Psychographic data, e.g. social class, lifestyle and personality characteristics etc.; (d) Attitudinal data, i.e. perceived data gathered from surveys that capture information about what people say they do in order to understand and interpret shoppers' behavior (Konuş, Verhoef, & Neslin, 2008; Woodside, 1973); (e) Sales data, that indicate shopping behaviors e.g. sales volume, number of visits, visit frequency, monetary volume, interarrival days, purchase time, visit recency etc., and customer necessities and preferences according to the mix of categories shoppers purchase; (f) Behavioral data, i.e. data that indicate behaviors other than shopping e.g. data derived from RFID-enabled carts demonstrating what shoppers put in their basket, data from RFID-enabled fitting rooms, demonstrating what clothes people try on, web-site or store navigation data, data derived from Bluetooth Low Energy (BLE) technologies such as beacons etc. For example, Larson, Bradlow, and Fader (2005) use RFID in shopping carts and then by implementing clustering they segment shopper paths and examine common travel behaviors in a grocery store.

Researchers utilize different data mining models for customer segmentation using sales data, such as models based on associations (e.g. association rules, Markov chains), classification (e.g. neural-networks, decision trees), clustering, sequence discovery, forecasting (e.g. neural-networks) (Ngai, Xiu, & Chau, 2009). In all these empirical works, researchers utilize *customer-level sales data* in order to segment shoppers and examine their purchase behavior. In other words, they either examine (a) shopping behaviors (e.g. sales volumes, visit frequency etc.) or (b) the mix of the products or product categories that shoppers have purchased in their whole purchase history i.e. during all their visits in a physical or web store of a retailer. For instance, a stream of studies that focuses on shopping behaviors, utilize sales data to segment shop-

pers based on their CLV (Customer Lifetime Value), mainly using RFM (Recency, Frequency, Monetary) and clustering analysis (Aeron et al., 2012; Chen et al., 2009; Cheng & Chen, 2009; Khajvand et al., 2011).

On the other hand, there are also studies that focus on the mix of the products or product categories that shoppers have purchased in their whole purchase history. For instance, Park et al. (2014) propose a modeling framework for customer base analysis in a multi-category context to predict customer purchase patterns. To this end, a beauty care company in Korea provided sales data that concern both shopping behavior and categories mix. Furthermore, statistical methods e.g. Markov chains, Euclidean distances, are utilized to model the time between a customer's purchases (interarrival time) at the firm and the product categories that comprise a shopping basket. In another study, Miguéis et al. (2012) propose a method for market segmentation in retailing based on a customer's lifestyle, supported by information extracted from a large transactional database. They analyze the product categories shoppers have purchased from a European retailing company. Using clustering, they propose promotional policies tailored to the customers of each segment, with the purpose to support loyal relationships and increase sales. In addition, Han, Ye, Fu, and Chen (2014) showcasing the role of categories in customer segmentation, they compared different techniques and performed clustering using k-means in customer-level sales data to segment shoppers. In their segmentation approach they identified customer segments e.g. customers who purchase routine, seasonal or convenience categories.

From a different perspective, Liao, Chen, and Hsieh (2011) utilized sales data and data collected via questionnaires from shoppers who purchase skincare and cosmetic products to segment customers into clusters, according to their lifestyle habits and purchasing behavior. By adopting clustering and association rules, they provide suggestions and solutions for direct marketing to design possible new services and sales for each customer segment. Boone and Roehm (2002) utilize sales data that concern shopping behaviors (e.g. orders, spending, days since last and first purchase etc.) provided by a retailer, and other artificial data, in order to examine the use of artificial neural networks (ANNs) as an alternative mean of segmenting retail databases. Their results indicate that ANNs may be more useful to retailers for segmenting markets because they provide more homogeneous segmentation solutions than mixture models and k-means clustering algorithms. In addition, Kitts, Freed, and Vrieze (2000) developed an algorithm to analyze a customer's purchasing history provided by an on-line and catalogue hardware retailer, in order to provide item-level recommendations and promotions. Liao and Chen (2004) combine various kinds of data, such as sales data regarding categories mix, demographics, and attitudinal data collected via questionnaires, and use a data mining approach to segment customers, in order to enhance the effectiveness of direct marketing and sales management in retailing, and more specifically to format electronic catalogues.

2.2. Market basket analysis

The second group of studies looks for associations between the items/products a shopper purchases during a visit. A famous example is that of diapers and beers in Wal-Mart stores. These studies perform *market basket analysis* (also known as *association rule mining*), which is a data mining method that examines large transactional databases to determine which items are most frequently purchased jointly (Agrawal et al., 1993; Srikant & Agrawal, 1995). Many extensions of the method have been proposed and it has been widely used in various domains, such as finance, telecommunications, retailing etc. (Chen, Tang, Shen, & Hu, 2005). For instance, Cil (2012) introduces a framework that identifies the associations among the purchased categories in a supermarket. These associations between product categories reveal "consumption universes" and are utilized to change the store's layout. For analysis purposes, he utilizes sales data and the categories mix provided by a Turkish supermarket. Furthermore, Tang et al. (2008) introduce a new approach to perform market basket analysis in a multiple-store and multiple-period environment. They use sales data provided by twenty stores of a supermarket chain in Taiwan and propose purchasing pattern analysis at a detailed level of time and place, such as a combination of days and stores. Although variations of association rule mining have been proposed, certain characteristics of real world data hinders the performance when the algorithms have been designed and evaluated with artificial data sets (Zheng, Kohavi, & Mason, 2001) thus, making the applicability in real world settings a crucial factor. The next section addresses a significant issue that is common in retail contexts.

2.3. Product taxonomy

We should not overlook the significant role of the product taxonomy (see Fig. 1 for an example) in such data analytics studies, since it may affect the knowledge discovery process and the data mining results (Cho et al., 2002). The study of the impact of product taxonomy on data mining is mainly found in the recommendation systems literature and the semantic web literature. Many researchers emphasize that it is critical to find the right product category granularity level, because it could affect association rule results and, thus, the whole recommendation system (Albadvi & Shahbazi, 2009; Cho & Kim, 2004; Cho et al., 2002; Han et al., 2014; Hung, 2005; Kim, Cho, Kim, Kim, & Suh, 2002; Srikant & Agrawal, 1995). Existing approaches handle this issue by examining the product items a customer purchases or interacts with at a stock-keeping-unit (SKU)/item level (Kim, Kim, & Chen, 2012); or examining product categories (e.g. beverages, breads, orange juices) based on the granularity level as indicated in the product taxonomy (e.g. level/height = 3 in Fig. 1) (Cil, 2012; Videla-Cavieres & Ríos, 2014). In addition, others utilize a cross-category level as indicated by marketers or domain experts (e.g. shaded nodes in Fig. 1) (Albadvi & Shahbazi, 2009). To the best of our knowledge, only Cho and Kim (2004) and Srikant and Agrawal (1995) propose an algorithmic logic to define the right granularity level of product taxonomy. On the one side, Cho and Kim (2004) define the right granularity level by selecting cross-category levels and merging some categories based solely on product purchases (e.g. merging socks and skincare). On the other side, Srikant and Agrawal (1995) propose producing associations between items at any level of the taxonomy and pruning redundant rules in order to address issues in the product taxonomy.

Apart from the recommendation systems literature, the problem of defining the right granularity level is also met in the semantic web literature. In this case, an ontology merging and mapping on products over the different product classification taxonomies is required. This is based solely on product semantics (e.g. merging of books and humor books) and it could be vital for product-comparison sites and recommender systems (Aanen, Vandic, & Frasincar, 2015; Park et al., 2014).

2.4. Identified research gaps

Overall, the aforementioned studies show that researchers have applied different data mining approaches to sales data collected per customer (customer-level) to produce customer segments. They divide the customers into groups based either on their complete shopping behavior in terms of sales volume, visit frequency etc. or on the mix of products or product categories recorded in their

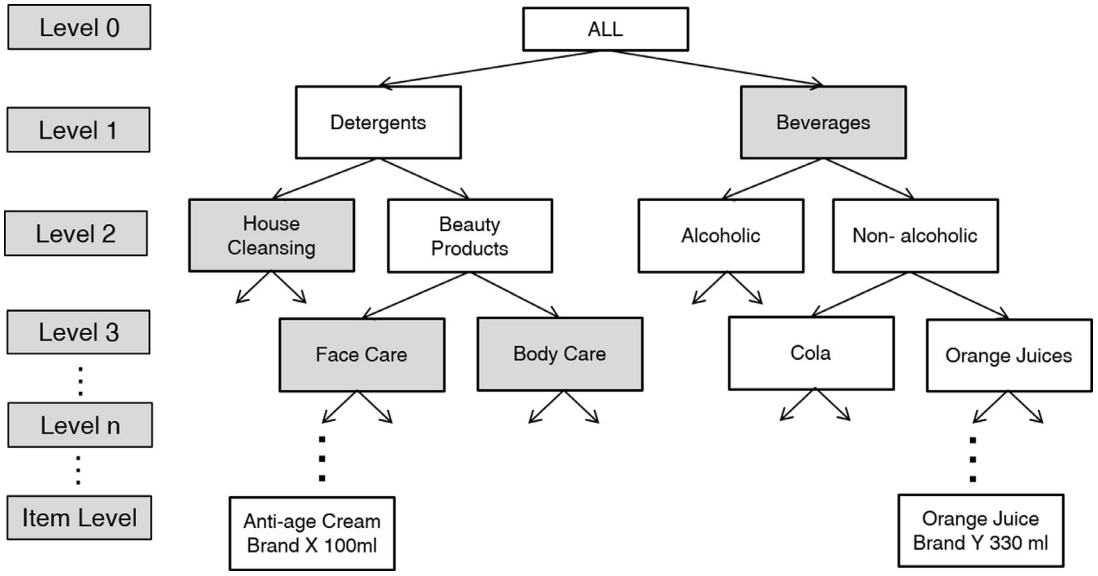

Fig. 1. Product taxonomy example.

total purchase history (Aeron et al., 2012; Boone & Roehm, 2002; Chen et al., 2009; Cheng & Chen, 2009; Han et al., 2014; Khajvand et al., 2011; Kitts et al., 2000; Liao & Chen, 2004; Liao et al., 2011; Park et al., 2014). In contrast, there is a group of scholars analyzing sales data per visit (basket-level) to identify associations between products (e.g. Agrawal et al., 1993; Cil, 2012; Srikant & Agrawal, 1995; Tang et al., 2008). In other words, they do not look for customer segments, but they focus on pairs of products the customers purchase together more frequently (e.g. diapers and beers in the famous Wal-Mart stores' study, or in another example e.g. egg →milk and dairy products).

In our retail data analytics approach we have chosen to analyze sales data at basket-level because we need to explore and characterize the customer shopping behavior per single visit. Both researchers (Bell et al., 2011; Walters & Jamil, 2003) and practitioners (ECR Europe, 2011) suggest that we should pay more attention to the single visits of shoppers because each single visit carries valuable insight on the shopper needs and, thus, can enable retailer actions to satisfy them. Working with the entire purchase history of the customers, namely the totality of shopping visits, would not allow to delve deeper into and comprehend the customers' shopping behavior and intentions per visit e.g. their shopping missions.

On the contrary, we apply clustering to shopping basket data and we produce groups of baskets, namely segments of customer visits, based on the product categories the baskets include. Let's assume that we identify a group of baskets that contain bread, cheese, milk, coffee, and butter products. So, we have found a segment of customer visits that took place because the respective customers needed food products for breakfast. In other words, the customers that held these baskets visited the retail shop with the intention to satisfy their "breakfast" needs. Thus, this retail analytics research first generates segments of customer visits, which then guide the identification of the customers' shopping intention per single visit.

Moreover, we do not work with basket data just to find pair of product categories that customers purchase more frequently in their shopping trips. We are interested in all the product categories found in each shopping basket, because we seek to understand the 'bigger picture' of each shopping visit, namely what shopping needs motivated the customer's shopping trip. The resulting mix of purchased product categories is not random but reflects the original shopping purpose of each single customer visit. Essentially, we generate the segments of customer visits to highlight the customers' shopping needs and intentions when visiting retail stores and, thus, enable retailers to satisfy them more effectively.

Moreover, regarding the product taxonomy, we chose to work on basket data at a customized product category level and not at item/SKU level. With a typical retail store, having more than 10.000 SKU's in its assortment, it is rather impossible to identify significant patterns at an SKU level and working at a higher level of analysis is required in order to avoid data sparsity problems. Besides, the main store retail activities (e.g. store replenishment, shelf space allocation, product assortment selection) and the relevant decisions mainly refer to product categories, as the shopper needs are often expressed at the category level (e.g. 'I need to buy milk') rather than at a specific SKU level (e.g. 'I need to buy this specific milk in a 250 ml bottle'). In addition, by working at the product category level we ensure that the results are more generic and may also apply to new products of a category.

In consensus with the aforementioned studies, we believe that identifying the right product category level, i.e. the right level of analysis in the product taxonomy tree, is crucial to the results of the study, it may affect the knowledge discovery process and the data mining results (Cho et al., 2002). The available studies show that there is no generic rule; the researchers select the product taxonomy level that better serves their research purposes (Cil, 2012; Videla-Cavieres & Ríos, 2014). However, researchers that have utilized a retailer's existing product taxonomy have often claimed very poor results in both the algorithms' accuracy and the business evaluation (Cho & Kim, 2004; Videla-Cavieres & Ríos, 2014). To avoid poor, not representative customer segmentation results, we propose formulating a customized product category level, via balancing a retailer's product taxonomy. In more detailer, we designed, developed and employed also suggest a semi-supervised feature selection method that uses a product taxonomy as an input and suggests the features as an output. This approach is used to balance retailer's product taxonomy tree, and it takes into account both the frequency of product purchases and the product semantics. Our approach differs from those utilized in the semantic web, which take into consideration only product semantics. In addition, our research differs from Cho and Kim (2004) that formulates categories based solely on product purchases without taking into account product semantics, leading to merging unrelated products such as skincare and socks. Last but not least, we also dif-

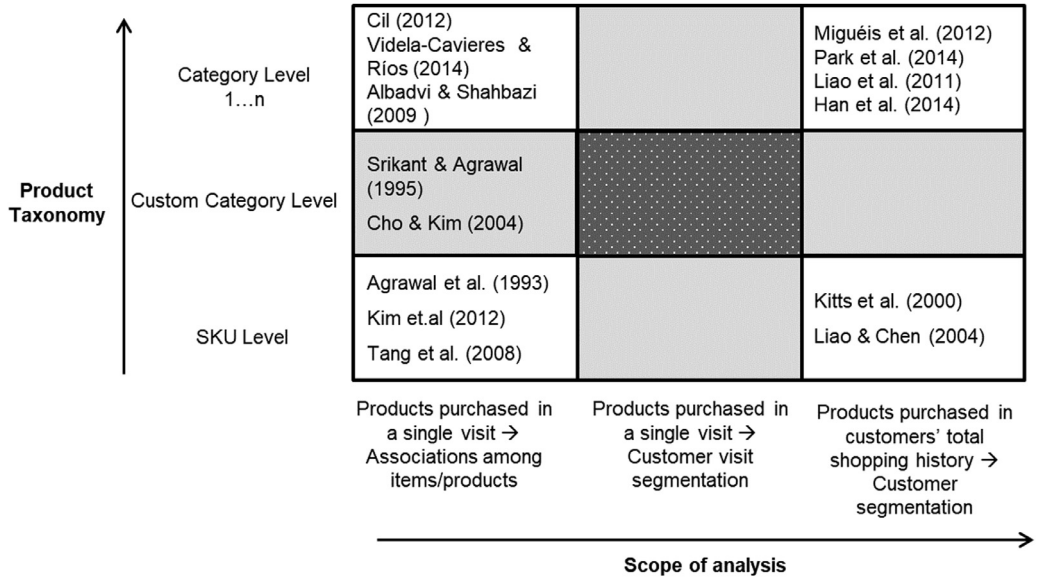

Fig. 2. Research gap.

fer from Srikant and Agrawal (1995), as in their approach they produce association rules for any product taxonomy level, and prune the redundant ones; however, they do not formulate customized product categories, also this approach is dependent on the data mining technique e.g. association rules, and hasn't been tested in other techniques such as clustering.

Fig. 2 depicts the research gap concerning the above two aspects i.e. the *scope of the analysis*, and the *product taxonomy*. The shaded areas declare the research gaps, and the dark grey rectangle in the middle is the area that our research contributes the most. More specifically, the scope of analysis describes the extent to which market baskets are utilized to study a specific issue. On the one hand, researchers study the associations between products that customers purchase during a visit. On the other hand, they study and group baskets using the entirety of a customer's shopping visits. In our perspective, this dimension is shaped with a view to study the shopping purpose/mission of a single customer visit. Also, regarding the product taxonomy, as mentioned earlier, researchers utilize the trees' internal nodes (product categories) or the tree leaves (SKU level) depending on the scope of analysis. In our work, we adjust the original product taxonomy, often defined by a retailer for operational purposes, and produce customized product-categories, which can adequately support the customer visit segmentation.

3. Business analytics approach for customer visit segmentation

We propose a business analytics approach that employs clustering technique to accomplish segmentation and characterization of customer visits, by analyzing the product categories a customer purchases while visiting a physical or even web store.

We have adjusted CRISP-DM, a cross industry standard process for data mining (Shearer, 2000), which includes these steps, to serve our research purpose. Starting from the six CRISP-DM steps: (a) Business Understanding, (b) Data Understanding, (c) Data Preparation, (d) Modeling, (e) Evaluation, and (f) Deployments, our approach includes four steps (see Fig. 3): (a) Business and Data Understanding, (b) Modeling, (c) Evaluation of the model and its results, and (d) Customer Visits Segmentation.

Overall, the originality of this approach is embodied in the "Modeling" phase (marked with red in Fig. 3) where we employ clustering for the customer visits segmentation. This phase includes: (a) product taxonomy adjustment, (b) cluster sampling and (c) adjustment of the input data in order to produce valid customer segments.

Next, we summarize the steps of our approach.

3.1. Business and data understanding

The business goal is to identify for the different segments of customer shopping visits which were the specific shopping missions, i.e. needs and preferences of the corresponding customers that motivated these shopping trips; and, then, offer them the appropriate service mix. We perform the segmentation by examining the product categories the customers purchase during their visits in physical or web retail stores. Apart from data referring to the product purchases per visit (i.e. basket data), the relative input dataset includes the product category tree and the product barcodes of the retailers' product assortment. More input data can be other interactions between shoppers and products during store visits. For instance, products that customers put in their physical or electronic basket, or garments that they try on in fitting rooms in fashion retail stores, but they do not purchase them. Such data may be captured by the standard point-of-sales devices or by RFID sensors, Bluetooth Low Energy (BLE) tracking devices (e.g. beacons), or by navigation data (e.g. google analytics) in the web environment. Extra data sources, e.g. customer demographics data, could enrich and enlighten the resulting visit segmentations.

Given the heterogeneity and the noisy nature of the data, it is not enough to just collect them and throw them into a data repository (Jagadish et al., 2014). Synchronizing and integrating the datasets derived from various sources for establishing data consistency is a major challenge. Thus, *data preparation* is required to support the comprehension of the data sources and the business context they originate from. In other words, we first perform data integration, which involves combining data residing in different sources (Lenzerini, 2002). Then, we apply data cleansing for detecting and correcting or removing errors and inconsistencies of the data to improve data quality (Rahm, 2000); and execute data transformations, such as transforming the obtained sales data in product categories included in each basket. Finally, we end up with data validation after each of the above steps to consolidate the data integrity of the available datasets based in ad hoc criteria selected by the researchers.

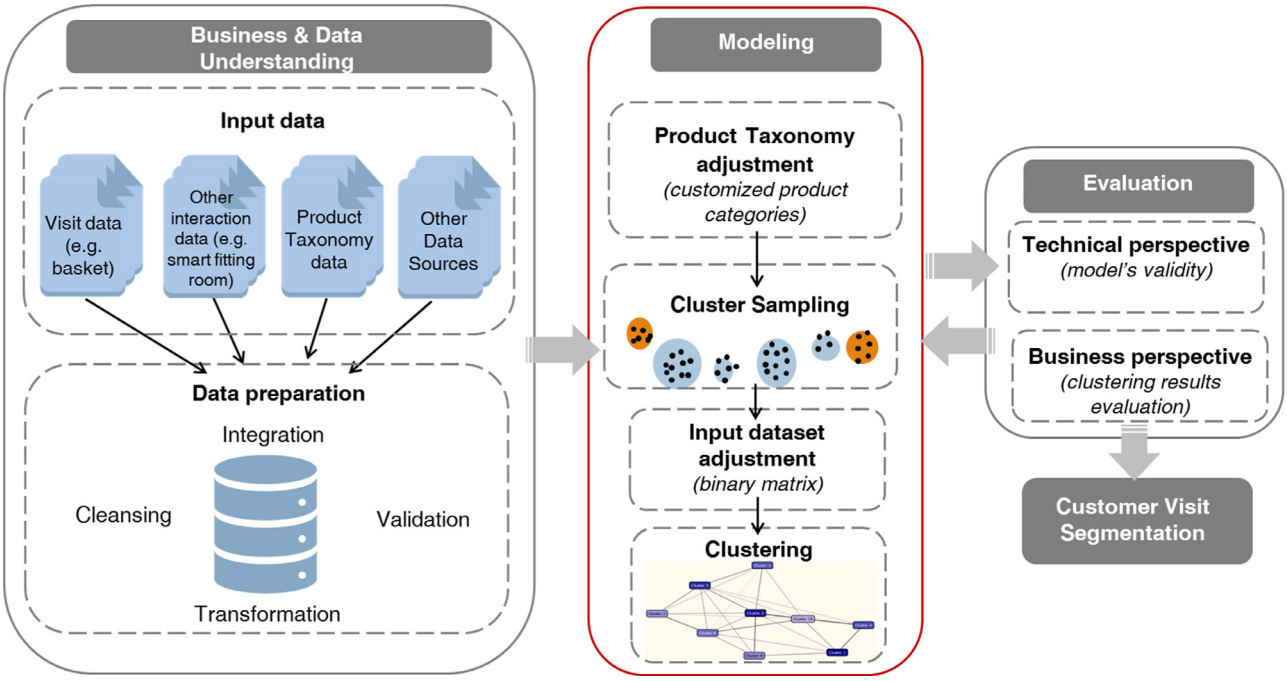

Fig. 3. Business analytics approach for customer visit segmentation.

3.2. Modeling

Essentially, this phase performs the following three prerequisite tasks: (a) product taxonomy adjustment, (b) cluster sampling and (c) input data adjustment and clustering, which ensure that the clustering analysis will produce meaningful results.

3.2.1. Product taxonomy adjustment

Each retail chain has designed and maintains a product hierarchy (often referred to as product-categories tree) that is necessary to conduct various business processes (e.g. store replenishment, shelf space allocation, product assortment selection). This tree corresponds to the product variety and market specialization to facilitate the operational activities in the best way possible. However, we suggest that it is not suitable "as-is" for data analytics purposes because it is often unbalanced and has characteristics hindering the performance of data mining algorithms. These characteristics, which we also came across in our study, are: (a) the height of the sub-trees is significantly different, indicating high product specialization in some product categories (sub-trees), (b) the product tree might be a forest from a data structure perspective meaning that the product categories are expanding separately and managed independently, and (c) the node's degree is varying significantly especially at the SKU/item level. Hence, we suggest that the underlying characteristics of the dataset might affect data mining activities due to the utilization of highly skewed data sets.

To see into the characteristics of the dataset and discover any signs of skewness, we examined the relationship between two variables, namely the number of SKUs/items classified at every branch of the Product Taxonomy (product variety) and the participation percentage of a branch in the baskets (basket frequency). In the next scatter plot (Fig. 4), we depict that the x-axis depicts the former variable and y-axis the latter for a product taxonomy tree with height = 3. Every point of the plot represents a single product's taxonomy branch and different colors are used to discriminate paths belonging to different product taxonomies (forest). The plot suggests significant positive skewness in both variables; therefore, we had to manage the dispersion either by merging nodes relying at the bottom right area or by splitting nodes found at the top left corner and produce an efficient balanced product taxonomy. According to Aggarwal (2016), our problem domain requires extreme-value analysis as it suffers from outliers and we adopted Aggrawal's suggestion that "the choice of the model depends highly on the analyst's understanding of the natural data patterns in that particular domain". In this spirit, we initially utilized relevant techniques (e.g. Box-Cox transformation) to manage outliers, but we finally came up with a semi-supervised feature (product category) selection approach that gets the product taxonomy as input and suggests the features as output.

More specifically, we propose an approach relying on the variety of the product categories in a shopping basket (product variety) that adjusts the retailer's original product taxonomy and produces a customized product-categories tree, which can adequately support the clustering analysis and the identification of the customer visit segments. The logic behind the balancing of the product-categories tree is mainly quantitative. The steps we follow to balance the product taxonomy tree are:

a. First, we identify the main product categories (e.g. initial level n in Fig. 5). Then, we utilize the other researchers' proposition (e.g. Bi, Faloutsos, & Korn, 2001) that retail sales data could be represented as a Discrete Gaussian Exponential (DGX) Distribution; thus, a relative small percentage of product categories contributes in most of sales (or Basket Frequency in our case). The role of DGX is to isolate product categories into two disjoint sets: (i) the *green* set includes product categories with high Basket Frequency and (ii) the *red* set assembles the remainder product categories. The proportion between green and red product categories empirically was found around 1:10.
b. Secondly, we adopt a bottom-up iterative approach and focus on the *red* set of product categories and merge nodes sharing the same parent. In other words, we shrink a sub-tree of red nodes and replace it with the parent node with respect to manage the long tails negative effects and the skewness of the data.

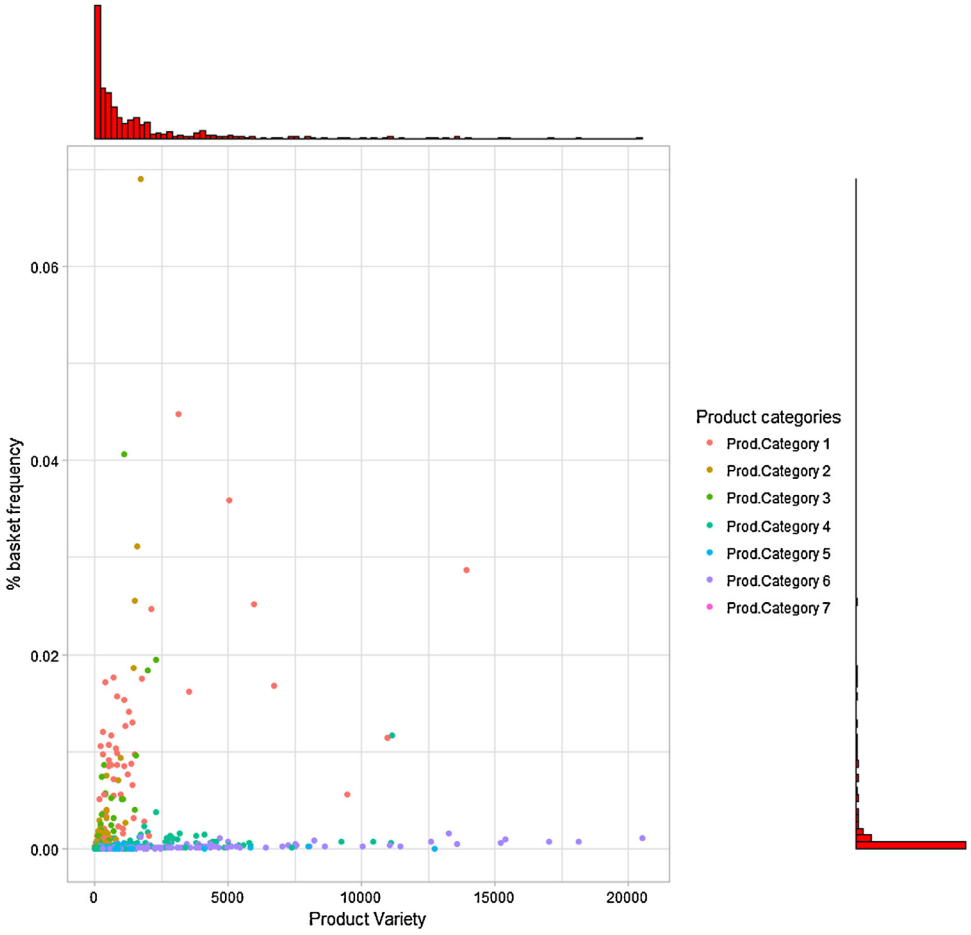

Fig. 4. The relationship between Product Variety and Basket Frequency.

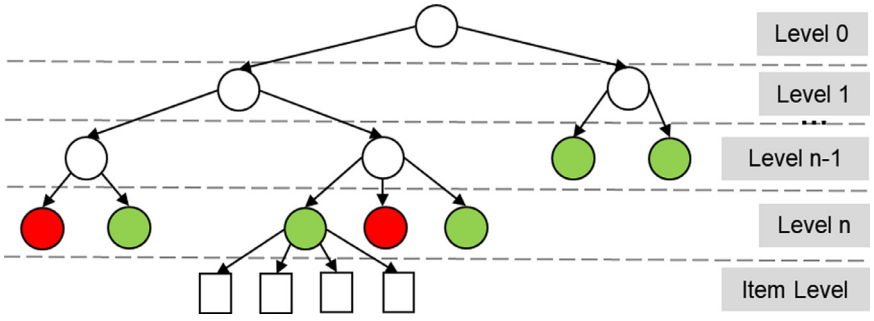

Fig. 5. Balancing product categories tree to formulate new product categories.

c. Regarding the height of the sub-trees, we revised the new product taxonomy and if a tree branch is shallow, e.g. see 'level n − 1' of Fig. 5, the last available nodes will also become *green*.
d. Finally, we reconsider the merged product categories in a qualitative manner taking into account the business context, the analysts' acquired knowledge of the context and experts' opinions (e.g. suppliers, retail managers etc.). Ultimately, we determine if we keep these *red* categories that had been merged as the algorithm indicated, or we split them, or we devise completely new categories that serve the data mining purposes by merging selected nodes.

We emphasize that we merge or split categories that belong to the same parent node, considering the experts' opinion and the product semantics in a way that we avoid merging unrelated categories e.g. skincare and socks.

3.2.2. Cluster sampling

We use *cluster sampling*, with equal sampling weights, to eliminate the outliers. We consider the visits during which a customer has purchased very few or too many products in terms of variety to be outliers. These visits correspond to too concrete or too abstract shopping trips (Bell et al., 2011). The concrete shopping trips are too targeted to extract any product affinities, whereas the abstract ones contain such a wide variety of purchased products, e.g. a monthly stock-out visit in a supermarket, which cannot highlight a specific shopping mission.

The cluster sampling technique groups a finite population into subpopulations-groups called clusters; then, a subset of these clusters is selected (Särndal, Swensson, & Wretman, 2003). We select the final meaningful clusters considering a basic criterion i.e. the percentage of baskets belonging to each cluster, as well as

Table 1
Fact table structure.

Visit/Basket UID	Custom Category 1	Custom Category 2	...	Custom Category n	Other relative data e.g. demographics and Meta- data e.g. basket size, revenue etc.
1	1	1		0	
2	0	0		1	
...					

other relevant descriptive statistics, e.g. revenues per cluster. For instance, in the case of a specific grocery retail store, we can eliminate the baskets that contain only one product (8% of total baskets) and those baskets with more than 80 products (2%). We will utilize the rest of the baskets which reflect 95% of the total revenues.

3.2.3. Input dataset adjustment and clustering

To perform clustering, we need to adjust/transform the dataset in order to form the *fact table* (Table 1) (Shearer, 2000), which represents the learning dataset of the clustering model and includes all the information about a customer shopping visit. Each row of our data table represents a visit (or basket) and the columns correspond to our customized product categories, as well as the visit attributes. The product categories' columns take a binary value, 1 or 0, indicating that the respective basket contains products of this product category or not, respectively. These categories-columns are the input to the data-mining model. In our case, we have selected clustering as the data mining technique to segment customer visits. More specifically, we have selected k-means. The basic idea of k-means is to discover k clusters, such that the objects within each cluster are similar to each other and dissimilar from the objects in other clusters. K-means is an iterative algorithm; thus, an initial set of clusters is defined, and the clusters are repeatedly updated until no further improvement is possible (Huerta-Muñoz, Ríos-Mercado, & Ruiz, 2017; You et al., 2015). The accuracy of this algorithm and the quality of the results depend also on the initial number of clusters (Mesforoush & Tarokh, 2013). Thus, it is critical to define a mechanism to determine the optimal number of clusters. Well-known methods to determine the optimal number of k are elbow, silhouette and gap statistic.

3.3. Evaluation

Here, we suggest that the resulted customer visit segments should be assessed in both business and technical terms. On the one hand, a group of industry experts should assess the validity of the results based on their accumulated experience. If they defy them, we should re-execute the analysis after changing the input dataset. For communicating the results of our approach to the business experts, we translate each found segment of customer visits to a shopping intention that motivated the segment's visits. More specifically, we characterize each group of shopping visits/trips by examining the prevailing product categories the customers purchased during the shopping visits of each segment. For example, if a cluster includes baskets that mainly include categories such as milk, cereals, coffee, sugar etc., we call this segment of visits as "breakfast", declaring that customers have visited the store to buy goods for their breakfast.

If we need to make changes to the original input data based on the experts' comments, we usually delete, merge, or split some of the customized product categories. Merging or splitting contiguous product categories is a widespread practice for increasing the internal consistency of clusters. Thus, after a first trial, it is more effective to reconsider custom categories level. For example, in some cases, merging two or more product categories has resulted in a generic category. On the contrary, disjoining results is the split of a custom category in its children categories/ nodes. From a data mining perspective, this decision decreases a sample's variability and consequently yields better performance results. Thus, the business evaluation constitutes a dialectic process between the experts and the data mining techniques. The researcher should calibrate the cluster model, to satisfy important data mining metrics and, at the same time, deliver a readable abstraction of the cluster to the experts.

On the other hand, in terms of technical evaluation, we need to test the model's validity. Since clustering is based on the similarity of the contained objects, metrics such as a cluster's compactness (e.g. how closely related the objects within the same cluster are) and separation (e.g. how separated the clusters are) could be calculated for the internal validation of the clusters (Liu et al., 2013).

3.4. Customer visit segmentation

Here, the clustering results are extracted, and the final visit segments are identified with the objective to give retailers new knowledge for decision support purposes. We suggest calculating some extra descriptive statistics/Key Performance Indicators (KPIs) per cluster, proposed by the experts, e.g. the basket variety and volume (i.e. the number of product categories and the average number of items it contains, respectively) and the revenues per cluster etc. (Fig. 6). Such measures can support the characterization of the final visit segments.

A drill-down analysis can further be applied to clusters that contain more abstract visits, namely to perform clustering within a single cluster. Then, an abstract cluster may contain more than one sub-clusters. For example, if we apply drill-down to a cluster with many products and product categories, such as wine, beverages, beers, chips, nuts, chocolates, ice, biscuits, orange juice etc., then the original cluster may split into two sub-clusters. Fig. 7 shows two new clusters, one with "beverages" and one with "snack" products. In other words, drilling-down can highlight hidden shopping purposes. However, the resulting sub-clusters are often the same with the original ones. An alternative option is executing a likelihood function between the well-defined and identified clusters with those that are more abstract.

4. Case: customer visit segmentation in a FMCG retailer

Here, we present our proposed business analytics approach in practice demonstrating how it achieves the original goal, i.e. to segment the customers' visits. We utilized original sales data from one European retailer with more than 300 stores, one of the major retailers in the national market.

4.1. Business and data understanding

The retailer has provided point-of-sale (POS) data, from January 2012 to May 2013, from six representative stores. The stores had common characteristics in pairs; two convenience stores, two supermarkets and two mini-hyper markets. We analyzed 36.797.639 records that correspond to 3.973.215 distinct baskets/store visits. Hence, we could retrieve all the baskets of a shopper for a year-and-a-half. Table 2 details our dataset. After integrating and cleansing the dataset, we kept 98% of the initial data. The data was already cleaned, and we only had to eliminate product returns, seasonal items and services provided by the retailer, e.g. product transfers to a shopper's home.

4.2. Modeling

We decided to study the stores and extract the outliers in pairs. For that reason, we created three different views of the data, one

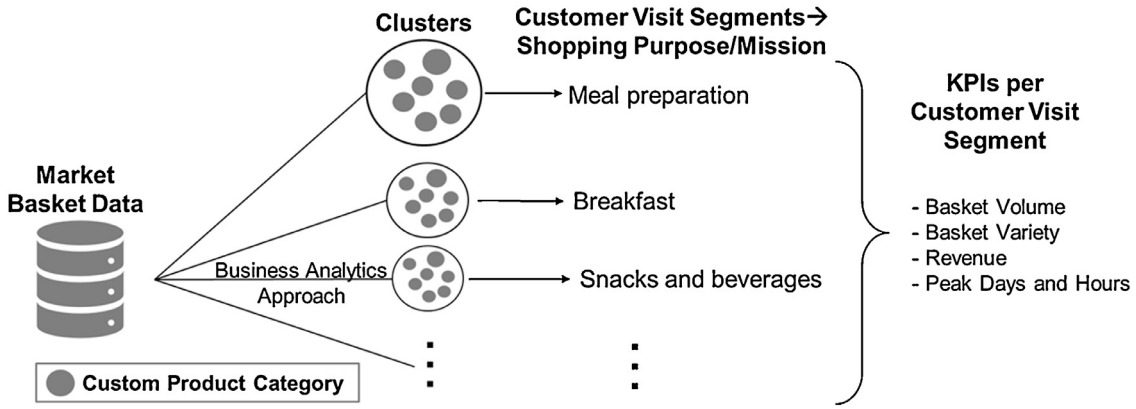

Fig. 6. Conceptualization of customer visit segmentation and characterization.

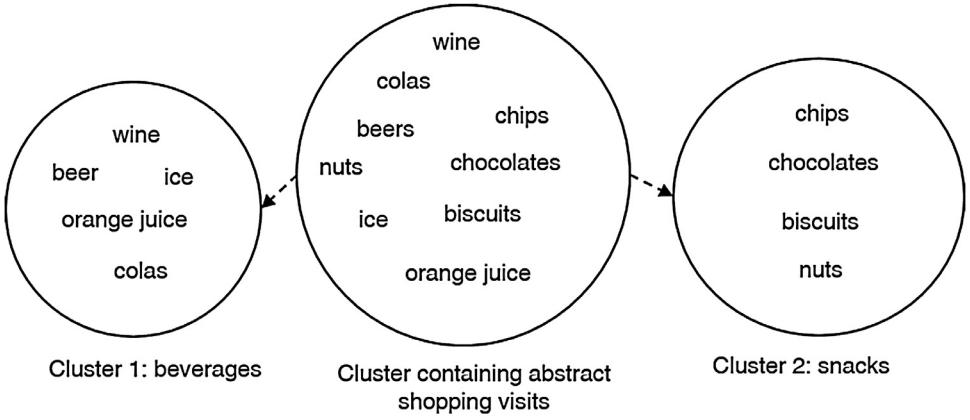

Fig. 7. Drill-down analysis in a cluster containing abstract shopping visits.

Table 2
Dataset identity.

Unique SKUs-Barcodes	126.402
No of Records	36.797.639
No of Baskets/Visits	3.973.215
Average Basket Size/Volume	13,3
Average Basket Variety (in SKU level)	9,4
Average Basket Revenue	29,50 €

Table 3
Summarized results of cluster sampling.

Store type	Basket size range sample	Percentage of the total baskets used	Percentage of revenue used
Convenient	2–24	78,64%	86,05%
Supermarket	3–40	75,49%	83,42%
Mini-Hyper	3–51	79,75%	84,54%

per store type. We utilized cluster sampling with equal sampling weights to eliminate outliers, namely very concrete (low limit) or very abstract (upper limit) shopping visits. After sampling, we calculated the actual number of baskets and the corresponding revenues per cluster to help us with the outliers' extraction. Table 3 summarizes the final dataset analyzed, according to cluster sampling results. The second row represents the range (from, to) of the basket size per store type and the last two columns include the percentage of baskets and their corresponding revenues that we have finally utilized to mine the customer segments.

Additionally, we performed product taxonomy adjustment beginning with rough balancing of the product tree on quantitative criteria. We balanced the product category tree by examining the participation of each tree node in the total purchases. Thus, we generated a first set of 104 customized product categories. Then, we consulted experts of the domain for the final fine-tuning of the product categories. Ultimately, we created 90 new-customized categories by merging some product categories-node. For example, the quantitative criteria highlighted that "lager" beers should be examined separately due to their high participation to the total beer purchases in all stores. Hence, we concluded that the other types of beer (such as stout, bock, ale etc.) should be grouped in one product category named "other beers". However, the experts prescribed to us that we should handle both "lager beers" and "other beers" as one product category named "beers", because the results should also correspond to how the retailers and suppliers handle and understand such products in reality.

We executed the k-means clustering method resulting in six models. We analyzed each store separately because identical visit segments will not necessary result from the same store type. For this reason, we developed Java code to create six fact tables, one per store.

4.3. Evaluation

One common method of choosing the appropriate cluster solution is to compare the sum of squared error (SSE) for various numbers of clusters (i.e. different numbers of K). SSE is defined as the sum of the squared distance between each object of a cluster and its cluster centroid. Hence, SSE can be seen as a global measure of error. It is common that the more the clusters the smaller the SSE, because clusters are smaller. A plot of the SSE against several values of K can provide a useful graphical way to choose an appropriate number of clusters. A suitable "K" value could be de-

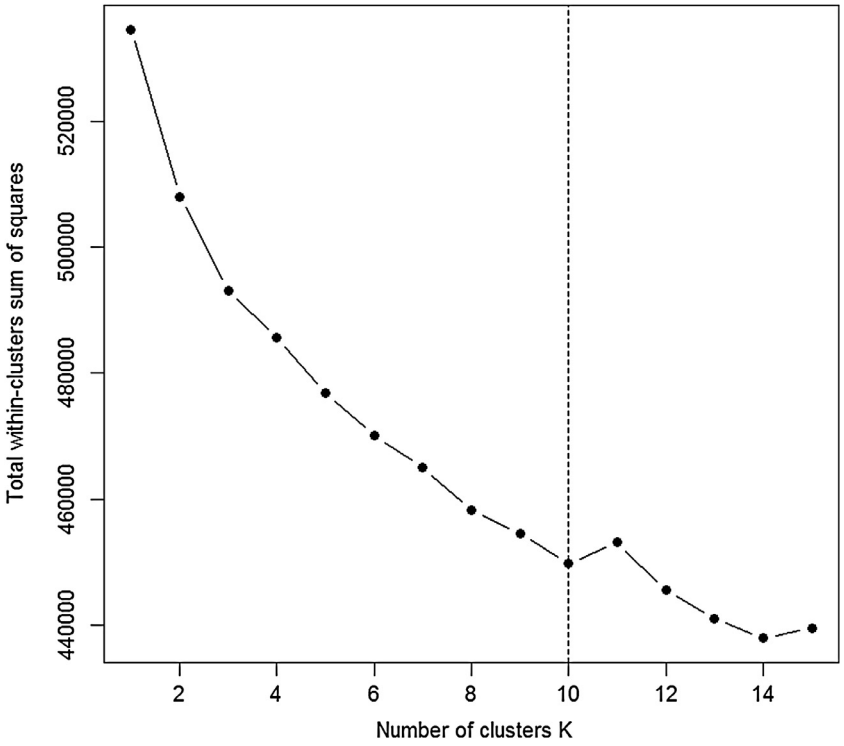

Fig. 8. Elbow method to determine the number of clusters for a supermarket.

fined as the one at which the reduction in SSE slows dramatically. This produces an "elbow" in the SSE plot against cluster solutions (Ketchen & Shook, 1996; Likas, Vlassis, & Verbeek, 2003). Fig. 8 depicts the Elbow Method for the fact table of a supermarket of our dataset. In our case, this plot doesn't show a very strong elbow. We do not have a substantial impact on the total SSE for "K" values between 6 and 10. Thus, we performed clustering several times experimenting with different "K" values ranging from 6 to 10. Again, we consulted domain experts to evaluate the results and depict the optimal number of clusters "K" from a business perspective.

Assessing the first clustering results, the industry people noticed an important product category absence. None of the clusters, in all the different trials, included product categories related to "meat". For that reason, we stepped back at the product taxonomy adjustment phase, and we merged meat-related categories, such as pork, beef, lamb etc., into one, to deliver a readable abstraction of the cluster to the experts. Finally, we ended up with 75 out of the initial 90 product categories. After re-executing clustering for different "K" values, the experts indicated to produce 10 clusters (K = 10) for the supermarket of Fig. 8. Alike, we found suitable K values for the rest of the stores.

We performed clustering using SQL Server data tools of Visual studio and we utilized R programming language to compute the SSE and determine the optimal number of clusters to split the dataset.

4.4. Customer visit segmentation

Fig. 9 shows the final cluster diagram for a supermarket. The more densely populated clusters have darker color. The intensity of the line's shading that connects one cluster to another represents the strength of the similarity of the clusters.

We have calculated the following descriptive statistics per customer visit segment and visit for translating the findings in a business meaningful way: (a) percentage of baskets (visits) per cluster (visit segment size), (b) average basket size in terms of items (visit volume), (c) average number of distinct product categories per basket (visit variety); and (d) average value in Euros per basket (visit value). For instance, Table 4 shows that cluster 2 includes shopper visits with 7,88 products (visit volume) that belong to 4,6 product categories (visit variety); and cost 14,67€ on average (visit value). Moreover, this cluster contains 12,04% (visit segment size) of the total shopping visits in this supermarket in a year and a half. Also, Table 4 depicts the percentage of visits that took place during each part of the day (morning, afternoon, evening) per each visit segment. For instance, 43,12% of the shopping visits, where customers entered the store to buy breakfast, took place in the morning. Similarly, Table 5 depicts the percentage of shopping visits per weekday, per each visit segment. For example, 23,47% of the shopping visits with the intention to buy snacks and beverages happens during Friday. At this point, we would like to mention that the percentage of baskets regarding Sunday is low, since stores are usually closed this day apart from some exceptions e.g. before public holidays.

Cluster 2 mainly contains product categories related to fresh vegetables, red meat, chicken, white cheese, pasta, eggs, bread, oil, vinegar etc. According to the contribution of these categories (i.e. frequency of appearance in the baskets), we infer that this cluster represents visits where the shopper's mission is "meal preparation". According to Table 4 shoppers enter the store to purchase products for meal preparation mainly during afternoon and evening. In addition, this visit segment -as shown in Table 5 is purchased almost equally each weekday. Alike, cluster 3 contains dominant categories, such as milk, baked goods, juice, coffee, tea, cereals, and oral care products. Thus, we can attribute these store visits to shoppers wishing to purchase for "breakfast". We see that according to the baskets in this cluster, shoppers purchase together the products to make their breakfast (e.g. coffee, cereals etc.) and the ones to wash their teeth (oral care category as a daily morning habit) in the same shopping trip. Such outcomes reveal hidden

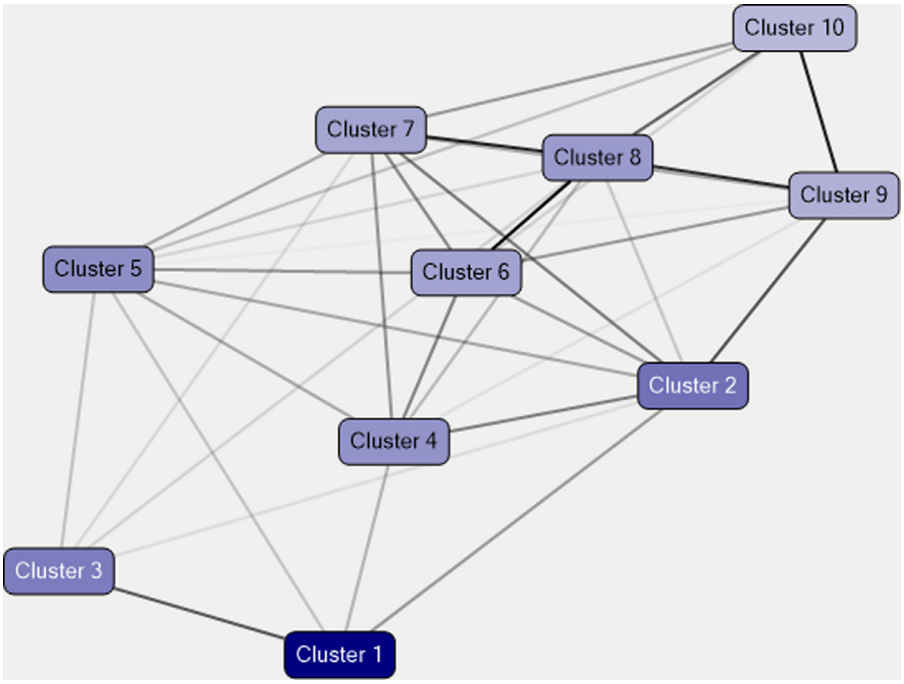

Fig. 9. Cluster diagram for a supermarket.

Table 4
Clustering descriptive statistics of a supermarket.

Cluster/segment	Visit segment name	Visit segment size	Visit volume	Visit variety	Visit value	Part of day		
						Morning	After-noon	Evening
1	Food and drink on-the-go	21,66%	6,3	3,3	13,66 €	27,75%	40,15%	32,10%
2	Meal preparation	12,04%	7,88	4,66	14,67 €	23,09%	36,92%	40,00%
3	Breakfast	11,06%	7,63	4,42	14,61 €	43,12%	23,01%	33,86%
4	Snacks and beverages	9,10%	9,47	5,31	17,59 €	26,43%	33,90%	39,67%
5	Detergents and hygiene	9,60%	10,03	5,77	20,59 €	29,22%	39,65%	31,12%
6	Sandwich with packed products	7,76%	11,97	7,3	24,53 €	28,52%	31,00%	40,48%
7	Light meal	7,58%	11,36	6,51	19,37 €	28,82%	39,08%	32,10%
8	Sandwich with fresh-cut products	8,50%	12,53	7,93	25,71 €	38,31%	33,66%	28,02%
9	Extended visits around food	6,69%	19,66	11,54	36,43 €	29,62%	38,27%	32,11%
10	Extended visits around non-food	6,02%	26,01	15,06	49,66 €	32,10%	38,14%	29,76%

Table 5
Percentage of visits that take place per visit segment per weekday.

Visit segment name	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Food and drink on-the-go	16,50%	16,40%	16,87%	17,49%	18,96%	12,37%	0,41%
Meal preparation	16,37%	15,87%	15,89%	16,29%	16,28%	18,87%	0,42%
Breakfast	19,76%	16,77%	14,94%	15,83%	16,21%	16,28%	0,20%
Snacks and beverages	15,03%	13,93%	14,24%	15,38%	23,47%	17,42%	0,52%
Detergents and hygiene	17,43%	15,97%	16,15%	16,37%	15,92%	17,86%	0,30%
Sandwich with packed products	17,87%	16,30%	14,85%	15,36%	16,70%	18,55%	0,37%
Light meal	17,61%	17,48%	15,91%	17,16%	16,21%	15,26%	0,37%
Sandwich with fresh-cut products	18,49%	16,12%	15,02%	14,86%	16,32%	18,83%	0,37%
Extended visits around food	17,38%	14,63%	13,85%	13,97%	17,37%	22,44%	0,35%
Extended visits around non-food	15,94%	13,65%	12,94%	13,38%	17,87%	25,68%	0,54%

shopper behavior insights that can be useful for marketing purposes. In addition, according to Table 4 shoppers enter the store to buy breakfast mainly during morning. Also, based on Table 5 Monday is the weekday that this visit segment scores the highest percentage. This is an interesting outcome as the marketing team of the collaborative retail chain informed as that each Monday they make discounts on milk, and thus, we realize that the discount on this category increased the rest visit segment.

Similarly, biscuits, chocolates, beverages, ice creams, beers, soft drinks, chips, nuts are the prominent categories in cluster 4. These shoppers visit the store to buy their “snacks and beverages”. Moreover, by examining the days that shoppers visit the store for “snacks and beverages”, we found that Friday and Saturday evening are the prevailing days. Also, this visit segment scores high at Sundays, as the stores are usually open during these days before public holidays (e.g. Christmas, Easter, Carnival). Cluster 1 also contains biscuits, chocolates, chips, coffee, soft drinks and water. The first impression is that it resembles a lot with the “breakfast” and the “snacks and beverages” visit segments, but a more thorough examination showed that it contains products only from 3,3 product categories on avg. The domain experts came again in our assistance and we, finally, recognized in this segment shoppers that

pass by and pick up some food products for immediate consumption. The descriptive statistics in Table 4 shows that this cluster has the biggest size. We can attribute this fact to the position of this supermarket; it is nearby many companies and, perhaps, many employees buy products that can eat and drink quickly during breaks. In addition, according to Tables 4 and 5 the most visits regarding food and drink on-the-go segment, take place during working days and mostly during noon probably at employee's lunch break.

Cluster 5 represents store visits with non-food products, mostly "detergents and hygiene". More specifically, the dominant categories are powders, dish washing, bathroom cleaners, paper rolls, shampoos, body creams, oral care etc., and these visits happens mostly during afternoon. Cluster 6 appears to involve visits where shoppers look for products to prepare a sandwich with "packed products", as the dominant categories are packed cheese, packed cold cuts and packed bakery products. These visits take place mainly during evening and with a more thorough examination we found out that there is a peak the hour before the stores close. Cluster 8 contains almost the same categories with cluster 6, but this time the products are fresh-cut instead of packed. Thus, we refer to this cluster as "sandwich with fresh-cut products". These two segments look a lot alike, but their shoppers have distinct shopping behaviors. In the first one, we may assume that shoppers have time restrictions and they choose not to spend time at the deli counter. On the contrary, the second segment concerns shoppers who value freshness more and are willing to wait at the queues. Also, these visits happen mostly during the morning, thus, probably here we have shoppers that have more time e.g. more elder housewives. Cluster 7 represents shoppers who visit the store with the intention to buy products for a "light meal". More precisely, they pick pasta, rice, pulses, vegetables, white cheese and canned food, but not meat. These visits take place mainly during working days and mostly during afternoon. Finally, clusters 9 and 10 indicate more abstract shopper visits. The first one concerns visits for food products, meaning that they visit the store to purchase and store food in general, and the second one visits containing many non-food products. So, both segments refer to more abstract shopping missions/purposes of visits that take place mainly during Saturday afternoon. We performed drill-down in both clusters, but the results did not reveal any further hidden visit segments. In more detail, the occurring sub-clusters either contained the same segments as those that were mentioned before, or they didn't indicate a certain shopping purpose.

Overall, the customer visit segments per store type shared similarities. We observed that the visit segments –and, thus, the customer shopping missions- are becoming more abstract, as the size of the store grows. Hence, the segments of the mini hyper-store type were more abstract than those of the other two store types. In the mini-hyper stores, we mined many shopping visits related to food-oriented missions, such as "meal preparation", "breakfast", "snack" etc. Still, we identified some segments with different shopping purpose, such as "biological products", "sweet preparation", "snacks and animal feed" and "semi-prepared food". In turn, the convenient stores gave us smaller visit segments in terms of items and revenues, and the visits were more targeted, mainly around food, and there was a lack of extended visits. Some of the identified segments were "snacks", "soft drinks and alcohols", "snacks and beverages", "sandwich with packed products", "light meal", "breakfast" etc.

5. Conclusions and discussion

The ultimate goal of this research is to delve deeper into and comprehend the customers' shopping behavior and intentions per visit and, thus, enable retailers to provide customers' satisfactory services tailored to their requirements per visit. Towards that end, we propose a business analytics approach that utilizes clustering techniques to identify segments of customer visits. We analyze retail basket data and we produce groups of customer visits based on the product categories the customers have purchased during each visit to a physical retail or web store. We suggest that the resulting mix of product categories that prevails each visit segment reflects the shopping intentions of the customers that held the baskets included in each visit segment. In other words, we generate segments of customer visits and, then, we attribute to each segment the shopping intention behind the visits. Let's assume that the prevailing product categories purchased during the shopping visits of a mined segment are biscuits, chocolates, beverages, ice creams, beers, soft drinks and chips. Then, we conclude that the shopping intention of the respective customers was to buy "snacks and beverages".

An extensive review of the relevant literature has revealed that researchers have analyzed sales data per customer (customer-level data) utilizing different methods e.g. clustering, Markov chains, etc. (e.g. Chen et al., 2009; Cheng & Chen, 2009; Han et al., 2014; Kitts et al., 2000; Liao et al., 2011). They examine the complete sales history per customer in terms of sales volumes, visit frequency, mix of products or product categories etc. and, thus, generate customer segments that provide a generic characterization of consumers in terms of the products they prefer in combination with other characteristics such as available budget, demography etc. Indicatively, in these works we find segments with customers who purchase routine, seasonal or convenience categories (Han et al., 2014) or indicating that customers who buy serums are more likely to buy makeups. Alternatively, there is a customer segment including: consumers that buy high proportion of delicatessen, hygiene, grocery and butchery products throughout their purchase history (Miguéis et al., 2012), or customers who have purchased cosmetics, have budget restraints and they have also bought nail-scrub with cleanser-polish during their purchase history (Park et al., 2014), or high spenders and frequent buyers (Aeron et al., 2012). Respectively, other studies employ association rules mining on basket-level data (e.g. Agrawal et al., 1993; Cil, 2012; Srikant & Agrawal, 1995; Tang et al., 2008) and extract the pairs of product categories the customers buy more frequently in their shopping trips e.g. egg → milk and dairy products. In turn, we analyze basket-level data and extract neither customer segments, nor pairs of product categories customers prefer to purchase together. We focus on each shopping visit and assign visits to groups, which are characterized by the product categories purchased during the visits of each segment. The purchased mix of product categories per visit segment supports us in finding the shopping intention that motivated each segment's visits. Different mixes of product categories per visit segment reflect different shopping needs of customers that conducted the visits of each segment. Ultimately, we start with groups of shopping visits to infer distinct customer shopping missions the retail store satisfies e.g. light meal, breakfast, snacks and beverages, food and drink on-the-go. This research's contribution is not attributed to the data mining method we have utilized (i.e. clustering). In contrast with the relevant studies, the novelty of this research is ascribed to its effort to move the attention to each single customer visit instead of dealing with the general shopping behavior of customers. We provide shopping behavior insight per shopping visit and, thus, we urge the retail industry to handle their customers' needs per visit. We believe that working with the entire sales per customer or focusing only on pairs of purchased product categories, we risk overlooking the knowledge behind each shopping visit and underestimating which shopping needs the customers want to satisfy per visit.

In brief, the essence of our approach for generating customer visit segments is summarized in three prerequisite actions: (a) we adjust the product taxonomy and select the right level of analy-

sis at the product-categories hierarchy tree, (b) we perform cluster sampling to eliminate the data outliers; and (c) we adjust the original input data accordingly to support clustering and produce valid customer visit segments and, consequently, sensible customer shopping intentions per visit. We have demonstrated the utility of our approach by applying it to sales data, at basket-level, provided by a major FMCG retailer for a period of more than one year.

In our proposed approach, the adjustment of the product taxonomy (i.e. the first required action) contributes significantly to shaping the feature space of the problem, as it determines the main input (i.e. fact table) of the clustering model. Thus, it is an important research decision to select the level of analysis at the product taxonomy that is efficient to obtain meaningful clustering results (Albadvi & Shahbazi, 2009; Cho & Kim, 2004; Cho et al., 2002; Han et al., 2014; Hung, 2005; Kim et al., 2002; Srikant & Agrawal, 1995). For example, whether we should refer to a can of sparkling orange juice of brand XYZ as sparkling beverage, as beverage, or as orange juice etc. To this end, selecting the lowest level at the product categories hierarchy tree (height = 4) could introduce more than 2100 candidate features, whereas by selecting a higher level (e.g. height = 2), then, an approximate number of 50 candidate features is produced. Taking also into account that customers' visits are unlabeled, it was rather straightforward to utilize existing techniques regarding feature selection (or more precise dimension reduction) for unsupervised learning (e.g. Principal Component Analysis). Although the particular technique is highly adaptive, it appeared having two major drawbacks in our case. Firstly, the produced principal components (product categories in our case) were not comprehensible by the expert and secondly, and most important from a technical perspective, the proposed components had a poor performance in terms of variance explanation, indicating that either the skewed data (existence of latent variables) or the 0/1 values of product categories feature space (distance metric) are performing poorly. To this end, these findings confirm the role of dimensionality reduction in clustering, as well as the capabilities of PCA as discussed in the existing literature (e.g Lawrence, 2005).

Hence, we decided to adjust the product taxonomy adopting the suggestions provided by Dy and Brodley (2004) and Guyon and Elisseeff (2003) regarding feature engineering and the role of the domain problem on the features. In the current study we designed, developed and employed a semi-supervised feature selection approach that uses the product taxonomy as input and suggests the features as output. Specifically, it parses bottom-up the product categories tree and acts in a two-fold manner: (i) it merges nodes (product categories) with low percentage in total baskets and (ii) splits a node with high percentage in total baskets. The proposed approach has not yet been thoroughly optimized and compared to similar methods, yet it supports the extraction of high quality clusters regarding shopper visit segmentation and, thus, we consider it as a first step towards contributing to the existing feature engineering literature. Moreover, we consider the major advantage of our suggested approach to be that we preserve the semantic information of features (product categories) because we deploy it according to expert's intervention.

The practical value of this work is stressed when considering the consumer-oriented business decisions it can support. Overall, this research capacitates the retailers to see customers' shopping behavior from an alternative point of view, not as their general shopping needs and intentions but focusing on their requirements and motives per shopping trip/ visit. More specifically, our approach of detecting customer visit segments can evolve to a tool for designing innovative marketing campaigns and bundled promotions for product categories that belong to the same shopping visit segment. For example, retailers may plan cross-coupon programs for addressing the needs of customers visiting the store with a specific purpose in mind. Alternatively, our approach may become the cornerstone of a recommendation system for real-time purchases in retail stores. It will propose to the customers more products that they may have forgotten to buy, considering their prior or current visit(s). Apart from recommending products from the same cluster/visit segment, marketing managers could also exploit the knowledge extracted from the more abstract clusters to make cross-cluster promotions that fit the specific needs of shoppers. For instance, considering a shopper that visits the store to buy beverages and the recommender promotes to her/him products from the snacks cluster aiming to increase the basket value and variety. This promotion is based on the detected new knowledge that beverages and snacks are sometimes co-purchased in a broader shopping visit.

In the same spirit, we can create offline and online product catalogues. For instance, we have detected women that enter a fashion store to purchase professional clothes and baby clothes. Thus, to promote the new collection, it could be more effective to send them product catalogues that meet their specific preferences, instead of including all the available new garments. The extracted knowledge could also be valuable for advertising purposes; for instance, instead of making advertisements of specific product categories, retailers could advertise bundled product categories that correspond to a shopping mission, e.g. breakfast products advertisements.

On the other hand, the suggested approach of behavioral segmentation and characterization of customer visits may support more business decisions that have a more indirect, but not less significant, impact on customer satisfaction. The customer visit segments can dictate a new redesigned retail store layout where product categories in the same visit segment are positioned in nearby store aisles and shelves. Considering the bigger picture, we can move from a category-based layout to a mission-based layout that can help customers locate products in the store more easily and buy more in less time (Cil, 2012; Sarantopoulos, Theotokis, Pramatari, & Doukidis, 2016).

Further, the store manager could reengineer store operations management and replenishment strategies by ordering groups of products based on the identified visit segments. Or even change the shelves replenishment by time of day and day of week given taking into consideration the peaks of each customer visit segment as shown for example in Tables 4 and 5. Additionally, this approach could be even utilized to rearrange and modify a retailer's warehouse, by placing in nearby aisles products matching online orders to decrease order-picking time. This kind of rearrangement has been previously examined in the literature using solely association rule mining (Chan & Pang, 2011; Chuang, Chia, & Wong, 2014).

We acknowledge that large companies, such as SAS, IBM, SAP etc., have developed commercial tools and suites e.g. IBM WATSON, IBM COGNOS, SAP HANA etc. to ease companies perform different data analyses varying from reporting to data mining (e.g. clustering). Companies utilize these suites to perform customized analyses e.g. produce customer segmentations (Chen, 2014). Our approach is complementary to such solutions. It may be treated as an additional layer of functionality on top of such software tools, for generating customer visit segmentations and consequently deducing customers' shopping intentions per visit.

Further research may address some limitations of this study. We can use more complex interaction data derived from alternative technological means (e.g. RFID, beacons) from other retail contexts to evaluate and validate the proposed approach. For instance, data that indicate the products a customer puts in his RFID-enabled shopping cart during a shopping visit in a grocery store. It would also be a challenge to use different interaction data of the same retailer and compare the results. For instance, we can examine the visit segments derived via combining different interaction data of

the same retailer. In addition, we can identify the selling gaps via comparing the visit segments stemming from data of products in the shopping carts and products that are finally purchased. Moreover, we can examine whether the scope of the analysis used in the literature, i.e. product items in a single visit, or the whole shopper history, are sufficient and applicable in every retail context, or if there are contexts where we need to examine groups of "X" continuous visits in order to identify the purpose behind a shopper's visit. For example, a shopper usually visits many times a retail store that sells products for home improvement, and purchases few materials each time (Wolf & McQuitty, 2011), thus how can we treat these visits to identify shopping purposes, habits and behaviors?

From a technical perspective, we can apply more data mining techniques, such as association rules, and compare the resulting visit segments with those that have been derived from clustering. Or even other techniques and algorithms could also be examined in order to cope with the difficulty to identify core visit segments at the hyper-stores. As already mentioned, a limitation of the proposed approach is that it works only when the number of product categories in a basket is adequate (e.g. more than two and less than 50), which is why the rest of the baskets are considered as outliers.

References

Aanen, S. S., Vandic, D., & Frasincar, F. (2015). Automated product taxonomy mapping in an e-commerce environment. *Expert Systems with Applications, 42*(3), 1298–1313.

Aeron, H., Kumar, A., & Moorthy, J. (2012). Data mining framework for customer lifetime value-based segmentation. *Journal of Database Marketing and Customer Strategy Management, 19*(1), 17–30.

Aggarwal, C. C. (2016). *Outlier analysis second edition*. Cham: Springer.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD International conference on management of data* (pp. 207–216).

Albadvi, A., & Shahbazi, M. (2009). A hybrid recommendation technique based on product category attributes. *Expert Systems with Applications, 36*(9), 11480–11488.

Anderson, J. L., Jolly, L. D., & Fairhurst, A. E. (2007). Customer relationship management in retailing: A content analysis of retail trade journals. *Journal of Retailing and Consumer Services, 14*(6), 394–399.

Bell, D. R., Corsten, D., & Knox, G. (2011). From point of purchase to path to purchase: How preshopping factors drive unplanned buying. *Journal of Marketing, 75*(1), 31–45.

Bi, Z., Faloutsos, C., & Korn, F. (2001). The "DGX" distribution for mining massive, skewed data. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining - KDD '01* (pp. 17–26).

Boone, D. S., & Roehm, M. (2002). Retail segmentation using artificial neural networks. *International Journal of Research in Marketing, 19*(3), 287–301.

Chan, H. L., & Pang, K. W. (2011). Association rule based approach for improving operation efficiency in a randomized warehouse. In *Proceedings of the 2011 international conference on industrial engineering and operations management* (pp. 704–710).

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems, 63*, 67–80.

Chen, J. (2014). Retail Customer Segmentation using SAS. SAS Users Group meeting, Calgary. https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Calgary-User-Group/Chen-RetailSegmentation-Apr2014.pdf. Accessed 15.12.2017.

Chen, Y.-L., Tang, K., Shen, R.-J., & Hu, Y.-H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems, 40*(2), 339–354.

Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications, 8*(5), 241–251.

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176–4184.

Cho, Y. H., & Kim, J. K. (2004). Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications, 26*(2), 233–246.

Cho, Y. H., Kim, S. H., & Kimb, J. K. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications, 23*, 329–342.

Chuang, Y. F., Chia, S. H., & Wong, J. Y. (2014). Enhancing order-picking efficiency through data mining and assignment approaches. *WSEAS Transactions on Business and Economics, 11*(1), 52–64.

Cil, I. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling. *Expert Systems with Applications, 39*(10), 8611–8625.

Cui, G., Wong, M., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science, 52*(4), 597–612.

Dy, J. G., & Brodley, C. E. (2004). Feature Selection for unsupervised learning. *Journal of Machine Learning Research, 5*, 845–889.

Europe, E. C. R. (2011). The Consumer and Shopper Journey Framework. https://www.ecrireland.ie/uploadedfiles/shopper/projects-ecr-eu/the-consumer-and-shopper-journey-framework-.pdf. Accessed 15.12.2017.

European Commission, Directorate-general for research and innovation, expert group on retail sector innovation. Six perspectives on retail innovation. (2014). https://ec.europa.eu/research/innovation-union/pdf/Six_perspectives_on_Retail_Innovation_EG_on%20Retail_Sector_Innovation_web.pdf. Accessed 15.12.2017.

Guyon, I., & Elisseeff, A. (2003). An Introduction to variable and feature selection. *Journal of Machine Learning Research (JMLR), 3*(3), 1157–1182.

Han, S., Ye, Y., Fu, X., & Chen, Z. (2014). Category role aided market segmentation approach to convenience store chain category management. *Decision Support Systems, 57*(1), 296–308.

Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications, 39*(2), 2127–2131.

Huerta-Muñoz, D. L., Ríos-Mercado, R. Z., & Ruiz, R. (2017). An iterated greedy heuristic for a market segmentation problem with multiple attributes. *European Journal of Operational Research, 261*(1), 75–87.

Hung, L. P. (2005). A personalized recommendation system based on product taxonomy for one-to-one marketing online. *Expert Systems with Applications, 29*(2), 383–392.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., et al. (2014). Big data and its technical challenges. *Communications of the ACM, 57*(7), 86–94.

Ketchen, D. J., & Shook, L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal, 17*(6), 441–458.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science, 3*, 57–63.

Kim, H. K., Kim, J. K., & Chen, Q. Y. (2012). A product network analysis for extending the market basket analysis. *Expert Systems with Applications, 39*(8), 7403–7410.

Kim, J. K., Cho, Y. H., Kim, W. J., Kim, J. R., & Suh, J. H. (2002). A personalized recommendation procedure for Internet shopping support. *Electronic Commerce Research and Applications, 1*(3–4), 301–313.

Kitts, B., Freed, D., & Vrieze, M. (2000). Cross-sell: A fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities. In *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining - KDD '00* (pp. 437–446).

Konuş, U., Verhoef, P. C., & Neslin, S. A. (2008). Multichannel shopper segments and their covariates. *Journal of Retailing, 84*(4), 398–413.

Kowalczyk, M., & Buxmann, P. (2015). An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study. *Decision Support Systems, 80*, 1–13.

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing, 22*(4), 395–414.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal ofMachine Learning Research, 6*, 1783–1816 Retrieved from.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 233–246).

Liao, S.-H., & Chen, Y.-J. (2004). Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications, 27*(4), 521–532.

Liao, S., Chen, Y., & Hsieh, H. (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems with Applications, 38*(5), 6059–6069.

Likas, A., Vlassis, N., & Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451–461.

Linoff, G., & Berry, M. (2011). *Data mining techniques: For marketing, sales, and customer relationship management* (third ed.). Wiley (Chapter 2).

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). *Clustering Validation Measures, 43*(3), 982–994.

Mesforoush, A., & Tarokh, M. J. (2013). Customer profitability segmentation for SMEs case study: Network equipment company. *International Journal of Research in Industrial Engineering, 2*(1), 30–44.

Miguéis, V. L., Camanho, A. S., Falcão, e., & Cunha, J. (2012). Customer data mining for lifestyle segmentation. *Expert Systems with Applications, 39*(10), 9359–9366.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602.

Park, C. H., Park, Y.-H., & Schweidel, D. A. (2014). A multi-category customer base analysis. *International Journal of Research in Marketing, 31*(3), 266–279.

Phan, D. D., & Vogel, D. R. (2010). A model of customer relationship management and business intelligence systems for catalogue and online retailers. *Information & Management, 47*(2), 69–77.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Data Science and Big Data, 1*(1), 51–59.

Rahm, E. (2000). Data cleaning : Problems and current approaches. *Informatica*, 1–11.
Sarantopoulos, P., Theotokis, A., Pramatari, K., & Doukidis, G. (2016). Shopping missions: An analytical method for the identification of shopper need states. *Journal of Business Research, 69*(3), 1043–1052.
Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling (Springer series in statistics)*. Springer (Chapter 4).
Shearer, C. (2000). The CRIS-DM model: The new blueprint for data mining. *Journal of Data Warehousing 14, 5*(4), 13–22.
Srikant, R., & Agrawal, R. (1995). Mining Generalized Association Rules. In *VLDB '95 Proceedings of the 21th international conference on very large data bases* (pp. 407–419).
Tang, K., Chen, Y.-L., & Hu, H.-W. (2008). Context-based market basket analysis in a multiple-store environment. *Decision Support Systems, 45*(1), 150–163.
Videla-Cavieres, I. F., & Ríos, S. a. (2014). Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications, 41*, 1928–1936 4 PART 2.
Walters, R. G., & Jamil, M. (2003). Exploring the relationships between shopping trip type, purchases of products on promotion, and shopping basket profit. *Journal of Business Research, 56*(1), 17–29.
Wolf, M., & McQuitty, S. (2011). Understanding the do-it-yourself consumer: DIY motivations and outcomes. *AMS Review, 1*, 154–170.
Woodside, A. G. (1973). Patronage motives and Marketing Strategies.pdf. *Journal of Retailing*.
You, Z., Si, Y. W., Zhang, D., Zeng, X., Leung, S. C. H., & Li, T. (2015). A decision-making framework for precision marketing. *Expert Systems with Applications, 42*(7), 3357–3367.
Zheng, Z., Kohavi, R., & Mason, L. (2001). Real world performance of association rule algorithms. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining – KDD '01* (pp. 401–406).

Anastasia Griva is currently PhD Candidate at the ELTRUN Research Center of the Athens University of Economics and Business. She holds a BSc in Management Science and Technology with a specialization in Quantitative Methods in Economics and Management, and a MSc (with honors) in Information Systems, both from the Athens University of Economics and Business. She has long-term experience in working on international research and industrial projects in the retail sector and strong data analytics background. Her research interests lie in the areas of business analytics, retail analytics, data mining, internet of things and customer behavior. Her research has already been presented and awarded in international conferences.

Cleopatra Bardaki is a Senior Researcher at ELTRUN Research Center of Athens University of Economics and Business (AUEB). She has a long experience in the research coordination and project management of leading edge research & development projects funded mostly from the EU. She holds a PhD in Information Systems enabled by Internet-of-Things technologies in the Supply Chain from AUEB. She also holds an MSc in Information Systems from AUEB and a BSc (with honors) in Informatics and Telecommunications from University of Athens. Her research interests include: Design & Evaluation of Internet of Things applications, Data analytics to support Decision making, Information Systems (IS) Evaluation with an emphasis on Information Quality (IQ) evaluation and Supply Chain Management. Her doctoral research has been funded by a 4-year scholarship from Bodossaki Foundation and she has received more academic distinctions and scholarships during her studies. She has published over thirty papers in academic journals, proceedings of international academic conferences and edited books.

Katerina Pramatari is Associate Professor at the Department of Management Science and Technology of the Athens University of Economics and Business and scientific coordinator of the ELTRUN/SCORE research group. Her research interests lie in the areas of retail analytics, supply chain information systems, digital innovation and entrepreneurship. She has received various academic distinctions and scholarships and has published more than 100 papers in scientific journals, peer-reviewed academic conferences and book chapters, amongst them in Decision Support Systems, Journal of Information Technology, European Journal of Information Systems, Journal of Retailing, Journal of Business Research, Journal of Strategic Information Systems etc.

Dimitris Papakiriakopoulos is Lecturer in Management Information Systems at the Department of Business Administration of the Technological Educational Institute of Athens. He holds a BSc in Informatics, MSc in Information Systems from Athens University of Economics and Business and a PhD in Information Systems and Supply Chain Management also from Athens University of Economics and Business. He has worked as researcher in various EU-funded projects. His research areas are management information systems, supply chain management, data mining and machine learning. He has published in several scientific journals including Decision Support Systems, Expert Systems with Applications, Industrial Management and Data Systems, Electronic Markets etc.