Semantic Customers' Segmentation

Jocelyn Poncelet, Pierre-Antoine Jean, François Trousset, Jacky Montmain

Semantic Customers' Segmentation

Jocelyn Poncelet[1,2(✉)], Pierre-Antoine Jean[1(✉)], François Trousset[1(✉)], and Jacky Montmain[1(✉)]

[1] LGI2P - IMT Mines Als - Université de Montpellier, Alès, France
{jocelyn.poncelet,pierre-antoine.jean,franois.trousset, jacky.montmain}@mines-ales.fr
[2] TRF Retail - 116 Alle Norbert Wiener, Nîmes, France
jocelyn.poncelet@trfretail.com

Abstract. Many approaches have been proposed to allow customers' segmentation in retail sector. However, very few contributions exploit the existing semantics links that may exist between objects and resulting groups. The aim of this paper is to overcome this drawback by using semantic similarity measures (SSM) in customers' segmentation to provide clusters based on product' topology instead of numerical indicators usually used (*i.e.* monetary indicators). More precisely, we intend to show the main advantage of SSM with a product taxonomy in the retail field. Usually, traditional approaches consider as similar three customers buying respectively apple, orange and beer. However, human intuition tends to group customers who buy orange and apple because both are fruits. Our approach is defined to identify this kind of grouping through SSM and abstract concepts belonging to product taxonomy. Experiments are conducted on real data from a French Retailer store and show the relevance of the proposed approach.

Keywords: Customers segmentation · Semantic clustering · Semantic similarity measures · Retail

1 Introduction

As a recurrent issue, outcomes of statistical analyses lack of semantics and their interpretability for decision-making purposes remains challenging. Data summarizing tools are required to provide more explicit aggregated or global indicators [1]. However, they do not improve the knowledge base (Customer Relationship Management - CRM) about customers of a given supermarket brand. As a result, they provide too much macroscopic insights which remains useless to determine information to advantage retailers in attracting and retaining customers. For a long time, retailers know the importance of data driven decision-making to capture customers behaviors and then use results to propose a customer driven approach in the retail industry. However, they focus on identifying good customers by mainly considering numerical values such as revenue, margin and the

frequency of customers [2,3,6,23]. Furthermore they do not consider the semantics that can be associated to products. Indeed, a customer that regularly buys fish and vegetable has obviously a closer behavior to a customer that usually buys meat and salad than to a customer that only buys sanitary products or seasonal items.

This type of inference clearly refers to approximate reasoning based on product similarity that symbolic artificial intelligence can automatically carry out. In this paper, we intend to introduce such reasoning on products purchased to provide semantic group of customers, i.e. groups derived from the hierarchical abstraction structure of products into classes (*e.g.*, *Greek yogurt* is a kind of *yogurt* which is in return a type of *dairy product*, etc.). Our objective is to identify customers' clusters depending on their conceptual purchasing behavior. Products could be organized within a taxonomic partial order defining an abstraction hierarchy. Products sold are the most specific classes of the partial order. We make the following hypothesis: the more specific products customers share, the closer they are. Assessment of the similarity of two customers is based on approximation over this product taxonomy (Sect. 3.1). Such reasoning clearly induces new semantic clustering techniques of customers based on the hierarchical taxonomy. Some approaches also consider a taxonomy [2,23], but do not consider nor semantic measures nor information content to define compared sets of items. Thereby, this study proposes a semantic clustering approaches where customers are considered as digraphs of product classes and then clustered. The rest of the paper is organized as follows. In Sect. 2 we address related works concerning customer segmentation. Section 3 introduces preliminary definitions to carry out our semantical clustering proposal.

2 Related Works on Customer Segmentation

In the retail world, the clustering of customers, usually called 'Customer Segmentation', consists in dividing heterogeneous customers groups based on common attributes. It requires to handle a large variety of customers [9,10]. Usually, the following kinds of data are considered: (a) Demographic data (e.g. gender, age, marital status); (b) Psychographic data (e.g. social class, lifestyle and personal characteristics); (c) Geographic data (e.g. area of residence or work); (d) Attitudinal Data (e.g. perceived data gathered from surveys); etc. Many different segmentation approaches are also applied. To name a few, RFM Recency, Frequency, and Monetary or CLV Customer Lifetime Value criteria are mostly used [3–5], for clustering [6], classification (e.g. neural-networks, decision trees) [4], models based on associations (association rules, Markov chain) [4], sequence discovery, forecasting [5]. Other researches focus on the mix of the products or product categories that customers have bought in their whole purchase history [11]. Even if our goal is also customer segmentation, the data used are the products bought by the clients and our approach is driven by the product classes.

3 Preliminary Definitions and Proposal

3.1 Hierarchical Abstraction and Taxonomies

Our main objective is to provide interpretable clusters of customers using similarity measure guided by the product organization. The similarity measure relies on the taxonomical structure of concepts. A taxonomical structure defines a partial order of the key concepts of a domain by generalizing and specializing relationships between concepts. Taxonomies give access to consensual abstraction of concepts with hierarchical relationships, *e.g.*, *Vegetables* defines a class or concept that includes *beans*, *leeks*, *carots* and so on, that are more specific concepts. Taxonomies are central components of a large variety of applications that rely on computer-processable domain expert knowledge, e.g. medical information and clinical decision support systems [8,12]. They are largely used in Artificial Intelligence systems, Information Retrieval, Computational Linguistics... [13]. In our case, customers clustering will be based on product taxonomy that defines a partially ordered set (poset) of products. An example of partial product taxonomy is shown in Fig. 1. In retail world, product taxonomy can be achieved by different means. Retailers or other experts can build this commodity structure. Most approaches usually introduce the Stock Keeping Unit (SKU) per item [14] or product categories (e.g. *Meat*, *Vegetables*, *Drinks*, etc.). Some researchers adopt the cross-category level indicated by domain experts and/or marketers [15].

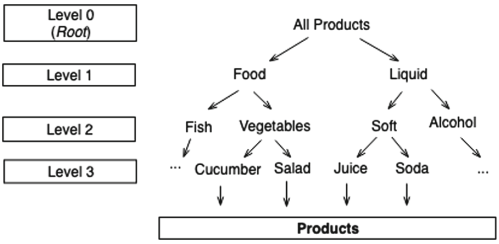

Fig. 1. Example of partial product taxonomy

3.2 Similarity and Informativeness Based on Taxonomy

Thanks to a product taxonomy, semantic similarity measures can be applied to define *similarity/dissimilarity* between customers. Those similarities compare sets of concepts associated to customers with *groupwise* measures [7]. These measures are themselves based on *pairwise* measures allowing the calculation of similarity between two concepts of the taxonomy [8]. Some of *pairwise* measures required Information Content (IC) associated to concepts, that is the amount of information associated to a concept (more a concept is specific, higher the information content of the concept is). There are several *groupwise* measures, in the same way there are several *pairwise* measures and several definitions of IC.

Groupwise measures allow comparison of sets of concepts related to objects (customers). There are two main categories: Direct and Indirect measure. Direct *groupwise* measures compare sets of concepts irrespective of their position in the taxonomy (*i.e.* Jaccard). Indirect *groupwise* measures aggregate similarity between concepts achieved by *pairwise* measures (*i.e.* BMA which stands for Best Match Average [16]). State-of-the-art approaches are divided in two kinds of *pairwise*: the first one based on the Information Content (IC) (*i.e.* Resnik *pairwise* measure [17]) and the other one based on the shortest path in the taxonomy (*i.e.* Wu & Palmer *pairwise* measure [18]). Finally, we discern two kinds of Information Content (IC): intrinsic and extrinsic. Intrinsic IC (*i.e.* Seco IC [19]) take only topological properties of the taxonomy into consideration while Intrinsic IC (*i.e.* Resnik IC [17]) used in addition frequency of concept in a observation bases (*i.e.* an ordinary corpus/data-set).

3.3 Methodology

Our goal is to identify sets of customers (*clusters*) with similar purchase behaviors. The main issue is to define appropriate similarity between customers. In practice, all customers' baskets are different whereas we intuitively make the difference between the two following customers: customer that mainly buys food product and customer who only purchases household laundry, cleaning products and textile.

Thanks to the use of the product taxonomy to compare customers, we obtain semantic clusters that are more interpretable by retailers. To this end, we used SSM with the Best Match Average [20] for the *groupwise* measure combined with the Resnik *pairwise* measure and Resnik IC [17]. This combination provides a matrix of *similarity/dissimilarity* between customers. Finally, this allows us to perform Hierarchical Clustering (HC) with Ward's method, which minimize distance inside each cluster (intra-clusters) while maximize distance between clusters (inter-clusters) [21]. The study tends to propose an approach of semantic clusters of customers close to the human intuition. In the following section, we present the application of semantic clustering on a real case. For more information about Semantic Similarity Measure, the reader is invited to refer to Harispe et al. studies ([8]).

4 Experiments and Results

Experiments have been conducted with real world data from one store, located in Paris (France) with 32 500 sales for 1 025 customers over a month. Table 1 describes some statistics about the dataset.

It contains data with different perspective: an overall vision, an average vision per customer and an average per sales receipt. For example, we can notice that the 1025 customers (fidelity cards) bought in average 25.36 different product categories (*Level 1 on the product taxonomy*). In other words, each customer probably purchase 25 different kinds of products. As explained above, to gather

Table 1. Some statistics about the dataset

	Number	Average per customer	Average per sale receipt
Number of customers	1025	-	-
Number of sale receipt	3692	5.75	-
Number of product bought	32552	48.54	8.63
Number of different categories	692	25.36	6.78
Number of different products	4979	31.70	7.47
Revenue	66590	102.45	18.22

customers, we used HAC method that cluster set of customers two by two, depending on their similarity. The key feature of the HAC is the identification of the ideal number of cluster. For the experiments, we used Ward's method besides a defined limit. Indeed, ideal clusters could be one customer per cluster but retailers try to identify a limited set of clusters. They try to figure it out "fictitious customers" that stand for as many customers as possible without loosing capital information. So, we varied the number of clusters from 2 to 25 and analyzed which number of clusters minimize the intra-class inertia and maximize the inter-class inertia. For the experiment, we obtained an optimal number of **4 clusters**, nominated respectively "C1" to "C4".

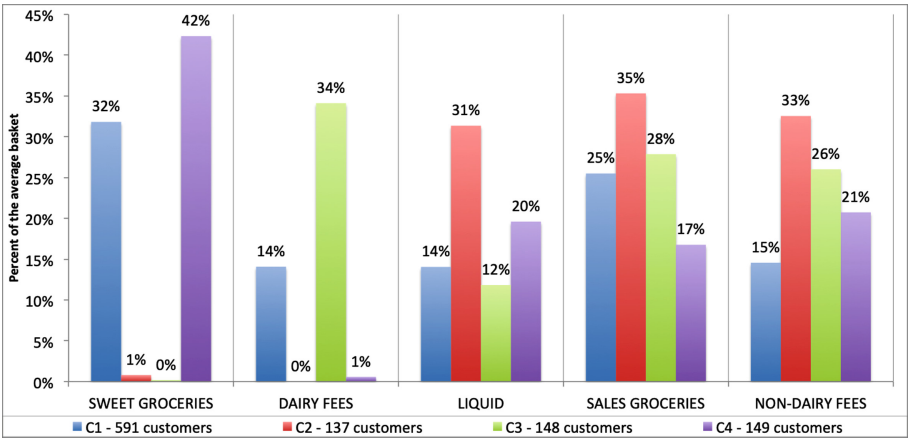

Fig. 2. Purchase frequency per product categories and clusters

The Fig. 2 represents the average frequency per product category level 1 (cf. Fig. 1) and per cluster. This information corresponds to the average basket of customers for each cluster. First of all, the analysis of the clusters' specificities underline that more than half of customers (591 customers) are gather in the

cluster "C1". It seems that people from this "set of customers" came to purchase equally likely products from all different categories (*Level 1 on the product taxonomy*). The three others clusters "C2", "C3", "C4" have respectively 137, 148 and 149 customers which is approximately the same order of magnitude (≈15%). However, each of them has its own particularity. Customers from clusters "C2" mainly purchase products from *LIQUID*, *SALES GROCERIES* and *NON-DAIRY FEES*. This is the cluster that only purchase products from three categories (*Level 1 on the product taxonomy*). Clusters "C3" and "C4", for their part, are opposed by categories *SWEET GROCERIES* and *DAIRY FEES*. Indeed, customers from cluster "C3" will mainly purchase *DAIRY FEES* products while customers from cluster "C4" will mainly purchase *SWEET GROCERIES* products. We can notice that customers from those clusters ("C3" and "C4") will never purchase product from the discriminant category of the "opposite" cluster. Note that those customers, from clusters "C3" and "C4" shared their purchase between four categories Level 1 on the product taxonomy.

To make cluster more understandable by retailers, we give a label to each cluster. We used the most discriminant product categories to label clusters "C3" and "C4". Thereby, they have respectively the label "*DAIRY FEES* Customers" and "*SWEET GROCERIES* Customers". Specificities of Cluster "C2" came from the lack of purchase in previous category used. That's why we agreed to label it as "**N0**-*DAIRY FEES* & **N0**-*SWEET GROCERIES* Customers". Finally, cluster "C1" does not have any specific category. That's why we labeled it as "*ALL PRODUCTS* Customers".

5 Conclusion

The aim of the paper is to identify similar customers depending on their purchase behavior. Semantic clustering is based on the product taxonomy and should brings results more understandable for retailers. This approach allows retailers the identification of abstract purchase behavior (*i.e.*) thanks to the taxonomy. The final goal of customers' segmentation (or customer clustering) is to identify "good" customers based on retailers' preferences. With this semantic approaches results are based on their own business what underline the added value of the proposed approach.

Note that we used HAC to cluster customers, but other clustering methods and/or other configurations of Semantic Similarity Measures (IC, Pairwise, Groupwise) could be used. In this paper, our objective was to introduce semantic approach in retail. We believe that a comparison of clusters from different stores may allow improvement by defining generic "pattern" of customers. To go further, we can suppose that clusters' trajectory analyses could be done to identify changes in customers' behaviors. This will allow stores to be more preventive than reactive towards new trends (i.e. Vegan Customers). After all, if retailers validate resulting clusters, classification methods can be used to associate any new customers into thus clusters [22].

References

1. Berrah, L., Mauris, G., Montmain, J.: Monitoring the improvement of an overall industrial performance based on a Choquet integral aggregation. Int. J. Manag. Sci. OMEGA **36**, 340–351 (2008)
2. Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D.: Retail business analytics: customer visit segmentation using market basket data. Expert Syst. Appl. **16**(1), 1–16 (2018)
3. Ching-Hsue, C., You-Shyang, C.: Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst. Appl. **36**(3), 4176–4184 (2009)
4. Chen, Y.L., Kuo, M.H., Wu, S.Y., Tang, K.: Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electron. Commer. Res. Appl. **8**(5), 241–251 (2009)
5. Khajvand, M., Zolfaghar, K., Ashoori, S., Alizadeh, S.: Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study. Procedia Comput. Sci. **3**, 57–63 (2011)
6. Lingras, P., Elagamy, A., Ammar, A., Elouedi, Z.: Iterative meta-clustering through granular hierarchy of supermarket customers and products. Inf. Sci. **257**, 14–31 (2014)
7. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: Semantic similarity from natural language and ontology analysis. Synthesis Lectures on Human Language Technologies (2015)
8. Harispe, S., Snchez, D., Ranwez, S., Janaqi, S., Montmain, J.: A framework for unifying ontology-based semantic similarity. Study in the biomedical domain. J. Biomed. Inform. **48**, 38–53 (2014)
9. Hong, T., Kim, E.: Segmenting customers in online stores based on factors that affect the customers intention to purchase. Expert Syst. Appl. **39**(2), 2127–2131 (2012)
10. Aeron, H., Kumar, A., Moorthy, J.: Data mining framework for customer lifetime value-based segmentation. Expert Syst. Appl. **19**(1), 17–30 (2012)
11. Park, C.H., Park, Y.H., Schweidel, D.A.: A multi-category customer base analysis. Int. J. Res. Mark. **31**(3), 266–279 (2014)
12. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics **30**, 740–742 (2014)
13. Harispe, S., Imoussaten, A., Trousset, F., Montmain, J.: On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2015)
14. Kim, H.K., Kim, J.K., Chen, Q.Y.: A product network analysis for extending the market basket analysis. Expert Syst. Appl. **39**(8), 7403–7410 (2012)
15. lbadvi, A., Shahbazi, M.: A hybrid recommendation technique based on product category attributes. Expert Syst. Appl. **36**(9), 11480–11488 (2009)
16. Pesquita, C., Faria, D., Bastos, H., Falcao, A., Couto, F.: Evaluating go-based semantic similarity measures. In: Proceedings of 10th Annual Bio-Ontologies Meeting, vol. 37, p. 38 (2007)
17. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of IJCAI 1995, pp. 448–453 (1995)
18. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133–138 (1994)

19. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In: 16th European Conference on Artificial Intelligence, pp. 1–5 (2004)
20. Schlicker, A., Domingues, F.S., Rahnenfhrer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinform. **7**, 302 (2006)
21. Murtagh, F.: Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? J. Classif. **31**, 274–295 (2014)
22. Silla Jr., C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. **22**, 31–72 (2011)
23. Srikant, R., Agrawal, R.: Mining generalized association rules. Future Gener. Comput. Syst. **13**, 161–180 (1997)