Application of data mining and complex networks in the representation of purchasing associations: a case study in supermarket purchases

Maicon Lima[1], **Melque Henrique Castro**[1], **Thiago Leite**[1], **Douglas Cordeiro**[2], **Núbia Rosa Da Silva**[1]

[1]Institute of Biotechnology (IBiotec), Federal University of Goiás (UFG)
Av. Dr. Lamartine Pinto de Avelar, 1120, 75704-020, Catalão, Brazil.

[2]Faculty of Information and Communication, Federal University of Goiás
Campus Samambaia, 74690-900, Goiânia, Brazil.

{maicon_lima,melque.henrique,thiago_leite}@discente.ufg.br

{cordeiro,nubia}@ufg.br

Abstract. *The development of solutions for the analysis of purchasing associations and consumer behavior is of great interest. One of the main challenges is to provide assertive results, with high accuracy, which enable the generation of strategic information to aid decision making. This paper presents an analysis of a supermarket sales data set by applying a hybrid solution based on data mining and complex networks. The results achieved reveal the potential of using complex networks in the information generation process.*

1. Introduction

The development of analytical solutions based on data exploration to support business intelligence is necessary and fundamental in a Big Data scenario. Processes related to the inference of purchasing associations are of great interest, since they allow the generation of strategic information that more assertively supports decision-making in marketing projects related to consumer behavior. Several studies present strategies for the use of data mining to analyze consumer behavior from purchasing data [Setiawan et al. 2017, Gull and Pervaiz 2018].

In this context, the use of complex networks can assist in enhancing the quality of the information generated. In [Zanin et al. 2016], ways of correlating complex networks with data mining are presented. The work shows that data mining techniques can be used to identify nodes and important connections in the system. The representation of data through complex networks, provides a wealth of quantitative variables in which systems can be described.

The idea of analyzing the retail market as a complex system provides results that complement those obtained through methods specifically of data mining. Several studies use data mining and association tasks to identify purchasing patterns. In [Fontanella 2010] a proposal is presented for the use of data mining to identify patterns of purchases made in a supermarket for a period of twelve months. The results presented demonstrate the potential of using data mining to identify sectors with the highest sales volume, as well as the most relevant associations between the products sold. One of the problems is the

limitation of the proposal in associations of only two items, as well as the identification of multiple associations in the same instantiation of data. Although the literature presents solutions for this type of problem based on the use of data mining, the use of complex networks is an alternative that allows the achievement of several advantages, such as the calculation of specific measures, as well as data visualization. Based on that, this paper aims to present a hybrid solution, based on data mining, following the proposal of [Fontanella 2010] and the use of complex networks, in order to generate a more targeted analysis on consumer behavior in purchasing association terms.

2. Metodology

For the experiments, the same dataset explored by [Fontanella 2010] (304,522 transactions) was considered. The output obtained through the application of the association rules solution was then used for the modeling of networks, where the nodes represent the products, and the edges the associations between them. The generated networks allow the generation of analyzes that contemplate the possibilities of association between all products. In addition, they also provide the calculation of measures that promote a better understanding of consumer behavior in relation to their purchases. For visualization, the Gephi[1] software was used. The edges thickness were defined according to the support value of each association, that is, the number of times that an association occurred in transactions. The description of product names in Portuguese has not been changed.

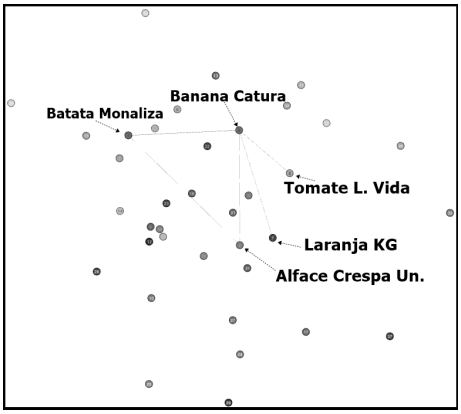
3. Results

Once the methodological steps were applied, visualizations of complex networks were obtained (Figure 1). From these networks, it is possible to observe the associative variations of bimonthly sales for the period considered. We show the associations between products in the fruit and vegetable sector, which observed the most relevant associations among all those observed in the period under analysis. During the first two months (Figure 1(a)), it can be seen that the greatest association found was *Batata Monaliza KG* and *Banana Catura KG* with 714, followed by *Alface Crespa UN.* and *Banana Catura KG* with 565 occurrences, *Tomato L. Vida KG* and *Banana Catura KG* with 537 associations, *Laranja KG* and *Banana Catura KG* 536, *Alface Crespa UN.* and *Batata Monaliza KG* with 522.

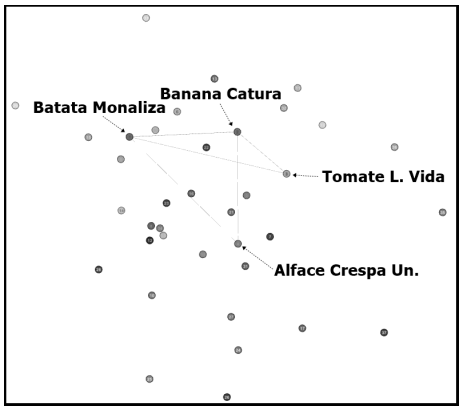
In Figure 1(b) it is possible to observe changes in the network edges, showing that in the second two months customers stopped buying some associated products. The product *Laranja KG* is no longer associated with other products with a lot of occurrence, as shown in the previous two months. The third bimester (Figure 1(c)) shows the appearance of the product *Sal Moc 1 KG* as one of the products most associated with the product *Banana Catura KG* with 412 occurrences. The fourth bimester (Figure 1(d)) has as its particular feature, the appearance of the product *Erva Verdland 1 KG* as one of the products most associated with the purchase of *Batata Monaliza KG*. This association occurred 432 times among all sales in this period. The fifth bimester (Figure 1(e)) showed changes in the support values of the most relevant associations of this period and the product *Sal Moc 1 KG* is no longer highly associated with the product *Batata Monaliza KG* as in the previous two months.

[1]https://gephi.org/.

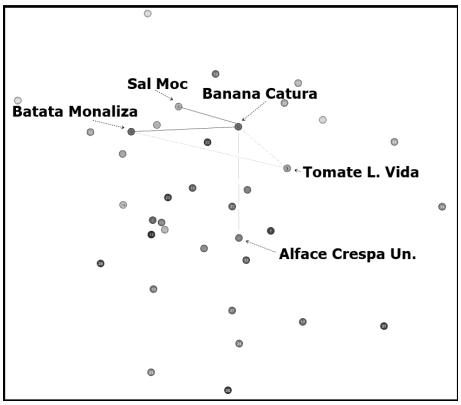
Finally, the sixth bimester (Figure 1(f)) presents a change in relation to all the previous ones. In this period, the presence of the product *Melancia Especial KG*, associated with the product *Banana Catura KG*, is noted as being the second largest association in this two-month period with support of 548. The appearance of this product can be justified by the month of December being marked by all the end festivities year and fruit consumption by customers is culturally common in Brazil.



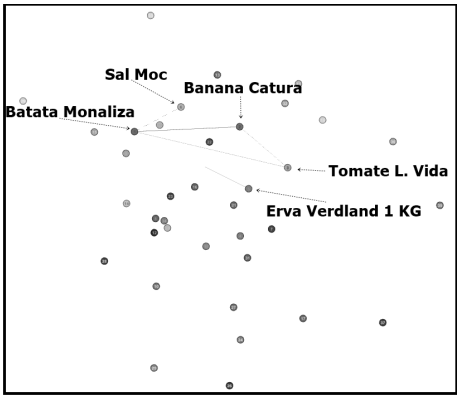
(a) Bimonthly 1.



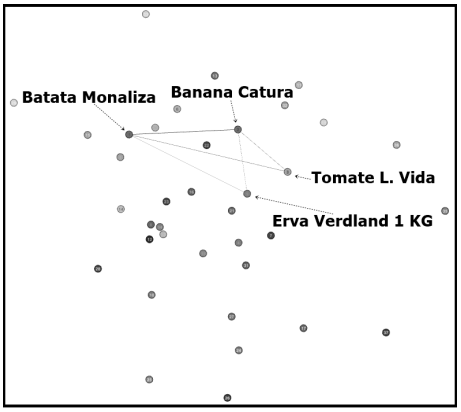
(b) Bimonthly 2.



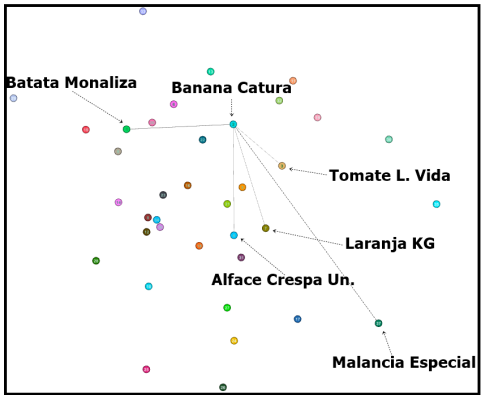
(c) Bimonthly 3.



(d) Bimonthly 4.



(e) Bimonthly 5.



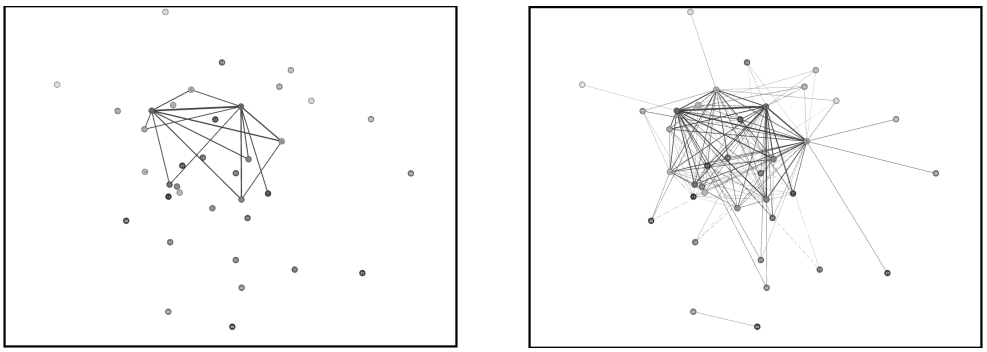
(f) Bimonthly 6.

Figure 1. Bimonthly networks.

The most common association is between the products *Batata Monaliza KG* and

Banana Catura KG, the only change in relation to this association is due to the variation of the support value depending on the analyzed period. Climate change can lead to changes in the price of these products, and thus directly influencing their sales. The networks represent the variations of these support values through the their edges thickness.

The product *Melancia Especial KG* does not stand out in this chain (Figure 2(a)), because no association in relation to this product has relevant support, this is justified due to the seasonality of the product, since there is a greater consumption of this in the last two months of the year. Figure 2(b) shows the Network that represents most the associations considered most relevant during the analysis period.



(a) Relevance graph. (b) Complete graph.

Figure 2. Relevant association networks.

4. Conclusions

From the behavioral simulation of the network it was possible to obtain relevant data on consumption trends, including seasonality issues, in addition to common purchasing associations. This information is useful for the strategic business sector. In conjunction with the use of data mining, complex networks proved to be a potential solution for providing information structuring, providing a wealth of quantitative variables, through which systems can be described.

References

Fontanella, P. (2010). Associações de compra em supermercado utilizando o data mining. Dissertação de mestrado, Universidade Federal do Paraná, Curitiba.

Gull, M. and Pervaiz, A. (2018). Customer behavior analysis towards online shopping using data mining. In *2018 5th International Multi-Topic ICT Conference (IMTIC)*, pages 1–5.

Setiawan, A., Budhi, G. S., Setiabudi, D. H., and Djunaidy, R. (2017). Data mining applications for sales information system using market basket analysis on stationery company. In *2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)*, pages 337–340.

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., and Boccaletti, S. (2016). Combining complex networks and data mining: why and how. *Physics Reports*, 635:1–44.