# Análise de Situações de Canibalização de Produtos em Sistemas de Retalho

## *Situation Analysis of Product Cannibalization in Retail Systems*

Rui Castro e Orlando Belo

Centro de I&D ALGORITMI, Escola de Engenharia, Universidade do Minho

Campus de Gualtar, 4710-057 Braga, PORTUGAL

ruipcastro@gmail.com, obelo@di.uminho.pt

### Resumo

Neste trabalho procuramos avaliar o desempenho de dois algoritmos de mineração de dados orientados especialmente para o estabelecimento de regras de associação positivas e negativas. Para isso utilizámos um conjunto de dados de um retalhista, relativos às vendas realizadas durante um dado período. Os resultados alcançados neste trabalho de análise e comparação revelou um grande potencial de aplicação deste tipo de soluções no sector do retalho, bem como evidenciou que a utilização do conhecimento adquirido com base num conjunto de regras de associação positivas e regras de associação negativas, apesar da sua grande diferença, quando conjugado, constitui um grande fator de diferenciação em termos de qualidade na personalização de clientes, apoiadas por relacionamentos entre produtos estabelecidas numa venda.

**Palavras-chave:** Técnicas de Mineração de dados, Regras de Associação Positivas, Regras de Associação Negativas, Canibalização de Produtos.

### *Abstract*

*In this work we evaluate the performance of two data mining algorithms geared especially for the establishment of positive and negative association rules. For this we used a fairly comprehensive set of data from a retailer about the sales of products of a given period. The results achieved in this analysis and comparison work revealed a great potential for this type of solutions in the retail sector, where the use of knowledge acquired based on a set of positive association rules and negative association rules, despite its big difference, when combined is a major differentiating factor in terms of quality for customer personalization based on product relationships.*

***Keywords:*** *Data Mining Techniques, Positive Association Rules, Negative Association Rules, Product Cannabilization.*

## 1. INTRODUÇÃO

Muitas das aplicações desenvolvidas suportadas por regras de associação em sistemas convencionais de retalho têm como objetivo preverem, de alguma forma, o comportamento futuro dos clientes nas lojas, de forma a se estabelecer, com algum grau de certeza, os produtos que ele vão adquirir com base no comportamento que revelaram no passado. Um dos conceitos mais utilizados para fazer previsão de vendas de um dado produto é o *Halo Effect* [Rosenzweig 2011], isto é, o impacto positivo que as vendas acrescidas de um produto vão ter sobre um segundo produto. Essa influência pode, também, ser negativa. Quando esta acontece, usualmente, é designada por canibalismo. Agrawal et al. [2015] definiram esta situação como:

*"Negative rules consider items that conflict with each other. In other words, negatives rules are used to represent that if product A is purchased, then product B will not be purchased"*. Hoje a previsão de vendas é uma ferramenta importantíssima em sistemas de retalho, uma vez que permite prever os padrões de compra de clientes e, assim, entender melhor as suas intenções comerciais a médio e longo prazo, bem como fazer uma melhor gestão de stocks ou realizar melhores e mais eficazes campanhas promocionais [Fayyad et al. 1996].

Neste artigo analisámos o impacto das regras de associação negativas na previsão das vendas de produtos em sistemas de retalho, um pouco à semelhança do que acontece com a aplicação de regras de associação positivas nesse tipo de cenários. Além disso, pretendemos obter um conjunto de regras de associação, positivas e negativas, sobre as vendas de produtos realizadas que pudessem revelar o que acontece com as vendas de produtos concorrentes, situações que são usualmente referidas como de "*halo effect*" para associações positivas e de "canibalismo" para associações negativas. A justificação para a realização deste estudo prendeu-se com o desenvolvimento de um sistema que permitisse suportar essas análise num dado domínio de negócio. Nas secções seguintes, apresentaremos as ferramentas e os algoritmos selecionados para este trabalho (secção 2), o caso de estudo selecionado (secção 3) e uma breve análise sobre alguns dos resultados que obtivemos (secção 4). Por fim, na última secção, apresentaremos algumas breves conclusões e algumas linhas de orientação para trabalho futuro.

## 2. CAPTURA E ANÁLISE DE REGRAS DE ASSOCIAÇÃO POSITIVAS E NEGATIVAS

Até há bem pouco tempo, os algoritmos de regras de associação disponíveis apenas eram capazes de extrair regras de associação positivas, sem serem capazes de prestarem grande atenção às regras de associação negativas [Silverstein et al. 1998]. Estas últimas analisam precisamente os mesmos *itemsets* analisados pelas regras de associação positivas, com o acréscimo de analisarem também a negação de um *itemset*, isto é, a sua inexistência. Se aplicarmos a aplicação das regras de associação negativas a uma lógica de retalho, poderemos ter casos como os seguintes [Brin et al. 1997]: uma situação na qual a compra do produto X implica que o produto Y não seja comprado, a ocorrência de uma situação na qual a não compra do produto X implica a compra do produto Y, ou uma situação na qual a não compra do produto X implica que o produto Y também não seja comprado. A seleção de um algoritmo para análise de regras de associação positivas (*halo effect*) e de um outro para análise de regras de associação negativas (canibalismo) foi desde o início deste trabalho uma das principais ações estabelecidas na linha de investigação traçada. Basicamente, quisemos identificar um par de algoritmos que nos pudesse suportar um processo de previsão envolvendo a análise dos efeitos de negócio revelados por um conjunto de regras de associação positivas e negativas. Depois de um estudo sobre vários algoritmos, para capturar e analisar regras de associação positivas optámos pela solução oferecida pelo RapidMiner [RapidMiner n.d.], que para esse tipo de aplicações usa o algoritmo FP-Growth [Han et al. 2000], e para capturar e analisar regras de associação negativas, selecionámos o

algoritmo MOPNAR [Martin et al. 2014], que para além de analisar regras de associação negativas, também analisa regras de associação positivas. Para suporte à execução do algoritmo MOPNAR utilizámos a ferramenta KEEL [Alcalá-Fdez et al. 2011].

## 3. ANÁLISE DE CASOS DE CANIBALIZAÇÃO DE PRODUTOS

No nosso estudo utilizámos dados fornecidos por um retalhista brasileiro, de grande dimensão, com várias lojas espalhadas pelo Brasil, relativos a um dado período das suas vendas. Nessa informação estavam incluídos dados relativos a produtos que foram vendidos durante um período de cerca de dois anos. No conjunto de dados inicial figuravam referências relativas a 76454 produtos diferentes, organizados por várias categorias. Os registos de vendas recebidos apresentam uma estrutura suportada por cinquenta e nove atributos, um nível de detalhe tão grande que nos permite aceder a dados como a semana, o dia ou a hora da venda, a identificação e a descrição do produto, ou a estrutura de mercado ao qual o produto pertence, entre muitos outros elementos. Para adequarmos os registos disponíveis ao nosso trabalho, aplicámos alguns critérios de seleção que produziram uma base de dados com os registos de todas as vendas realizadas dos produtos pertencentes às categorias selecionadas, num total de 2860049 registos de produtos vendidos, na qual figuram 860 produtos diferentes, agrupados em 83 categorias diferentes de nível 4, correspondendo a 1576179 carrinhos de compras diferentes, e contendo uma média de 1.81 produtos por carrinho de compras. Após a constituição desta base de dados procedemos à preparação dos dados para posterior mineração. Aqui tivemos que aplicar alguns mecanismos de limpeza de dados, devido à existência de algumas anomalias. Por exemplo, ao iniciarmos o nosso processo de análise dos dados deparámo-nos com vários registos de produtos repetidos, dentro do mesmo carrinho de compras.

| Algoritmo | Modelos | Caracterização | Registos |
|---|---|---|---|
| **FP-Growth** | R1 | Categoria de Produtos: Culinários | --- |
| | R1.1 | Categoria de Produtos: Doces | --- |
| | R1.2 | Categoria de Produtos: Sanduiches | --- |
| | (...) | (...) | (...) |
| | R2 | --- | 250000 |
| | R2.1 | --- | 500000 |
| | R2.2 | --- | 750000 |
| | (...) | (...) | (...) |
| **MOPNAR** | K1 | Categoria de Produtos: Culinários | --- |
| | K1.1 | Categoria de Produtos: Doces | --- |
| | K1.2 | Categoria de Produtos: Sanduiches | --- |
| | (...) | (...) | (...) |
| | K2 | Categoria de Produtos: Culinários (c/ data e hora) | --- |
| | K2.1 | Categoria de Produtos: Doces (c/ data e hora) | --- |
| | K2.2 | Categoria de Produtos: Sanduiches (c/ data e hora) | --- |
| | (...) | (...) | (...) |

Tabela 1 – Extrato da caracterização dos modelos criados.

Numa fase seguinte, projetámos e desenvolvemos vários modelos de mineração (Tabela 1), fazendo variar, por exemplo, os tipos de dados selecionados ou a sua dimensão. Dessa forma, criámos 31 modelos

diferentes, que organizámos em 6 grupos distintos. O primeiro grupo (R1) utilizou o algoritmo FP-Growth, contendo cada um dos seus modelos os produtos transacionados de uma categoria específica, nomeadamente, produtos culinários, doces, sanduiches, bolos, cafés, ovos e pão. O segundo grupo (R2) também utilizou o algoritmo FP-Growth. Porém, neste grupo cada um dos modelos contém uma quantidade de registos diferente, nomeadamente 250 mil registos, 500 mil registos, 750 mil registos, 1 milhão de registos, 1 milhão e 250 mil registos e, finalmente, a totalidade dos registos. Os grupos de modelos K1 e K2 utilizaram o algoritmo MOPNAR. São modelos bastante semelhantes ao grupo R1. Tal como os modelos R1, os modelos K1 e K2 analisam categorias de produtos diferentes, mas estes além dos registos das transações dos produtos contêm atributos relativos ao dia da semana e à hora a que os produtos foram transacionados. Apesar de não figurarem na Tabela 1, organizámos outros grupos de modelos, nomeadamente os grupos K3 e K4, que foram definidos com base no modelo K1.3 (produtos da categoria bolos, sem atributos de data). Os modelos K3 integraram um conjunto de produtos diferente, contendo apenas 15, 20, 25 ou 30 dos produtos da categoria de bolos, enquanto que o grupo de modelos K4, com uma lógica semelhante à dos modelos K3, consideraram, respetivamente, apenas 20%, 40%, 60% e 80% das transações de K1.3.

## 4. ANÁLISE DE RESULTADOS

O processo de captura das regras de associação, tanto positivas como negativas, não foi de difícil implementação. Porém, as regras de associação negativas não são de angariação simples. Vejamos, então, alguns dos resultados obtidos com o estudo que realizámos. Por exemplo, ao observarmos os dados apresentados na Tabela 2, verificamos que em algumas das categorias escolhidas o processo de aquisição de regras de associação não foi muito produtivo, mesmo com a utilização de valores baixos para o suporte (1%) e para a confiança (30%). Em particular, essa evidência verifica-se claramente ao nível dos modelos R1.1, R1.3, R1.5 e R1.6. Isto indica-nos que poderá não haver regras de associação entre os produtos destas categorias ou então que os dados analisados não são em número suficiente para que se possa encontrar regras de associação com algum peso significativo. O modelo que apresentou melhores valores de suporte foi o R1.4, para os produtos da categoria "Cafés", o que pode ser justificado por ser a categoria com a menor diversidade de produtos – esta categoria contém apenas 10 produtos.

| Modelos | Categoria | Registos | Transações | Produtos | Suporte | Confiança |
|---|---|---|---|---|---|---|
| **R1** | Culinários | 224 | 110 | 13 | 0.01 | 0.3 |
| **R1.1** | Doces | 4070 | 1814 | 29 | 0.01 | 0.3 |
| **R1.2** | Sanduiches | 2972 | 1477 | 15 | 0.01 | 0.3 |
| **R1.3** | Bolos | 13021 | 6110 | 30 | 0.01 | 0.3 |
| **R1.4** | Cafés | 16527 | 8196 | 10 | 0.01 | 0.3 |
| **R1.5** | Ovos | 13027 | 6418 | 21 | 0.01 | 0.3 |
| **R1.6** | Pães | 96108 | 43385 | 47 | 0.01 | 0.3 |

Tabela 2 - Resultados gerais dos modelos R2.

| Modelos | Categoria | Regras Obtidas | Suporte (m) | Confiança (m) | Tempo (s) |
|---|---|---|---|---|---|
| **K2** | Culinários | 45 | 0.23 | 0.61 | 11 |
| **K2.1** | Doces | 29 | 0.2 | 0.71 | 102 |
| **K2.2** | Sanduiches | 40 | 0.23 | 0.78 | 106 |
| **K2.3** | Bolos | 35 | 0.18 | 0.71 | 374 |
| **K2.4** | Cafés | 45 | 0.29 | 0.91 | 239 |

Tabela 3 - Resultados gerais dos modelos K2.

Para apresentarmos outro exemplo de resultados obtidos, escolhemos os resultados obtidos para os modelos K2 (Tabela 3). De referir que, os modelos K2 atuaram sobre os mesmos dados que os modelos K1 (Tabela 1), com a diferença de estarem a atuar sobre um outro conjunto de atributos (hora e dia da transação). A título de curiosidade, nestes modelos, uma das regras que se destacou foi a seguinte:

$$\text{“CAFE EXPRESSO C/LEITE MEDIO (LANCH)”} == \text{True} \Longrightarrow \text{“CAFE ANG ORGANICO UN”} == \text{False}$$

com um suporte de 0.65 e uma confiança de 1. Ao observarmos esta regra, facilmente vemos que sempre que o produto antecedente está no carrinho (`CAFE EXPRESSO C/LEITE MEDIO (LANCH)`) o produto consequente (`CAFE ANG ORGANICO UN`) nunca está presente.

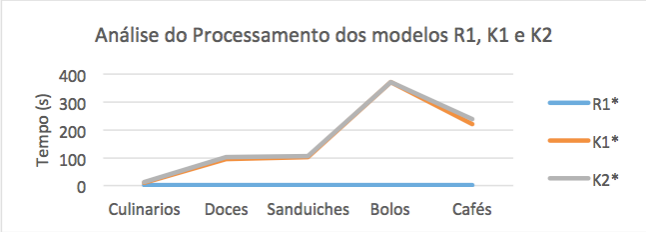


Figura 1 - Processamento dos modelos R1, K1 e K2.

Em termos de desempenho dos modelos produzidos, verificámos que os modelos R1, K1 e K2 apresentam o mesmo tipo de comportamento, uma vez que todos eles analisam os dados das mesmas categorias de produtos, nomeadamente “Culinários”, “Doces”, “Sanduiches”, “Bolos” e “Cafés” (Figura 1). Facilmente verificamos que os modelos R1 são mais rápidos do que os modelos K1 e K2 em todos os casos analisados. Isto deve-se ao facto dos modelos R1 analisarem apenas regras de associação positivas, que é, usualmente, um processo consideravelmente menos complexo e menos exigente do que o processo de analisar regras de associação negativas.

## 5. Conclusões e Trabalho Futuro

Com a realização deste trabalho pretendemos fazer a seleção de dois algoritmos de associação e avaliar o seu desempenho quando aplicados sobre uma base de dados de registos de compras reais de um retalhista, de forma a extrair um conjunto de regras de associação, tanto do tipo positivo como do tipo negativo. Os algoritmos escolhidos foram o FP-Growth e o MOPMAR. O primeiro foi utilizado na ferramenta

RapidMiner com o objetivo de extrair regras de associação positivas, enquanto que o segundo na ferramenta KEEL para extrair tanto regras de associação positivas como regras de associação negativas. Depois de realizarmos este trabalho, temos presente que alguns aspetos dos algoritmos utilizados podem ser estudados em iniciativas semelhantes num futuro próximo. Em particular, seria interessante alargar o âmbito do estudo realizado e abordar outras características da transação na análise das vendas realizadas, com vista a analisar as características da transação e também do próprio cliente.

## REFERÊNCIAS


Agrawal, J., Agrawal, S., Singhai, A., Sharma, S:, "SET-PSO-based approach for mining positive and negative association rules". *Knowledge and Information Systems*, Volume 45, Issue 2, pp. 453-471, 2015.

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F., "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework". *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3), pp. 255–287, 2011.

Brin, S., Motwani, R., Silverstein, C., "Beyond Market Baskets: Generalizing Association Rules to Correlations". *SIGMOD Rec.*, 26(2), pp. 265–276, 1997.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., "From data mining to knowledge discovery in databases". *AI magazine*, 17(3), pp.37–54, 1996.

Han, J., Pei, J., Yin, Y., "Mining Frequent Patterns Without Candidate Generation". *SIGMOD Rec.*, 29(2), pp. 1–12, 2000.

Martín, D., Rosete, A., Alcalá-Fdez,J., Herrera, F., "A New Multi-Objective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules". *IEEE Transactions on Evolutionary Computation*, 18(1), pp. 54–69, 2014.

RapidMiner, "The core of RapidMiner is open source". Available at: https://rapidminer.com/the-core-of-rapidminer-is-open-source/ [Accessed October 12, 2015].

Rosenzweig, P., "The Halo Effect … and the Eight Other Business Delusions that Deceive Managers", Free Press, 2007.

Silverstein, C., Brin, S., Motwani, R., "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules". *Data Min. Knowl. Discov.*, 2(1), pp. 39–68, 1998.