Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Market Basket Analysis e aplicação de Regras de Associação Hierárquica: um caso de estudo numa empresa de retalho portuguesa

Maria Alexandra Rodrigues Fonseca Baptista Verde

Orientada por:

João Manuel Portela da Gama

Carlos Gomes Ferreira

Setembro de 2016

Nota Biográfica

Alexandra Fonseca nasceu a 22 de Junho de 1991.

Minhota, natural de Caminha, foi desde cedo influenciada pelas tradições e costumes da região desenvolvendo uma aptidão para comunicar com os outros e tomando o gosto pela arte, história e economia local.

O seu avô, José da Costa Fonseca, advogado e diretor do Externato de Santa Rita, foi o grande impulsionador da sua paixão pelo estudo e vontade de querer conhecer e saber sempre mais.

Em 2009 completa o seu estudo secundário na Ancorensis Cooperativa de Ensino, na área de Ciências Socioeconómicas e ingressa na Faculdade de Economia da Universidade do Porto para aprofundar os seus estudos em Economia, uma área onde reconhece o cerne de todas as questões do dia-a-dia e onde pretende seguir carreira.

Em 2014 inicia a sua carreira profissional na área comercial de uma grande empresa de retalho portuguesa e ingressa no Mestrado em Modelação, Análise de Dados e Sistemas de apoio à Decisão com o objetivo de aprofundar conhecimentos analíticos através de técnicas inovadoras propensas à aplicação prática em casos reais.

Agradecimentos

Ter uma ideia é simples, pensamos todos os dias, estamos programados para isso.

Ter uma ideia inovadora é relativamente simples, basta visão e alguma criatividade.

Transformar a ideia inovadora num estudo com resultados práticos aplicados a um caso real já não é assim tão simples e, acrescida a responsabilidade de não só ser o fim de um ciclo académico como também representar um compromisso profissional, foi apenas possível pela influência, apoio e compreensão de algumas pessoas.

Em primeiro lugar aos meus avós a quem devo o que sou hoje pela educação e carinho com que me criaram e à minha irmã que, apesar de mais nova, é um exemplo de luta e superação todos os dias.

Aos meus orientadores, Professor Doutor João Gama e Professor Doutor Carlos Ferreira por todo o apoio, disponibilidade e compreensão quando as palavras não saíam e suporte quando o sistema não avançava. Obrigada por me ajudarem a combater a alergia à programação e demonstrarem que um economista também pode ter uma veia de engenharia, ainda que muito pequenina.

Ao Gonçalo e à Bárbara por acreditarem, bem antes de mim, que esta não era uma missão impossível e por me darem espaço sem nunca deixarem de me apoiar em todos os projetos em que me envolvo.

Aos meus amigos pela fonte de energia positiva, paciência e apoio incondicional.

À minha equipa, em particular à Mariana e à Vânia, pelo interesse e apoio desde o início, pela compreensão do meu estado físico e psicológico e por, na arte do malabarismo, nunca me deixarem fazer cair nenhum projeto.

Por último e mais importante, dedico o culminar do meu percurso académico à pessoa que o tornou uma realidade por me ter mostrado que não há impossíveis e que somos capazes de fazer tudo aquilo que nos propomos se tivermos as doses certas de motivação, perseverança e, sobretudo, coragem.

Obrigada Mãe.

Abstract

The present work studies the application of Hierarchical Association Rules in a retail company in Portugal.

Since each article is associated with a certain structure, this study was developed based on the existing hierarchy instead of an analysis at the level of the article in order to find the most common product types and derive association rules between them.

By applying Data Mining techniques, mainly Market Basket Analysis, it was explored a large volume of transactions in order to find common patterns of consumption.

This multilevel analysis provided a different view and culminated in very relevant and practical applications for business.

The knowledge discovered in the databases was explored and applied through the revision of the store layout, in order to bring products close together, and marketing techniques, through the construction of a campaign proposal tool based on the most relevant association rules, with the objective of increasing sales and improving costumer experience.

Keywords: *Data Mining; Association Rules with Taxonomies; Market Basket Analysis; Case Study*

Resumo

O presente trabalho estuda a aplicação de Regras de Associação Hierárquica numa empresa de retalho em Portugal.

Uma vez que cada artigo tem associado determinada estrutura, o estudo foi desenvolvido com base na hierarquia existente em detrimento de uma análise ao nível do artigo por forma a encontrar as tipologias de produto mais frequentes e derivar regras de associação entre elas.

Através da aplicação de técnicas de *Data Mining*, nomeadamente na área da *Market Basket Analysis*, foi explorado um grande volume de transações por forma a encontrar padrões frequentes de consumo.

Esta análise multinível permitiu uma visão diferente e culminou em aplicações práticas bastante relevantes para o negócio. Foi explorada a aplicação dos conhecimentos obtidos na revisão de *layout* de loja, por forma a aproximar produtos relacionados, e técnicas de marketing através da descrição de uma ferramenta de proposta de campanhas que tem por base as regras de associação mais relevantes e como objetivo a potenciação de vendas e atração do cliente.

Palavras-chave: *Data Mining; Regras de Associação Hierárquica; Market Basket Analysis; Case Study*

Índice

Nota Biográfica ii
Agradecimentos iii
Abstract iv
Resumo v
Índice de Figuras vii
Índice de Equações ix
1. Introdução 1
1.1. Definição do Tema 2
1.2. Motivação 3
1.3. Organização do Documento 4
2. Revisão de Literatura 5
2.1. Market Basket Analysis 5
2.2. Regras de Associação Hierárquica 12
3. Estudo do Caso 18
3.1. Descrição dos Dados 19
3.2. Tratamento e Leitura de Dados 21
3.3. *Itemsets* Frequentes e Regras de Associação 22
3.3.1. Definição do Valor dos Parâmetros 23
3.3.2. *Itemsets* Frequentes para cada Nível Hierárquico 25
3.3.3. Regras de Associação para cada Nível Hierárquico 29
3.4. Regras de Associação Hierárquica 40
3.4.1. Visualização do *Output* 47
4. Aplicações Práticas 53
4.1. Análise de *Layouts* de Loja 53
4.2. Ferramenta de Proposta de Campanhas 55
5. Conclusão 58
5.1. Limitações 60
5.2. Investigação Futura 61

Índice de Figuras

Figura 1: Base de dados com 5 transações 7
Figura 2: *Itemsets* frequentes das transações da figura 1 com suporte mínimo de 2 8
Figura 3: Regras de associação derivadas a partir dos *itemsets* frequentes da figura 2 9
Figura 4: Exemplo de hierarquia 13
Figura 5: Algoritmo Cumulate 15
Figura 6: Algoritmo Est Merge 16
Figura 7: Exemplo de estrutura de um artigo 19
Figura 8: Exemplo de duas transações da base de dados ao nível da categoria 20
Figura 9: Comando para leitura de dados com o *package arules* 22
Figura 10: Comando para identificação de *itemsets* frequentes e derivação de regras de associação 23
Figura 11: Teste à sensibilidade do parâmetro suporte para um nível de confiança de 0.001 24
Figura 12: Os 10 *itemsets* frequentes com maior suporte ao nível da categoria 25
Figura 13: Os 10 *itemsets* frequentes com maior suporte ao nível da subcategoria 26
Figura 14: Os 10 *itemsets* frequentes com maior suporte ao nível da unidade base 28
Figura 15: As 10 regras de associação com maior *lift* ao nível da categoria 30
Figura 16: As 10 regras de associação com menor *lift* ao nível da categoria 32
Figura 17: As 10 regras de associação com maior *lift* ao nível da subcategoria 33
Figura 18: As 10 regras de associação com menor *lift* ao nível da subcategoria 36
Figura 19: As 10 regras de associação com maior *lift* ao nível da unidade base 37
Figura 20: Teste à sensibilidade do parâmetro confiança para um nível de suporte de 0.01 41
Figura 21: Os 10 *itemsets* frequentes com maior suporte independentemente do nível da hierarquia 42
Figura 22: As 10 regras de associação hierárquica com maior *lift* para um suporte mínimo de 0.01 e uma confiança de 0.01 43
Figura 23: Regras de associação hierárquica para um suporte mínimo de 0.01 e uma confiança de 0.01 após eliminação de regras redundantes 44
Figura 24: Exemplo de regras de associação hierárquica redundantes 45

Figura 25: Regras de associação hierárquica relevantes para um suporte mínimo de 0.005 e um nível de confiança de 0.01 46
Figura 26: Comando para visualização das 20 estruturas presentes em mais regras 48
Figura 27: Representação das 20 estruturas presentes em mais regras 48
Figura 28: Comando para visualização das 50 regras de associação hierárquica com maior *lift* através de um gráfico de redes sociais 49
Figura 29: Representação das 50 regras de associação hierárquica com maior *lift* 50
Figura 30: Representação das 50 regras de associação hierárquica relevantes com maior *lift* 51

Índice de Equações

Equação 1: Confiança de uma regra de associação {X → Y} 9
Equação 2: *Lift* de uma regra de associação 10
Equação 3: Convicção de uma regra de associação 11

1. Introdução

O KDD (*Knowledge Discovery in Databases*) refere-se ao processo de aquisição de conhecimento relevante através de uma base de dados e implica vários passos, sendo *Data Mining* um deles.

Data Mining é uma técnica fundamental de extração de conhecimento de dados que, através da construção e análise de modelos de exploração, permite a identificação de padrões frequentes e a extração de informação útil e interessante de grandes conjuntos de dados (Silwattananusarn e Tuamsuk, 2012).

Data Mining responde a questões complexas que incidem em técnicas de consulta avançada, à descoberta automática de padrões, à criação de informação de fácil acesso e ao estudo de grandes volumes de dados através de algoritmos - composições matemáticas sofisticadas que permitem a segmentação de dados para avaliar a probabilidade de ocorrerem certos eventos futuros.

Durante muitos séculos a identificação de padrões relevantes foi feita de forma manual, algo que se tornou insustentável dada a quantidade de informação e crescente capacidade de armazenamento das bases de dados atuais, cada vez mais ricas e complexas. Assim, não só aumentou a importância de *Data Mining* como também a sua necessidade natural de evoluir apoiada em novos métodos computacionais, tais como redes neuronais, análise de *clusters*, algoritmos genéticos, árvores de decisão e regras de associação (Kantardzic, 2002).

A implementação das ferramentas de *Data Mining* pressupõe uma capacidade de análise abrangente através de um processamento rápido de grandes volumes de informação com o fim de explorar dados complexos e obter conclusões mais refinadas (Bastos, 2001).

De facto, as técnicas de extração de conhecimento de dados têm sido utilizadas com sucesso num grande número de problemas reais (Gama *et al.*, 2012). Neste trabalho será dado maior ênfase ao âmbito da *Market Basket Analysis* com foco nas Regras de Associação.

Considerada uma das áreas mais antigas de *Data Mining*, a *Market Basket Analysis* procura aplicar as relações relevantes encontradas em bases de dados de transações de retalho com o objetivo de potenciar vendas e atrair clientes (Raeder e Chawla, 2011).

De facto, a *Market Basket Analysis* constitui uma ferramenta extremamente importante no sistema de retalho organizacional que se foca nos cabazes de consumo dos clientes para monitorizar padrões de compra e potenciar a satisfação do cliente (Microstrategy, 2003), permitindo responder à questão principal que se coloca quando é analisado um conjunto de transações, "Que produtos são vendidos na mesma transação ou para o mesmo cliente?" e utiliza a capacidade de identificar padrões frequentes para potenciar vendas através de proposta de bens relacionados (Svetina e Zupančič, 2005).

O estudo em análise pressupõe a procura dos conjuntos de artigos mais frequentes derivando destes regras de associação que representam os padrões de compra dos clientes que, devido à dimensão das bases de dados, levam normalmente a um excesso de padrões encontrados.

Para combater este problema, e porque os artigos normalmente obedecem a uma determinada estrutura, surge uma variação das Regras de Associação: as Regras de Associação Hierárquica, que são extremamente importantes para a extração de conhecimento de dados, sendo frequentemente utilizadas em organizações comerciais, de saúde, em biologia molecular ou ligadas a outras áreas. As taxonomias (hierarquias) utilizadas refletem uma visão coletiva ou individual de como os artigos podem ser hierarquicamente classificados (Domingues e Rezende, 2011).

Em 1995, Srikant e Agrawal, estudam pela primeira vez as taxonomias para encontrar associações entre artigos de cada nível da sua hierarquia permitindo ainda afinar a pesquisa, eliminar regras menos interessantes e/ou redundantes e facilitar a identificação de padrões relevantes que tornam a interpretação dos resultados mais fácil e acessível.

Neste âmbito, e uma vez que os artigos em análise obedecem a uma determinada estrutura, serão estudadas as Regras de Associação Hierárquica.

1.1. Definição do Tema

Dentro das áreas e técnicas expostas acima, este trabalho propõe-se a desenvolver uma análise de padrões de consumo numa empresa portuguesa de retalho.

Através do estudo das Regras de Associação entre os artigos, e aproveitando a hierarquia existente para identificar as estruturas mais relacionadas entre si no momento de compra, pretende-se que a aplicação do conhecimento extraído neste caso prático

permita tirar conclusões relevantes que, quando aplicáveis, possam influenciar positivamente as vendas reais.

A base de dados em estudo conterá todas as transações de uma loja desportiva durante um ano, correspondendo cada transação a um determinado momento de compra em que são descriminados os artigos adquiridos e cada um destes, no seu código, contém intrínseca a estrutura hierárquica a que pertence.

Assim, uma vez que é possível identificar as estruturas hierárquicas compradas num mesmo momento, o estudo da base de dados, auxiliado pelas técnicas previamente descritas, tem como propósito identificar todas as estruturas compradas na mesma transação e retirar ilações sobre as tipologias de artigos que melhor se relacionam entre si.

O objetivo deste estudo passa por encontrar relações relevantes entre a compra de produtos de hierarquias diferentes considerando para isso a estrutura a que estão associados.

O conhecimento extraído permitirá desenvolver ações que potenciem as vendas através de duas áreas principais, organização de loja e mecânicas promocionais.

O estudo do *layout* de loja terá como fim o de promover a aproximação das tipologias de artigo com maior relação entre si enquanto que a construção de uma ferramenta de sugestão de campanhas permitirá propor mecânicas promocionais baseadas nas relações encontradas com o objetivo de atrair o cliente.

O estudo dos padrões de consumo para identificar os artigos com maior relação entre si permitirá ter uma visão dinâmica do comportamento dos clientes que colmatará numa aprendizagem contínua e preditiva do que será esperado do mercado. Esta abordagem facilitará o cumprimento do principal objetivo a que se propõe este trabalho: o de incrementar as vendas e melhorar a experiência do cliente em loja, o chamado *costumer engagement.*

1.2. Motivação

O tema em estudo surge da vontade de aplicar o conhecimento adquirido durante o percurso académico a casos reais.

Por um lado o Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão é bastante completo e rico quanto à variedade de temas que apresenta para estudo

tendo a disciplina *Data Mining* suscitado maior interesse pelo enorme potencial que apresenta quando se trata da sua aplicação a casos concretos da vida real.

Por outro lado, com a experiência obtida na área de retalho portuguesa é possível ter uma visão mais ampla e conhecer um vasto leque de áreas em que existem oportunidades para aplicação do conhecimento adquirido no Mestrado. Estas oportunidades são muito abrangentes uma vez que o mundo do retalho é bastante rico em áreas onde é necessário melhorar a tomada de decisões e antecipar ações das partes envolvidas, especialmente dos consumidores.

Assim, o presente trabalho permitirá não só aprofundar o estudo académico relacionado com as áreas de *Data Mining* e a sua aplicação prática, como também trará benefícios significativos para a empresa alvo de estudo, destacando-se o conhecimento adquirido acerca dos padrões de consumo dos clientes, a identificação das tipologias com maior presença nas transações e a derivação de relações de compra fortes entre artigos de diferentes níveis hierárquicos.

De um modo geral, este estudo permitirá satisfazer a necessidade e motivação de aplicar uma vertente analítica de forma prática numa área rica em oportunidades de melhoria.

1.3. Organização do Documento

Na sequência da introdução, é apresentada a revisão de literatura referente às técnicas que serão abordadas (*Data Mining*, *Market Basket Analysis* e Regras de Associação Hierárquica).

No capítulo 3 é efetuado o estudo do caso e apresentados os resultados obtidos. É descrita a metodologia adotada e a sua aplicação prática através das técnicas descritas anteriormente.

A obtenção de resultados precede a aplicação dos mesmos na empresa alvo de estudo e as conclusões finais.

Por último são indicadas as referências bibliográficas que foram utilizadas como suporte deste trabalho.

2. Revisão de Literatura

Data Mining representa o processo de aplicação de métodos de extração de conhecimento de dados com a intenção de descobrir padrões relevantes em grandes bases de dados. Através do estudo da forma como os dados são extraídos e armazenados, executa algoritmos de aprendizagem bastante completos para descobrir a forma mais eficiente de encontrar padrões relevantes em bases de dados que são cada vez maiores.

As técnicas de *Data Mining* derivam da evolução tecnológica inspirada na aprendizagem natural, que conhecemos no organismo humano e foram impulsionadas pelo avanço nas ciências de computação e engenharia, nomeadamente em *Machine Learning* onde se destacam as redes neuronais (StatSoft White Paper, 2007).

A extração de conhecimento de dados, amplamente aceite há cerca de duas décadas, tem vindo a ganhar cada vez mais importância e popularidade pela capacidade que tem de construir modelos e otimizar processos complexos e aos quais a Estatística não consegue dar uma resposta tão assertiva (StatSoft White Paper, 2007).

De facto, segundo Kantardzic (2002), *Data Mining* apresenta vantagens face à Análise Estatística e Inteligência Artificial sendo que, no entanto, depende delas uma vez que são a base matemática que permite a construção dos algoritmos que utiliza.

As técnicas de *Data Mining* têm cada vez mais impacto em áreas de procura de conhecimento e sistemas de informação, nomeadamente áreas organizacionais com maior interesse em tomar decisões de forma mais eficiente e assertiva. Assim, espera-se que a investigação em *Data Mining* e a sua aplicação na gestão de conhecimento aumente, tal como a crescente importância de técnicas de gestão de organizações, nomeadamente *Clustering* e Regras de Associação (Silwattananusarn e Tuamsuk, 2012).

Data Mining engloba diferentes áreas e aborda variadas técnicas, no entanto, aquela que mais se adequa ao estudo de uma empresa de retalho e, consequentemente, exploração dos padrões de compra dos clientes é a *Market Basket Analysis* (Brijs, 2002).

2.1. Market Basket Analysis

Market Basket Analysis é uma área baseada em modelos que explora a teoria de que se um determinado conjunto de artigos é comprado, existe uma maior ou menor probabilidade de ser adquirido outro conjunto de artigos (Albion Research Ltd., 2016). É

um tipo de análise de afinidade de entre as técnicas de *Data Mining* que se foca na identificação dos artigos mais frequentes e na probabilidade de serem comprados juntamente com outros (Qualls, 2013).

Estuda os padrões existentes entre atividades desenvolvidas por indivíduos ou grupos de pessoas e é uma técnica importante de análise no retalho uma vez que estuda os padrões de compra dos clientes e, assim, permite prever comportamentos futuros e desenvolver ações que promovam as vendas através de *cross* ou *up-selling*, promoções, alterações de *layout* de loja, entre outros (Gutierrez, 2006). De facto, a análise dos cabazes de compras e a sua monitorização permite estabelecer padrões de consumo por forma a explorar aspetos demográficos, a taxa de fidelização de clientes, a importância atribuída a marcas, os bens essenciais, entre outros (Svetina e Zupančič, 2005).

A *Market Basket Analysis* é bastante vantajosa na área do retalho uma vez que permite conhecer comportamentos dos consumidores, descobrir produtos com padrão de compra semelhante e aproxima-los dentro da loja, identificar produtos alvo de *cross-selling* e tomar decisões assertivas quanto a promoções e preços (Limitedbrands, 2004).

O uso comercial destas técnicas tem crescido significativamente com a evolução tecnológica que permitiu as vendas omni-canal levando a que muitas organizações tirem partido da informação sobre padrões de consumo, nomeadamente no âmbito da recomendação de produtos ou de música, na área da medicina em diagnósticos médicos e em otimização de conteúdos nomeadamente em *sites* de revistas ou *blogs* (Marafi, 2013). De facto, empresas como a Amazon utilizam a *Market Basket Analysis* como ferramenta de *cross-selling* recomendando produtos com base no histórico de pesquisa e vendas aos seus clientes.

As técnicas utilizadas não servem apenas de base para análises de organizações de retalho. Num estudo conduzido por McCormick *et al.* em 2011, foi analisado o comportamento de pacientes que visitavam regularmente os seus médicos por forma a identificar padrões e prever sintomas futuros dado o seu historial médico. Outro estudo dentro desta área foi conduzido por Jain e Gautam em 2014 com o objetivo de estudar a implementação do algoritmo Apriori no sector da saúde.

Na área de retalho, em que se foca este estudo, os cabazes de compra descrevem o comportamento de consumo dos clientes e dessa análise são retiradas ilações referentes

aos padrões de compra entre artigos e identificados os conjuntos de artigos frequentes, bem como retiradas conclusões acerca da relação entre eles (Agrawal *et al.,* 1993).

Sendo que os artigos em estudo são de desporto, apresenta-se o seguinte exemplo:

Seja X={x1,x2,…,xn} um conjunto de n artigos que podem ser comprados em conjunto numa determinada transação T, em que T={t1, … ,tn} para n transações, tendo cada uma um número de identificação próprio. Cada transação corresponde a um momento de compra efetuado por um determinado cliente e pode conter um ou vários artigos.

Transações	**Artigos**
t1	{Calções, T-Shirt, Camisola, Casaco}
t2	{T-shirt, Casaco}
t3	{Calções, Calças, Camisola}
t4	{Calções,T-Shirt, Camisola}
t5	{T-Shirt}

Figura 1: Base de dados com 5 transações

Fonte: Elaboração própria

Na figura 1 observa-se o exemplo de um conjunto de 5 transações que contêm 5 artigos, no entanto, as bases de dados estudadas atualmente contêm milhares ou milhões de transações e artigos pelo que é necessário fazer uma triagem das transações que se pretendem estudar.

Assim, numa primeira fase, para determinada base de dados de transações, são gerados conjuntos frequentes de dados para depois serem derivadas regras de associação. Os conjuntos frequentes de dados são encontrados através do suporte mínimo definido sendo que prevalecem aqueles que tiverem suporte igual ou superior ao mínimo estabelecido e as regras de associação são encontradas consoante a sua confiança seja superior à confiança mínima fixada (Ulas, 1999).

Introduzido por Agrawal *et al*. em 1993, o suporte absoluto de um *itemset* representa o número de transações onde se encontra esse *itemset*, sendo que o suporte relativo representa a proporção de transações que contêm esse mesmo *itemset* e é calculado através da divisão do suporte absoluto pelo número total de transações (Gama *et al.,* 2012). Utilizando o exemplo da figura 1 o suporte absoluto dos calções é 3, sendo o seu suporte relativo de $\frac{3}{5} = 0{,}6$ o que significa que 60% das transações contêm calções.

Por forma a encontrar as transações mais relevantes, fixa-se um suporte absoluto mínimo a ser cumprido pelos grupos de artigos. Para o exemplo da figura 1, ao fixar um suporte mínimo de 2, obtêm-se os *itemsets* frequentes da figura 2.

1 item	2 items	3 items
{Calções}: 3	{Calções, T-Shirt}: 2	{Calções, T-Shirt, Camisola}: 2
{T-Shirt}: 4	{Calções, Camisola}: 3	
{Camisola}: 3	{T-Shirt, Camisola}: 2	
{Casaco}: 2	{T-Shirt, Casaco}: 2	

Figura 2: *Itemsets* frequentes das transações da figura 1 com suporte mínimo de 2

Fonte: Elaboração própria

Tendo em conta o exemplo da figura 1, ao definir um suporte mínimo de 2 obtêm-se 9 conjuntos de artigos frequentes (com 1, 2 ou 3 artigos) sendo que calças é excluído dos conjuntos frequentes por não ter suporte mínimo (apenas faz parte de uma transação).

O suporte permite-nos aferir quão comum é determinado *itemset* na base de dados (Ulas, 1999) sendo que a definição de um determinado nível mínimo de suporte é utilizada para restringir o número de conjuntos gerados e, assim, garantir resultados mais relevantes (Wang *et al*., 2000).

Depois de encontrados os conjuntos frequentes de dados, procede-se à derivação de regras de associação.

Regras de Associação são uma das principais ferramentas para encontrar padrões relevantes na base de dados de transações dos clientes no âmbito da *Market Basket Analysis*, um dos campos mais antigos de *Data Mining* (Raeder e Chawla, 2011).

Introduzidas por Agrawal *et al.*, em 1993, são uma técnica bastante popular constituindo uma ferramenta muito poderosa na análise de cabazes de compra e sendo utilizadas em estudos de retalho alimentar, sistemas de recomendação ou outros tipos de negócio com o objetivo de aumentar as vendas de produtos e serviços (Rodrigues *et al.*, 2012).

As Regras de Associação apresentam a forma de "se antecedente então consequente". Por exemplo, se X então Y, sendo X e Y conjuntos de artigos (Gama *et al.*, 2012). A regra {X → Y} representa a relação de compra entre o *itemset* X e o *itemset* Y e indica a probabilidade de Y ser adquirido quando X foi comprado.

O estudo de padrões de compra entre artigos permite responder a questões relativas aos produtos que são vendidos em conjunto ou à dependência que poderá existir entre determinados artigos ou conjuntos de artigos. Se quisermos testar a dependência entre X e Y, sendo X e Y *itemsets*, deveremos encontrar as regras de associação do tipo {X → Y} ou {Y → X} (Ulas, 1999).

O grau de interesse de uma regra de associação é medido através da sua confiança, conceito que foi introduzido, juntamente com o suporte, por Agrawal et al. em 1993.

O cálculo da confiança é feito pelo quociente entre o suporte das transações que contêm o antecedente e o consequente e o suporte do antecedente, ou seja, o grau de interesse de uma regra mede a relação entre o número de transações que incluem todos os *itemsets* do conjunto e o número de transações que incluem todos os *itemsets* do antecedente (Gama *et al.*, 2012).

$$Conf\ (X \rightarrow Y) = \frac{supp\ (XUY)}{supp\ (X)}$$

Equação 1: Confiança de uma regra de associação {X → Y}

A equação 1 representa o cálculo da confiança de uma regra de associação (em que X é o antecedente e Y o consequente), ou seja, o grau de interesse que permite aferir a proporção de transações de X que também contêm Y.

Utilizando a base de dados da figura 1 e os *itemsets* frequentes da figura 2 é possível derivar regras de associação e calcular o seu grau de confiança.

Regra	Confiança
{Calções→ T-Shirt}	(2/3) = 66,7%
{Calções → Camisola}	(3/3) = 100%
{T-Shirt → Casaco}	(2/4) = 50%

Figura 3: Regras de associação derivadas a partir dos *itemsets* frequentes da figura 2

Fonte: Elaboração própria

A figura 3 engloba alguns exemplos de regras de associação derivadas dos conjuntos de artigos frequentes da figura 2.

Observa-se um exemplo com 100% de confiança, ou seja, todas as transações que contêm calções também contêm camisolas. No caso da regra {T-Shirt → Casaco}, concluímos que quando é comprada uma t-shirt a probabilidade de ser comprado um casaco na mesma transação é de 50%.

A definição de valores mínimos de suporte e confiança, apesar de permitir encontrar dados mais relevantes, não invalida a produção de regras de associação em excesso pelo que é necessário utilizar medidas de interesse adicionais (Hahsler e Hornik, 2008) que permitam validar e aferir a qualidade da regra estudada tais como o *lift* ou a convicção.

Estas métricas têm em conta a dependência entre variáveis considerando que, se duas variáveis são independentes a regra inferida dessas variáveis não terá interesse porque a associação entre os dois *itemsets* é aleatória e, por isso, não poderemos inferir nada acerca dela.

O *lift* é uma medida para avaliar níveis de associação. Introduzido por Brin *et al*. em 1997, mede em quanto é que a confiança de uma regra excede a sua confiança esperada, ou seja, a diferença entre o número de vezes em que o antecedente e o consequente são adquiridos juntos e a frequência esperada de serem adquiridos se fossem estatisticamente independentes (Hahsler, 2015).

O cálculo do *lift* é feito através do quociente entre a confiança da regra de associação e o suporte do consequente, como apresentado na equação 2.

$$Lift\ (X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} = \frac{supp\ (XUY)}{supp(X)\ x\ supp\ (Y)}$$

Equação 2: *Lift* de uma regra de associação

A equação 2 mede o grau de dependência dos conjuntos de artigos objeto de análise, variando no intervalo [0, +∞] (Azevedo e Jorge, 2007).

Valores de *lift* inferiores a 1 indicam que os artigos estão negativamente relacionados, valores superiores a 1 indicam que os artigos estão positivamente relacionados e valores de *lift* próximos de 1 indicam que os artigos têm uma fraca relação entre si e são, por isso, independentes (Hahsler e Hornik, 2008).

Por sua vez, a convicção mede a frequência com que a regra faz predições erradas, isto é, o rácio da frequência esperada de ocorrer o antecedente sem que ocorra o consequente (Hahsler, 2015).

A convicção foi introduzida em 1997 como alterativa à confiança, por Brin *et al*., e mede o quão convincente é uma regra de associação através do cálculo da sua frequência de erro, como demonstrado na equação 3.

$$Convicção(X \rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \rightarrow Y)}$$

Equação 3: Convicção de uma regra de associação

Tal como o *lift*, a convicção varia entre 0 e 1 sendo que os valores próximos de 1 representam *itemsets* independentes.

A convicção permite colmatar algumas falhas da confiança e do *lift* uma vez que é sensível à direção da regra e considera o suporte tanto do antecedente como do consequente (Azevedo e Jorge, 2007). No entanto, o *lift* considera a informação obtida através de métricas precisas como medida de cálculo estatístico.

As regras de associação são aplicadas em diversos contextos sendo que, para isso, foi desenvolvido um conjunto de algoritmos para as identificar em grandes bases de dados (Raeder e Chawla, 2011).

O algoritmo Apriori é uma das primeiras e mais rápidas ferramentas implementadas para identificação de padrões dependendo dele muitos métodos de exploração de regras de associação (Ulas, 1999). A sua base assenta no princípio de que qualquer subconjunto de *itemsets* frequentes deve ser um *itemset* frequente e utiliza uma estratégia de procura em largura (Gama *et al.*, 2012).

Este algoritmo começa por gerar os *itemsets* que têm suporte superior ao definido para cada nível. Gera os de nível 1 e tendo em conta os conjuntos frequentes de artigos gerados nesse nível gera os de nível 2 e assim sucessivamente. Depois de gerados os candidatos a *itemsets* frequentes, o algoritmo testa a sua frequência voltado a correr a base de dados da transação.

Numa segunda fase, o algoritmo Apriori gera as regras de associação derivadas dos *itemsets* frequentes considerando o valor mínimo definido para a confiança. Este algoritmo pode ser otimizado uma vez que, quando o *itemset* é movido do antecedente para o consequente, a confiança não pode aumentar (Rodrigues *et al.*, 2012).

As regras de associação são a forma mais tradicional de estudar grandes conjuntos de dados de consumo (transações) (Rodrigues *et al.*, 2012), no entanto, os métodos

tradicionais de exploração de padrões frequentes produzem demasiadas regras redundantes e este excesso aumenta o tempo de computação, a complexidade em tirar ilações e a dificuldade em tomar decisões assertivas (Zaki, 2000).

Para colmatar este problema, têm sido utilizadas diversas técnicas, nomeadamente *Social Network Analysis*, atribuição de prioridades a cada regra, *itemsets* frequentes fechados e Regras de Associação Hierárquica.

A *Social Network Analysis* ajuda a combater a perceção de regras de associação redundantes através de uma visualização em que os artigos são representados por vértices e as relações entre eles arestas. A ligação de vértices por arestas representa os produtos que foram comprados na mesma transação e facilita a análise dos padrões de consumo (Rodrigues *et al.*, 2012).

Num estudo de 2009, Zhao *et al.*, abordaram o problema de excesso de padrões atribuindo prioridades a cada regra por forma a priorizar as ações a serem tomadas. Introduziram também a questão da perceção de relevância das regras sendo este, não um processo estático mas sim dinâmico que varia ao longo do tempo.

Um *itemset* frequente é considerado fechado quando tem suporte superior ao mínimo definido e, para o mesmo conjunto de dados que o forma, não existem outras associações com suporte igual (Verma, 2009). Este conceito permite eliminar regras redundantes sem perder informação importante uma vez que apenas determina quais os conjuntos frequentes de dados a considerar (Zaki, 2000).

Outra proposta de solução ao grande número de regras de associação encontradas, e aproveitando a hierarquia dos artigos em análise, é a que vai ser estudada ao longo deste trabalho: as Regras de Associação Hierárquica.

2.2. Regras de Associação Hierárquica

As Regras de Associação Hierárquica surgiram, com Agrawal *et al.*, em 1993, devido à necessidade de afinar a pesquisa de padrões de consumo.

Através da utilização da estrutura das variáveis em análise, esta técnica permite encontrar regras de associação que tenham em conta o nível da hierarquia e, assim, eliminar regras redundantes e menos interessantes (Srikant *et al.*, 1997).

A utilização de hierarquias é uma abordagem muito utilizada em páginas web e representa a afinação progressiva do estudo de uma variável permitindo a organização da informação dos *itemsets* da base de dados em análise (Martin *et al.*, 2007).

É também frequente em organizações de retalho, que normalmente têm uma determinada estrutura de produtos, utilizar a hierarquia para distinguir diferentes tipologias de artigos e, dentro destas, variações do mesmo produto.

No exemplo apresentado anteriormente na figura 1 foram consideradas tipologias de artigos (Calções, T-Shirt, Camisola, entre outros) que, no âmbito da base de dados em análise, pertencem a uma determinada estrutura.

Na figura 4 observam-se as tipologias referidas anteriormente e respetiva estrutura hierárquica, tal como apresentado nos dados em estudo.

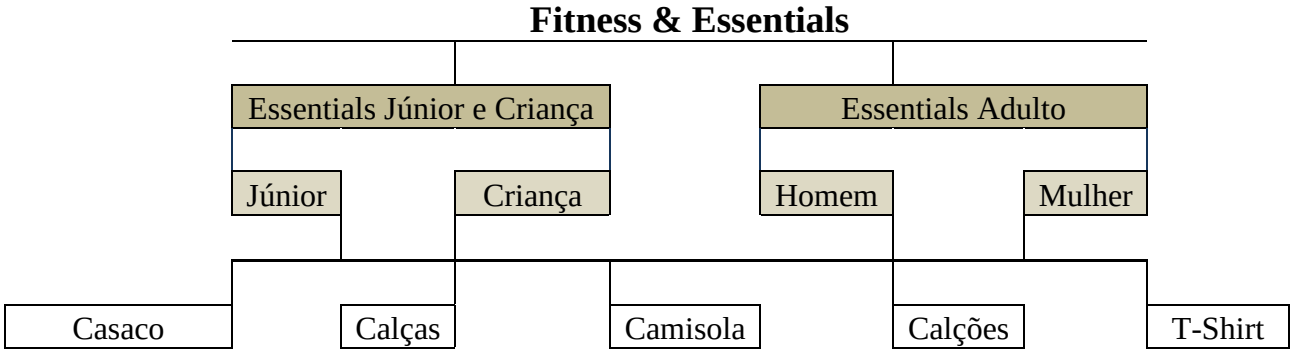


Figura 4: Exemplo de hierarquia

Fonte: Elaboração própria

Considerando a hierarquia apresentada na figura 4, é possível encontrar uma regra que infira que "pessoas que compram artigos de Essentials Adulto - Mulher costumam comprar Calças de Criança" e esta regra prevalece mesmo que regras como "pessoas que compram artigos de Essentials Adulto costumam comprar Calças de Criança" ou "pessoas que compram Calças de Essentials Adulto - Mulher costumam comprar Calças de Criança" não sejam válidas.

Esta abordagem permitirá explorar em profundidade os padrões existentes a diversos níveis da hierarquia possibilitando a descoberta de padrões mais específicos e relevantes devido à sua capacidade de estudar diferentes níveis da base de dados (Han e Fu, 1995).

No entanto, é necessário ter em conta algumas notas referentes ao estudo de regras de associação considerando a hierarquia dos *itemsets*.

É esperado que os níveis mais elevados tenham um suporte maior, pelo que o critério de suporte mínimo terá de ser ajustado e reduzido substancialmente para não excluirmos *itemsets* de níveis mais baixos da hierarquia (uma vez que ocorrem menos frequentemente). Por outro lado, será muito difícil encontrar regras de associação fortes para os níveis mais baixos da hierarquia como por exemplo ao nível do código de barras do produto (Han e Fu, 1995).

Só deverão ser estudados descendentes de *itemsets* frequentes visto que, se um *itemset* não for frequente, os seus descendentes sê-lo-ão ainda menos (Geyer, 2003).

Por forma a eliminar e afinar o estudo de padrões relevantes quando a hierarquia é tida em consideração, deverá ser observada uma medida introduzida por Agrawal *et al.*, em 1993, que preconiza que, se o suporte e confiança de uma regra é próxima do seu valor esperado baseado num antecessor da regra, a regra não deverá ser considerada pois é redundante.

O método utilizado para extração de padrões a vários níveis segue uma lógica descendente que, progressivamente, aprofunda a análise aos níveis mais baixos da hierarquia através de algoritmos desenvolvidos para este fim, sendo assim esperado encontrar *itemsets* frequentes e padrões mais fortes nos níveis superiores da hierarquia (Han e Fu, 1995).

O algoritmo Apriori é uma estratégia simples que pode ser aplicada a este problema mas conduz à produção de demasiadas regras desinteressantes. Depois de encontrar *itemsets* frequentes através da definição de um suporte mínimo, deriva regras de associação pela fixação de um nível mínimo de confiança. No entanto, ao aplicar-se estes métodos em níveis inferiores teriam de reduzir-se os critérios definidos, resultando daqui a identificação de regras redundantes e menos relevantes nos níveis intermédios e superiores da hierarquia (Geyer, 2003).

Assim, apesar do algoritmo Apriori dever ser utilizado numa fase inicial, é necessário abordar outros algoritmos ou técnicas para estudar as relações a vários níveis da estrutura.

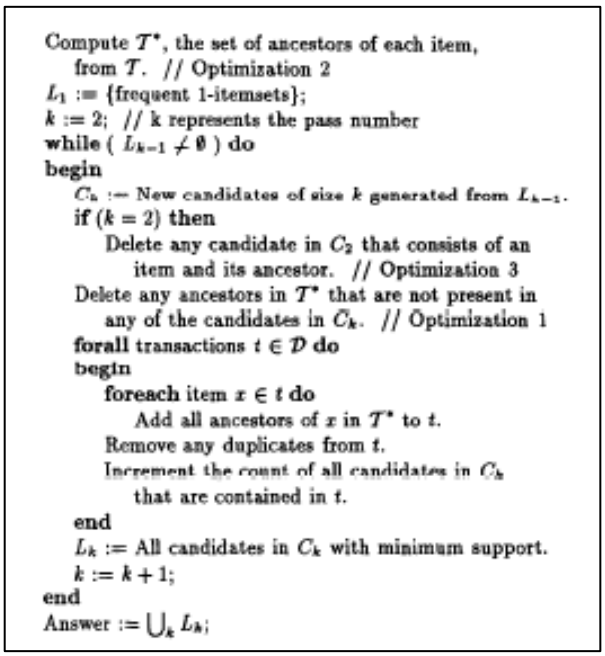
O algoritmo Basic, que apesar de simples resulta num processo demorado, apresenta uma forma fácil de incorporar a informação da hierarquia do artigo na derivação

de regras de associação através da adição de todos os antepassados do artigo na transação em que ele se encontra (Srikant e Agrawal, 1995).

Assim, em 1995, Srikant e Agrawal introduziram dois algoritmos por oposição ao algoritmo Basic que consideram mais lento: o Cumulate e o Est Merge. Ambos são versões otimizadas do algoritmo referido acima mas que chegam a correr entre 2 a 5 vezes mais rápido.

O algoritmo Cumulate acrescenta, em cada transação, os antecessores de cada artigo presente nessa transação, correndo depois, tal como no algoritmo Basic, o algoritmo Apriori sobre as transações alargadas. Esta operação otimiza a procura de regras de associação relevantes considerando apenas antecessores que estejam presentes em um ou mais *itemsets* da nova transação alargada e não considera conjuntos de dados que contenham tanto o artigo como o seu antecessor uma vez que o suporte desse *itemset* (antecessor e sucessor) seria o mesmo que o suporte do *itemset* sem o antecessor (Srikant e Agrawal, 1995).

O algoritmo Cumulate recorre a uma pré-computação de antecessores em vez de procurar antecessores para cada conjunto de dados, não considerando, portanto, antecessores que não estejam presentes em qualquer variável ao mesmo tempo. A Figura 5 representa o algoritmo Cumulate tal como apresentado por Srikant e Agrawal em 1995.



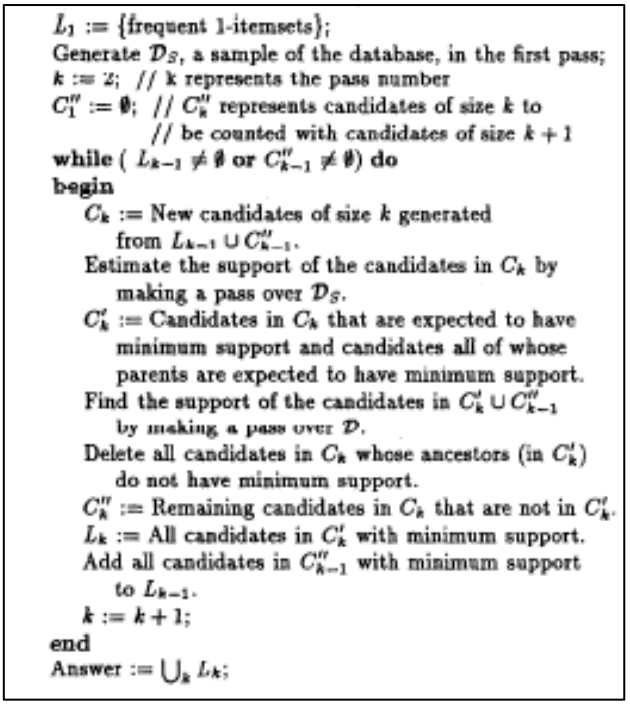
Compute $\mathcal{T}^*$, the set of ancestors of each item,
from $\mathcal{T}$. // Optimization 2
L_1 := {frequent 1-itemsets};
$k := 2$; // k represents the pass number
while ($L_{k-1} \neq \emptyset$) **do**
begin
C_k := New candidates of size k generated from L_{k-1}.
if ($k = 2$) **then**
Delete any candidate in C_2 that consists of an item and its ancestor. // Optimization 3
Delete any ancestors in $\mathcal{T}^*$ that are not present in any of the candidates in C_k. // Optimization 1
forall transactions $t \in \mathcal{D}$ **do**
begin
foreach item $x \in t$ **do**
Add all ancestors of x in $\mathcal{T}^*$ to t.
Remove any duplicates from t.
Increment the count of all candidates in C_k that are contained in t.
end
L_k := All candidates in C_k with minimum support.
$k := k + 1$;
end
Answer := $\bigcup_k L_k$;

Figura 5: Algoritmo Cumulate

Fonte: Srikant e Agrawal, 1995

Por sua vez, o algoritmo Est Merge utiliza a amostragem para melhorar a sua performance. Através deste método, estima o suporte das variáveis e conta os *itemsets* frequentes esperados e aqueles que não se espera que tenham um suporte mínimo mas cujos antecedentes o têm. Uma vez que a estimativa pode conter algum erro, os descendentes dos *itemsets* que acabam por ter suporte mínimo são considerados o que, apesar de poder gerar um aumento de *itemsets*, faz diminuir o número de etapas do algoritmo (Srikant e Agrawal, 1995).

A figura 6 representa o algoritmo Est Merge tal como apresentado por Srikant e Agrawal em 1995.



L_1 := {frequent 1-itemsets};
Generate $\mathcal{D}_S$, a sample of the database, in the first pass;
$k := 2$; // k represents the pass number
$C_1'' := \emptyset$; // C_k'' represents candidates of size k to
// be counted with candidates of size $k+1$
while ($L_{k-1} \neq \emptyset$ **or** $C_{k-1}'' \neq \emptyset$) **do**
begin
C_k := New candidates of size k generated
from $L_{k-1} \cup C_{k-1}''$.
Estimate the support of the candidates in C_k by
making a pass over $\mathcal{D}_S$.
C_k' := Candidates in C_k that are expected to have
minimum support and candidates all of whose
parents are expected to have minimum support.
Find the support of the candidates in $C_k' \cup C_{k-1}''$
by making a pass over $\mathcal{D}$.
Delete all candidates in C_k whose ancestors (in C_k')
do not have minimum support.
C_k'' := Remaining candidates in C_k that are not in C_k'.
L_k := All candidates in C_k' with minimum support.
Add all candidates in C_{k-1}'' with minimum support
to L_{k-1}.
$k := k+1$;
end
Answer := $\bigcup_k L_k$;

Figura 6: Algoritmo Est Merge

Fonte: Srikant e Agrawal, 1995

Os dois algoritmos apresentados são mais eficazes do que o algoritmo Basic, que tomam como base. Os testes efetuados por Srikant e Agrawal demostram que o algoritmo Est Merge tem uma performance ligeiramente superior ao do Cumulate (é 25% a 30% mais rápido). A diferença é ainda mais significativa com o aumento do número de

transações da base de dados visto que a precisão das estimativas do suporte das variáveis tem uma correlação positiva com a dimensão da base de dados (Srikant e Agrawal, 1995).

Considerando a performance dos dois algoritmos acima e a particularidade deste estudo e os seus objetivos, a melhor opção neste contexto seria o algoritmo Est Merge, no entanto, existem técnicas mais simples e acessíveis passíveis de serem adotadas no âmbito deste trabalho, nomeadamente a linguagem e *software* R que também permite a implementação de algoritmos.

Criado por Ross Ihaka, em 1993, o R é um *software* gratuito utilizado por mais de 2 milhões de pessoas que contém extensões desenvolvidas por milhares de cientistas.

> Para além de ser um *software* gratuito de análise de dados é também uma linguagem de programação que permite fazer análises estatísticas, visualizar dados e prever modelos de forma fácil e smples.

O *arules*, um dos *package*s do R, permite encontrar conjuntos de dados frequentes e derivar destes regras de associação através de diferentes algoritmos, entre eles o algoritmo Apriori mencionado anteriormente.

Oferece, também, suporte para a consideração da estrutura dos *itemsets* por forma a construir regras a vários níveis da hierarquia estudada.

A análise multinível, feita através da agregação, cria novos grupos de dados para cada nível mantendo-se cada *itemset* no grupo se um ou mais dos restantes também fizer parte do grupo de dados original. Caso um determinado grupo de dados se repita no antecedente e no consequente de uma regra, esse grupo será eliminado acontecendo o mesmo com todas as regras e *itemsets* que não sejam únicos após a agregação (Hahsler, 2016).

3. Estudo do Caso

Nesta secção será descrita a metodologia adotada bem como postas em prática as técnicas descritas nas secções anteriores com o objetivo de alcançar os resultados propostos na introdução.

As conclusões serão analisadas de forma correta e sustentada com o fim de propor aplicações práticas adequadas no âmbito das particularidades deste trabalho e respetivos objetivos.

Este trabalho será sustentado, do início ao fim, por um único *software* que permite realizar o estudo proposto.

O programa R, nomeadamente o *package arules*, permite a análise e o tratamento dos dados bem como a derivação dos resultados propostos anteriormente de forma fácil, simples e sem qualquer custo.

O *arules* providência uma estrutura que permite representar, tratar e analisar dados de transações bem como encontrar padrões frequentes tornando possível este estudo através da utilização dos algoritmos Apriori e Eclat (Hahsler *et al*., 2005) ao mesmo tempo que facilita o estudo das Regras de Associação tendo em conta a hierarquia existente na estrutura dos produtos através da agregação de dados. Por forma a completar a análise dos resultados, será utilizada uma extensão do R ligada ao *arules*, *arules Viz*, que proporciona uma melhor visualização dos resultados uma vez que a extração de conhecimento de dados pode resultar num grande número de regras e padrões e essa perceção tornará a sua análise mais simples e acessível (Hahsler e Chelluboina, 2011).

Este *package* permite visualizar os resultados obtidos de forma a facilitar a sua análise e tornar mais rápida a identificação de padrões relevantes, diminuindo assim o impacto que regras mais redundantes possam ter ao tornar a informação mais densa e confusa.

Será apresentada a base de informação a ser utilizada bem como o processo de tratamento de dados e, de seguida, serão explanadas as técnicas utilizadas para obtenção de *itemsets* frequentes e a derivação de Regras de Associação Hierárquica.

Ao longo deste estudo, será justificada a escolha da utilização destas técnicas pelas suas vantagens e apresentadas conclusões sustentadas no *output* alcançado.

3.1. Descrição dos Dados

Neste caso de estudo será utilizada uma base de dados de transações de uma empresa portuguesa que comercializa artigos de desporto.

Cada transação tem um número identificativo próprio e corresponde à compra de um ou mais artigos efetuada por um cliente num determinado momento de tempo.

Os artigos são identificados na transação através da sua estrutura mercadológica, ou seja, cada artigo tem um código que representa a estrutura hierárquica em que se encontra.

Os artigos obedecem a uma hierarquia que se divide nos 4 níveis apresentados abaixo, ordenados por ordem hierárquica:

- Unidade de Negócio
- Categoria
- Subcategoria
- Unidade Base

Cada um destes níveis representa 2 dos 8 dígitos que constituem a estrutura do artigo, sendo os da esquerda hierarquicamente superiores aos da direita.

A estrutura de cada nível acumula os dígitos de cada um dos seus antecedentes.

Unidade de Negócio	Categoria	Subcategoria	Unidade base
23 - Fitness e Essentials	2307 - Essentials Têxtil Adulto	230702 - Homem	23070214 - T-shirts

Figura 7: Exemplo de estrutura de um artigo

Fonte: Elaboração própria

No exemplo da figura 7 observamos um artigo com código 23070214 e, analisando os pares de dígitos, constatamos que se trata de um artigo da Unidade de Negócio Fitness e Essentials (**23**), da categoria Essentials Adulto (23**07**), da subcategoria – neste caso género – Homem (2307**02**) sendo que a unidade base – tipo de artigo – é uma T-Shirt (230702**14**).

A análise deste trabalho terá apenas em conta a categoria, subcategoria e unidade base uma vez que a unidade de negócio é demasiado ambígua para podermos retirar qualquer ilação dos resultados obtidos a esse nível.

Por forma a conciliar os dados cedidos pela empresa e a capacidade de analisar grandes bases de dados, os *tickets* (transações) a analisar representarão as compras efetuadas durante o ano de 2015 numa determinada loja, sendo esta uma loja de dimensão considerável e com um volume de tráfego acima da média.

A base de dados contém 288.199 transações e respetivos códigos adquiridos em cada uma. Uma vez que este estudo assenta nos diferentes níveis hierárquicos estabelecidos, teremos 3 bases de dados, cada uma com os artigos adquiridos por transação e identificados pelo seu respetivo nível hierárquico.

Ou seja, a base engloba 3 ficheiros sendo que, cada um, contém as 288.199 transações para determinado nível de estrutura: um ficheiro com as transações ao nível da categoria, outro ao nível da subcategoria e, por fim, um último ao nível da unidade base. Nos ficheiros, cada linha representa um momento de compra onde é identificada a transação em causa e os códigos dos artigos que a constituem.

Cada transação contém, em média, 1.35 categorias representadas por 1.41 subcategorias que, por sua vez, culminam em 1.52 unidades base. O número mínimo de códigos adquiridos por transação é 1 sendo o máximo de 15, 18 e 35 consoante se trate de uma base de dados ao nível da categoria, subcategoria ou unidade base, respetivamente.

Cada transação é representada por um conjunto de dígitos sob a forma x#y#w#z em que x representa o número do *ticket* (talão de compra), y o número do ponto de venda na loja que registou a compra, w a loja em que a compra foi efetuada e z a data em que a transação se efetuou.

A figura 8 representa 2 linhas (transações) com estruturas adquiridas ao nível da categoria.

11488#4#155010001#20150102	2401	2602	
114877#2#155010001#20151206	2102	2102	2303

Figura 8: Exemplo de duas transações da base de dados ao nível da categoria

Fonte: Elaboração própria

Como se pode observar na figura 8, existe uma transação com *ticket* 11488, processada pelo ponto de venda 4, na loja 155010001 no dia 2 de janeiro de 2015

(transação 11488#4#155010001#20150102) onde foram adquiridos artigos das categorias 2401 (Corrida Calçado) e 2602 (Calçado Casual Mulher).

No dia 6 de dezembro de 2015, na mesma loja, foi efetuada uma transação com *ticket* 114877, processada pelo ponto de venda 2 (transação 114877#2#155010001#20151206) onde foram adquiridos dois artigos da categoria 2102 (Futebol Têxtil) e um artigo da categoria 2303 (Ginásio Têxtil).

3.2. Tratamento e Leitura de Dados

Para que o *software* utilizado leia corretamente os dados é necessário adapta-los ao programa.

Neste caso, o R irá ler dados sob a forma {T;x1;x2;…;xn} sendo T a transação e x os códigos dos artigos comercializados. As estruturas deverão estar separadas por ponto e vírgula sendo que não há limite de estruturas para a mesma transação.

No caso de uma transação conter mais do que um artigo da mesma estrutura, esta não deverá ser repetida, uma vez que não é relevante o número total de vezes que a estrutura aparece. Apenas nos interessa o número de transações em que a estrutura é comprada face ao número total de transações existentes.

Não se considera a duplicação de códigos no mesmo *ticket* visto que o que se pretende estudar são padrões de compra passíveis de serem adotados por um grande número de clientes. Se fosse considerado o número de vezes que o artigo é adquirido e o cliente adquirisse numa só transação um número muito superior ao normal deste artigo, os padrões seriam afetados por um ato esporádico, seriam influenciados por *outliers*.

Utilizando o exemplo da figura 8 verifica-se que os dados repetem estruturas, ou seja, neste caso foram adquiridos 2 artigos de Futebol Têxtil pelo que a categoria respetiva (2102) aparece duas vezes na transação.

Assim, foram eliminadas todas as duplicações de estrutura para a mesma transação ficando a base de dados resumida a 3 ficheiros, cada um com transações a diferentes níveis da hierarquia:

- Categoria;
- Subcategoria;
- Unidade Base.

A primeira ação para leitura dos dados é selecionar o diretório em que estes estão guardados por forma a que, apenas com a identificação do nome do ficheiro, o R os consiga ler.

Depois de identificado o local de armazenamento dos dados, será utilizado o algoritmo Apriori através do comando observado na figura 9 para a sua leitura.

```
tr<-read.transactions("tickets 2015 CAT sd.csv",format="basket",sep=";")
```

Figura 9: Comando para leitura de dados com o *package arules*

Fonte: Elaboração própria

Os comandos apresentados permitem ler o ficheiro "*tickets* 2015 CAT sd.csv" que corresponde às transações ao nível da categoria e tem o formato especificado anteriormente.

Para além do apresentado na figura 9, existem outros comandos que permitem analisar a base de dados no software utilizado. Nomeadamente:

- O comando *inspect* permite visualizar no ecrã toda a base de dados;
- O comando *image* cria uma representação gráfica dos dados em análise;
- O comando *length* indica o número de linhas da base de dados.

Estes comandos são atalhos importantes para trabalhar com um grande volume de dados uma vez que permitem, de forma rápida, analisar o que foi inserido.

3.3. *Itemsets* Frequentes e Regras de Associação

Aproveitando a hierarquia existente na base de dados, serão usadas as Regras de Associação Hierárquica para encontrar e representar as relações entre os artigos comercializados.

Nesta primeira fase, através do algoritmo Apriori, pretende-se encontrar os grupos de artigos mais frequentes e derivar regras de associação para cada nível da hierarquia do produto consoante os valores definidos para os parâmetros suporte e confiança.

A figura 10 representa o comando que permite identificar *itemsets* frequentes e derivar regras de associação através do algoritmo Apriori conforme valores de suporte e confiança definidos.

```
rules<- apriori(a, parameter=list(supp=0.001, conf= 0.001))
```

Figura 10: Comando para identificação de *itemsets* frequentes e derivação de regras de associação

Fonte: Elaboração própria

Na figura 10 foi definido, como exemplo, um valor mínimo de 0.001 para os parâmetros suporte e confiança.

Neste caso, seriam considerados frequentes os códigos de artigos que estivessem presentes em, pelo menos, 0,01% das transações e, destes, seriam derivadas regras de associação que tivessem, pelo menos, 0,01% de confiança, ou seja, regras em que existisse uma probabilidade de, pelo menos, 0,01% de um determinado código ser comprado sendo que outro o foi.

Foi efetuado um teste de sensibilidade aos parâmetros uma vez que é bastante importante fixar valores que permitam encontrar os resultados mais relevantes dada a base de dados e o objetivo deste estudo.

3.3.1. Definição do Valor dos Parâmetros

Os valores a serem fixados deverão ter em conta os níveis mais baixos da hierarquia e, uma vez que é esperado que os níveis mais elevados tenham um suporte maior, se não reduzirmos o valor dos parâmetros a utilizar poderemos estar a excluir *itemsets* de níveis hierárquicos inferiores que, apesar de relevantes, não ocorrem tão frequentemente (Han e Fu, 1995).

Por outro lado a confiança terá um valor baixo uma vez que a escolha das regras de associação será feita através do *lift* cujo cálculo foi previamente apresentado na equação 2. O *lift* representa o rácio que mede em quanto a confiança de uma regra excede a confiança esperada dessa regra, ou seja, a predisposição de um cliente para comprar um artigo uma vez que comprou outro. Assim, por forma a encontrar as regras de associação

mais relevantes, será utilizada a métrica de validação l*ift* para eliminar regras cuja associação de artigos seja aleatória e, por isso, possam ser redundantes.

Considerando o acima exposto, foi efetuado um teste de sensibilidade aos parâmetros para definir os valores utilizados para o suporte e confiança. A confiança deverá ser fixada a um nível baixo e, dependendo do número de regras conseguidas por cada valor de suporte, é escolhido o nível a fixar neste parâmetro. Assim, garantem-se os parâmetros que permitem o máximo de regras para o máximo de confiança nos resultados extraídos.

Fixado um nível de confiança baixo, de 0.001, foi feito um teste de sensibilidade ao suporte, ou seja, para diferentes níveis de suporte (mesma confiança), verificamos o volume de resultados obtidos. Esta análise está resumida nas tabelas da figura 11.

***Itemsets* Frequentes**

Suporte	CAT	SUB	UB
0.001	55	164	318
0.005	42	77	72
0.01	35	38	31
0.05	10	1	0

Regras de Associação

Suporte	CAT	SUB	UB
0.001	218	126	54
0.005	20	2	2
0.01	6	0	0
0.05	0	0	0

Figura 11: Teste à sensibilidade do parâmetro suporte para um nível de confiança de 0.001

Fonte: Elaboração própria

Para cada nível de suporte e estrutura em análise (CAT - categoria, SUB - subcategoria e UB - unidade base), a primeira tabela apresenta o número de *itemsets* frequentes encontrados e a segunda o número de regras de associação derivadas.

Utilizando como exemplo o nível da categoria, observa-se que, para uma confiança de 0.001 e um suporte de 0.001, foram encontrados 55 *itemsets* frequentes e derivadas 218 regras de associação.

Para o nível mais baixo de suporte, o número de *itemsets* frequentes aumenta com a descida na hierarquia. Isto deve-se ao facto de existirem mais códigos com 6 dígitos do que com apenas 4, uma vez que códigos duplicados no mesmo *ticket* não são considerados. Assim, um *ticket* que teria dois artigos da categoria 2303 (Ginásio Têxtil),

apenas contabilizaria uma vez essa categoria. No entanto, descendo à subcategoria os dois códigos poderiam ser diferentes não sendo, por isso, duplicados e consequentemente eliminado um deles. A mesma análise pode ser feita em relação à unidade base (com 8 dígitos) face à subcategoria (com 6 dígitos).

No que respeita às regras de associação, existem mais regras nos níveis altos da hierarquia uma vez que a probabilidade de um código de 4 dígitos se relacionar com outro é maior do que isso acontecer entre códigos de 6 ou 8 dígitos.

Tendo em consideração os resultados obtidos foi escolhido um suporte de 0.001 que, apesar de baixo, é aquele que nos permite obter mais *itemsets* frequentes e regras de associação.

3.3.2. *Itemsets* Frequentes para cada Nível Hierárquico

Utilizando o algoritmo Apriori para um suporte e confiança de 0.001 foram encontrados os conjuntos de *itemsets* frequentes para cada nível da hierarquia, representados nas figuras 12 (ao nível da categoria), 13 (ao nível da subcategoria) e 14 (ao nível da unidade base).

Categoria	Descrição da Categoria	Suporte
2307	Essentials Têxtil Adulto	0.088
2309	Essentials Interiores	0.074
2614	Praia	0.070
2402	Corrida Têxtil	0.064
2103	Futebol Equipamentos	0.060
2401	Corrida Calçado	0.055
2303	Ginásio Têxtil	0.055
2601	Casual Calçado Homem	0.054
2310	Alimentação Desportiva	0.053
2702	Outdoor Casual Têxtil	0.050

Figura 12: Os 10 *itemsets* frequentes com maior suporte ao nível da categoria

Fonte: Elaboração própria

Analisando a informação obtida e apresentada na tabela da figura 12, verificamos que a categoria 2307 – Essentials Têxtil Adulto, está presente em 8,8% (0.088) das transações efetuadas, sendo a categoria com maior suporte.

A segunda categoria com maior suporte é Essentials Interiores, uma categoria maioritariamente composta por meias. Este tipo de artigos são bastante fáceis de encontrar no mercado e, normalmente, não apresentam tecnicidade específica que permita incentivar a sua compra num determinado sítio. Assim, será interessante perceber porque é que esta tipologia de artigos alcança uma percentagem tão alta no bolo total de transações. A análise de regras de associação permitirá despistar se estes resultados se devem a uma relação forte com outras tipologias de artigos bastante frequentes, como por exemplo o calçado.

Importa também salientar o facto de a categoria Praia estar presente em 7% das transações totais da loja uma vez que é uma categoria sazonal com destaque apenas 2 a 3 meses por ano. Uma análise complementar que esta abordagem e técnica permite é a de verificar como é que o valor do suporte variou ao longo dos meses do ano. Por outro lado, verificou-se que os chinelos de praia se mantêm todo o ano em loja junto da secção de natação. Seria interessante perceber se esta tipologia de artigo está relacionada com os produtos de natação e se poderá ser esse o fator que potencia o valor de suporte da categoria Praia.

Sub-Categoria	Descrição da Sub-Categoria	Suporte
230702	Essentials Têxtil Adulto – Homem	0.056
230303	Ginásio Têxtil - Mulher	0.048
220876	Natação Equipamentos - Acessórios	0.043
210374	Futebol Equipamentos - Bolas	0.039
231076	Alimentação Desportiva - Bebidas	0.038
230960	Essentials Interiores - Meias Invisíveis	0.035
210278	Futebol Têxtil - Treino e Jogo	0.034
230703	Essentials Têxtil Adulto - Mulher	0.034
240202	Corrida Têxtil - Homem	0.034
261402	Praia - Homem	0.033

Figura 13: Os 10 *itemsets* frequentes com maior suporte ao nível da subcategoria

Fonte: Elaboração própria

À semelhança do que foi feito ao nível da categoria, na tabela da figura 12, a tabela da figura 13 representa os conjuntos frequentes de dados com maior suporte ao nível da subcategoria.

A análise destes resultados permite concluir que é novamente a categoria de Essentials Têxtil Adulto que apresenta maior suporte. Neste caso, a subcategoria de Essentials Têxtil Adulto do género masculino está presente em 5,6% de todas as transações efetuadas na loja em análise.

Verificamos que a categoria 2309 – Essentials Interiores, antes com o segundo maior suporte, tem agora representatividade na 6ª posição devendo-se isto ao facto da categoria em causa ter muitas subcategorias. Ainda assim, meias invisíveis fazem parte de 3,5% das transações da loja.

Praia, pelos mesmos motivos (ter muitas subcategorias) tem representatividade através dos artigos de homem, em 3,3% das transações.

Destacam-se também subcategorias que obtiveram melhor posição no *ranking* do que a sua categoria, como é o caso de: Ginásio Têxtil (devido à subcategoria Mulher) e Alimentação Desportiva (devido à subcategoria Bebidas). Esta questão prende-se com o facto de serem subcategorias bastante frequentes mas pertencerem a categorias que não são tão frequentes:

- Ginásio Têxtil tem apenas duas subcategorias: Ginásio Têxtil-Homem e Ginásio Têxtil-Mulher, no entanto, é uma categoria com bastante mais peso ao nível de volume de vendas em mulher. Quando analisada apenas a categoria, o suporte da subcategoria Ginásio Têxtil-Homem puxa o suporte da categoria para baixo mas quando analisadas as subcategorias, verificamos que Mulher consegue atingir um nível de suporte superior à maior parte das outras subcategorias.
- Alimentação Desportiva comporta-se da mesma forma sendo que a subcategoria Bebidas, devido a uma unidade base em particular, destaca-se a nível de vendas o que permite que o seu suporte garanta uma posição superior à que a categoria atingiu na tabela da figura 12.

Podemos então concluir que a maior parte das subcategorias com maior suporte tinham a respetiva categoria representada no top 10 do respetivo nível, apesar de existirem categorias que conseguem aparecer no top 10 apenas no detalhe à subcategoria como é o caso dos equipamentos de Natação (através dos Acessórios) e Futebol Têxtil (através de Treino e Jogo) que se destacam apenas a este nível uma vez que ambas as categorias têm poucas subcategorias.

Unidade Base	Descrição da Unidade Base	Suporte
23107677	Alimentação Desportiva - Bebidas – Águas	0.036
21037476	Futebol Equipamentos - Bolas - Futebol 11	0.030
22087676	Natação Equipamentos - Acessórios – Toucas	0.029
26018076	Casual Calçado Homem - Urban – Sapatilhas	0.027
23070214	Essentials Têxtil Adulto - Homem - T-shirts	0.024
24010276	Corrida Calçado - Homem – Neutro	0.022
22116901	Porta artigos - Sacos de Desporto – Adulto	0.018
23070234	Essentials Têxtil Adulto - Homem – Calções	0.018
22087667	Natação Equipamentos - Acessórios - Óculos de Natação	0.017
23096002	Essentials Interiores - Meias Invisíveis – Homem	0.016

Figura 14: Os 10 *itemsets* frequentes com maior suporte ao nível da unidade base

Fonte: Elaboração própria

Como referido na análise da tabela da figura 13, a categoria Alimentação Desportiva, devido a uma unidade base em particular (águas), consegue atingir valores elevados de suporte. Tal como analisado quando se abordou a questão de Essentials Interiores (meias), o produto não é suficientemente forte nem especial que obrigue o consumidor a entrar na loja apenas para o adquirir. Assim, as regras de associação permitirão verificar se a posição desta unidade base no *ranking* se deve à relação com outras unidades base frequentes ou à sua localização estratégica dentro da loja.

Conclui-se, portanto, que 3,6% das transações da loja durante todo o ano de 2015 incluíam, pelo menos, uma água.

A subcategoria Essentials Têxtil Adulto - Homem acaba por cair de posição dentro do top10 o que se deve ao facto de, dentro da subcategoria, existirem diversas unidades base. Ainda assim, 2,4% das vendas incluíram t-shirts de homem e 1,8% calções.

Destacam-se novamente os acessórios de natação, nomeadamente toucas e óculos. Será interessante verificar se existem regras de associação que incluam estas duas unidades base e se o mesmo acontece para as duas unidades base de Essentials Têxtil Adulto - Homem presentes no top10: t-shirts e calções.

Os equipamentos de futebol, devido à unidade base Bolas - Futebol de 11 está presente em 3% de todas as transações efetuadas. Uma análise complementar a fazer seria a de verificar as condicionantes nesta unidade base em 2015 ao nível de acordos com as marcas. Por exemplo, se determinada marca é a responsável pela bola de futebol da

primeira liga e apenas a vender à empresa em estudo, as transações destes artigos aumentam, uma vez que particulares e clubes apenas poderiam comprar estes artigos exclusivos nesta companhia.

Analisando a unidade base retiramos conclusões o mais próximo possível da tipologia dos artigos e, consequentemente, do próprio artigo.

De facto, uma análise efetuada ao nível da unidade base acaba por ser melhor e mais interessante do que uma análise ao nível do código de barras do próprio artigo. Ou seja, por exemplo, analisar a tipologia calções é melhor do que analisar um determinado calção específico.

Por um lado, a maior parte dos artigos comercializados tem um ciclo de vida curto, que obedece a estações (Primavera-Verão ou Outono-Inverno) o que obriga a uma maior frequência de análise e a uma rápida resposta face aos resultados sob pena das alterações propostas entrarem em vigor depois do ciclo de vida dos produtos. Por outro lado, a análise ao nível do artigo seria tão redutora que o nível de suporte teria de ser bastante baixo e seria muito difícil encontrar regras de associação fortes que não fossem redundantes (Han e Fu, 1995).

Assim, uma análise de padrões frequentes ao nível da hierarquia é bastante vantajosa. De facto, para além do já referido não necessita de grandes volumes de dados (como os que teria se a análise descesse ao nível do artigo) pelo que a sua computação é mais fácil e permite retirar conclusões mais claras, simples e objetivas.

A estrutura do artigo permite ter uma visão mais geral que potencia ganhos de vendas por associação de determinadas tipologias de produto e não por artigos específicos em si, o que seria bastante complicado de gerir (dado o elevado número de artigos) e se tornaria confuso para os clientes.

Os *itemsets* frequentes encontrados permitem ter uma primeira visão sobre quais as tipologias de artigos mais comercializadas. Resta então derivar destes as associações de compra mais fortes entre produtos.

3.3.3. Regras de Associação para cada Nível Hierárquico

Corrido o algoritmo Apriori para um suporte de 0.001 e confiança de 0.001 foram encontradas regras de associação para cada nível de hierarquia.

O *lift* foi utilizado para escolher os padrões de associação mais fortes e, como uma regra {A→B} tem sempre o mesmo *lift* que a regra {B→A}, apesar de ter valores de confiança diferentes, foi desconsiderada a segunda em prol da primeira.

As 10 regras de associação com maior e menor *lift* consoante critérios já definidos são apresentados nas tabelas das figuras 15, 17 e 19 consoante digam respeito à categoria, subcategoria e unidade base, respetivamente.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
2197	Serviços	2102	Futebol Têxtil	0.667	14.278
2206,2208	Natação Calçado e Natação Equipamentos	2207	Natação Têxtil	0.543	14.097
2206,2207	Natação Calçado e Natação Têxtil	2208	Natação Equipamentos	0.568	13.225
2207,2208	Natação Têxtil e Natação Equipamentos	2206	Natação Calçado	0.153	11.638
2801	Bicicletas	2802	Ciclismo	0.312	7.678
2207	Natação Têxtil	2208	Natação Equipamentos	0.305	7.103
2206	Natação Calçado	2207	Natação Têxtil	0.241	6.250
2206	Natação Calçado	2208	Natação Equipamentos	0.252	5.864
2101,2103	Futebol Calçado e Futebol Equipamentos	2102	Futebol Têxtil	0.246	5.258
2102,2103	Futebol Têxtil e Futebol Equipamentos	2101	Futebol Calçado	0.222	5.050

Figura 15: As 10 regras de associação com maior *lift* ao nível da categoria

Fonte: Elaboração própria

Analisando a informação obtida verificamos que as regras derivadas apenas contêm *itemsets* frequentes e, visto que todos os valores de *lift* são superiores a 1, podemos afirmar que todas as regras de associação, com menor ou maior confiança, representam fortes relações de dependência entre as categorias apresentadas.

Destacam-se as categorias Serviços e Futebol Têxtil como tendo a associação mais forte, uma vez que cerca de 67% das transações que contêm serviços também contêm artigos de Futebol Têxtil. Esta é uma conclusão esperada dado que a categoria Serviços tem, como principal produto, a estampagem de equipamentos que é frequentemente requerida por clubes e particulares.

Apesar da categoria Bicicletas se relacionar com a de Ciclismo (que engloba a parte têxtil e de acessórios), o maior destaque encontra-se nas categorias de Futebol e Natação.

No futebol, para além do já observado em relação a Serviços, verifica-se que as três categorias existentes na unidade de negócio - têxtil, calçado e equipamentos, têm uma forte relação entre si. Estes são valores esperados dado tratar-se de um desporto bastante praticado em Portugal e com características próprias o que implica que seja necessário adquirir material específico.

Por sua vez as categorias de natação estão todas representadas neste top 10. Todas as categorias se relacionam, quer em regras de dois antecedentes quer em regras de apenas um, destacando-se principalmente a regra {Natação Calçado; Natação Equipamentos→Natação Têxtil}. Esta regra, apesar de ter maior *lift*, levanta algumas dúvidas visto ser normal assumir que o tipo de artigo que leva um cliente à loja é o têxtil enquanto o calçado e acessórios funcionam como um complemento ao fato de banho ou calção e não ao contrário. De facto, a regra {Natação Calçado; Natação Têxtil→Natação Equipamentos}, apesar de ter menor *lift* é aquela que tem maior grau de confiança. Concluímos, portanto, que 57% dos clientes que compram calçado e têxtil de natação compram também acessórios de natação.

Tal como no futebol, a forte relação entre todas as categorias da natação deve-se ao facto de se tratar de um desporto específico que necessita de material exclusivamente desenhado para a sua prática.

Nestes dois casos, a criação de *packs* que englobem todos os artigos essenciais à prática do respetivo desporto poderá ser uma mais-valia uma vez que, sabendo a forte relação que têm, o facto de existir uma campanha ou promoção alusiva a essa relação torna ainda mais expectável a sua compra pelo consumidor sendo que esta, quando efetuada, abrangerá todas as categorias.

Depois de apresentadas as regras mais fortes, serão analisadas aquelas que têm *lift* inferior a 1 e, por isso, representam categorias que estão negativamente relacionadas.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
2614	Praia	2402	Corrida Têxtil	0.018	0.02735
2103	Futebol Equipamentos	2307	Essentials Têxtil Adulto	0.027	0.02996
2601	Casual Calçado Homem	2614	Praia	0.023	0.03228
2103	Futebol Equipamentos	2614	Praia	0.024	0.03415
2601	Casual Calçado Homem	2307	Essentials Têxtil Adulto	0.033	0.03693
2303	Ginásio Têxtil	2614	Praia	0.027	0.03839

2208	Natação Equipamentos	2307	Essentials Têxtil Adulto	0.035	0.03986
2310	Alimentação Desportiva	2614	Praia	0.029	0.04212
2702	Outdoor Casual Têxtil	2402	Corrida Têxtil	0.027	0.04233
2802	Ciclismo	2402	Corrida Têxtil	0.028	0.04324

Figura 16: As 10 regras de associação com menor *lift* ao nível da categoria

Fonte: Elaboração própria

A maior parte dos valores encontrados na tabela da figura 16 não são surpreendentes uma vez que se tratam de associações entre categorias bastante diferentes, como é o caso das regras {Praia → Corrida Têxtil}, {Casual Calçado Homem → Praia}, {Natação Equipamentos → Essentials Têxtil Adulto}, entre outras.

Existem também regras que, numa primeira fase poderiam surpreender por terem *lift* inferior a 1 como é o caso concreto da regra {Futebol Equipamentos → Essentials Têxtil Adulto}, no entanto, tal como já referido, este resultado é devido à questão do futebol ser um desporto específico que requer material próprio. O mesmo acontece com o material de ciclismo na regra {Ciclismo → Corrida Têxtil}. Por outro lado, a regra {Futebol Equipamentos → Praia} poderia ser alavancado pela compra de bolas de futebol para jogar na praia, no entanto, os artigos da categoria Futebol Equipamentos são bastante específicos e direcionados para o campo existindo outras categorias dentro da empresa com acessórios de praia.

É também importante destacar uma regra que, com um *lift* de 0.098, representa uma associação negativa entre as duas categorias em causa: Essentials Têxtil Adulto e Essentials Têxtil Júnior e Criança. Estes são dados surpreendentes dada a tendência existente para pensar que os pais, quando compram roupa para si, tendem também a comprar roupa para os filhos.

Um valor de *lift* inferior a 1 representa uma relação negativa entre *itemsets* não se traduzindo necessariamente em canibalismo entre artigos, ou seja, uma vez que apenas analisamos conjuntos frequentes de dados, regras de associação com *lift* inferior a 1 não significam que o cliente apenas compra o antecedente ou o consequente, ou seja, não significa que compra um em detrimento do outro.

Aumentando o grau de análise é apresentado o estudo ao nível da subcategoria.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
260879	Surf Porta Artigos – Estojos	260871	Surf Porta Artigos - Mochilas	0.561	60.220
219776	Serviços - Futebol	210276	Futebol Têxtil - Réplica de Clubes	0.521	47.929
280277	Ciclismo – Têxtil	280278	Ciclismo – Acessórios de Ciclista	0.148	13.123
220709	Natação Têxtil - Criança Rapariga	220876	Natação Equipamentos - Acessórios	0.507	11.797
280280	Ciclismo – Acessórios de Manutenção e Reparação	280279	Ciclismo - Acessórios de Ciclismo	0.150	10.441
280278	Ciclismo – Acessórios de Ciclista	280279	Ciclismo - Acessórios de Ciclismo	0.147	10.212
220706	Natação Têxtil - Júnior Rapariga	220876	Natação Equipamentos - Acessórios	0.368	8.563
220702	Natação Têxtil – Homem	220876	Natação Equipamentos - Acessórios	0.306	7.121
220703	Natação Têxtil - Mulher	220876	Natação Equipamentos - Acessórios	0.299	6.976
220705	Natação Têxtil - Júnior Rapaz	220876	Natação Equipamentos - Acessórios	0.290	6.743

Figura 17: As 10 regras de associação com maior *lift* ao nível da subcategoria

Fonte: Elaboração própria

Analisando a informação obtida na figura 17 pode-se concluir que, tal como acontecia com os *itemsets* frequentes, as regras encontradas para este nível contêm subcategorias cujas categorias também se destacaram na análise a um nível superior. É o caso do Futebol, Natação e Ciclismo, destacando-se, mais uma vez, os dois primeiros.

Com o nível de *lift* mais elevado surge uma associação entre duas subcategorias de Surf Porta Artigos que antes não tinha visibilidade ao nível da relação com outras categorias.

De facto, ao aprofundar o nível de análise observa-se fortes relações entre subcategorias da mesma categoria o que nos leva a um estudo mais profundo que não é afetado pelos resultados do nível anterior. Isto é, o facto de uma categoria não ter uma forte relação com outras não impede que continue a ser analisada podendo até ser encontradas, dentro da mesma, fortes relações entre os seus níveis inferiores. É o caso da regra {Surf Porta Artigos – Estojos→ Surf Porta Artigos – Mochilas}.

A regra referida poderá ser potenciada pela forte atividade promocional que utiliza a associação entre estas duas tipologias, nomeadamente em campanhas de regresso às aulas em que, na compra da mochila existe normalmente a oferta do estojo. Tal como foi

referido acerca de resultados anteriores, uma análise complementar seria a de estudar a sazonalidade desta associação, ou seja, verificar se esta relação mantem os mesmos niveis de confiança e *lift* ao longo do ano ou se depende de determinadas vagas promocionais (regresso às aulas). No entanto, ao contrário dos exemplos anteriores, o que poderia ter maior variação seria o suporte e não propriamente os valores da regra. Neste caso o interesse estaria em perceber se a regra é forte devido a determinados períodos por atividade promocional ou se se mantém constante ao longo do tempo.

Em relação às restantes regras, com subcategorias cujo nível superior já se encontrava no top 10 da análise ao nível da categoria, continuam a destacar-se três áreas principais: Futebol, Natação e Ciclismo.

Em relação ao Futebol, a única regra que se destaca é a que associa Serviços a Equipamentos, neste caso plenamente justificada com os comentários anteriores feitos à categoria uma vez que se trata de uma associação entre Serviços de Estampagem e Réplicas de Clubes. Assim, verificamos que grande parte da associação entre as duas categorias se deve a particulares que requisitam o serviço de estampagem para colocar o nome e/ou número nas camisolas dos clubes de futebol.

Em relação ao Ciclismo, verificam-se fortes relações entre os vários tipos de acessórios comercializados porque, sendo esta uma prática muito específica, é natural que determinados acessórios impliquem a compra de outros, quer para reparação, quer para a manutenção das bicicletas. Assim, encontra-se aqui uma consistente relação que, numa análise extrapolada a diferentes níveis, levariam à descoberta de fortes relações entre a categoria de Bicicletas e subcategorias de acessórios essenciais à sua manutenção e aconselhados no momento de compra.

A Natação continua a destacar-se em todas as suas categorias com particular ênfase nas subcategorias que representam o género feminino tendo em conta a regra vista anteriormente de relação entre o têxtil e os acessórios. Uma análise complementar permitiria observar se é o género feminino quem mais pratica este desporto e, assim, adaptar a gama oferecida ao cliente conforme as conclusões retiradas desse estudo. Paralelamente, poder-se-ia ir mais além e explorar o segmento de idades uma vez que, no caso da subcategoria de Criança Rapariga, cerca de 51% das transações que contêm têxtil também contêm acessórios. Partindo do princípio que, em escalões mais jovens e de iniciação ao desporto são necessários mais acessórios, uma medida a tomar seria a de

identificar quais os produtos essenciais à iniciação da Natação e criar um *pack* que aproveite as fortes relações encontradas e potencie as vendas.

Importa ainda destacar uma regra que, apesar de não ser apresentada como tendo um dos 10 maiores valores de *lift*, acaba por obrigar a retomar a análise de uma conclusão obtida anteriormente, a de que Essentials Têxtil Adulto e Essentials Têxtil Júnior e Criança tinham uma relação negativa ao nível da categoria. Quando a análise desce ao nível da subcategoria verifica-se que existe uma regra de associação entre Essentials Têxtil Adulto-Homem e Essentials Têxtil Júnior e Criança - Rapaz que, apesar de próxima de 0, demonstra que é possível existirem determinados comportamentos num nível e verificar-se o contrário noutro mais baixo.

Em relação à categoria Essentials Interiores, nomeadamente subcategorias de meias, verifica-se a existência de relações positivas com uma variedade de subcategorias onde não só se incluem categorias com quem já estava relacionada, como também outras cuja tipologia de artigo, quando comparado com meias, não faz propriamente sentido: relações entre meias e natação, praia, alimentação desportiva. Por outro lado, na análise anterior, foi verificada uma relação negativa entre a categoria Essentials Interiores e determinados tipos de calçado, nomeadamente calçado de futebol, casual e *outdoor*.

Assim, e uma vez que existem muitas relações em que tanto o antecessor como o consequente são subcategorias de Essentials Interiores (meias), podemos concluir que se tratam de compras de conveniência motivadas pela localização do produto dentro da loja. Por estes artigos se situarem, normalmente, perto da frente de caixa, nas linhas de pagamento e saída da loja representam compras por impulso e não compras por associação a outras tipologias de artigo.

Tal como foi feito para a categoria, de seguida, são analisadas as regras de associação encontradas ao nível da subcategoria com *lift* menos elevado, neste caso, inferior a 1.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
230303	Ginásio Têxtil – Mulher	230702	Essentials Têxtil Adulto – Homem	0.035	0.06192
231076	Alimentação Desportiva – Bebidas	220876	Natação Equipamentos – Acessórios	0.028	0.06574
230960	Essentials Interiores - Meias Invisíveis	220876	Natação Equipamentos – Acessórios	0.029	0.06735

240202	Corrida Têxtil – Homem	230303	Ginásio Têxtil – Mulher	0.033	0.06762
231076	Alimentação Desportiva – Bebidas	230702	Essentials Têxtil Adulto – Homem	0.043	0.07588
261402	Praia – Homem	230702	Essentials Têxtil Adulto – Homem	0.044	0.07753
231076	Alimentação Desportiva – Bebidas	230303	Ginásio Têxtil – Mulher	0.038	0.07859
210278	Futebol Têxtil - Treino e Jogo	231076	Alimentação Desportiva – Bebidas	0.031	0.08135
230703	Essentials Têxtil Adulto – Mulher	231076	Alimentação Desportiva – Bebidas	0.031	0.08193
231076	Alimentação Desportiva - Bebidas	230960	Essentials Interiores - Meias Invisíveis	0.031	0.08617

Figura 18: As 10 regras de associação com menor *lift* ao nível da subcategoria

Fonte: Elaboração própria

Ao analisar as regras com *lift* inferior a 1 e, portanto, representantes de relações negativas, constatamos que alguns dos padrões encontrados já tinham sido verificados ao nível da categoria como é o caso de associações entre Praia e categorias de têxtil desportivo.

Uma primeira análise do *output* alcançado permite verificar que existe uma relação negativa entre a compra de têxtil de mulher e a compra de têxtil de homem, nomeadamente em atividades mais específicas como é o caso da Corrida e de Ginásio.

Por outro lado, apesar de ser um dos *itemsets* mais frequentes, a Alimentação Desportiva, nomeadamente as bebidas, estão negativamente relacionadas com a maior parte das categorias. A análise seguinte, a um nível inferior, permitirá averiguar a questão colocada anteriormente acerca da razão pela qual a água é uma unidade base tão frequente.

É ainda de referir e destacar a regra {Corrida Calçado – Homem→Essentials Têxtil Homem} que se deve ao facto da corrida ser, cada vez mais, praticada por especialistas que preferem adquirir produtos de têxtil com tecnologias direcionadas e intrínsecas à própria atividade.

A análise que se segue representa as regras de associação entre artigos ao nível da unidade base como observadas na tabela da figura 19.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
21027877	Futebol Têxtil - Treino e Jogo – T-Shirts Júnior	21027879	Futebol Têxtil - Treino e Jogo - Calções Júnior	0.605	184.579
26140348	Praia - Mulher - Bikinis Partes Baixo	26140347	Praia - Mulher - Bikinis Partes Cima	0.503	90.617
21977676	Serviços - Futebol - Transfers Oficial	21027601	Futebol Têxtil - Réplica de Clubes - Adulto	0.325	64.021
23107778	Alimentação Desportiva - Alimentação - Proteica	23107776	Alimentação Desportiva - Alimentação – Energética	0.353	57.150
21027876	Futebol Têxtil - Treino e Jogo - T-Shirts Homem	21027878	Futebol Têxtil - Treino e Jogo - Calções Homem	0.401	57.070
21027879	Futebol Têxtil - Treino e Jogo - Calções Júnior	21027859	Futebol Têxtil - Treino e Jogo – Meias	0.394	31.047
21027877	Futebol Têxtil - Treino e Jogo - T-Shirts Júnior	21027859	Futebol Têxtil - Treino e Jogo – Meias	0.371	29.243
24020234	Corrida Têxtil - Homem - Calções	24020214	Corrida Têxtil - Homem - T-Shirts	0.300	21.901
27020327	Outdoor Casual Têxtil - Mulher - Polares Meio Fecho	27020227	Outdoor Casual Têxtil - Homem - Polares Meio Fecho	0.183	19.828
23030316	Ginásio Têxtil - Mulher - No Sleeve	23030335	Ginásio Têxtil - Mulher - Corsários	0.207	19.186

Figura 19: As 10 regras de associação com maior *lift* ao nível da unidade base

Fonte: Elaboração própria

A análise a este nível da hierarquia é a mais próxima possível à análise ao artigo, algo que não poderemos fazer uma vez que estudar relações ao nível do código de barras torna muito difícil encontrar regras de associação fortes.

Assim, esta é a análise mais fiel à tipologia do artigo e, apesar dos baixos valores de parâmetros fixados devido à exaustão da análise, é a que nos permite retirar as conclusões mais detalhadas acerca dos padrões de compra identificados.

À semelhança das análises em níveis anteriores, mantém-se a forte presença do futebol por oposição à natação que, para o nível da unidade base, não apresenta resultados tão elevados no *ranking*. Isto deve-se ao facto de todas as unidades base funcionarem bem como um todo elevando os valores de subcategorias e categorias, mas não destacando, em particular, nenhuma tipologia de nível mais baixo.

No caso de futebol, mantem-se o pressuposto de ser um desporto específico que requer material apropriado e similar, como é o caso da regra entre os calções e t-shirts de

júnior bem como todas as outras que dizem respeito a associações entre diferentes unidades base de têxtil. Por outro lado, continua a verificar-se a forte relação entre os serviços de estampagem e as camisolas dos clubes. Seria interessante analisar esta relação ao longo do tempo. Ou seja, seria interessante verificar se num ano de Mundial ou Europeu de Futebol esta relação se torna mais forte ou se se mantém estável. O caso particular de 2016 em que Portugal se sagrou Campeão Europeu de Futebol seria um bom barómetro para medir esta relação, tendo sempre presente que o impacto desta análise seria muito superior no estudo da frequência de compra, apesar de também permitir uma visão da estabilidade desta regra ao longo do tempo.

Uma regra que se destaca é a que diz respeito á categoria de Praia, nomeadamente {Praia – Mulher – Bikinis Partes Baixo → Praia – Mulher – Bikinis Partes Cima} uma vez que, ao contrário de outras empresas no mercado, este tipo de produto é comercializado em separado podendo o cliente escolher diferentes peças. É, assim, uma relação bastante óbvia, apesar do valor do nível de confiança ser surpreendente na medida em que seria esperado que mais de 50% das transações que contêm partes de baixo contivessem também partes de cima. Este valor confirma que a decisão de vender bikinis como duas peças separadas é acertada na medida em que apenas metade dos clientes compra os dois artigos em conjunto.

Ao nível da unidade base destacam-se ainda regras de produtos complementares como é o caso de calções e t-shirts na corrida e corsários e camisolas sem mangas no segmento feminino de ginásio.

De uma forma geral, ao nível do têxtil, verifica-se que as regras de associação que contêm partes de baixo no antecedente e partes de cima no consequente, por exemplo {Corrida Têxtil - Homem - Calções → Corrida Têxtil - Homem - T-Shirts}, têm níveis de confiança mais elevados do que as regras com a situação contrária. Estas regras fazem sentido uma vez que, normalmente, qualquer pessoa possui mais partes de cima do que de baixo.

A análise ao nível da unidade base permite abordar a questão colocada anteriormente acerca da compra de água. Uma vez que não foram encontradas fortes relações entre este artigo e outros, conclui-se que o nível de suporte da unidade base Alimentação Desportiva – Bebidas – Água se deve à sua localização estratégica dentro

da loja. Localizadas no corredor de saída para as caixas são, na sua maioria, compras por impulso e não planeadas.

As águas são um artigo de conveniência que os clientes compram, não por necessidade, mas por força do momento aliado ao baixo preço exercido neste tipo de produtos. Tal como referido na análise da categoria de Praia, seria interessante analisar se o suporte desta tipologia aumenta em determinadas alturas do ano em que a temperatura ambiente é bastante elevada permitindo, assim, ações de reaprovisionamento de mais unidades do que o normal em loja e, por conseguinte, potenciar ainda mais este tipo de compra por impulso.

No caso da análise à unidade base não existem tipologias de artigos com *lift* inferior a 1 pelo que se depreende que, para os parâmetros definidos, não existem associações negativamente relacionadas.

O estudo de padrões frequentes de compra ao nível da hierarquia permite identificar de uma forma simples quais as tipologias de artigos que mais se relacionam entre si, algo que não aconteceria se a análise descesse ao detalhe do próprio artigo.

Tal como já foi mencionado, o facto de se estudar relações entre códigos de barras tornaria quase impossível encontrar regras de associação mesmo que a base de dados fosse mais alargada e os critérios fossem fixados a um nível muito reduzido.

Por outro lado, a análise ao nível da estrutura permite tomar medidas direcionadas a tipologias de produtos que se mantêm constantes ao longo do tempo e, por isso, pressupõe um menor esforço de revisão e adaptação. Isto é, se o estudo fosse ao nível do artigo, seriam tomadas medidas que potenciassem apenas a sua venda e que, para além de serem mais difíceis de encontrar, se não fossem postas em prática rapidamente correriam o risco de entrar em vigor muito perto da data de saída de linha do artigo, uma vez que a maior parte dos produtos em análise é sazonal.

Assim, o estudo de regras de associação ao nível da tipologia do produto tem grandes vantagens, mas também algumas limitações uma vez que apenas é feita dentro de cada nível, isto é, categoria com categoria, subcategoria com subcategoria e unidade base com unidade base.

O número de tipologias dentro de cada nível hierárquico é distribuído de forma diferente dependendo da categoria. Por exemplo, a categoria de Essentials Têxtil Adulto

tem duas subcategorias que são o género: homem e mulher e a unidade base representa o tipo de artigo - calções, t-shirts, calças, casacos, entre outros. Por sua vez, a categoria de Essentials Interiores define o tipo de artigo em 14 subcategorias (meias curtas, meias invisíveis, t-shirts, etc.) enquanto o género está definido ao nível da unidade base.

Outro exemplo é o de Praia que tem apenas uma categoria, no entanto os acessórios de praia são uma das muitas subcategorias de Desportos Aquáticos. Assim, mesmo que exista uma forte relação entre a categoria Praia e a subcategoria Equipamentos de Praia esta não é visível dado que, logo no primeiro nível (análise entre as categorias de Praia e Desportos Aquáticos), as muitas subcategorias dos Desportos Aquáticos acabam por ter um impacto negativo no suporte desta categoria. Por outro lado, os Equipamentos de Praia poderão ter uma forte relação com Praia mas isso não quer dizer que o tenham com uma sua subcategoria em particular e, por isso, não fora encontradas, nesta análise, regras fortes que o demonstrem.

Uma forma de colmatar esta falha é fazer uma análise de padrões frequentes considerando os diferentes níveis que compõem a estrutura do produto através do estudo das Regras de Associação Hierárquica apresentadas de seguida.

3.4. Regras de Associação Hierárquica

Terminada a primeira abordagem, a identificação de padrões relevantes será afinada através da consideração da hierarquia dos artigos, ou seja, serão encontrados padrões frequentes entre tipologias de artigos independentemente do nível hierárquico a que pertençam.

Por forma a continuar a análise nos mesmos moldes e linha de raciocínio seguidos anteriormente, as Regras de Associação Hierárquica serão encontradas tendo por base os resultados previamente apresentados e utilizando o mesmo *software*.

Apesar de existirem outras técnicas, nomeadamente os algoritmos testados por Srikant e Agrawal em 1995, será continuado o estudo utilizando o programa R e o algoritmo Apriori. Este *software*, apesar de permitir utilizar comandos específicos para derivação destas regras, como é o caso da função *Agregatte*, possibilita também a utilização da mesma lógica usada na primeira fase do estudo sendo necessário apenas alterar a base de dados.

Os 3 ficheiros iniciais foram fundidos por forma a que, em cada transação, estejam identificadas as categorias, subcategorias e unidades base adquiridas.

Antes de correr o algoritmo Apriori utilizando os códigos apresentados anteriormente nas figuras 9 e 10, foi efetuado um teste de sensibilidade aos parâmetros por forma a definir os valores de suporte e confiança a serem utilizados.

Verificou-se que se podia aumentar os dois parâmetros face aos valores utilizados anteriormente sem, no entanto, aumentar demasiado o suporte uma vez que existem várias hierarquias na mesma transação e hierarquias mais baixas pressupõem um nível de suporte mais baixo. Assim, foi fixado um suporte mínimo de 0,01 (superior ao definido anteriormente) e efetuado um teste de sensibilidade ao parâmetro que permite definir a confiança cujo resultado é apresentado na figura 20.

Confiança	*Itemsets* frequentes	Regras de Associação
0.01	106	302
0.05	12	302

Figura 20: Teste à sensibilidade do parâmetro confiança para um nível de suporte de 0.01

Fonte: Elaboração própria

Dado os resultados obtidos foi escolhida uma confiança de 0.01 uma vez que, para o mesmo número de regras, proporciona um maior número de *itemsets* frequentes.

Assim, serão encontrados conjuntos de dados frequentes que tenham um suporte superior a 1% e derivadas regras de associação com um nível de confiança mínimo de 1%, ou seja, serão encontrado códigos de artigos presentes em, pelo menos, 1% das transações e derivadas regras de associação em que a probabilidade do consequente ser adquirido sendo que o antecedente foi comprado é de, pelo menos, 1%.

Ao executar o algoritmo Apriori para o novo conjunto de dados foram encontrados os conjuntos de *itemsets* frequentes que analisados de seguida.

Estrutura	Descrição da Estrutura	Suporte
2307	Essentials Têxtil Adulto	0.088
2309	Essentials Interiores	0.074
2614	Praia	0.070
2402	Corrida Têxtil	0.064
2103	Futebol Equipamentos	0.060
230702	Essentials Têxtil Adulto - Homem	0.056
2401	Corrida Calçado	0.055
2303	Ginásio Têxtil	0.055
2601	Casual Calçado Homem	0.054
2310	Alimentação Desportiva	0.053

Figura 21: Os 10 *itemsets* frequentes com maior suporte independentemente do nível da hierarquia

Fonte: Elaboração própria

Analisando a informação obtida, verificamos que os níveis representados no Top 10 pertencem maioritariamente ao nível da categoria visto que, tal como já foi referido anteriormente, códigos de 4 dígitos são mais frequentes que os restantes porque todos os códigos de 6 ou 8 dígitos implicam a presença dos seus níveis superiores.

Importa destacar a presença da subcategoria Essentials Têxtil Adulto - Homem que tem representatividade em 5,6% de todas as transações. Tendo em conta a primeira estrutura da tabela, este era um resultado de esperar uma vez que, caso o top 10 apresentasse uma subcategoria, esta deveria pertencer a um nível mais baixo de Essentials Têxtil Adulto.

Os restantes resultados que estão de acordo com o que foi analisado anteriormente são importantes mas não tanto como a análise das regras de associação.

Com o objetivo de suprimir regras redundantes, foram eliminadas todas as regras com um nível de confiança de 1, dado tratar-se de regras em que o antecedente é uma subcategoria ou unidade base do consequente. Ou seja, foram eliminadas regras em que adquirir o antecedente implica sempre adquirir o consequente como é o caso do exemplo {Essentials Têxtil Adulto - Homem – Calças → Essentials Têxtil Adulto - Homem} em que, sempre que um cliente compra calças da subcategoria Essentials Têxtil Adulto - Homem estará a comprar essa mesma subcategoria.

Após eliminação das regras com 100% de confiança, foi elaborado o top 10 de regras de associação com maior *lift*, no entanto e como podemos observar, todas as regras dizem respeito a tipologias dentro da mesma categoria de artigos.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
260480	Casual Calçado Criança - Urban	26048076	Casual Calçado Criança - Urban - Sapatilhas	0.961	90.457
2604;260480	Casual Calçado Criança e Casual Calçado Criança - Urban	26048076	Casual Calçado Criança - Urban - Sapatilhas	0.961	90.457
260280	Casual Calçado Mulher - Urban	26028076	Casual Calçado Mulher - Urban - Sapatilhas	0.923	87.999
2602;260280	Casual Calçado Mulher e Casual Calçado Mulher - Urban	26028076	Casual Calçado Mulher - Urban - Sapatilhas	0.923	87.999
260179	Casual Calçado Homem - Surfwear	26017976	Casual Calçado Homem - Surfwear - Sapatilhas	0.746	71.477
2601;260179	Casual Calçado Homem e Casual Calçado Homem - Surfwear	26017976	Casual Calçado Homem - Surfwear - Sapatilhas	0.746	71.477
260582	Casual Calçado Bebé - Estruturado	26058276	Casual Calçado Bebé - Estruturado - Sapatilhas	0.927	69.546
2605;260582	Casual Calçado Bebé e Casual Calçado Bebé - Estruturado	26058276	Casual Calçado Bebé - Estruturado - Sapatilhas	0.927	69.546
240103	Corrida Calçado - Mulher	24010376	Corrida Calçado - Mulher - Neutro	0.780	57.160
2401;240103	Corrida Calçado e Corrida Calçado - Mulher	24010376	Corrida Calçado - Mulher - Neutro	0.780	57.160

Figura 22: As 10 regras de associação hierárquica com maior *lift* para um suporte mínimo de 0.01 e uma confiança de 0.01

Fonte: Elaboração própria

Todas as regras apresentadas na tabela anterior são redundantes na medida em que, apesar de apresentarem antecedentes e consequentes de níveis hierárquicos diferentes, são pouco relevantes porque apenas dizem respeito à mesma categoria.

Estas regras apenas permitem aferir quais as subcategorias e unidades base mais importantes de cada categoria o que não é o que se pretende com este estudo. É, no entanto uma análise complementar interessante suscetível de ser aprofundada.

Por forma a encontrar as regras de associação hierárquica mais relevantes, foram apenas selecionadas regras que tenham pelo menos duas categorias diferentes. As regras encontradas são as que constam na figura 23.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
2207	Natação Têxtil	2208	Natação Equipamentos	0.305	7.103
2207	Natação Têxtil	220876	Natação Equipamentos - Acessórios	0.305	7.103
2208	Natação Equipamentos	2207	Natação Têxtil	0.274	7.103
220876	Natação Equipamentos – Acessórios	2207	Natação Têxtil	0.274	7.103
2208;220876	Natação Equipamentos e Natação Equipamentos – Acessórios	2207	Natação Têxtil	0.274	7.103
2303	Ginásio Têxtil	2307	Essentials Têxtil Adulto	0.210	2.373
2307	Essentials Têxtil Adulto	2303	Ginásio Têxtil	0.129	2.373

Figura 23: Regras de associação hierárquica para um suporte mínimo de 0.01 e uma confiança de 0.01 após eliminação de regras redundantes

Fonte: Elaboração própria

Dados os valores definidos para os parâmetros e a eliminação de regras redundantes, apenas foram encontradas 7 regras de associação hierárquica consideradas relevantes, no entanto, quando analisadas mais profundamente concluímos que os critérios poderiam ser mais afinados.

Algumas das regras já foram analisadas anteriormente uma vez que, apesar de dizerem respeito a categorias diferentes, se situam no meso nível hierárquico como é o caso da regra {Natação Têxtil → Natação Equipamentos} que diz respeito a uma associação entre duas tipologias ao nível da categoria. Dado que esta análise já foi feita, apenas nos interessa estudar casos que associam artigos situados em diferentes níveis hierárquicos.

Por outro lado, verificou-se também que algumas regras poderiam ser desconsideradas visto que acabam por induzir conclusões muito semelhantes às retiradas por outras regras. Utilizando o exemplo anterior, poder-se-ia concluir que a regra {Natação Têxtil → Natação Equipamentos} permite retirar as mesmas conclusões que a regra {Natação Têxtil → Natação Equipamentos - Acessórios} e, uma vez que o objetivo é ser o mais minucioso possível e descer a análise o mais perto possível do artigo deve considerar-se a segunda regra em detrimento da primeira.

Para além das razões acima enumeradas, as regras apresentadas acabam por perder interesse devido aos valores de *lift* alcançados e, uma vez que são poucas, a sua análise não permite tirar vantagem suficiente o que torna inócuo o aprofundamento do seu estudo com o fim de aplicar o conhecimento obtido em casos práticos.

Assim, manteve-se o nível de confiança de 0.01 e foi ajustado o suporte para 0.005 com o objetivo de encontrar um maior número de regras de associação sem, no entanto, manter os mesmos níveis da primeira análise. Os novos valores definidos para cada parâmetro permitiram a obtenção de 106 *itemsets* frequentes e 735 regras de associação, sendo que os 10 conjuntos frequentes de dados com maior suporte se mantêm face à análise anterior e podem ser revisitados na figura 21.

O processo de eliminação de regras redundantes foi conduzido à semelhança do anterior. Foram eliminadas todas as regras cujo antecedente faz parte do consequente (nas quais estão as regras com 100% de confiança), foram desconsideradas regras que apenas diziam respeito a uma categoria e regras que já foram estudadas anteriormente, uma vez que analisam estruturas com o mesmo nível. Foram também desconsideradas regras que se repetiam através da troca de uma estrutura pela estrutura do seu grau superior. A figura 24 apresenta um conjunto de regras redundantes.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição Do Consequente	Confiança	Lift
2207;220876	Natação Têxtil E Natação Equipamentos – Acessórios	22087676	Natação Equipamentos - Acessórios - Toucas	0.831	28.776
2207;2208	Natação Têxtil E Natação Equipamentos	22087676	Natação Equipamentos - Acessórios – Toucas	0.831	28.776
2207;2208;220876	Natação Têxtil E Natação Equipamentos E Natação Equipamentos – Acessórios	22087676	Natação Equipamentos - Acessórios - Toucas	0.831	28.776

Figura 24: Exemplo de regras de associação hierárquica redundantes

Fonte: Elaboração própria

No exemplo da figura anterior é facilmente percetível a repetição de padrões através da substituição de um nível inferior pelo nível superior e, uma vez que as 3 regras têm os mesmos valores de confiança e *lift* apenas se considera a primeira visto ser a mais simples

e próxima da análise ao nível do artigo. Assim, pode-se concluir que cerca de 83% dos clientes que compram Natação Têxtil e Natação Equipamentos – Acessórios compram Toucas. Sendo o consequente parte de um dos antecedentes, poderemos afirmar, de forma mais simples, que os clientes que compram artigos de Natação Têxtil tendem a comprar também toucas de natação.

Eliminando todas as regras consideradas redundantes obtiveram-se os resultados apresentados no quadro abaixo e que dizem respeito a todas as regras de associação hierárquicas encontradas que foram consideradas interessantes.

Código do Antecedente	Descrição do Antecedente	Código do Consequente	Descrição do Consequente	Confiança	Lift
2207;220876	Natação Têxtil e Natação Equipamentos - Acessórios	22087676	Natação Equipamentos - Acessórios - Toucas	0.831	28.776
230303;2307	Ginásio Têxtil - Mulher e Essentials Têxtil Adulto	230703	Essentials Têxtil Adulto - Mulher	0.864	25.575
2303;230703	Ginásio Têxtil e Essentials Têxtil Adulto - Mulher	230303	Ginásio Têxtil - Mulher	0.989	20.500
2303;2309	Ginásio Têxtil e Essentials Interiores	230303	Ginásio Têxtil - Mulher	0.854	17.727
2307;2309	Essentials Têxtil Adulto e Essentials Interiores	230702	Essentials Têxtil Adulto - Homem	0.648	11.512
22087676	Natação Equipamentos - Acessórios - Toucas	2207	Natação Têxtil	0.338	8.776
230303	Ginásio Têxtil - Mulher	2307	Essentials Têxtil Adulto	0.199	2.251
2303;230303	Ginásio Têxtil e Ginásio Têxtil - Mulher	2307	Essentials Têxtil Adulto	0.199	2.251
230702	Essentials Têxtil Adulto - Homem	2309	Essentials Interiores	0.115	1.559
230303	Ginásio Têxtil - Mulher	2309	Essentials Interiores	0.105	1.420

Figura 25: Regras de associação hierárquica relevantes para um suporte mínimo de 0.005 e um nível de confiança de 0.01

Fonte: Elaboração própria

Numa primeira análise, os resultados da figura 25 permitem destacar 3 tipologias: artigos de Natação, roupa de Ginásio Mulher e roupa de Essentials Homem. O relevo destas 3 áreas era de esperar uma vez que são tipologias bastante frequentes, como foi referido anteriormente. A Natação tem regras de associação bastante fortes devido às suas

características particulares e Ginásio Mulher e Essentials Homem são subcategorias bastante frequentes, com um grande nível de suporte.

A Natação destaca-se pela associação entre toucas e têxtil, sendo que 83% das transações que incluíram Natação Têxtil também incluíram Toucas e 34% dos clientes que compram Toucas também compraram Natação Têxtil.

Todas as outras categorias em análise pertencem à área de Fitness onde se faz notar a importância do Têxtil Ginásio no segmento feminino e do Têxtil mais básico de Essentials no segmento Masculino, ambas subcategorias muito frequentes.

Concluímos, pois, que as categorias de têxtil no segmento feminino estão bastante relacionadas entre si uma vez que cerca de 87% das mulheres que compram Ginásio Têxtil compram também artigos mais básicos de Essentials e cerca de 99% das mulheres que compram Essentials também compram Ginásio Têxtil. Estes resultados devem-se ao facto de Essentials Adulto, alavancada pelo segmento masculino, ser uma categoria mais frequente do que Ginásio Têxtil. Assim, a regra que contém o nível da categoria diminui o valor da confiança apesar do valor elevado de *lift*, enquanto que a regra que contém a subcategoria (segmento feminino), aumenta a confiança mas tem menor *lift*.

No segmento masculino não é visível uma forte relação entre Ginásio e Essentials, no entanto verifica-se uma forte associação entre Essentials Têxtil Adulto – Homem e Essentials Interiores. Esta categoria, já mencionada na análise anterior devido à sua frequência, tem também uma forte relação com Ginásio e Essentials Mulher.

Um dos resultados que seria esperado obter desta análise seria uma relação forte entre a categoria de Praia e a subcategoria de Desportos Aquáticos – Equipamentos de Praia, no entanto, o suporte destas estruturas não permitiu aferir o grau desta relação.

As Regras de Associação Hierárquica derivadas são menos de que as que foram obtidas na análise feita por níveis. No entanto, o conhecimento adquirido permitiu uma visão interessante e diferente do negócio através dos padrões relevantes encontrados o que se revela bastante importante e vantajoso para uma aplicação prática em áreas de foco no cliente e potenciação de vendas.

3.4.1. Visualização do *Output*

Quando se trabalha um grande volume de dados e se obtêm bastantes resultados, a sua visualização torna-se muito importante para facilitar a análise e para retirar

conclusões a partir dos padrões encontrados. Por outro lado e uma vez que se trata de encontrar relações fortes entre tipologias de produto, este tipo de estudo requer uma análise mais dinâmica que a visualização de resultados ajuda a alcançar

Assim, por forma a compreender melhor os resultados alcançados acima, foi utilizado o *sub-package* do *arules* (*arules Viz)*, que permite a reprodução gráfica dos resultados obtidos com o algoritmo Apriori.

Numa primeira fase, foram identificadas as estruturas mais frequentes e com representação num maior número de regras. Os resultados encontrados constam da matriz representada na figura 27 criada pelo comando da figura 26.

```
plot(rules, method="grouped", control=list(k=20))
```

Figura 26: Comando para visualização das 20 estruturas presentes em mais regras

Fonte: Elaboração própria

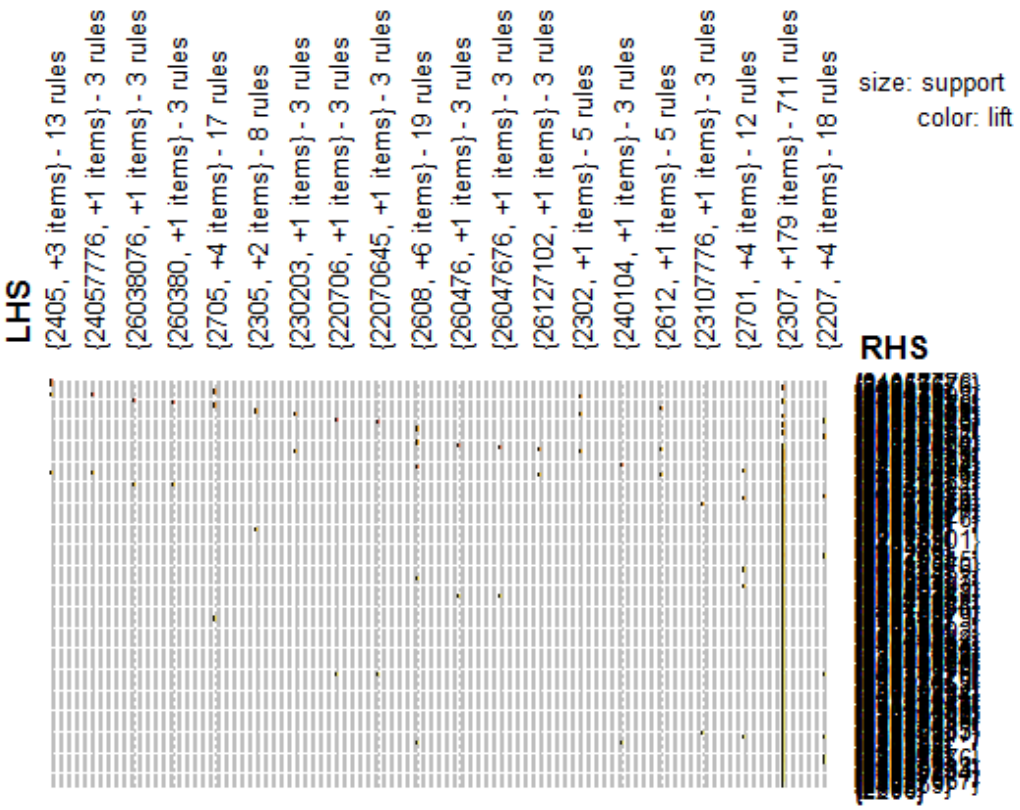


Figura 27: Representação das 20 estruturas presentes em mais regras

Fonte: *Output* do R

A figura 27 representa uma matriz que indica as relações que existem entre os antecedentes das regras (LHS – na linha superior horizontal) e os consequentes dessas mesmas regras (RHS – na coluna vertical).

Apresenta, no eixo superior horizontal, as 20 estruturas mais frequentes sendo indicado em quantas regras são o antecedente e o número de *itemsets* com que estão associados.

Não é possível distinguir os códigos que figuram no eixo vertical uma vez que existem bastantes consequentes a fazer parte das regras que incluem os 20 antecedentes mais frequentes.

A categoria com maior destaque é a 2307 – Essentials Têxtil Adulto uma vez que, pela análise feita anteriormente, é a que tem maior suporte estando presente em 711 regras. As restantes estruturas têm suporte inferior ao desta categoria, no entanto é importante frisar que os gráficos não eliminam regras redundantes uma vez que foram feitos com base no universo total de regras derivadas.

A matriz indica o suporte através do tamanho de cada estrutura sendo o *lift* da regra representado por cores. No entanto, devido à grande quantidade de resultados, estes valores não são percetíveis.

Por forma a colmatar as falhas desta visualização foi gerado um gráfico de redes sociais através do comando da figura 28.

```
subrules2 <- head(sort(rules, by="lift"), 50)
plot(subrules2, method="graph")
```

Figura 28: Comando para visualização das 50 regras de associação hierárquica com maior *lift* através de um gráfico de redes sociais

Fonte: Elaboração própria

O comando da figura 28 permite a criação de um gráfico de redes sociais, utilizado no âmbito da *Social Network Analysis*, que identifica as estruturas por pontos e as relações entre elas por arestas.

De forma a facilitar a análise, foram apenas selecionadas as 50 regras com maior *lift*. O gráfico em questão encontra-se na figura 29.

Figura 29: Representação das 50 regras de associação hierárquica com maior *lift*

Fonte: *Output* do R

A figura 29 permite uma visão mais simples dos resultados alcançados anteriormente. Porém também considera o universo total de regras não eliminando, por isso, regras redundantes.

Utilizando, como exemplo, as associações do canto superior esquerdo verifica-se que existem relações entre a categoria 2302 – Ginásio Calçado, a subcategoria 230203 – Ginásio Calçado – Mulher e a unidade base 23020303 – Ginásio Calçado – Mulher – Treino. Estas são regras redundantes uma vez que, sempre que é comprada uma unidade base os seus níveis superiores também fazem parte da transação, ou seja, comprar Ginásio Calçado – Mulher – Treino implica comprar Ginásio Calçado – Mulher que, por sua vez, implica comprar Calçado Ginásio.

Todos os resultados apresentados no gráfico representam relações existentes entre níveis da mesma categoria e, uma vez que na sua maioria são categorias com poucos níveis inferiores, o grau de relação entre eles é maior.

Sendo esta a forma mais simples de representar os resultados obtidos, foi alterada a base de dados para que contenha apenas transações que resultem nas regras não redundantes que foram encontradas na figura 25. A representação destas regras é feita através do gráfico da figura 30.

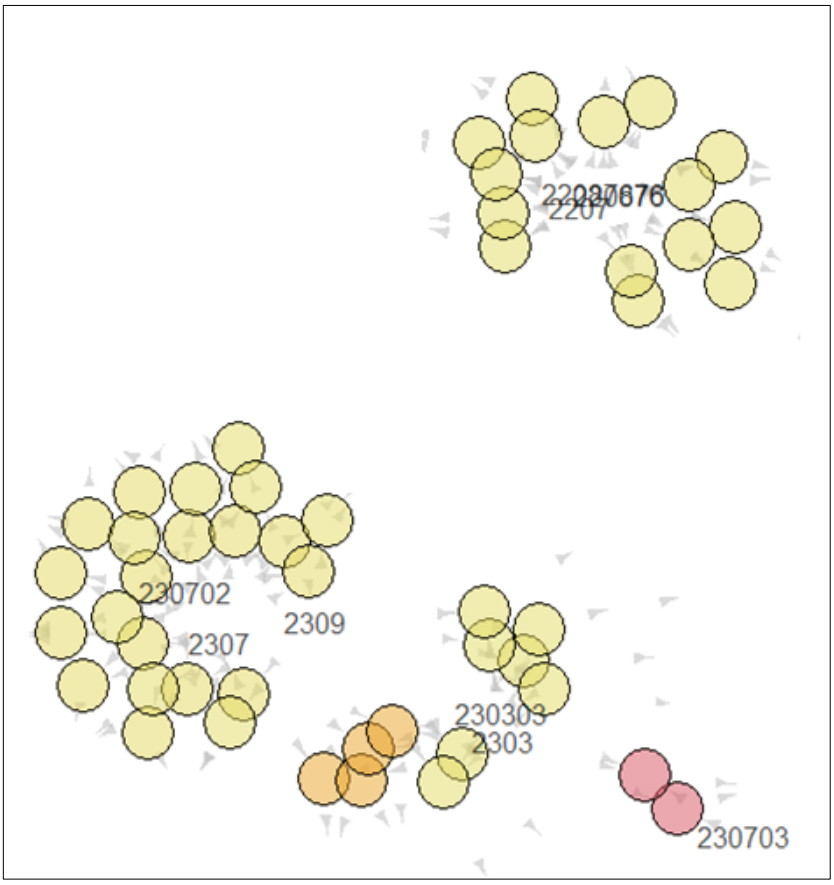


Figura 30: Representação das 50 regras de associação hierárquica relevantes com maior *lift*

Fonte: Elaboração própria com base no *output* do R

No gráfico da figura 30 é facilmente visível o elevado suporte que a categoria 2307 – Essentials Adulto e as suas subcategorias têm através do número de regras de que fazem parte. A associação entre esta categoria e as categorias 2303 – Ginásio Têxtil, sobretudo devido a Mulher, e 2309 – Essentials Interiores destaca-se pelo número de regras que representam e pela sua forte proximidade e relação.

Verificamos, também, as associações entre os artigos de Natação, representados pelas categorias 2207 – Natação Têxtil e 2208 – Natação Equipamentos.

De uma forma geral, as associações derivadas entre os diferentes níveis hierárquicos respeitam os *itemsets* frequentes encontrados quer na análise anterior, entre níveis, quer na análise multinível. Algumas das regras mais fortes estudadas em cada nível culminaram em regras de associação hierárquica pela relação com outras hierarquias mas, geralmente, de categorias semelhantes.

A análise multinível permite encontrar relações não identificadas anteriormente devido à forma como as estruturas estão construídas, nomeadamente, a quantidade de níveis inferiores que têm por cada nível superior e a tipologia de artigos em cada nível (em alguns casos a subcategoria é o género e noutros é o tipo de produto).

No entanto, a dimensão da base de dados não permitiu aprofundar algumas questões referidas anteriormente como é o caso da relação entre Praia e Acessórios de Praia, Bicicletas e Câmaras de Ar, entre outras. De facto, a maior parte dos clientes que compra uma bicicleta compra também uma câmara de ar, mas como as câmaras de ar são uma unidade base de Acessórios de Ciclismo e esta subcategoria tem muitas unidades base, o suporte mínimo definido e a extensão da base de dados não permite retirar conclusões sobre esta relação.

O exemplo de Praia demonstra a limitação provocada pela dimensão da base de dados e a falta de informação sobre as relações de compra com bicicletas podem dever-se ao facto de não serem artigos de aquisição muito frequente devido ao seu elevado preço e durabilidade.

Uma análise complementar que poderá ser feita para colmatar as falhas acima apontadas é a realização deste estudo focado em determinados grupos. Ou seja, identificados os tipos de artigo que se quer analisar e que, devido às suas particularidades, têm baixo nível de suporte, podem criar-se subgrupos e analisar esses produtos só com as categorias que poderão ter alguma relação. No exemplo das bicicletas, poder-se-iam isolar as categorias de Bicicletas e Ciclismo e considerar este o universo total de transações de forma a conseguir retirar conclusões relevantes com elevado nível de suporte e confiança.

4. Aplicações Práticas

O estudo elaborado no capítulo anterior, uma vez que foi conduzido ao nível da estrutura dos artigos, permite extrair conhecimento relevante e retirar conclusões facilmente aplicáveis na empresa em análise através da utilização de técnicas com o propósito de atrair os clientes e potenciar vendas.

O estudo da associação de produtos com uma forte relação de dependência, efetuado por técnicas de *Market Basket Analysis*, permite aferir os padrões de compra dos consumidores conduzindo a diversas análises com aplicações práticas que serão aprofundadas no subcapítulo de investigação futura da conclusão.

As aplicações do conhecimento obtido podem ser inúmeras dada a tipologia da empresa de estudo uma vez que se trata de uma organização de retalho onde a informação sobre o cliente, principalmente sobre os seus padrões de consumo, proporciona bastantes oportunidades de melhoria com impacto nos resultados da empresa. No entanto, este estudo será focado em duas áreas em particular que, apesar da aparente simplicidade de análise que requerem, têm um grande impacto nas vendas da companhia.

A ênfase das aplicações deste estudo passará, então, por uma análise e possível proposta de alteração de *layout* de loja e pela construção de uma ferramenta de potenciação de vendas.

4.1. Análise de *Layouts* de Loja

A proposta de alteração de *layout* passa por uma análise da alocação dos artigos integrantes das regras de associação com maior nível de *lift* e confiança por forma a perceber se artigos que influenciam a presença de outros na mesma transação, estão próximos uns dos outros na loja.

Uma vez que o objetivo é o de potenciar vendas pretende-se verificar se os artigos estão próximos por forma a aumentar, ainda mais, a probabilidade de um artigo ser adquirido sendo que o outro é comprado. Utilizando como exemplo a regra {Essentials Têxtil Adulo – Homem – Calções → Essentials Têxtil Adulto – Homem – T-Shirts} pretende-se verificar se os calções e as t-shirts de homem estão no mesmo espaço dentro da loja por forma a que, se um cliente se deslocar à loja para comprar uns calções, a conveniência de ter t-shirts no mesmo local, aumente a probabilidade do cliente adquirir também este artigo.

Por outro lado, para regras com elevado nível de confiança, em que os *itemsets* que as integram são bastante dependentes, o raciocínio seria o oposto, ou seja, o de afastar os artigos dentro da loja por forma a obrigar o cliente a percorrer uma maior área e ver mais produtos.

Como as regras encontradas obedecem ao primeiro exemplo, a análise do *layout* da loja será no sentido de verificar a aproximação dos artigos pertencentes a regras de associação com elevado nível de confiança.

Analisando as regras de associação e conjuntos de dados frequentes, verificamos que o *layout* da loja em análise se encontra perfeitamente ajustado aos resultados alcançados uma vez que artigos com elevado grau de dependência estão próximos dentro da loja.

As categorias de Natação estão juntas dentro da loja e os chinelos de praia, que são permanentes em loja, também se encontram dentro do espaço de natação.

Categorias mais específicas, como é o caso do Futebol e da Corrida, têm as suas categorias aproximadas dentro das lojas uma vez que o tipo de cliente que adquire estes produtos pretende comprar produtos técnicos inerentes ao respetivo desporto.

Por outro lado, as bebidas (águas) e os interiores (meias) localizam-se junto ao *check out* da loja visto serem produtos de conveniência que, a maior parte das vezes, acabam por ser compras por impulso.

Por fim, todas as áreas de ginásio e roupa desportiva mais básica estão no mesmo espaço da loja sendo que, em algumas situações, o têxtil das categorias de Essentials e Ginásio encontra-se misturado. Parte dos interiores, pela forte relação que tem com a categoria de Essentials Adulto, não só está no *check out* mas também junto ao têxtil e calçado de homem.

A análise da organização da loja é uma questão bastante complexa uma vez que acaba por poder ser tanto uma aplicação prática dos resultados como também a sua origem. Assim, uma vez que o espaço está perfeitamente ajustado às conclusões obtidas, é difícil perceber se é a organização da loja que potenciou os resultados ou se, pelo contrário, foram os resultados que levaram a uma mudança estrutural.

A questão anterior pode ser respondida através de um estudo complementar. Uma vez que a loja em análise sofreu uma remodelação no final de 2014, realizar o mesmo estudo com dados anteriores à remodelação permitiria perceber se os dados se

mantiveram constantes, e por isso a remodelação já os teve em conta ou se, pelo contrário, os resultados obtidos pelos dados de 2015 refletem a nova disposição dos produtos na loja.

4.2. Ferramenta de Proposta de Campanhas

O estudo de regras de associação revela-se muito importante na tomada de decisões de Marketing, nomeadamente no processo de opções assertivas no âmbito da configuração de campanhas promocionais cujo objetivo é alcançarem um elevado nível de adesão por parte dos clientes.

Neste caso em particular, o conhecimento obtido no estudo de padrões frequentes, será utilizado para alavancar vendas de produtos que não estejam a cumprir os seus objetivos através de campanhas associadas a outros com quem tenham forte relação e que estejam a alcançar as vendas pretendidas.

Apoiada numa análise de histórico, *targets* de vendas e conjugada com os padrões de consumo encontrados, a ferramenta em análise permitirá a proposta de mecânicas promocionais que utilizem a informação obtida na extração de padrões frequentes com o objetivo de aumentar vendas e alcançar uma maior taxa de adesão, bem como torná-las mais eficazes e lucrativas. Se, por exemplo, existir uma forte relação X → Y, a ferramenta sugerirá uma promoção em Y se as vendas deste forem inferiores ao critério definido (sugeriria uma promoção do tipo: na compra de X, desconto em Y).

A ferramenta em análise será desenvolvida para o nível mais baixo da hierarquia, a unidade base, uma vez que é o que permite uma maior aproximação à tipologia de produto e, assim, obter resultados mais relevantes e com maior impacto sobre o cliente.

A construção da ferramenta de potenciação de vendas será desenvolvida em Excel e terá como apoio a base de vendas por unidade base, as regras de associação relevantes encontradas na secção anterior e os *targets* de vendas que cada estrutura deverá atingir.

As regras de associação em análise serão ordenadas pelo grau de confiança e não pelo *lift*, como definido até agora. Uma vez que esta aplicação pretende atrair os clientes para uma determinada compra, é mais importante considerar a probabilidade existente desse cliente comprar um artigo sabendo que comprou outro produto.

Assim, a análise começará com um estudo do desvio da unidade base face ao *target* de vendas. Se o objetivo de vendas não for atingido, serão analisadas as regras de associação de que a estrutura faz parte no consequente da regra.

Encontrada a regra com maior nível de confiança, verifica-se se o antecedente da regra atinge as vendas pretendidas e, em caso positivo, será proposta uma campanha em que, na compra do antecedente se oferece um desconto na compra do consequente.

Por exemplo, se ao analisar a unidade base Natação Equipamentos – Acessórios – Toucas se verificar que os objetivos não estão a ser cumpridos, analisar-se-á o antecedente da regra {Natação Têxtil; Natação Equipamentos – Acessórios → Natação Equipamentos – Acessórios – Toucas}. Neste caso, uma vez que as toucas fazem parte dos Equipamentos de Natação – Acessórios interessa apenas analisar as vendas de Natação Têxtil e, se estas cumprirem o objetivo, poderá então propor-se a campanha "Na compra de artigos de Natação Têxtil oferece-se um desconto de x% na compra de toucas de natação".

A percentagem de desconto poderá ser definida pelo grau de distância a que a estrutura se encontra do seu objetivo de vendas. Para diferenças maiores face ao *target*, poderá ser oferecido um desconto mais elevado para aumentar a probabilidade do cliente adquirir o consequente da regra.

Se, por outro lado, o antecedente analisado também não cumprir o *target* de vendas poderá ser proposto um *pack* com desconto.

No exemplo da regra {Futebol Têxtil - Treino e Jogo - T-Shirts Júnior → Futebol Têxtil - Treino e Jogo – Calções Júnior}, se as duas unidades base não cumprirem os objetivos definidos, poderá ser criado um *pack* promocional em que, na compra dos dois artigos, o cliente terá determinado desconto sendo este ajustado ao *gap* que exista entre as vendas conseguidas e as que deveriam acontecer.

Esta é uma aplicação bastante simples de implementar e que irá permitir colmatar falhas, principalmente no âmbito da tomada de decisão de campanhas. Mesmo os profissionais experientes acabam por se basear naquilo que pensam que está associado e esta análise, apoiada em factos e valores de relações, poderá desmistificar regras que as equipas comercias tomavam como certas e alertar para outras que não julgavam tão interessantes.

A consideração da informação do cartão de cliente, no âmbito desta análise, tornaria mais rica a sua aplicação prática.

Uma vez que cerca de 50% dos clientes desta empresa são fidelizados, uma análise individual de padrões de consumo permitiria um estudo complementar à abordagem referida anteriormente. Com esta informação, seria possível estudar padrões frequentes de clientes específicos e criar *clusters* de clientes com o fim de dirigir, de forma mais assertiva e direta, campanhas específicas para cada tipo de cliente.

Estas campanhas segmentadas teriam um elevado impacto visto que incidiriam em clientes já fidelizados e, por isso, com maior probabilidade de aderir à atividade promocional da empresa. Por outro lado criariam, também, oportunidade de aliciar clientes não tão assíduos a visitar a loja em questão.

As aplicações práticas referidas dizem respeito à loja que foi analisada podendo, no entanto, ser aplicadas em todas as outras lojas, individualmente ou no universo total, da companhia.

5. Conclusão

As tomadas de decisão com que nos deparamos diariamente deverão ser suportadas por estudos sustentados na informação disponível, muitas vezes armazenada em grandes volumes de dados difíceis de ler e explorar.

Data Mining permite a exploração de padrões frequentes e respetiva extração de conhecimento com o objetivo de o aplicar no âmbito de melhorias contínuas em casos reais. De entre as suas áreas destaca-se a *Market Basket Analysis* que se revela fundamental no estudo de transações de retalho e se foca na aplicação do conhecimento das relações relevantes encontradas com o objetivo de potenciar vendas e atrair o cliente.

Neste âmbito foram estudadas as transações de uma empresa portuguesa de artigos de desporto para identificar os conjuntos de dados mais frequentes e, destes, derivar regras de associação relevantes que tivessem em conta a estrutura a que os artigos pertencem. O estudo de Regras de Associação Hierárquica teve como objetivo a aplicação do conhecimento adquirido através de propostas de organização em loja que aproximem os produtos com associação mais forte e de sugestões de campanhas que aproveitem as relações encontradas para promover vendas de artigos com piores resultados.

O suporte deste estudo teve como base a linguagem e *software* R, uma vez que é acessível, gratuito e bastante completo permitindo desenvolver toda a análise com recurso ao mesmo algoritmo, o Apriori.

Foi escolhido o *lift* como métrica de validação, fixou-se um nível de confiança baixo de 0,001 e, após realização de um teste de sensibilidade ao suporte, foi definido um suporte mínimo de 0.001 e encontrados os *itemsets* frequentes para cada nível hierárquico.

De um modo geral, categorias consideradas bastante frequentes conduziram a subcategorias frequentes e, consequentemente, estas culminaram em unidades base com suporte também elevado. De entre estes resultados destacaram-se as categorias de Essentials Têxtil Adulto com um peso de 8,8% no número total de transações e Praia que, apesar de ser uma categoria sazonal apresenta o terceiro maior suporte. De entre as subcategorias destaca-se o segmento masculino de Essentials Têxtil Adulto e o segmento feminino de Ginásio Têxtil, alcançando também bons resultados a Natação Equipamentos, nomeadamente acessórios. Descendo a uma análise o mais próxima possível do artigo, a tipologia de artigo mais comercializada foi a água seguida de bolas de futebol de 11 e toucas de natação.

Quando considerada a hierarquia dos artigos para um suporte de 0.01, os resultados foram semelhantes aos dos obtidos ao nível da categoria, uma vez que é mais frequente aparecerem níveis mais altos nas transações do que níveis mais baixos. Ainda assim, destaca-se a presença de uma subcategoria no top 10 de estruturas com maior suporte, Essentials Têxtil Adulto – Homem com presença em 5,6% das transações.

Dos *itemsets* frequentes foram derivadas regras de associação para cada nível. As regras de associação mais fortes dizem respeito a tipologias com a mesma unidade de negócio, destacando-se na natação as relações entre o têxtil, calçado e acessórios e no futebol a associação entre os serviços e as réplicas de clubes. Existe também uma relação muito forte entre os estojos e as mochilas e a compra de partes de baixo e partes de cima em que a regra {Parte de Baixo → Parte de Cima} apresenta maior confiança que a regra {Parte de Cima → Parte de Baixo} dado que, em média, cada pessoa possui mais partes de cima do que de baixo. Dentre estas, destacam-se as regras {Bikinis Partes de Baixo → Bikinis Partes de Cima} e {Calções → T-shirts}.

Alguns dos conjuntos de dados frequentes, nomeadamente água e meias, não integraram regras de associação relevantes pelo que se depreendeu que o seu grau de frequência se deve ao local onde se encontram em loja e não à compra de artigos complementares visto serem produtos comprados por impulso.

Quando considerada a hierarquia dos produtos, as regras de associação derivadas foram na sua maioria irrelevantes para a análise em estudo. Fixado um suporte mínimo de 0,005 e definido um nível de confiança de 0.01, foram eliminadas todas as regras redundantes, ou seja, todas as regras que apresentavam 100% de confiança por questões de estruturação da hierarquia, regras entre tipologias do mesmo nível, que incidissem sobre a mesma categoria e que repetissem regras anteriores substituindo apenas níveis inferiores por níveis superiores.

As regras de associação hierárquica obtidas focaram principalmente duas áreas, natação e têxtil desportivo.

À semelhança das relações estudadas na primeira fase, as regras encontradas na área de natação dizem respeito a relações entre têxtil e acessórios, nomeadamente toucas.

Por sua vez, o têxtil desportivo fez-se representar em várias categorias: Ginásio Têxtil, Essentials Têxtil Adulto e Essentials Interiores. As regras encontradas permitiram observar a forte relação {Ginásio Têxtil - Mulher → Essentials Têxtil Adulto - Mulher}

e {Essentials Têxtil Adulto – Homem → Essentials Interiores} sendo que, de uma forma geral, os interiores – representados na sua maioria por meias – são um complemento à compra de têxtil desportivo e as categorias de Ginásio Têxtil e Essentials Têxtil Adulto complementam-se entre si.

O estudo efetuado, ao considerar a estrutura dos artigos, permitiu retirar ilações passíveis de aplicação, ao contrário do que aconteceria se a análise tivesse sido feita ao nível do artigo e não da sua tipologia. Assim, este trabalho apresenta resultados e conclusões que se manterão para além do ciclo de vida dos produtos.

A limitação que o estudo em cada hierarquia implica foi ultrapassada através do estudo das Regras de Associação Hierárquica. A análise multinível colmatou o facto do número de tipologias dentro de cada nível não ser uniforme e a informação em cada nível variar de categoria para categoria, isto é, existem categorias em que o género é definido à subcategoria e outras em que a subcategoria representa a tipologia do produto.

De entre as aplicações práticas possíveis, os resultados obtidos foram utilizados para analisar a disposição dos artigos em loja e construir uma ferramenta de sugestão de campanhas com base nas regras encontradas.

Verificou-se que o *layout* da loja em análise está de acordo com os resultados obtidos e a utilização da informação obtida como suporte de campanhas é a principal aplicação deste estudo, uma vez que a sua aplicação prática é simples, flexível e com resultados fáceis de monitorizar, permitindo ainda desmistificar ideias pré-concebidas acerca de como os artigos se relacionam entre si.

5.1. Limitações

Apesar das imensas vantagens e aplicações deste estudo, a dimensão da base de dados bem como a classificação dos produtos em categorias abrangentes impediu o alcance de conclusões mais profundas acerca das relações multinível. Exemplo disto é o facto de não terem sido encontradas relações entre a categoria Praia e a subcategoria de Desportos Aquáticos - Acessórios de Praia.

Por outro lado, o suporte fixado não permitiu considerar categorias que, apesar de importantes, não são frequentes devido aos elevados preços que os seus artigos têm. É o caso da categoria de Bicicletas que, considerado outro suporte, teriam uma grande probabilidade de fazer parte de regras juntamente com câmaras de ar.

Para além dos constrangimentos da base de dados, a visualização das Regras de Associação Hierárquica tornou-se difícil uma vez que o programa utilizado, só por si, não eliminou regras redundantes, tendo sido estas desconsideradas manualmente.

5.2. Investigação Futura

Os resultados obtidos servirão de base a trabalhos futuros que permitirão aprofundar as aplicações práticas focadas nesta análise.

Em relação ao *layout* de loja será estudada a remodelação por que esta passou em 2014 por forma a identificar se a disposição que hoje se observa em loja foi fruto de uma análise semelhante a esta ou se, pelo contrário, os resultados obtidos neste trabalho foram influenciados pelo *layout* já existente.

A ferramenta de proposta de campanhas será colocada em prática na empresa em estudo aprofundada pela possibilidade de utilizar a informação de fidelização de clientes para alargar a análise a um âmbito inter-transação que permita desenvolver campanhas segmentadas e direcionadas a determinado tipo de consumidores.

Uma análise complementar passaria por realizar este estudo em determinados períodos de tempo para perceber a influência que as estações ou determinados eventos têm nos resultados. De facto, considerando distintos períodos de tempo, o estudo da frequência de conjuntos de dados conduzirá a uma análise sazonal que poderá incidir sobre o aprovisionamento desses conjuntos em determinadas épocas sazonais. Por exemplo a Praia, que é uma categoria sazonal, terá um maior grau de aprovisionamento durantes os meses de Verão.

Por forma a colmatar a limitação de categorias importantes mas que, por serem menos frequentes pelos preços que praticam, não figurarem nos resultados, poderão ser criados subgrupos de análise que englobem apenas essas categorias e outras com que tenham determinado tipo de relação. Seria o caso de considerar o sub-grupo de bicicletas e ciclismo como o universo total por forma a encontrar relações como a da categoria Bicicletas com a da unidade base Ciclismo – Acessórios de Manutenção e Reparação – Câmaras de Ar.

Uma análise de dependência entre tipologias poderá proporcionar um estudo de substituição entre artigos. No caso das fortes relações entre Essentials Adulto - Mulher e Ginásio Têxtil – Mulher, se as t-shirts forem sempre adquiridas na primeira categoria e

as *leggings* na segunda, poderá não haver necessidade de incluir t-shirts em Ginásio e *leggings* em Essentials. Se as duas subcategorias se complementam, as compras poderão ser feitas tendo como base a procura do cliente e não a procura individual em cada categoria.

O estudo da frequência dos artigos permite, também, uma análise de *slow* e *fast movers*, ou seja, permite a identificação dos artigos em que não devemos ou devemos apostar, respetivamente. A aplicação prática deste conhecimento é fundamental para oferecer ao cliente uma seleção de produtos adequados, diminuir as sobras de produtos que não são procurados e rentabilizar o espaço em loja.

Outro estudo a implementar passaria pela análise da combinação entre a tipologia e as marcas dos produtos por forma perceber se os clientes que compram determinado tipo de marca apenas compram tipologias dessa marca ou se também adquirem outras.

Do presente trabalho resultará um *paper* a ser submetido à Conferência Internacional de Inteligência Artificial a realizar-se no Porto, em Setembro de 2017.

Referências Bibliográficas

Agrawal, R., Imielinski, T e Swami, A. (1993), "Mining Association Rules between Sets of Items in Large Databases", *Acm sigmod record*, Vol. 22, Nº 2, pp. 207-216.

Albion Research Ltd. (2016), "Market Basket Analysis: What is it?", http://www.albionresearch.com/data_mining/market_basket.php, acedido em 5 de Junho de 2016.

Azevedo, P. e Jorge, A. (2007), "Comparing Rule Measures for Predictive Association Rules", *European Conference on Machine Learning*, pp. 510-517.

Bastos, G. (2001), "Algumas Aplicações Práticas da Tecnologia Data Mining", http://livrozilla.com/doc/1021304/algumas-aplica%C3%A7%C3%B5es-pr%C3%A1ticas-da-tecnologia-data-mining, acedido em 15 de Agosto de 2016.

Brijs, T. (2002), "Retail market basket analysis: a quantitative modelling approach", dissertação para obtenção do grau de doutor em ciências económicas aplicadas, Limburg University Center.

Brin, S., Motwani, R., Ullman, J. e Tsur, S. (1997), "Dynamic itemset counting and implication rules for market basket data", *ACM SIGMOD Record*, Vol. 26, Nº 2, pp. 255-264.

Domingues, M. e Rezende, S. (2011), "Using Taxonomies to Facilitate the Analysis of the Association Rules", *arXiv preprint arXiv:1112.1734.*

Gama, J., Carvalho, A., Faceli, K., Lorena, A., Oliveira, M. (2012), *Extração de Conhecimento de Dados –Data Mining*, Lisboa: Edições Sílabo.

Geyer, C (2003), "Regras de Associação Aplicadas aos Filtros de Mensagens e Canais de Informação do Projeto Direto", dissertação de doutoramento, Universidade Federal do Rio Grande do Sul.

Gutierrez, N. (2006), "Demystifying Market Basket Analysis", *DM Review Special Report*.

Hahsler (2016), M, "Support for Item Hierarchies", http://finzi.psych.upenn.edu/library/arules/html/aggregate.html, acedido em 21 de Agosto de 2016.

Hahsler, M. (2015), "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules",

http://michael.hahsler.net/research/association_rules/measures.html, acedido em 10 de Agosto de 2016.

Hahsler, M. e Hornik, K. (2008), "New Probabilistic Interest Mesures for Association Rules", *Intelligent Data Analysis*, Vol. 11, Nº 5, pp. 437-455.

Hahsler, M. e Karpienko, R. (2011), "Visualizing Association Rules in Hierarchical Groups", *Journal of Business Economics*, pp. 1-19.

Hahsler, M., Chelluboina, S. (2011), "Visualizing Association Rules: Introduction to the R-extension Package arulesViz", *R project module*, pp.223-238.

Hahsler, M., Hornik, K., Grun, B. e Buchta, C. (2005), "A computational environment for mining association rules and frequent item sets ", *Journal of Statistical Software*. Vol. 14, Nº 15, pp. 1-25.

Han, J. e Fu, Y. (1995), "Discovery of Multiple-Level Association Rules from Large Databases", *VLDB*, Vol. 95, pp. 420-431.

Jain, D. e Gautam, S. (2013) "Implementation of Apriori Algorithm in Health Care Sector: A Survey", *International Journal of Computer Science and Communication Engineering*, Vol. 2, Nº 4.

Kantardzic, M. (2002), *Data Mining: Concepts, Models, Methods, and Algorithms*, IEEE Press & John Wiley.

Limitedbrands (2004), "Achieving Greater Efficiencies with Market Basket Analysis", Microstrategy World Conference, Miami.

Marafi, S. (2014), "Market Basket Analysis with R", http://www.salemmarafi.com/code/market-basket-analysis-with-r/, acedido em 21 de Agosto de 2016.

Martin, T., Shen, Y. e Azvine, B. (2007), "A Mass Assignment Approach to Granular Association Rules for Multiple Taxonomies", *Proceedings of the Third International Conference on Uncertainty Reasoning for the Semantic Web*, Vol. 327, pp. 25-36.

McCormick, T., Rudin, C. e Madigan, D. (2011), "A Hierarchical Model for Association Rule Mining or Sequential Events: an Approach to Automated Medical Sympto Prediction", *Annals of Applied Statistics.*

Microstrategy (2003) "Business Inteligence in the Retail Industry", Microstrategy World Conference, Las Vegas.

Qualls, B. (2013), "Introduction to Market Basket Analysis", AA.

Raeder, T. e Chawla, N. (2011), "Market Basket Analysis with Networks", *Social Network. Analysis and Mining*, Vol. 1, Nº 2, pp. 97-113.

Rodrigues, M., Gama, J. e Ferreira, C. (2012), "Identifying Relationships in Transactional Data", *Ibero-American Conference on Artificial Intelligence,* pp. 81-90.

Silwattananusarn. T. e Tuamsuk, K (2012), "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, Vol. 2, Nº 5.

Srikant, R. e Agrawal, R. (1995), "Mining Generalized Association Rules", *VLDB*, Vol. 95, pp. 407-419.

Srikant, R., Vu, Q. e Agrawal, R. (1997), "Mining Association Rules with Item Constraints", *KDD*, Vol. 97, pp. 67-73.

StatSoft (2007), "What is Data Mining, and How is it Useful for Power Plant Optimization? (and How is it Different from DOE, CFD, Statistical Modeling)", https://www.statsoft.com/Portals/0/Support/Download/White-Papers/What_Is_Data_Mining.pdf, acedido em 10 de Janeiro de 2016.

Svetina, M. e Zupančič, J. (2005), "How to Increase Sales in Retail with Market Basket Analysis", *Systems Integration*, pp. 418-428.

Ulas, M. (1999), "Market Basket Analysis for Data Mining", dissertação para obtenção do grau de mestre em ciências e engenharia computacional, Bogaziçi University.

Verma, R. (2009), *The Data Mining Hypertextook*, http://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter002/section004/blue/page002.html, acedido em 20 de Julho de 2016.

Wang, K., He, Y. e Han, J. (2000), "Mining Frequent Itemsets Using Support Constraints", *VLDB*, pp. 43-52.

Zaki, M. (2000), "Generating non-redundant association rules", *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 34-43.

Zhao, Y., Zhang, C. e Cao, L. (2009), *Post-Mining of Association Rules:Techniques for Effective Knowledge Extraction*, IGI Global.