UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Gustavo Bonnard Schonhorst

Mineração de Regras de Associação Aplicada à Modelagem dos Dados Transacionais de um Supermercado

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção como requisito parcial à obtenção do título de ***Mestre em Ciências em Engenharia de Produção***.

Área de Concentração: Qualidade e Produto.

Orientador: Prof. Pedro Paulo Balestrassi, Dr.

Março de 2010

Itajubá - MG

Ficha catalográfica elaborada pela Biblioteca Mauá –
Bibliotecária Margareth Ribeiro- CRB_6/1700

S371m

Schonhorst, Gustavo Bonnard
Mineração de regras de associação aplicada à modelagem dos dados transacionais de um supermercado / Gustavo Bonnard Schonhorst. -- Itajubá, (MG) : [s.n.], 2010.
68 p. : il.

Orientador: Prof. Dr. Pedro Paulo Balestrassi.
Dissertação (Mestrado) – Universidade Federal de Itajubá.

1. Mineração de dados. 2. Regras de associação. 3. Market Basket Analysis. 4. Supermercado. I. Balestrassi, Pedro Paulo, orient. II. Universidade Federal de Itajubá. III. Título.

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Gustavo Bonnard Schonhorst

Mineração de Regras de Associação Aplicada à Modelagem dos Dados Transacionais de um Supermercado

Dissertação aprovada por banca examinadora em 22 de março de 2010, conferindo ao autor o título de ***Mestre em Ciências em Engenharia de Produção***

Banca Examinadora:

Prof.ª Dra. Marcela Aparecida Guerreiro Machado

Prof. Dr. Anderson Paulo de Paiva

Prof. Dr. Pedro Paulo Balestrassi (Orientador)

Março de 2010

Itajubá - MG

EPÍGRAFE

“A vida só pode ser
compreendida olhando-se
para trás; mas só pode
ser vivida olhando-se
para a frente.”

Soren Kierkegaard

DEDICATÓRIA

Dedico este trabalho ao meus pais Erno Schonhorst e Joycelem Bonnard Schonhorst por todo carinho e apoio.

AGRADECIMENTOS

Ao Professor Pedro Paulo Balestrassi pelo incentivo, orientação e amizade;

À Professora Marcela Aparecida Guerreiro Machado por gentilmente aceitar o convite para avaliação deste trabalho, bem como suas sugestões de melhoria do mesmo;

Ao amigo Rubens pelas incontáveis horas de discussão sobre mineração de dados e trocas de conhecimentos;

Ao Ricardo da *Intelliware* pela paciência e auxílio no processo de obtenção dos dados utilizados neste trabalho;

À minha família e principalmente aos meus pais por todo o carinho, apoio e por me proporcionarem oportunidades para que eu viesse a ter a formação pessoal e profissional que tenho hoje;

Ao meu irmão e também engenheiro Bruno Bonnard Schonhorst. Seu carinho e apoio a mim e à nossa família foram fundamentais para a minha permanência em Itajubá;

À minha namorada Izabela, por todo carinho e compreensão;

Finalmente aos meus amigos de república e aqueles que não moram, mas estão sempre lá, pela amizade e por transformarem aquela casa em um ambiente agradável, ou melhor, familiar.

SUMÁRIO

RESUMO

A descoberta de regras de associação é uma tarefa de mineração de dados que vem sendo estudada desde o início da década de 1990. Uma das principais aplicações destas regras é na análise de cesto de compras. Neste tipo de problema busca-se por padrões no comportamento dos consumidores, adquirindo o conhecimento de quais produtos costumam ser levados juntos em uma mesma compra. O conhecimento adquirido pode ser utilizado como suporte a tomada de decisões tanto em nível operacional como em nível estratégico. O crescente interesse de pesquisadores nesta área se deve tanto a utilidade prática como as dificuldades e limitações presentes neste tipo de análise. Mesmo assim aplicações reais ainda são poucas. O objetivo deste trabalho é realizar uma análise de cesto de compras através da mineração de regras de associação sobre os dados transacionais de um supermercado de forma a proporcionar maior conhecimento do negócio. Foram consideradas na análise centenas de milhares de compras relativas a quatro meses. A técnica utilizada neste trabalho mostrou-se capaz de gerar grande quantidade de conhecimento útil à tomada de decisões. A definição de um foco específico mostrou-se fundamental para o sucesso da análise.

ABSTRACT

The discovery of association rules is a data mining task that has been studied since the early 1990s. A major application of these rules is the market basket analysis. In this type of problem the goal is to search for patterns in consumer behavior, acquiring knowledge of what products are usually brought together in a single purchase. The knowledge can be used to support decision making at both operational and strategic level. The growing interest of researchers in this area is due to both practical use and the difficulties and limitations present in this type of analysis. Yet real applications are still few. The objective of this work is to conduct a market basket analysis through association rules mining on transactional data from a supermarket in order to provide greater business insight. There were considered 256,096 purchases of four months. The technique used in this study proved capable of generating large amount of useful knowledge for decision making. The definition of a specific focus proved to be crucial to the success of the analysis.

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. Introdução

1.1 Considerações Iniciais

As duas últimas décadas foram marcadas pelo avanço de tecnologias de armazenamento e processamento de dados. Atualmente, armazenar e processar grandes quantidades de dados não são mais tarefas complexas e de alto custo. O desafio hoje está em outra questão: Como transformar estes dados em informações e estas informações em conhecimento a fim de prover suporte à tomada de decisões? O dinamismo do mercado aumentou a demanda por informações precisas, detalhadas e atualizadas. As organizações estão cada vez mais interessadas em extrair conhecimento útil de suas bases de dados com o objetivo de obter maior vantagem competitiva em relação aos seus concorrentes.

Durante anos, métodos predominantemente manuais foram utilizados com o objetivo de transformar dados em conhecimento. Tratando-se de grandes bases de dados, estes métodos são dispendiosos (em termos financeiros e de tempo), subjetivos e na maioria das vezes impraticáveis (Fayyad et al., 1996a). A busca por métodos mais rápidos e eficientes de transformação de dados em conhecimentos incentivou a realização de pesquisas nesta área, que hoje é conhecida na literatura como *Knowledge Discovery in Database* (KDD), *data mining* ou mineração de dados.

Segundo Rezende et al. (2003), a mineração de dados tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo de tomada de decisão. A descoberta de regras de associação é uma técnica de mineração de dados que já vem sendo estudada desde o início da década de 1990. A tarefa de associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens no mesmo registro (Agrawal e Srikant, 1994).

Uma das principais e mais estudadas aplicações das regras de associação é na análise de cesto de compras, mais conhecida na literatura por M*arket Basket Analysis* (MBA). Neste tipo de problema busca-se por padrões no comportamento dos consumidores, adquirindo o conhecimento, por exemplo, de quais produtos costumam ser levados juntos em uma mesma compra (Brin et al., 1998). O conhecimento adquirido através de uma análise de cesto de compras pode ser utilizado como suporte à tomada de decisões tanto em nível operacional como em nível estratégico. Segundo Chen et al. (2005), o MBA pode ajudar os gerentes de organizações no projeto de *layout*, *web* sites, *mix* de produtos e outras estratégias de marketing.

Pretende-se com este trabalho realizar uma análise de cesto de compras através da mineração de regras de associação sobre os dados transacionais de um supermercado localizado no sul do estado de Minas Gerais. A seguir são apresentados: objetivos, justificativa, limitações, e a metodologia de pesquisa utilizada.

1.2 Objetivo

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é realizar uma análise de cesto de compras através da mineração de regras de associação sobre os dados transacionais de um supermercado de forma a proporcionar maior conhecimento do negócio. Quando gerado, o conhecimento deverá ser representado por padrões no comportamento dos consumidores, isto é, por regras que demonstrem quais itens geralmente são comprados em conjunto pelos clientes.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Descobrir padrões no comportamento de compra dos consumidores;
- Descobrir oportunidades de venda cruzada de produtos (*cross-selling*);
- Discutir a aplicabilidade do conhecimento descoberto no auxílio à tomada de decisões.

1.3 Justificativa

Segundo Brijs et al. (2004), enquanto muitos pesquisadores têm contribuído significativamente com o desenvolvimento de algoritmos eficientes para geração de regras de associação, a literatura no que diz respeito ao uso desta técnica em aplicações reais ainda é limitada. Segundo o autor, uma ampla aceitação de regras de associação como uma valiosa técnica para a solução de problemas de negócio irá depender do sucesso de aplicações futuras em dados concretos, reais.

Para Piatetsky-Shapiro (2007), grande sucesso foi alcançado na melhoria de desempenho dos algoritmos de busca de regras de associação e muitos deles foram amplamente aceitos, mas aplicações em problemas reais ainda são poucas.

Ferramentas analíticas de auxílio à tomada de decisões estão ajudando varejistas a conquistarem novos clientes e manter fiéis àqueles que já possuem. Segundo Grewal e Levy (2007), a utilização de técnicas de análise de dados para a tomada de decisões no varejo

relacionadas ao CRM (*Customer Relationship Management*) em geral e à programas de fidelidade em particular representam novas e promissoras áreas para pesquisas acadêmicas.

Um outro fator que valoriza esta pesquisa é o fato de que no Brasil, pesquisas relacionadas à mineração de regras de associação ainda são poucas. Os principais trabalhos se resumem a algumas dissertações e teses de mestrado e doutorado respectivamente, do curso de Ciências da Computação da Universidade de São Paulo (USP). Mesmo assim estes trabalhos têm foco mais voltado para a complexidade computacional deste tipo de problema, diferentemente desta dissertação. Uma das explicações da falta de aplicações reais de regras de associação é a dificuldade que existe em se obter uma base de dados real para a análise.

Este trabalho pretende contribuir cientificamente com uma aplicação real de mineração de regras de associação para extração de padrões no comportamento de compra dos consumidores. A utilização de regras de associação na descoberta de padrões no comportamento dos consumidores de um supermercado se caracteriza como um método estatístico aplicado à melhoria de Qualidade e Produtividade, assuntos inerentes a Engenharia de Produção.

Enquanto dissertação de mestrado, esta pesquisa não tem o objetivo de esgotar todas as questões sobre regras de associação e análise de cesto de compras já que são áreas muito amplas e delas derivaram diversas subáreas devido à intensidade das pesquisas científicas.

1.4 Limitações

A principal limitação desta dissertação de mestrado é relativa ao objeto de estudo que é o banco de dados de um supermercado do sul do estado de Minas Gerais. Qualquer padrão encontrado através de análises neste banco não necessariamente se confirmará em um outro supermercado. Além disso, mesmo os padrões encontrados no supermercado estudado podem não ser os mesmos em tempos diferentes, já que o comportamento dos consumidores pode variar de acordo com o tempo.

1.5 Metodologia de Pesquisa

A pesquisa realizada para este trabalho é de natureza aplicada devido ao interesse prático em seus resultados. Quanto aos seus objetivos, é uma pesquisa descritiva, pois busca descrever as características de determinada população e estabelecer relações entre variáveis. Quanto à forma de abordar o problema é uma pesquisa combinada. Quantitativa por tratar de análise de informações numéricas e qualitativa em virtude da análise dos resultados com fonte no ambiente natural pesquisado. O método de pesquisa empregado é a modelagem (Bertrand e Fransoo, 2002; Lakatos e Marconi, 2001).

Esta pesquisa consiste na utilização de uma técnica de mineração de dados conhecida como regras de associação, para modelar o comportamento de compra de consumidores de um supermercado localizado no sul do estado de Minas Gerais, gerando conhecimento útil à tomada de decisões. A Figura 1.1 mostra as etapas envolvidas na modelagem.

Figura 1.1 – Etapas da Modelagem

A seta da Figura 1.1 mostra que existe a possibilidade de realimentação devido às limitações e desafios referentes à capacidade de processamento e de armazenamento, presentes em qualquer processo de mineração de dados. Estas limitações são discutidas no Capítulo 2. Cada etapa da modelagem é explicada em detalhes no Capítulo 3.

1.6 Estrutura do Trabalho

Este trabalho está dividido em 4 capítulos, incluindo-se a introdução.

No **Capítulo 2** é realizada uma revisão de literatura sobre Mineração de Dados e Regras de Associação. São apresentadas as etapas envolvidas no processo de mineração de dados, as definições e os conceitos de regras de associação, os desafios envolvidos na utilização desta técnica, as medidas de avaliação do conhecimento e finalmente é apresentado o problema da análise de cesto de compras, mais conhecido como *market basket analysis*. O **Capítulo 3** trata de como foi conduzida a pesquisa. Neste capítulo são apresentados os resultados, são realizadas análises sobre estes e uma discussão sobre a aplicabilidade da técnica. Finalmente, no **Capítulo 4** são apresentadas as conclusões obtidas, bem como sugestões para trabalhos futuros.

2. Mineração de Dados e Regras de Associação

2.1 Considerações Iniciais

O processo de coletar e armazenar grande quantidade de dados tem se tornado cada vez mais fácil e barato devido ao avanço de tecnologias de processamento e armazenamento. O montante de dados existente hoje nas bases de dados da maioria das organizações excede em muito a capacidade humana de realizar análises. Segundo Witten e Frank (2005), ao mesmo tempo em que o volume de dados cresce em taxas, há apenas alguns anos, consideradas inimagináveis, a proporção destes dados que as pessoas conseguem entender diminui drasticamente. Diversas pesquisas têm sido direcionadas ao desenvolvimento de tecnologias de extração automática de conhecimento a partir dos dados. Esta área de pesquisa é conhecida hoje como mineração de dados.

Atualmente existe uma grande variedade de técnicas de busca de tendências, padrões e correlações em grande volume de dados. A descoberta de regras de associação é uma técnica de mineração de dados que tem recebido grande atenção de pesquisadores. Segundo Hipp et al. (2002), a associação se tornou uma técnica de mineração muito popular em função de sua aplicabilidade a problemas de negócio juntamente com sua compreensibilidade inerente, pois até mesmo não especialistas em mineração de dados conseguem compreendê-las.

Neste capítulo é apresentada uma visão geral do processo de mineração de dados, de cada uma das etapas que o constitui, e de uma de suas técnicas que é a descoberta de regras de associação. São apresentadas as definições e os conceitos referentes a esta técnica, os desafios presentes na aplicação da mesma, medidas de interesse para avaliação das regras, e finalmente é apresentado o problema da análise de cesto de compras, mais conhecido como *Market Basket Analysis*.

2.2 Mineração de Dados

Muitas pesquisas têm sido direcionadas para o desenvolvimento de técnicas com objetivo de extrair informações a partir de um grande volume de dados e transformar estas informações em conhecimento útil. Esta área é conhecida na literatura como *Knowledge Discovery in Database* (KDD), *data mining* ou mineração de dados. Alguns autores consideram os termos KDD e mineração de dados referentes a processos distintos (Fayyad et al., 1996b). Entretanto, neste trabalho, estes termos serão tratados indistintamente referenciando o processo de extrair conhecimento a partir de dados.

Ao longo dos últimos 10 anos a mineração de dados passou por enormes transformações, influenciado por forças externas, tais como o crescimento do comércio eletrônico, grandes

progressos em biologia molecular e o seu freqüente e controverso uso para segurança nacional. Novas áreas de pesquisa surgiram como a *web mining* e softwares de código aberto foram desenvolvidos como o WEKA, o que levou a uma maior "popularização" da mineração de dados (Piatetsky-Shapiro, 2007).

Fayyad et al. (1996a) definem mineração de dados como sendo: "Processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados". Os autores dividem o processo de mineração de dados em nove etapas. Existem outras abordagens para a divisão deste processo. A divisão adotada por este trabalho é a de Rezende et al. (2003), que dividem o processo em três grandes etapas: Pré-Processamento, Extração de Padrões e Pós-Processamento. Os autores também propõem uma fase anterior ao processo de mineração de dados que se refere ao conhecimento do domínio e identificação do problema e uma fase posterior que se refere à utilização do conhecimento obtido. A Figura 2.1 ilustra estas etapas.

Figura 2.1 – Etapas do processo de mineração de dados (Rezende et al., 2003)

2.2.1 Identificação do Problema

O conhecimento do domínio de aplicação é fundamental para o sucesso da mineração de dados. Este conhecimento geralmente é fornecido por um especialista da área. Nesta fase são definidos os objetivos e metas a serem alcançados e são identificados e selecionados os dados a serem utilizados para a extração do conhecimento.

A compreensão do domínio é importante em todas as etapas do processo de mineração de dados e é um dos fatores críticos para o sucesso da aplicação. Antes da etapa de pré-

processamento, é necessário entender o domínio de aplicação e identificar o objetivo do processo de mineração de dados do ponto de vista do usuário final (Fayyad et al., 1996b).

Além dessa análise inicial para definição das principais metas, objetivos e restrições, o conhecimento sobre o domínio deve ser utilizado em todas as etapas do processo de extração de conhecimento. Na etapa de pré-processamento, esse conhecimento auxilia na escolha do melhor conjunto de dados para se realizar a extração de padrões, saber quais valores são válidos para os atributos, os critérios de preferência entre os possíveis atributos e as restrições de relacionamento ou informações para geração de novos atributos.

Na etapa de extração de padrões, o conhecimento sobre o domínio pode ajudar os analistas na escolha de um critério de preferência entre os modelos gerados, no ajuste dos parâmetros do processo de indução, ou mesmo na geração de um conhecimento inicial a ser fornecido como entrada do algoritmo de mineração para aumentar a eficiência no aprendizado dos conceitos e melhorar a precisão ou a compreensibilidade do modelo final.

Na etapa de pós-processamento, o conhecimento extraído pelos algoritmos de extração de padrões deve ser avaliado. Alguns critérios de avaliação utilizam o conhecimento do especialista para saber, por exemplo, se o conhecimento extraído é interessante ao usuário (Geng e Hamilton, 2006).

2.2.2 Pré-Processamento

Na maioria das vezes não é possível a aplicação de algoritmos de extração de conhecimento diretamente sobre a base de dados. Normalmente os dados disponíveis para análise não estão em formato adequado e há também limitações relacionadas à memória ou tempo de processamento. Além disso, a maioria dos bancos de dados não está livre de ruídos, dados faltantes, incompletos ou inconsistentes. A qualidade dos dados é fundamental para o sucesso do processo de extração de conhecimento. Torna-se necessária então a aplicação de métodos para o tratamento, limpeza e redução do volume de dados.

Segundo Rezende et al. (2003), diversas tarefas podem ser executadas na etapa de pós-processamento, entre elas: extração e integração, transformação, limpeza, seleção e redução dos dados. Estas tarefas são descritas a seguir.

Extração e Integração – Os dados selecionados para a análise podem estar em diferentes formatos e fontes como arquivos-texto, arquivos no formato de planilhas, banco de dados ou *data warehouse*. Há a necessidade de extrair e integrar estes dados em um só formato, adequado para a mineração de dados. Na maioria dos casos este formato é o de exemplo-atributos como mostra a Tabela 2.1. No caso, os exemplos são representados pela variável

Indivíduo e os atributos são representados pelas seguintes características dos indivíduos: *Sexo*, *Idade*, *Cidade* e *Estado*. Cada linha constitui um exemplo e cada coluna constitui um atributo.

Tabela 2.1 – Formato exemplo-atributos

Indivíduo	Sexo	Idade	Cidade	Estado
Pedro	Masculino	13	Itajubá	MG
Izabela	Feminino	21	Taubaté	SP
João	Masculino	53	Resende	RJ
Mariana	Feminino	20	Salvador	BA

Transformação – Após a extração e integração dos dados algumas transformações podem ser necessárias como: resumo, generalização, normalização e construção de atributo (Han e Kamber, 2006). O resumo é aplicado quando, por exemplo, as vendas diárias devem ser agrupadas para serem computadas como o total de vendas no mês ou no ano. A generalização ocorre quando um dado de baixo nível deve ser substituído por um conceito de alto nível através do uso de hierarquias ou taxonomias. Por exemplo, o atributo *rua* pode ser substituído pelo atributo *bairro* ou *cidade*. A normalização deve ser aplicada quando se deseja transformar os valores de um atributo contínuo para um intervalo definido, como entre 0 e 1. Finalmente a construção de atributo é realizada para a inclusão de um novo atributo que irá ajudar no processo de mineração de dados.

Limpeza – O processo de coleta de dados pode gerar problemas como dados faltantes, inconsistentes ou incompletos e a presença de ruídos. Estes problemas podem ser originados por erros de digitação ou erros de leitura dos dados pelos sensores. Para garantir a qualidade dos dados é necessária a aplicação de técnicas de limpeza. Por exemplo, algumas estratégias para lidar com dados faltantes são: substituí-los por uma constante especificada pelo analista; pela média (atributos numéricos) ou moda (atributos categóricos); ou por valores gerados aleatoriamente de acordo com a distribuição dos dados observada.

Redução de Dados – A memória e o tempo de processamento são fatores limitantes de uma aplicação de mineração de dados. Muitas vezes, dependendo da quantidade de dados a ser analisada, a aplicação se torna impraticável. Técnicas de redução podem ser aplicadas a fim de se obter uma representação reduzida do conjunto de dados, mas que ainda assim mantenham a sua integridade. Assim, a mineração de dados deve ser mais eficiente e prover os mesmos ou quase mesmos resultados analíticos que seriam encontrados através do conjunto de dados original. Segundo Weiss e Indurkhya (1998), a redução de dados pode ser realizada de três maneiras: reduzindo o número de exemplos, de atributos ou de valores de um atributo. A redução do número de exemplos pode ser realizada por meio de amostragem significativa. Já na redução do número de atributos deve-se eliminar aqueles menos

significativos. Finalmente, na redução do número de valores de um atributo, geralmente substitui-se valores de um atributo contínuo por intervalos discretos ou realiza-se uma suavização, onde os valores de um atributo são substituídos por um valor numérico representativo como a média, mediana ou mesmo valores extremos.

2.2.3 Extração de Padrões

Esta é a etapa principal do processo de descoberta de conhecimento. Nela é realizada a escolha da tarefa de mineração de dados a ser empregada, a escolha do algoritmo e a extração dos padrões propriamente dita.

A escolha da tarefa é muito importante e deve ser realizada de acordo com os objetivos desejáveis para a solução a ser encontrada, pois sua escolha determina o tipo de conhecimento extraído. As possíveis tarefas de um algoritmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas (Fayyad et al 1996a).

A predição envolve o uso de algumas variáveis ou atributos presentes no banco de dados para prever valores desconhecidos ou futuros de outras variáveis de interesse. Já as atividades descritivas buscam descrever os dados através de padrões de fácil interpretação. Alguns modelos preditivos podem ser descritivos também à medida que são compreensíveis e vice-versa. Porém a distinção é importante para a compreensão do objetivo global da descoberta.

As duas principais tarefas preditivas são classificação e regressão. A classificação consiste na predição de um valor categórico como, por exemplo, predizer se o cliente é bom ou mau pagador. Na regressão, o atributo a ser predito consiste em um valor contínuo como, por exemplo, predizer o lucro ou a perda em um empréstimo (Weiss & Indurkhya 1998).

Algumas das tarefas de descrição são *clustering* e associação. Também conhecida como segmentação, o *clustering* realiza o agrupamento dos dados em subconjuntos, de acordo com características em comum. Estes subconjuntos são definidos pelos dados ao contrário da tarefa de classificação onde as classes são predefinidas. A tarefa de associação busca regras que estabelecem que algumas combinações de valores ocorrem com outras combinações de valores com uma certa freqüência. Uma aplicação comum é na análise de cesto de compras onde busca-se conhecer quais produtos são comprados juntos com outros produtos (Bradley et al., 1998).

Uma grande variedade de algoritmos pode ser utilizada para executar a tarefa escolhida. A escolha do algoritmo é feita de maneira subordinada à linguagem de representação dos padrões a serem encontrados. Segundo Witten e Frank (2005), existem muitas formas de representação do conhecimento que é extraído através da aplicação da mineração de dados, e cada uma delas que irá determinar os algoritmos que podem ser utilizados. Algumas formas

de representação do conhecimento extraído são: árvores de decisão, regras de produção, modelos lineares, modelos não-lineares (redes neurais artificiais), *clusters* e modelos probabilísticos (redes bayesianas). Não existe um algoritmo ideal para cada tipo de tarefa ou representação dos padrões. Não existe um método universal e a escolha de um algoritmo para uma determinada aplicação é algo como uma arte (Fayyad et al., 1996b).

Escolhido o algoritmo a ser utilizado e ajustados seus parâmetros, realiza-se a extração de padrões. Pode ser escolhido mais de um algoritmo gerando diversos modelos que, na etapa de pós-processamento serão tratados para prover o conjunto de padrões ideal ao usuário. Assim como todo o processo de mineração de dados, essa etapa também é um processo iterativo e pode ser necessária sua execução diversas vezes para ajustar o conjunto de parâmetros visando à obtenção de resultados mais adequados aos objetivos preestabelecidos.

2.2.4 Pós-Processamento

Geralmente técnicas de mineração de dados tendem a gerar uma quantidade grande de padrões, sendo que muitos deles não são interessantes. São necessárias então técnicas de pós-processamento a fim de prover apenas o conhecimento interessante e útil ao usuário final. Nesta etapa o conhecimento extraído pode ser simplificado, avaliado, visualizado ou apenas documentado para sua utilização em processos de tomada de decisões. Bruha e Famili (2000) agrupam os métodos de pós-processamento nas seguintes categorias:

Filtragem do Conhecimento – Pode ser realizada por meio de mecanismos de pós-poda para o caso de árvores de decisão, de truncagem no caso de regras de decisão, de restrição de atributos ou ordenação de regras por meio de métricas (medidas de avaliação) para regras de associação.

Interpretação e Explanação – O conhecimento extraído pode ser comparado com o conhecimento prévio do domínio de aplicação.

Avaliação – Pode ser realizada por meio de critérios como precisão, compreensibilidade, complexidade computacional, grau de interesse, entre outros.

Integração do Conhecimento – O conhecimento pode ser integrado a um sistema inteligente de auxílio à tomada de decisões. Estes sistemas podem ser constituídos de uma única técnica ou podem combinar os resultados de vários modelos a fim de se obter maior precisão.

Caso o conhecimento extraído não seja interessante, pode ser preciso alterar os parâmetros do algoritmo, selecionar um outro, selecionar um novo conjunto de dados, entre outras ações, retornando a alguma etapa do processo de mineração.

2.2.5 Utilização do Conhecimento

Esta fase é posterior ao processo de mineração de dados. Após a avaliação e validação do conhecimento, este pode ser utilizado diretamente pelo usuário final para apoio a algum processo de tomada de decisão, incorporado a um sistema inteligente, documentado e reportado a outras partes interessadas ou ainda utilizado para resolver eventuais conflitos entre o conhecimento pré-existente (fornecido pelo especialista) e o conhecimento obtido com o processo de mineração de dados (Fayyad et al. 1996a).

2.3 Regras de Associação

A tarefa de regras de associação foi desenvolvida inicialmente por Agrawal et al. (1993) a partir da observação dos itens presentes em uma compra de supermercado, com o objetivo de descobrir relações do tipo: "Um cliente que compra o item *A* frequentemente compra também o item *B*". Esta tarefa busca caracterizar o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens no mesmo registro (Agrawal e Srikant, 1994).

A mineração de regras de associação tem atraído grande interesse de pesquisadores tanto na área acadêmica como em aplicações práticas. Apesar de aplicações em problemas reais ainda serem poucas, regras de associação têm sido utilizadas em diversas áreas como no comércio eletrônico, navegação *WEB*, na medicina, em serviços bancários, detecção de fraudes em cartões de crédito, gerenciamento de projetos, entre outras áreas (Metwally et al., 2005; Kazienko, 2009; Ribeiro et al., 2008; Karabatak e Ince, 2009; Aggelis, 2004; Sánchez et al., 2009; García et al., 2008).

2.3.1 Definições e Conceitos

Uma regra de associação é representada como uma implicação na forma $LHS \Rightarrow RHS$, em que *LHS* e *RHS* são respectivamente o antecedente (*Left Hand Side*) e o conseqüente (*Right Hand Side*) da regra. As regras de associação são definidas por Agrawal e Srikant (1994):

Seja D uma base de dados composta por um conjunto de itens $A = \{a_1,...,a_m\}$, ordenados lexicograficamente, e por um conjunto de transações $T = \{t_1,...,t_m\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens (chamado *itemset*) tal que $t_i \subseteq A$. É dito que uma transação t_i suporta o *itemset X* se $X \subseteq t_i$. O suporte *P(X)* de um *itemset X* representa a probabilidade da ocorrência do evento *X*.

A regra de associação é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações *T* com confiança *conf* e suporte *sup*, onde *P(LHS RHS)* representa o suporte da regra (a probabilidade da ocorrência da transação $LHS \cup RHS$) e *P(RHS|LHS)* a confiança da regra (a probabilidade condicional de *RHS* dado *LHS*).

Segundo Zhang e Zhang (2002), o suporte e a confiança são as medidas mais utilizadas em regras de associação, onde o suporte representa a freqüência dos padrões e a confiança a força da implicação, isto é, em pelo menos *c%* das vezes que o antecedente ocorrer nas transações, o conseqüente também deve ocorrer.

Em termos algébricos, o suporte pode ser definido da seguinte maneira:

$$\sup(LHS \Rightarrow RHS) = \sup(LHS \cup RHS) = \frac{n(LHS \cup RHS)}{N} \qquad (2.1)$$

- $n(LHS \cup RHS)$ é o número de transações em que *LHS* e *RHS* ocorrem juntos;
- *N* é o número total de transações.

Já a confiança é definida como:

$$conf\,(LHS \Rightarrow RHS) = \frac{\sup(LHS \cup RHS)}{\sup(LHS)} = \frac{n(LHS \cup RHS)}{n(LHS)} \qquad (2.2)$$

- $n(LHS)$ é o número de transações nas quais *LHS* ocorre.

Agrawal et. al (1993) definem o problema de extração de regras de associação em duas etapas:

1) Encontrar todos os *k-itemsets* (conjuntos de *k* itens) que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário (*sup-min*). Os *itemsets* com suporte maior ou igual ao suporte mínimo são definidos como os *itemsets* freqüentes;
2) Utilizar os *k-itemsets* freqüentes, com $k \geq 2$, para gerar as regras de associação. Para cada itemset freqüente $l \subseteq A$, encontrar todos os subconjuntos *ã* de itens de *l*, não vazios e diferente de *l*. Para cada subconjunto $ã \subseteq l$, gerar uma regra na forma $ã \Rightarrow (l - ã)$ se a razão de *sup(l)* por *sup(ã)* for maior ou igual a confiança mínima especificada pelo usuário (*conf-min*). Com o conjunto de *itemsets* freqüentes *{a, b, c, d}* e um subconjunto *{a, b}*, por exemplo, pode-se gerar a regra $a,b \Rightarrow c,d$, desde que $conf(a,b \Rightarrow c,d) \geq conf\text{-}min$, em que $conf(a,b \Rightarrow c,d) = sup(a, b, c, d)/sup(a, b)$.

Segundo Li (2008), o algoritmo mais utilizado para geração dos *intemsets* frequentes é o algoritmo Apriori, que satisfaz o suporte mínimo definido pelo usuário. Este algoritmo foi

desenvolvido por Agrawal e Srikant (1994) e está presente na *survey* de Wu et al. (2008) que apresenta os 10 algoritmos de mineração de dados mais difundidos na comunidade acadêmica.

Desde a introdução de regras de associação por Agrawal et al. (1993), muitos esforços têm sido concentrados no desenvolvimento de algoritmos eficientes para a geração de regras. Diversos novos algoritmos combinando diferentes características (formato do banco de dados, técnica de decomposição, procedimento de busca) foram introduzidos por Zaki (2000). Vários autores realizaram estudos comparativos entre os desempenhos de diferentes algoritmos (Hipp et al., 2000; Zheng et al., 2001; Goethals e Zaki, 2004). Regras de associação têm sido adaptadas para o tratamento de atributos quantitativos e também têm sido aplicadas em novas áreas de pesquisa como mineração espaço-temporal, multimídia e mineração da *web* (*web mining*). Han et al. (2007) fornecem uma visão geral do estado da arte em regras de associação e discutem algumas direções para pesquisas futuras na área.

2.3.2 Desafios da Aplicação de Associação

O tempo de processamento com certeza é um desafio presente na aplicação de regras de associação, mas pode-se dizer que este "desafio" se caracteriza mais como uma limitação presente em qualquer processo de mineração de dados. Além disso, segundo Piatetsky-Shapiro (2007), grande progresso foi alcançado no desenvolvimento de algoritmos mais eficientes para a geração de regras de associação e alguns deles como o SVM (*Support Vector Machine*) foram amplamente aceitos.

Uma problemática associada à descoberta de regras de associação é o número de regras geradas. Segundo Cheng et al. (2008), um problema intrínseco na descoberta de regras de associação é o grande número de regras geradas facilmente, o que torna a análise complexa e muitas vezes restringe o seu uso na prática.

Um outro problema está no fato de o modelo suporte-confiança não ser capaz de mensurar a dependência entre dois *itemsets*. Por exemplo, este modelo não é capaz de identificar implicações negativas do tipo: "Um cliente que compra o item *A* geralmente não compra o item *B*". Talvez em uma análise de cesto de compras este tipo de regra não seja importante, mas em outros casos ela pode representar conhecimento valioso. Um piloto, por exemplo, poderia querer saber se a ausência de determinado equipamento no carro está relacionada à ocorrência de acidentes. Um problema mais sério está ilustrado no exemplo abaixo.

Exemplo: Suponha *n* transações em um supermercado. Considerando somente as compras de café e chá, temos a tabela de contingência abaixo (Tabela 2.2), onde *x* representa a presença

do item e $\bar{x}$ representa a ausência do mesmo e os números representam porcentagens de transações.

Tabela 2.2 – Tabela de contingência (Adaptado de Brin et al., 1998)

	Chá	$\overline{Chá}$	Soma
Café	20	70	90
$\overline{Café}$	5	5	10
Soma	25	75	100

Considerando a regra $Chá \Rightarrow Café$, temos um suporte de 0.2, ou seja, em 20% das compras, os dois itens foram comprados juntos. Este suporte é relativamente alto. A confiança desta regra é de 0.8, isto é, em 80% das compras em que o chá esteve presente, o café também foi comprado. Este valor é um valor alto também. Conclui-se então que a regra $Chá \Rightarrow Café$ é uma regra interessante.

Agora considere o fato de que a probabilidade de uma pessoa comprar café é de 90%. Um cliente que compra chá tem uma probabilidade 10% menor de comprar café do que um cliente sobre o qual não se tem informação. Pode ser que seja interessante saber que muitas pessoas que compram chá, também compram café, mas esta regra por si só, está omitindo informações e na pior das hipóteses é um engano. Existe uma dependência negativa entre a ocorrência de chá e café. Uma forma de calcularmos esta dependência é a seguinte:

$$\frac{\sup(Chá \Rightarrow Café)}{\sup(Chá) * \sup(Café)} = \frac{0.2}{0.25 * 0.9} = 0.89.$$

O fato é que esta medida é menor que 1 e indica que há uma dependência negativa entre chá e café. Por outro lado temos que:

$$\frac{\sup(Chá \Rightarrow \overline{Café})}{\sup(Chá) * \sup(\overline{Café})} = \frac{0.05}{0.25 * 0.1} = 2.00.$$

Sendo esta medida maior que 1, pode-se afirmar que há uma dependência positiva entre a ocorrência de chá e a não ocorrência de café. Esta dependência é mais significativa do que aquela entre a ocorrência dos dois itens. Então um gerente poderia não colocar café em suas prateleiras de chá. A medida de confiança, por não considerar a dependência entre os itens, pode gerar um número muito grande de regras que apresentam relacionamentos falsos e ilusórios.

Diante das limitações apresentadas aqui, muitos trabalhos têm sido realizados com o objetivo de desenvolver metodologias para identificação somente das regras que realmente

sejam interessantes ao usuário final. Os objetivos destes trabalhos envolvem a eliminação de redundâncias, uso de taxonomias, uso de medidas de avaliação das regras, entre outros métodos (Cheng et al., 2008**;** Melanda, 2004; Carvalho, 2007; Tan et al., 2004; Geng e Hamilton, 2006).

2.3.3 Medidas de Interesse Objetivas

As medidas de interesse ou medidas de avaliação do conhecimento gerado pela mineração de regras de associação têm o objetivo de mensurar o quanto uma regra é interessante ao usuário. Segundo Tan et al. (2004), a tarefa central das regras de associação é a descoberta de pares de itens que co-ocorrem a uma determinada freqüência em um banco de dados transacional, enquanto as medidas de interesse buscam identificar grupos de variáveis as quais estão altamente correlacionadas entre si ou com relação a uma variável-alvo específica.

Segundo Geng e Hamilton (2006) as medidas de avaliação podem ser utilizadas de três maneiras diferentes dentro do processo de mineração de dados, como ilustra a Figura 2.2. Primeiramente, as medidas de interesse podem ser utilizadas durante a mineração como mecanismo de poda dos padrões não interessantes. Podem ser utilizadas também para ordenar os padrões encontrados de acordo com os valores de interesse. E finalmente, estas medidas podem ser utilizadas durante o pós-processamento para selecionar apenas os padrões mais interessantes. Os autores realizam uma *survey* das medidas de avaliação, comparando suas propriedades e fornecendo estratégias para a seleção das medidas adequadas de acordo com a aplicação.

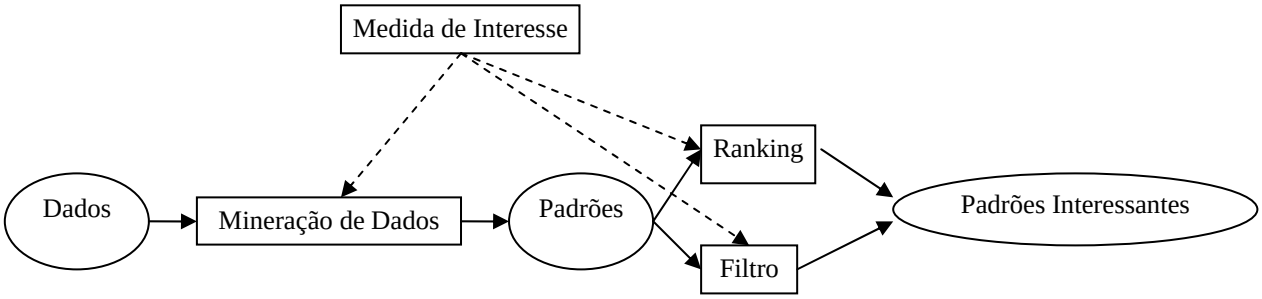

Figura 2.2 – Objetivos das medidas de interesse no processo de mineração de dados (Geng e Hamilton, 2006)

As medidas de avaliação de regras de associação podem ser classificadas como objetivas e subjetivas. Medidas objetivas são aquelas que dependem apenas da estrutura do conjunto de dados e dos padrões. Não é necessário o conhecimento do domínio ou da aplicação. A maioria destas medidas é baseada em teorias da probabilidade e estatística. Já as medidas subjetivas consideram tanto a estrutura de uma regra e os dados utilizados no processo de descoberta,

como também o conhecimento que o usuário possui e seu interesse no momento da análise dos padrões.

Muitas pesquisas têm sido realizadas a fim de estudar medidas que não dependam do usuário ou do domínio da aplicação. A Tabela 2.3 mostra 38 medidas objetivas de interesse estudadas por Tan et al. (2004), Lenca et al. (2004), Ohsaki et al. (2004) e Lavrac et al. (1999) e que estão presentes na *survey* de Geng e Hamilton (2006).

Tabela 2.3 – Medidas de interesse objetivas

Medida	Fórmula
Support	$P(AB)$
Confidence/ Precision	$P(B/A)$
Coverage	$P(A)$
Prevalence	$P(B)$
Recall	$P(A/B)$
Specificity	$P(\overline{B}/\overline{A})$
Accuracy	$P(AB)+P(\overline{A}\overline{B})$
Lift/Interest	$P(B/A)/P(B)$ ou $P(AB)/(P(A)*P(B))$
Leverage	$P(B/A)-P(A)*P(B)$
Added Value/Change of Support	$P(B/A)-P(B)$
Relative Risk	$P(B/A)/P(B/\overline{A})$
Jaccard	$P(AB)/(P(A)+P(B)-P(AB))$
Certainty Factor	$(P(B/A)-P(B))/(1-P(B))$
Odds Ratio	$(P(AB)*P(\overline{A}\overline{B}))/(P(A\overline{B})*P(\overline{B}A))$
Yule's Q	$\frac{P(AB)*P(\overline{A}\overline{B})-P(A\overline{B})*P(\overline{A}B)}{P(AB)*P(\overline{A}\overline{B})+P(A\overline{B})*P(\overline{A}B)}$
Yule's Y	$\frac{\sqrt{P(AB)*P(\overline{A}\overline{B})}-\sqrt{P(A\overline{B})*P(\overline{A}B)}}{\sqrt{P(AB)*P(\overline{A}\overline{B})}+\sqrt{P(A\overline{B})*P(\overline{A}B)}}$
Klosgen	$\sqrt{P(AB)}*(P(B/A)-P(B))$, $\sqrt{P(AB)}*\max(P(B/A)-P(B),P(A/B)-P(A))$
Conviction	$(P(A)*P(\overline{B}))/P(A\overline{B})$
Interestingness Weighting Dependency	$\left(\left(\frac{P(AB)}{P(A)*P(B)}\right)^{k}-1\right)*P(AB)^{m}$, onde *k* e *m* são coeficientes de dependência e generalidade respectivamente. São pesos que mensuram a importância relativa dos dois fatores.
Collective Strength	$\frac{P(AB)+P(\overline{B}/\overline{A})}{P(A)*P(B)+P(\overline{A})*P(\overline{B})}*\frac{1-P(A)*P(B)-P(\overline{A})*P(\overline{B})}{1-P(AB)-P(\overline{B}/\overline{A})}$

Laplace Correction	$(N(AB)+1)/(N(A)+2)$
Gini Index	$P(A)*\left\{P(B/A)^2+P(\overline{B}/A)^2\right\}+P(\overline{A})*\left\{P(B/\overline{A})^2+P(\overline{B}/\overline{A})^2\right\}-P(B)^2-P(\overline{B})^2$
Goodman and Kruskal	$\frac{\sum_i \max_j P(A_iB_j)+\sum_j(\max_j P(A_iB_j)-\max_i P(A_i)-\max_i P(B_j))}{2-\max_i P(A_i)-\max_i P(B_j)}$
Normalized Mutual Information	$\sum_i\sum_j P(A_iB_j)*\log_2\frac{P(A_iB_j)}{P(A_i)*P(B_j)}/(-\sum_i P(A_i)*\log_2 P(A_i))$
J-Measure	$P(AB)*\log\frac{P(B/A)}{P(B)}+P(A\overline{B})*\log\frac{P(\overline{B}/A)}{P(\overline{B})}$
One-Way Support	$P(B/A)*\log_2(P(AB)/(P(A)*P(B)))$
Two-Way Support	$P(AB)*\log_2(P(AB)/(P(A)*P(B)))$
Two-Way Support Variation	$P(AB)*\log_2\frac{P(AB)}{P(A)*P(B)}+P(A\overline{B})*\log_2\frac{P(A\overline{B})}{P(A)*P(\overline{B})}+$ $P(\overline{A}B)*\log_2\frac{P(\overline{A}B)}{P(\overline{A})*P(B)}+P(\overline{A}\overline{B})*\log_2\frac{P(\overline{A}\overline{B})}{P(\overline{A})*P(\overline{B})}$
ϕ*−Linear Correlation Coefficient*	$\frac{P(AB)-P(A)*P(B)}{\sqrt{P(A)*P(B)*P(\overline{A})*P(\overline{B})}}$
Piatetsky-Shapiro	$P(AB)-P(A)*P(B)$
Cosine	$P(AB)/\sqrt{P(A)*P(B)}$
Loevinger	$1-(P(A)*P(\overline{B}))/P(A\overline{B})$
Information gain	$\log(P(AB)/(P(A)*P(B)))$
Sebag-Schoenauer	$P(AB)/P(A\overline{B})$
Least Contradiction	$(P(AB)-P(A\overline{B}))/P(B)$
Odd Multiplier	$(P(AB)*P(\overline{B}))/(P(B)*P(A\overline{B}))$
Example and Counterexample Rate	$1-P(A\overline{B})/P(AB)$
Zhang	$\frac{P(AB)-P(A)*P(B)}{\max(P(AB)*P(\overline{B}),P(B)*P(A\overline{B}))}$

Piatetsky-Shapiro (1991) propôs 3 propriedades-chave (*P1, P2* e *P3*) que uma boa medida de avaliação *M* de um padrão de associação do tipo $A \rightarrow B$ deve satisfazer:

P1: $M = 0$ se *A* e *B* são estatisticamente independentes;

P2: *M* cresce monotonicamente com *P(AB)* quando *P(A)* e P*(B)* permanecem constantes;

P3: *M* decresce monotonicamente com *P(A)* (ou *P(B)*) quando os parâmetros restantes (*P(AB)* e *P(A)* ou *P(B)*) permanecem constantes.

Onde: *P(A)*, *P(B)* e *P(AB)* são as probabilidades de ocorrência de *A*, *B* e *A* e *B* respectivamente.

Tan et al. (2004) propõem mais 5 propriedades (*O1, O2, O3, O4* e *O5*) além das 3 já apresentadas aqui. Estas propriedades são descritas através de uma notação de matrizes. Nesta notação, cada tabela de contingência 2x2 é representada por uma matriz $M = [f_{11} f_{10}; f_{01} f_{00}]$, enquanto cada medida objetiva é um operador de matriz, *O*, que transforma a matriz *M* em um valor escalar *k*, isto é, $O(M) = k$. Então as propriedades de uma medida podem ser analisadas realizando diversas operações com as tabelas de contingência, como é mostrado na Figura 2.3. As propriedades são:

O1: Simetria sob permutação de variável. Uma medida *O* é simétrica sob a permutação de variável (Figura 2.3 (a)), $A \leftrightarrow B$, se $O(M^T) = O(M)$ para todas as matrizes de contingência *M*. Se não, trata-se de uma medida assimétrica. Na prática, medidas assimétricas são utilizadas quando não existe a necessidade de distinguir a regra $A \rightarrow B$ da regra $B \rightarrow A$, isto é, o sentido da implicação não faz diferença.

O2: Invariância escalar de linha e coluna. Considerando a matriz 2x2 $R = C = [k_1 0; 0 k_2]$, onde k_1 e k_2 são constantes positivas. Na operação *R*M*, a primeira linha da matriz *M* é multiplicada por k_1 e a segunda por k_2, enquanto na operação *M*C*, a primeira coluna da matriz *M* é multiplicada por k_1 e a segunda por k_2 (Figura 2.3(b)). A medida *O* satisfaz a propriedade de invariância escalar de linha e coluna se $O(R * M) = O(M)$ e $O(M * C) = O(M)$ para todas as matrizes de contingência *M.*

O3: Anti-simetria sob permutação de linha e coluna. Considerando a matriz 2x2 $S = [01;10]$. A medida normalizada *O* é anti-simétrica sob permutação de linha se $O(S * M) = -O(M)$, e anti-simétrica sob permutação de coluna se $O(M * S) = -O(M)$ para todas as matrizes de contingência *M* (Figura 2.3(c)). Na prática, medidas que são simétricas sob permutação de linha e coluna não fazem distinção entre correlação positiva e negativa.

O4: Invariância sob inversão. Considerando a matriz 2x2 $S = [01;10]$. A medida *O* é invariante sob inversão (Figura 2.3(d)) se $O(S * M * S) = O(M)$ para todas as matrizes de contingência *M.*

O5: Invariância nula. A medida *O* satisfaz esta propriedade se $O(M + C) = O(M)$, onde $C = [00;0K]$ e *k* é uma constante positiva.

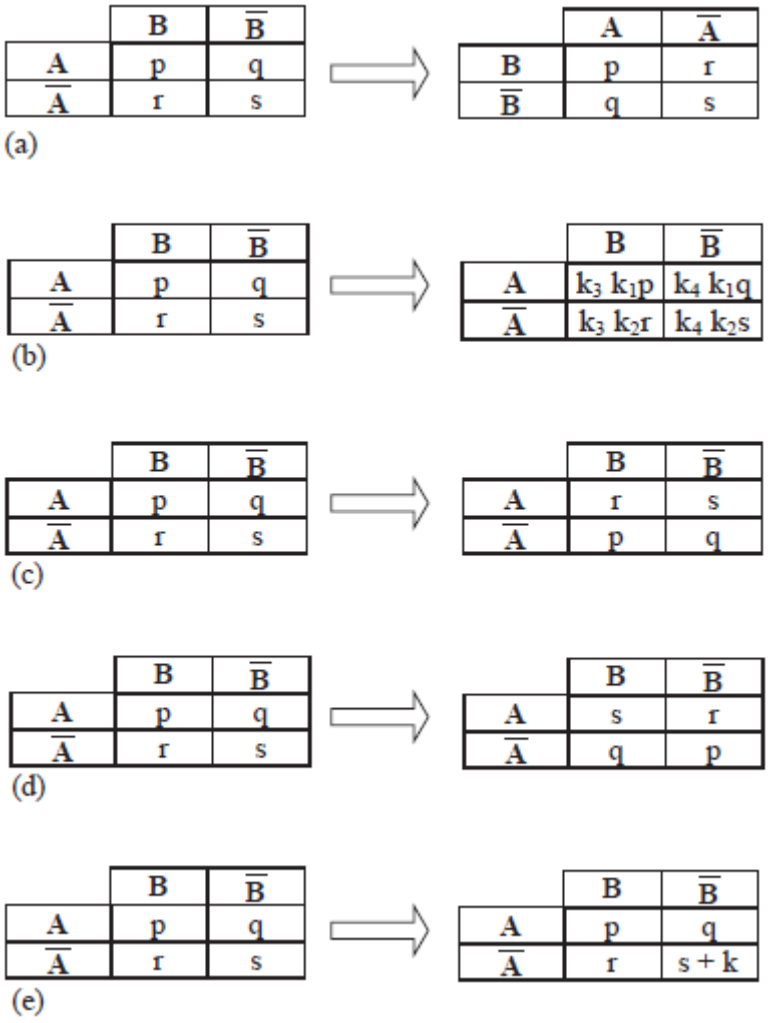

Figura 2.3 – Operações com a tabela de contingência. (a) Simetria sob permutação de variável; (b) Invariância escalar de linha e coluna; (c) Anti-simetria sob permutação de linha e coluna; (d) Invariância sob inversão; (e) Invariância nula (Tan et al., 2004).

Os autores descrevem 21 medidas de avaliação de regras de associação, realizam um estudo comparativo entre estas medidas, agrupam-nas de acordo com suas propriedades e mostram que não existe uma medida ideal para todos os tipos de aplicações. A aplicabilidade de cada medida varia de acordo com suas propriedades.

Lenca et al. (2004) introduzem mais 5 propriedades (*Q1, Q2, Q3, Q4* e *Q5*) para avaliação das medidas de interesse:

Q1: A medida de avaliação *M* é constante se não existe um contra-exemplo para a regra. Esta propriedade afirma que as regras com confiança igual a 1 devem ter o mesmo valor de interesse, independentemente do suporte, o que contradiz a sugestão de Tan et al. (2004), que diz que uma medida deve variar de acordo com o suporte;

Q2: A medida *M* decresce com $P(A\overline{B})$ de forma linear, côncava, ou convexa em torno de 0+. Esta propriedade descreve a maneira a qual o valor de interesse decresce quando alguns contra-exemplos são adicionados. Se o usuário pode tolerar alguns contra-exemplos, um decréscimo côncavo é desejável. Se a aplicação requer rigorosamente uma confiança igual a 1, um decréscimo convexo é desejável;

Q3: A medida *M* aumenta quando o número total de registros aumenta. Esta propriedade descreve as mudanças que ocorrem no valor de interesse quando o número de registros aumenta e *P(A), P(B)* e *P(AB)* permanecem constantes;

Q4: É fácil de se estabelecer um limite. Quando um valor limite para a medida de interesse é fixado para separar as regras interessantes das não interessantes, a escolha deste limite deve ser uma tarefa simples;

Q5: A semântica da medida deve ser de fácil expressão.

As propriedades 1, 4 e 5, são desejáveis para as medidas de avaliação de regras. Já as propriedades 2 e 3 podem ou não ser desejadas pelos usuários de acordo com a aplicação. Os autores ainda propõem um método de ranking para ordenar as medidas de interesse de acordo com vários critérios, onde marcas e pesos são atribuídos a cada propriedade que o usuário julgue interessante a sua aplicação.

Vaillant et al. (2004) agrupam as medidas de acordo com a similaridade existente entre suas propriedades e também de acordo com a intensidade das mesmas, isto é, os valores destas para diferentes conjuntos de regras.

Geng e Hamilton (2006) propõem mais 2 propriedades (*S1* e *S2*) desejáveis de uma medida de avaliação de regras de associação:

S1: Uma medida de interesse *M* deve ser uma função crescente do suporte se as margens da tabela de contingência se mantêm constantes;

S2: Uma medida de interesse *M* deve ser uma função crescente da confiança se as margens da tabela de contingência se mantêm constantes.

Para a propriedade *S1*, assume-se que as margens na tabela de contingência se mantêm constantes, isto é, $n(A) = a$, $n(\overline{A}) = N - a$, $n(B) = b$ e $n(\overline{B}) = N - b$. Sendo o suporte representado por *x*, então $P(AB) = x$, $P(\overline{A}B) = \frac{b}{N} - x$, $P(A\overline{B}) = \frac{a}{N} - x$ e $P(\overline{A}\overline{B}) = 1 - \frac{a+b}{N} + x$. Substituindo essas fórmulas nas medidas, são obtidas as medidas em função do suporte *x*. Por exemplo, o *Lift* é dado por $\frac{P(AB)}{P(A) * P(B)} = \frac{x}{\frac{a}{n} * \frac{b}{n}}$. Observa-se

claramente que o *Lift* é uma função crescente do suporte. Da mesma maneira, é possível analisar as outras medidas em função do suporte. A propriedade *S2* pode ser relacionadda com a propriedade *Q2* de Lenca et al. (2004), ainda que de forma inversa, pois se uma medida diminui com $P(A\overline{B})$, a mesma aumenta com *P(AB)*. No entanto a propriedade *Q2* descreve a relação existente entre a medida e $P(A\overline{B})$, sem restringir os outros parâmetros (P(AB), $P(\overline{A}B)$ e $P(\overline{A}\overline{B})$. Esta falta de restrições torna a análise difícil. Com a propriedade *S2*, restrições são aplicadas às margens das tabelas de contingência, o que facilita a análise. Geng e Hamilton (2006) consideram 38 medidas diferentes em sua *survey*. A Tabela 2.5 mostra as medidas estudadas pelos autores e as propriedades que cada medida satisfaz.

Tabela 2.4 – Medidas de interesse e propriedades (Geng e Hamilton, 2006)

Medida	P1	P2	P3	O1	O2	O3	O4	O5	Q1	Q2	Q3	S1
Support	*N*	*Y*	*N*	*Y*	*N*	*N*	*N*	*N*	*N*	1	N	0
Confidence	*N*	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	*Y*	1	N	0
Coverage	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	3	N	1
Prevalence	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	1	N	1
Recall	*N*	*Y*	*N*	*N*	*N*	*N*	*N*	*Y*	*N*	2	N	0
Specificity	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	3	N	0
Accuracy	*N*	*Y*	*Y*	*Y*	*N*	*N*	*Y*	*N*	*N*	1	N	1
Lift/Interest	*N*	*Y*	*Y*	*Y*	*N*	*N*	*N*	*N*	*N*	2	N	0
Leverage	*N*	*Y*	*Y*	*N*	*N*	*N*	*N*	*Y*	*N*	1	N	0
Added Value	*Y*	*Y*	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	1	N	0
Relative Risk	*N*	*Y*	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	1	N	0
Jaccard	*N*	*Y*	*Y*	*Y*	*N*	*N*	*N*	*Y*	*N*	1	N	0
Certainty Factor	*Y*	*Y*	*Y*	*N*	*N*	*N*	*Y*	*N*	*N*	0	N	0
Odds ratio	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*Y*	0	N	4
Yule's Q	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*Y*	0	N	4
Yule's Y	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*Y*	0	N	4
Klosgen	*Y*	*Y*	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	0	N	0
Conviction	*N*	*Y*	*N*	*N*	*N*	*N*	*Y*	*N*	*Y*	0	N	0
Interestingness Weighting Dependency	*N*	*Y*	*N*	*N*	*N*	*N*	*N*	*Y*	*N*	6	N	0
Collective Strength	*N*	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*N*	*N*	0	N	0
Laplace Correction	*N*	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	1	N	0
Gini Index	*Y*	*N*	*N*	*N*	*N*	*N*	*Y*	*N*	*N*	0	N	4
Goodman and Kruskal	*Y*	*N*	*N*	*Y*	*N*	*N*	*Y*	*N*	*N*	5	N	3
Normalized Mutual Information	*Y*	*Y*	*Y*	*N*	*N*	*N*	*Y*	*N*	*N*	5	N	3
J-Measure	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*Y*	0	N	4
One-Way Support	*Y*	*Y*	*Y*	*N*	*N*	*N*	*N*	*Y*	*N*	0	N	0
Two-Way Support	*Y*	*Y*	*Y*	*Y*	*N*	*N*	*N*	*Y*	*N*	0	N	0
Two-Way Support Variation	*Y*	*N*	*N*	*Y*	*N*	*N*	*Y*	*N*	*N*	0	N	4
ϕ −Linear Correlation Coefficient	*Y*	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*N*	*N*	0	N	0
Piatetsky-Shapiro	*Y*	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*N*	*N*	1	N	0
Cosine	*N*	*Y*	*Y*	*Y*	*N*	*N*	*N*	*Y*	*N*	2	N	0
Loevinger	*Y*	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	*Y*	4	N	2
Information gain	*Y*	*Y*	*Y*	*Y*	*N*	*N*	*N*	*Y*	*N*	2	N	0
Sebag-Schoenauer	*N*	*Y*	*Y*	*N*	*N*	*N*	*N*	*Y*	*Y*	0	N	0
Least Contradiction	*N*	*Y*	*Y*	*N*	*N*	*N*	*N*	*Y*	*N*	2	N	0
Odd Multiplier	*N*	*Y*	*Y*	*N*	*N*	*N*	*N*	*N*	*Y*	0	N	0
Example and Counterexample Rate	*N*	*Y*	*Y*	*N*	*N*	*N*	*N*	*Y*	*Y*	2	N	0
Zhang	*Y*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	0	N	4

Y - Sim, a medida satisfaz a propriedade; N - Não, a medida não satisfaz a propriedade; 0 - Aumenta com o suporte; 1 - Não varia com o suporte; 2 - Diminui com o suporte; 3 - Não aplicável; 4 - Depende dos parâmetros.

Os trabalhos realizados sobre medidas de avaliação do conhecimento descoberto através da mineração de regras de associação mostram que esta área de pesquisa ainda está bastante aberta a novos trabalhos, publicações, etc. Existem muitas medidas diferentes, cada medida satisfaz determinadas propriedades e a seleção de uma medida depende das propriedades que são interessantes para a aplicação em questão. Ohsaki et al. (2007) avaliam 40 medidas de interesse para avaliação de conhecimento na área médica comparando cada medida com o conhecimento de um médico especialista e chegam a um conjunto de cinco medidas que segundo os autores, são as mais eficientes para aquela aplicação. O estudo dos autores, assim como este, é realizado sobre um banco de dados específico e seus resultados não podem ser generalizados.

2.3.4 Análise de Cesto de Compras

Uma das principais aplicações das regras de associação é na análise de cesto de compras, mais conhecida na literatura por *market basket analysis* (MBA). Neste tipo de problema busca-se por padrões no comportamento de compra dos consumidores, adquirindo o conhecimento de quais produtos costumam ser levados juntos em uma mesma compra.

Pesquisas em marketing têm sido direcionadas para a análise da co-incidência de múltiplas categorias de produtos em diferentes compras, a fim de planejar as atividades de marketing de forma que o lucro máximo seja obtido. Um varejista tipicamente tem que tomar decisões sobre quais produtos colocar em promoção, como e quando. Segundo Li (2008) os dados transacionais contêm valiosas informações sobre a associação de produtos. Estas informações contribuem para a uma coordenação eficiente das atividades de marketing, pois mudanças nos preços ou promoções de determinadas categorias de produtos podem afetar não apenas as vendas destas, mas de outras também. Por exemplo, uma promoção de cervejas pode aumentar a venda de amendoins.

Diversas técnicas têm sido desenvolvidas para modelar a venda cruzada de produtos (*cross-selling*). Durante a década de 1990 modelos multivariados foram desenvolvidos com o objetivo de estudar o comportamento de escolha de múltiplas categorias de produtos por parte dos consumidores (*multi-category choice models*). Estes modelos buscam a maximização do lucro através da coordenação das atividades de marketing em diversas categorias. Comparados com a mineração de dados, provêem medidas mais precisas dos efeitos da venda cruzada, embora apresentem muitas vezes um alto custo computacional, desde que envolvem tipicamente modelos *Markov chain Monte Carlo* (MCMC) **(**Manchanda et al., 1999; Russell e Petersen, 2000**).** Alguns trabalhos com regras de associação também foram desenvolvidos com o objetivo de levar em consideração o lucro, onde o valor de interesse da medida de

avaliação é dependente também do lucro que pode ser obtido com determinada associação. Wang et al. (2002) sugerem um modelo integrado para minerar regras de associação e recomendar as melhores de acordo com o lucro que o usuário pode obter.

Para mensurar a freqüência em que pares de produtos são levados juntos em uma loja com milhares de produtos a disposição dos consumidores, tem-se que checar a freqüência de milhões de pares de produtos. O número de produtos ou categorias em bases de dados reais faz com que a identificação de todas as associações por modelos puramente estatísticos se torne impraticável. É necessária a seleção a priori de apenas alguns produtos, selecionados por especialistas. Além disso, os modelos estatísticos assumem usualmente relações lineares e geralmente não consideram interações. Aqui, a técnica de regras de associação mostra-se como uma ótima alternativa se tratando de grandes quantidades de dados. Esta técnica não considera o conhecimento a priori de produtos que podem estar associados. Não é objetivo deste trabalho estudar diferentes técnicas de modelagem da escolha de múltiplas categorias.

Manchanda et al. (1999) afirmam que duas categorias podem ser compradas juntas por uma variedade de razões. São elas:

Complementaridade – Esta razão está relacionada aos efeitos que as atividades de marketing podem ter sobre categorias correlacionadas. Em categorias complementares, espera-se que ações de marketing sobre uma categoria tenham influência não apenas nesta, mas em outra também. Duas categorias podem ser positivamente relacionadas, isto é, complementos positivos, quando a promoção de uma das categorias aumenta as vendas da outra também. Podem ser negativamente relacionadas ou substitutas, quando a promoção de uma das categorias diminui a venda da outra. Ou ainda podem ser independentes, quando atividades de marketing sobre uma não exercem influência nenhuma sobre a outra.

Heterogeneidade – Duas categorias de produtos podem ser compradas juntas por questão de preferência dos consumidores. Por exemplo, fatores demográficos podem influenciar a compra de duas categorias diferentes de produtos. Em Resende-RJ, uma pessoa pode escolher levar arroz e feijão preto, enquanto em Itajubá-MG, um outro consumidor pode preferir levar o feijão marrom.

Co-incidência – É definida como o conjunto de todas as razões que não sejam relacionadas à complementaridade e a heterogeneidade dos consumidores. Algumas delas são: hábito dos consumidores, ambiente físico da loja, conhecimento do cliente sobre um determinado supermercado, humor do consumidor, entre outras.

Percebe-se que a complementaridade está sob o controle dos gerentes. Já a heterogeneidade e a co-incidência são mais difíceis de serem controladas.

Segundo Groth (2000), o MBA pode ser aplicado em: análise de vendas cruzadas; definição de *layout*; projeto de catálogos de produtos; análise de perda de liderança; definição de preço e promoções de produtos; dentre outros. Estas aplicações são baseadas na crença de que as vendas entre categorias de produtos diferentes são correlacionadas.

De acordo com o relatório “Leading Practices in Market Basket Analysis: How Top Retailers are Using Market Basket Analysis to Win Margin and Market Share” publicado em 2008 pela renomada empresa de pesquisa FactPoint Group, os grandes varejistas estão utilizando o MBA para:

- Desenvolver campanhas de publicidade e promoções mais lucrativas. Varejistas estão utilizando o MBA para elaborar promoções mais previsíveis, entendendo melhor como os consumidores respondem a diferentes ofertas e veículos de comunicação. Por exemplo, através do MBA varejistas podem evitar descontos desnecessários entendendo quando e onde uma redução no preço será realmente eficiente e aumentará as vendas significativamente;
- Aumentar a precisão das promoções. O MBA é utilizado para otimizar campanhas e promoções elevando a margem de vendas com maior precisão. Atingir o objetivo das promoções com maior precisão resulta em maior retenção de clientes e permite que os varejistas ofereçam o mix de produto certo, para o cliente certo e na hora certa. Por exemplo, um varejista poderia querer saber se enviar 5 milhões de e-mails com ofertas variadas para clientes diferentes seria mais eficiente que enviar 10 milhões de e-mails iguais.
- Melhorar as promoções através de análise longitudinal sobre os dados dos cartões de fidelidade. O MBA está sendo aplicado sobre os dados de cartões de fidelidade para entender quem, o que, como, quanto, onde e quando o cliente está comprando. Dados pessoais como sexo, idade, estado civil, bairro, entre outros, estão sendo utilizados para segmentação dos clientes e personalização dos serviços;
- Aumentar o tráfego dentro das lojas. Com este objetivo, primeiramente o MBA é utilizado para analisar o que realmente atrai os consumidores para o interior das lojas. Então é analisado posteriormente o que mantêm os consumidores dentro das lojas por mais tempo. Mantê-los por mais tempo dentro das lojas resulta em compras maiores.

- Aumentar o tamanho e o valor das compras. Com os dados de cartões de fidelidade, varejistas podem saber quantas vezes o cliente freqüentou a loja e o que foi comprado e aproveitar este conhecimento para aumentar o tamanho e o valor das compras deste cliente. Por exemplo, promoções podem ser direcionadas para quem compra todos os produtos de mercearia, com exceção da ração, ou para aqueles que compravam papel e não o fazem mais;

- Testar e aprender utilizando o ponto de venda como um laboratório. Alguns varejistas estão selecionando um conjunto de lojas que é chamado de grupo de controle e aplicando as análises em outro conjunto chamado grupo de teste. Por exemplo, o MBA pode ajudar na determinação de como aumentar o valor médio das compras sem a necessidade de gastar mais e sacrificar o lucro;

- Determinar a faixa de preço ideal para cada loja. O MBA está sendo utilizado para se determinar "zonas" de preço e fronteiras respondendo a questões como, por exemplo: Quanto se pode aumentar o preço da mercadoria *X* sem que os clientes comecem a optar pela mercadoria *Y*?

- Adequar o inventário a necessidade de cada loja. Os padrões de comportamento de compra de consumidores podem variar de local para local. A oferta ideal de produtos varia de acordo com fatores demográficos, climáticos, econômicos etc. O MBA pode ser utilizado para definir quais produtos devem fazer parte do inventário de cada loja.

- Otimizar o *layout* das lojas. Varejistas utilizam o MBA para otimizar o layout de cada loja maximizando a probabilidade de venda cruzada dos produtos e respondendo a perguntas como: Produtos expostos na ponta dos corredores realmente são mais vendidos e se sim, o aumento da venda destes produtos está resultando no aumento da venda de produtos complementares?

2.4 Considerações Finais

Neste capítulo foi apresentada uma visão geral do processo de mineração de dados. Foram apresentados conceitos e definições de regras de associação, os desafios envolvidos na aplicação da técnica, medidas de avaliação utilizadas para avaliar o conhecimento extraído, bem como uma das principais aplicações da associação que é a análise de cesto de compras.

O próximo capítulo trata do objeto de estudo, explica como a pesquisa foi conduzida e mostra os resultados atingidos através da mineração de regras de associação aplicada a

modelagem dos dados transacionais de um supermercado do sul do estado de Minas Gerais e traz uma discussão sobre a aplicabilidade desta técnica.

3. Modelagem dos Dados do Supermercado

3.1 Considerações Iniciais

O supermercado onde a pesquisa foi realizada está localizado no sul do estado de Minas Gerais. A escolha da empresa foi motivada pelo fato de esta já possuir um banco de dados estruturado, apesar de ainda estar “engatinhando” no que diz respeito à disponibilidade de informações para a tomada de decisões. Não existe no supermercado nenhum processo de tomada de decisão com base na extração de conhecimento a partir de grande quantidade de dados. Além disso, trata-se de uma grande loja, trabalhando com aproximadamente 15000 itens, o que faz com que seu banco de dados possivelmente seja uma fonte interessante de conhecimento.

Uma das prioridades de varejistas hoje é o foco no cliente. Diante da necessidade de conhecer melhor o comportamento de seus clientes, o pesquisador e os donos do supermercado chegaram à conclusão de que a mineração de dados seria uma ferramenta útil. Os proprietários da empresa já possuem uma maneira de consolidar e agregar os seus dados para entender os fundamentos do negócio: o que eles estão vendendo, quantas unidades estão em movimento e a quantidade total de vendas. Contudo, assim como a maioria dos varejistas, não se aventuraram longe o suficiente para analisar as informações em seu mais baixo nível de granularidade: a análise de cesto de compras.

A análise de cesto de compras é o processo de análise dos dados transacionais com o objetivo de gerar valor ao negócio. O conhecimento extraído fornece aos usuários uma visibilidade direta da composição da compra de cada um dos clientes tornando possível a compreensão não só da quantidade de itens que é levada, mas também de como estes itens são comprados em conjunto com os outros.

3.2 O Processo de Modelagem

A Figura 3.1 mostra cada uma das 4 etapas da modelagem dos dados através da mineração de regras de associação bem como as entradas e saídas destas etapas.

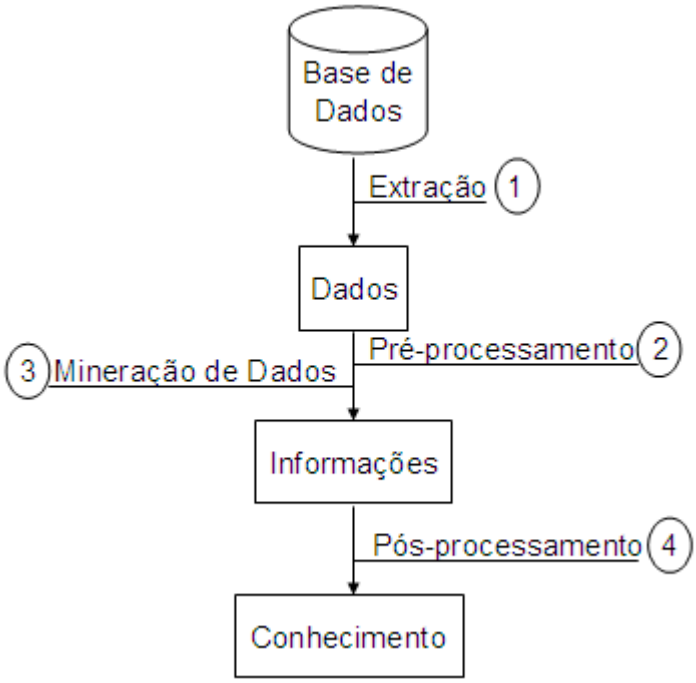

Figura 3.1 – Processo de modelagem dos dados

Base de Dados

O supermercado estudado possui um banco de dados relacional onde todas as transações são armazenadas. Através da linguagem SQL (*Structured Query Language*) é possível realizar consultas avançadas e obter as informações necessárias à análise como, por exemplo, o ano, o mês, o dia do mês, o dia da semana e o horário em que a compra foi realizada bem como quais foram os produtos levados, qual a quantidade comprada, a forma de pagamento, etc. Alguns supermercados possuem cartões de fidelidade. Com estes cartões é possível também identificar o cliente que está comprando, sua idade, sexo, estado civil, renda, endereço, entre outros. Como o supermercado estudado está atualmente iniciando sua política de cartões de fidelidade, não foi possível trabalhar com dados pessoais de clientes.

Extração

A seleção e extração dos dados foram realizadas através da linguagem SQL. Como o mercado realizou mudanças recentemente em sua base de dados, foram selecionados os dados a partir destas mudanças. Foi selecionado um período de 4 meses (Janeiro, Fevereiro, Março e Abril) do ano de 2009. A Tabela 3.1 mostra as informações extraídas de cada compra realizada neste período.

Tabela 3.1 – Informações extraídas da base de dados do supermercado

Informação	Descrição
Mês	Janeiro, Fevereiro, Março ou Abril
Dia do Mês	Um, Dois, Três, ... ou 31
Dia da Semana	Domingo, Segunda, Terça, ... ou Sábado
Período do Dia	Manhã, Tarde ou Noite
Tipo de Dia	Normal, Véspera de Feriado, Feriado ou Pós-feriado
Valor da Compra	Valor total da compra em R$
Produtos Levados	Ex: Maçã, Nescau
Categorias	Ex: Frutas, Achocolatados
Promoção	Produto em promoção – Sim ou Não

Seria também considerada uma variável que representasse o retorno de cada produto para o supermercadista, mas o supermercado estudado não continha esta informação de forma correta e atualizada em sua base de dados.

Através da linguagem SQL foi gerado um arquivo texto (.txt) com todas as informações extraídas, onde cada linha deste arquivo representa uma compra realizada durante o período selecionado e cada informação extraída é separada por ponto e vírgula da informação seguinte. A Figura 3.2 mostra uma parte deste arquivo.

```
JANEIRO;DOIS;SEXTA;MANHÃ;POS;34,53;RAID PROTEC.AP.45N.GTS.15NOITES;INSETICIDAS;NÃO;MASC.DOVE
JANEIRO;DOIS;SEXTA;MANHÃ;POS;23,46;SALS.SADIA 500GR;CONGELADOS;NÃO;FILE FGO P.PAF 500DESFIAD
JANEIRO;DOIS;SEXTA;MANHÃ;POS;80,03;PEPSI TWIST PET 2LT;REFRIG.PET ACIMA 600ML;NÃO;CERV.SKOL
JANEIRO;DOIS;SEXTA;MANHÃ;POS;132,97;ARROZ ECCO 5KG T1;ARROZ;NÃO;AC.CRISTAL CAETE 5KG;ACUCAR
JANEIRO;DOIS;SEXTA;MANHÃ;POS;14,15;AROMAT.RODAB.GEL 60CAR.NOVO;PROD.VEICULOS;NÃO;OLEO LIZA 9
JANEIRO;DOIS;SEXTA;MANHÃ;POS;202,11;LARANJA;FRUTAS;NÃO;TOMATE;VERDURAS;NÃO;CEBOLA BAND.SSJOS
JANEIRO;DOIS;SEXTA;MANHÃ;POS;5,73;MACA GALA;FRUTAS;NÃO;TOMATE;VERDURAS;NÃO;ALFACE;VERDURAS;N
JANEIRO;DOIS;SEXTA;MANHÃ;POS;44,77;SUCO MAGUAR.500MARACUJA PET;SUCO LIQUIDO;NÃO;SAPOL.RADIUM
JANEIRO;DOIS;SEXTA;MANHÃ;POS;162,52;AZEITE ANDOR.500LATA;AZEITE;NÃO;AZEITE ANDOR.500LATA;AZE
JANEIRO;DOIS;SEXTA;MANHÃ;POS;103,14;COND.SEDA 350MELAN.UV/PTO LUM;PROD.CABELO CONDICIONADOR;
JANEIRO;DOIS;SEXTA;MANHÃ;POS;213,65;BATATA;VERDURAS;NÃO;ABOB.ITALIA;VERDURAS;NÃO;CENOURA;VER
JANEIRO;DOIS;SEXTA;MANHÃ;POS;16,77;LEITE SAC.MATOSA;LEITES LIQUIDOS;NÃO;BATATA;VERDURAS;NÃO;
JANEIRO;DOIS;SEXTA;MANHÃ;POS;1,89;LA DE ACO BOMBRIL 08UN.;ESPONJAS AÇO/SINTETICA;NÃO;
JANEIRO;DOIS;SEXTA;MANHÃ;POS;51,44;MANGA KEIT;FRUTAS;NÃO;PIMENT.VERD;VERDURAS;NÃO;MARACUJA;F
JANEIRO;DOIS;SEXTA;MANHÃ;POS;350,33;BATATA;VERDURAS;NÃO;MILHO JUREMA 200VAPOR;ERVILHA/MILHO;
JANEIRO;DOIS;SEXTA;MANHÃ;POS;94,41;FILE PEITO AC;FRANGOS ACOUGUE;NÃO;PERNIL S/OSSO CONG.;CAR
JANEIRO;DOIS;SEXTA;MANHÃ;POS;110,21;BANANA PRATA;FRUTAS;NÃO;LARANJA;FRUTAS;NÃO;FILE PEITO FC
JANEIRO;DOIS;SEXTA;MANHÃ;POS;4,59;COCA COLA 2,5L PET;REFRIG.PET ACIMA 600ML;NÃO;BIS.VISC.WF
JANEIRO;DOIS;SEXTA;MANHÃ;POS;64,45;BROCOLIS;VERDURAS;NÃO;CENOURA;VERDURAS;NÃO;INHAME;VERDURA
JANEIRO;DOIS;SEXTA;MANHÃ;POS;46,28;MAMAO FORM.;FRUTAS;NÃO;ALHO GRANEL;VERDURAS;NÃO;BANANA PR
JANEIRO;DOIS;SEXTA;MANHÃ;POS;51,46;BIS.BAUD.WF.MORAN.165GR;BISCOITOS;NÃO;BIS.BAUD.WF.MORAN.1
JANEIRO;DOIS;SEXTA;MANHÃ;POS;64,13;COCA COLA 2L PET;REFRIG.PET ACIMA 600ML;NÃO;COCA COLA 1,5
JANEIRO;DOIS;SEXTA;MANHÃ;POS;9,74;CERV.SKOL 473LATA;CERVEJAS LATA;NÃO;CHICORIA;VERDURAS;NÃO;
JANEIRO;DOIS;SEXTA;MANHÃ;POS;20,5;CONTRA FILE;CARNE BOVINA 1;NÃO;PATINHO;CARNE BOVINA 1;NÃO;
JANEIRO;DOIS;SEXTA;MANHÃ;POS;5,9;IOG.NINHO 180MACA/BANANA;IOGURTES;NÃO;IOG.NINHO 180MACA/BAN
JANEIRO;DOIS;SEXTA;MANHÃ;POS;214,23;LEITE PIRACANJUBA INT.1L;LEITES LIQUIDOS;NÃO;REQ.VENTANI
```

Figura 3.2 – Arquivo texto gerado a partir da consulta SQL

Todas as etapas seguintes do processo de modelagem dos dados do supermercado através da mineração de regras de associação foram realizadas através da manipulação deste arquivo texto. A intervenção do pesquisador no banco de dados do supermercado estudado foi única. Esta forma de abordagem fez com que a pesquisa permanecesse sob o controle do pesquisador.

Pré-processamento

O pesquisador optou por uma manipulação dos dados com o objetivo de extrair algumas informações básicas que caracterizem as compras que são realizadas no supermercado estudado e também de verificar a qualidade dos dados que foram selecionados e extraídos. Para isso foi criada uma tabela a partir do arquivo texto da Figura 3.2, onde cada linha representa uma compra e cada coluna as variáveis desta compra (mês, dia, valor, produtos, categorias, etc). Uma parte desta tabela é mostrada na Tabela 3.2.

Tabela 3.2 – Tabela criada a partir do arquivo texto

Mês	Dia	D.Semana	Período	TipoDia	Valor	Prod.1	Cat1	Promo
JANEIRO	DOIS	SEXTA	MANHÃ	POS	23,84	COCA COLA 2L PET	REFRIG.PET ACIMA 600ML	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	16,94	SALS.HOT DOG GRANEL	PADARIA	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	169,2	OLEO SOJA COCAMAR 900ML PET	OLEOS	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	118,53	REF.MANTIQ.2LT LARANJA	REFRIG.PET ACIMA 600ML	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	67,21	SUCRILHOS KELLOG.730TRAD.	CEREAL MATINAL	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	95,24	CAFE FLORESTA 500EX.FORTE	CAFES	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	156,56	CEBOLA	VERDURAS	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	163,98	SUCO ADES 1L.GTS150ML PESSEGO	SUCO LIQUIDO	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	2,35	OVOS SET.CEU VERMELHO	PADARIA	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	10,47	COND.FRUCTIS 300QUIMI-RESIST	PROD.CABELO CONDICIONADOR	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	3,82	PAO DE SAL (FRANCES)	PADARIA PILAR	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	1,79	BALA PARATI ABACAXI 150G	BALAS	NÃO
JANEIRO	DOIS	SEXTA	MANHÃ	POS	76,76	TORR.BAUD.160LIGHT INTEGRAL	TORRADAS	NÃO
JANEIRO	DOIS	SEXTA	TARDE	POS	210,06	PATINHO	CARNE BOVINA 1	NÃO

Alguns filtros foram aplicados sobre esta tabela. A primeira informação buscada foi o volume de compras realizadas neste período. A Tabela 3.3 mostra a quantidade de compras em cada mês e o total.

Tabela 3.3 – Quantidade de compras realizadas

Mês	Volume
Janeiro	65.795
Fevereiro	60.554
Março	65.456
Abril	64.291
Total	256.096

Buscou-se também a informação de como as compras se comportam semanalmente e foi obtido o gráfico da Figura 3.3, onde o eixo *x* representa os dias dos meses considerados e o eixo *y* representa o volume de compras.

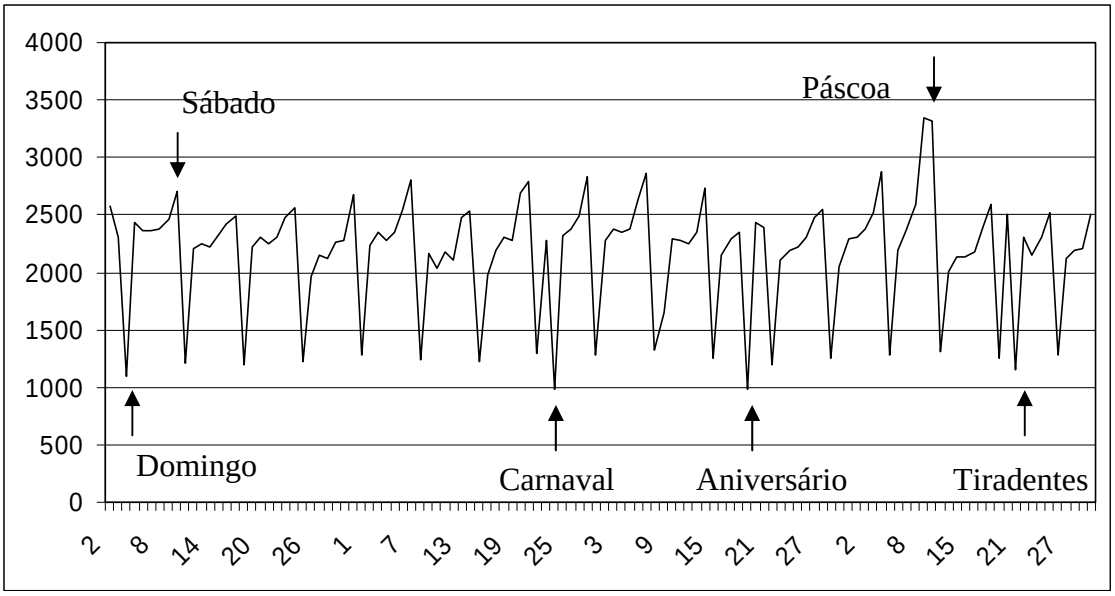

Figura 3.3 – Comportamento semanal do volume de compras

O gráfico da Figura 3.3 mostra claramente que há um padrão no volume de compras que se repete semanalmente. O dia em que o volume de vendas é mais baixo é domingo. Faz sentido, pois no domingo o supermercado funciona somente pela manhã. E o dia em que o volume de vendas é mais alto é sábado. As anomalias presentes no gráfico representam feriados. A primeira delas é no dia 25 de fevereiro, quarta-feira de cinzas. A segunda é no dia 19 de março, aniversário da cidade. A terceira é no dia 21 de abril, feriado de Tiradentes. Além disso, é possível observar um pico de vendas nos dias 9 e 11 de abril que se referem à quinta-feira e sábado da Semana Santa.

Observou-se também a variação do volume de compras de acordo com o período do dia: manhã, tarde e noite. Foi verificado que existe um maior volume de compras no período da manhã. Em segundo lugar vem o período da tarde. Finalmente o período da noite com um menor volume.

Através da observação do valor total das compras realizadas nestes 4 meses considerados, foi possível perceber que aproximadamente 96% destas (245.042) possuem valor menor ou igual a R$ 200,00 e são responsáveis por aproximadamente 62% do faturamento total neste período. Trata-se de um supermercado de “sacolinha”, isto é, a maioria das compras realizadas ali são compras pequenas, diárias ou semanais. Aproximadamente 90% das compras possuem valor menor ou igual a R$ 100,00 e são responsáveis por aproximadamente 43% do faturamento.

Em relação a categorias e produtos, neste período foram comprados:

- 2.398.050 produtos;
- 13.114 produtos diferentes;
- Produtos de 184 categorias diferentes.

Buscou-se também entender como as promoções funcionam no supermercado, isto é, qual o tempo médio em que um produto permanece em promoção, se existem promoções “relâmpago”, quantos produtos geralmente estão em promoção ao mesmo tempo, se é possível existir mais de um produto por categoria em promoção, etc. Ao ser realizada a verificação de quais produtos estiveram alguma vez em promoção e por quanto tempo, foi encontrado um problema. As promoções não estavam sendo registradas da maneira correta na base de dados do supermercado. Basicamente só havia o registro do dia que o produto entrava em promoção e muitas promoções nem registro de lançamento possuíam. Este problema tornou impossível a análise do efeito das promoções, isto é, analisar a correlação entre as promoções e o aumento da venda de produtos em promoção ou produtos complementares, pois mesmo realizando uma nova busca no banco de dados, não há registros corretos de quando uma promoção foi lançada e de quando ela foi finalizada.

Devido a limitações de capacidade de processamento e armazenamento, presentes em qualquer processo de mineração de dados foi necessário “rodar” o algoritmo várias vezes variando seus parâmetros de entrada e o conjunto de dados utilizado. O pré-processamento continuou de forma diferente para cada uma das análises realizadas.

Mineração de Dados

Na etapa de mineração dos dados foi utilizado o algoritmo Apriori para a geração de regras de associação. O algoritmo é utilizado para encontrar todos os *k-itemsets* freqüentes contidos em uma base de dados. Esse algoritmo gera um conjunto de *k-itemsets* candidatos e então percorre a base de dados para determinar se os mesmos são freqüentes, identificando desse modo todos os *k-itemsets* freqüentes. Utiliza-se a notação L_k para representar o conjunto de *k-itemsets* freqüentes e C_k para representar o conjunto de *k-itemsets* candidatos.

Seja D uma base de dados composta por um conjunto de itens $A = \{a_1,...,a_m\}$, ordenados lexicograficamente, e por um conjunto de transações $T = \{t_1,...,t_m\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens (chamado *itemset*) tal que $t_i \subseteq A$, o algoritmo Apriori é dado por:

1) L_1 := {*1-itemsets* freqüente};
2) for (k := 2; $L_{k-1} \neq \phi$; k + +) do
3) C_k := *apriori-gen* (L_{k-1}); //Gera novos conjuntos candidatos
4) for all (transações $t \in T$) do
5) C_k := *subset* (C_k, *t*); //Conjuntos candidatos contidos em *t*
6) for all candidatos $c \in C_t$ do
7) *c.count* + +;
8) end for
9) end for
10) L_k := $\{ c \in C_k \mid c.count \geq sup\text{-}min \}$;
11) end for
12) Resposta := $\bigcup_k L_k$

Inicialmente o algoritmo conta a ocorrência de itens, determinando os *1-itemsets* freqüentes que são armazenados em L_1. O passo seguinte (passo *k*) é dividido em duas etapas. Na primeira (linha 3) o conjunto de *itemsets* freqüentes L_{k-1} obtido no passo ($k - 1$) é utilizado para gerar o conjunto de *k-itemsets* candidatos C_k usando a função *apriori-gen*. A

seguir (linhas 4 a 9), a base de dados é percorrida para determinar o valor do suporte dos *k-itemsets* candidatos em C_k. Finalmente, são identificados os *k-itemsets* freqüentes em cada passo (linha 10). A solução final é dada pela união dos conjuntos L_k de *k-itemsets* freqüentes.

A principal característica deste algoritmo é sua propriedade de linha de fronteira (*downward closure*). Através da função *apriori-gen*, o algoritmo percorre toda a base de dados a procura dos *1-itemsets* freqüentes, isto é, aqueles *itemsets* com apenas 1 item e que satisfazem o suporte mínimo. O próximo passo é a descoberta dos *2-itemsets* que satisfazem o suporte mínimo. Agora, ao invés do algoritmo percorrer toda a base de dados, ele percorre apenas os *1-itemsets* freqüentes descobertos na etapa anterior, pois o suporte é sempre o mesmo e então os *2-itemsets* só poderão surgir dos anteriores. Este procedimento é baseado na fato de que se um *x-itemset* tem o suporte mínimo, então todos os subconjuntos dele também o terão. Da mesma forma são gerados os *3-itemsets* e assim sucessivamente. Esta propriedade faz com que não seja necessário percorrer o conjunto de dados inteiro e otimiza a tarefa de geração dos *itemsets* freqüentes.

Na aplicação do algoritmo Apriori, podem ser definidos como parâmetros de entrada, o suporte e a confiaça mínima ou mesmo um intervalo de suporte e um de confiança. Desta forma, as regras descobertas possuirão como valores de suporte e confiança um valor maior ou igual ao suporte e confiança mínimos especificados pelo usuário. Para cada análise diferente estes parâmetros de entrada receberam valores diferentes para que um resultado satisfatório fosse alcançado.

Informações

Os algoritmos de mineração de regras de associação geram como resultado regras do tipo:

- $cerveja \rightarrow amendoim$; suporte = 20%; confiança = 40%; Em 20% de todas as compras realizadas, cerveja e amendoim foram comprados juntos. Considerando agora apenas as compras nas quais ocorreu cerveja, em 40% destas também ocorreu amendoim.

O número de regras geradas depende da quantidade de compras consideradas, quantidade de atributos considerados e do suporte e confiança mínimos especificados. Na grande maioria das vezes este número torna inviável a observação de todas as regras geradas para a obtenção de um conhecimento que possa auxiliar algum processo de tomada de decisão.

Este problema pode ser observado claramente através de Melanda (2004), que aplicou algoritmos de regras de associação a algumas bases de dados e obteve os resultados mostrados na Tabela 3.4.

Tabela 3.4 – Problemática de regras de associação (Melanda, 2004)

Nº de atributos	Nº de exemplos	Nº de regras	Suporte (%)	Confiança (%)
10	339	8827	2	50
19	120	24713	2	50
10	18539	2147	6	25
1000	1000	18150	6	25
100	25000	596626	6	25
995	60000	1459070	6	25

Devido ao grande número de informações geradas há a necessidade de um pós-processamento destas informações com o objetivo de transformá-las em conhecimento.

Pós-processamento

Para o pós-processamento das informações a fim de transformá-las em conhecimento, utilizou-se a medida de avaliação *lift*. A medida de interesse *lift*, também conhecida como *interest*, é uma das mais utilizadas para avaliar dependências (Brin et al, 1998). Dada uma regra de associação $A \Rightarrow B$, esta medida indica o quanto mais freqüente torna-se *B* quando *A* ocorre. O valor do *lift* é computado pela equação 3.1:

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{\sup(B)} \tag{3.1}$$

Se $lift(A \Rightarrow B) = 1$, então *A* e *B* são independentes. Se $lift(A \Rightarrow B) > 1$, então *A* e *B* são positivamente dependentes. Se $lift(A \Rightarrow B) < 1$, *A* e *B* são negativamente dependentes. Esta medida varia entre 0 e ∞ e possui interpretação bastante simples: quanto maior o valor do *lift*, mais interessante a regra, pois *A* aumentou ("*lifted*") *B* numa maior taxa. Se $lift(A \Rightarrow B) = 5$ por exemplo, significa que *B* tem 5 vezes mais chances de ocorrer quando *A* ocorre.

Utilizou-se como *lift = 2* como parâmetro de corte das regras não interessantes, isto é, só foram consideradas as regras em que *B* tem no mínimo duas vezes a mais de chances de acontecer quando *A* acontece. Isto já reduz o número de regras interessantes, mas ainda houve a necessidade de se observar o suporte e a confiança da regra para escolher um conjunto de regras interessantes.

Conhecimento

Como já foi falado, o conjunto de regras final de cada análise foi gerado através de uma poda, especificando um *lift* mínimo igual a 2 e também através da observação dos valores de suporte e confiança das regras. O conhecimento gerado foi constituído de regras que mostram algum comportamento até então desconhecido por especialistas e também por regras que representam um comportamento óbvio ou já conhecido, mas que ainda sim seja interessante mensurá-lo. Todos sabem, por exemplo, que o fato de um consumidor comprar uma cerveja provavelmente aumenta a probabilidade de este comprar amendoim, mas poucos sabem o quanto exatamente aumenta.

3.3 Resultados

Primeiramente, para saber quais categorias ocorrem juntas nas compras com determinada freqüência, foram consideradas na análise apenas as variáveis que representam quais categorias que foram levadas em cada compra, sem repetição, isto é, considerou-se se aquela categoria ocorreu ou não na compra, independente do número de vezes.

Buscou-se identificar quais as categorias que aparecem em mais compras diferentes. Para isso foram consideradas todas as 256.096 compras e foi necessário adotar um suporte de 8% (20.486 compras). O tempo de mineração foi de aproximadamente 5 minutos. A Tabela 3.5 mostra as 24 categorias que mais aparecem em compras diferentes. Foi adotado um valor alto de suporte para gerar apenas as categorias mais vendidas e o mais rápido possível, sem que fosse preciso gerar nenhuma regra. A confiança não tem importância neste caso, já que gerar regras ainda não era o objetivo.

Tabela 3.5 – Categorias mais freqüentes

Categoria	Freqüência	Suporte (%)
PADARIAX	90723,00	35,42539
VERDURAS	66748,00	26,06366
FRUTAS	55933,00	21,84064
LEITESLIQUIDOS	49617,00	19,37438
REFRIG.PETACIMA600ML	49252,00	19,23185
BISCOITOS	47581,00	18,57936
PADARIA	44427,00	17,34779
QUEIJOS	44408,00	17,34037
CONGELADOS	39667,00	15,48911
IOGURTES	37046,00	14,46567
PAPELHIGIENICO	30479,00	11,90140
FRIOS/FATIADOS	27858,00	10,87795
CAFES	25612,00	10,00094
SABONETES	25000,00	9,76196
CERVEJASLATA	24448,00	9,54642
MARGARINAS/MANTEIGA	23896,00	9,33088
CARNEBOVINA1	23629,00	9,22662
SABAOEMPO	23574,00	9,20514
SUCOLIQUIDO	23026,00	8,99116
OLEOS	22837,00	8,91736
MASSAS	22392,00	8,74360
DETERGENTES	21311,00	8,32149
FRANGOSACOUGUE	21204,00	8,27971
CREMESDENTAIS	20668,00	8,07041

A freqüência na tabela indica em quantas compras a categoria apareceu e o suporte indica a percentagem que esta freqüência representa considerando que existe um total de 256.096 compras. Percebe-se que a categoria mais comprada é PADARIAX que representa os produtos feitos pelo supermercado como o pão francês que individualmente é o produto mais vendido. A categoria PADARIA representa produtos de padaria que não são produzidos pelo supermercado como, por exemplo, o pão de forma.

O próximo passo foi identificar quais são os conjuntos de 5 categorias que mais ocorrem e identificar as regras mais interessantes. Devido a limitações de armazenamento e processamento, foram consideradas somente as compras menores que R$ 200,00 (245.042 compras). Para esta análise foi adotado um suporte de 0,8% (1.959 compras), uma confiança de 10% e o tempo de mineração de dados foi de aproximadamente 40 minutos. Foram consideradas combinações de até 5 categorias. A Tabela 3.6 mostra os conjuntos de 5 categorias que mais aparecem em compras diferentes.

Tabela 3.6 – Conjuntos de 5 categorias mais freqüentes (< R$200,00)

Conjunto	Freqüência	Suporte (%)
(FRUTAS, VERDURAS, PADARIA, QUEIJOS, IOGURTES)	2556,00	1,04309
(FRUTAS, VERDURAS, PADARIA, BISCOITOS, QUEIJOS)	2535,00	1,03452
(FRUTAS, VERDURAS, PADARIA, QUEIJOS, LEITESLIQUIDOS)	2508,00	1,02350
(FRUTAS, VERDURAS, PADARIA, CONGELADOS, QUEIJOS)	2438,00	0,99493
(FRUTAS, VERDURAS, PADARIA, QUEIJOS, QUEIJOS/FRIOS)	2368,00	0,96636
(FRUTAS, VERDURAS, PADARIA, BISCOITOS, IOGURTES)	2354,00	0,96065
(FRUTAS, VERDURAS, PADARIAX, PADARIA, QUEIJOS)	2319,00	0,94637
(PAPELHIGIENICO, ACUCARCRISTAL, OLEOS, ARROZ, CAFES)	2236,00	0,91250
(FRUTAS, VERDURAS, PADARIA, BISCOITOS, LEITESLIQUIDOS)	2235,00	0,91209
(FRUTAS, VERDURAS, BISCOITOS, QUEIJOS, IOGURTES)	2202,00	0,89862
(PAPELHIGIENICO, MASSAS, OLEOS, ARROZ, CAFES)	2196,00	0,89617
(SABONETES, PAPELHIGIENICO, OLEOS, ARROZ, CAFES)	2192,00	0,89454
(FRUTAS, VERDURAS, PADARIA, IOGURTES, LEITESLIQUIDOS)	2106,00	0,85944
(PAPELHIGIENICO, FEIJAO, OLEOS, ARROZ, CAFES)	2087,00	0,85169
(PAPELHIGIENICO, OLEOS, ARROZ, CAFES, SABAOEMPO)	2076,00	0,84720
(DETERGENTES, PAPELHIGIENICO, OLEOS, ARROZ, CAFES)	2071,00	0,84516
(FRUTAS, VERDURAS, PADARIA, QUEIJOS, REFRIG.PETACIMA600ML)	2060,00	0,84067
(MASSAS, ACUCARCRISTAL, OLEOS, ARROZ, CAFES)	2056,00	0,83904
(FRUTAS, VERDURAS, BISCOITOS, IOGURTES, LEITESLIQUIDOS)	2054,00	0,83822
(SABONETES, PAPELHIGIENICO, MASSAS, OLEOS, ARROZ)	2031,00	0,82884
(PAPELHIGIENICO, MASSAS, ACUCARCRISTAL, OLEOS, ARROZ)	2029,00	0,82802
(FRUTAS, VERDURAS, PADARIAX, PADARIA, BISCOITOS)	2022,00	0,82516
(FRUTAS, VERDURAS, QUEIJOS, IOGURTES, LEITESLIQUIDOS)	2019,00	0,82394
(SABONETES, CREMESDENTAIS, PAPELHIGIENICO, OLEOS, ARROZ)	2014,00	0,82190
(FRUTAS, VERDURAS, CONGELADOS, QUEIJOS, IOGURTES)	2007,00	0,81904
(FRUTAS, VERDURAS, PADARIAX, PADARIA, LEITESLIQUIDOS)	2006,00	0,81864
(FRUTAS, VERDURAS, BISCOITOS, QUEIJOS, LEITESLIQUIDOS)	2005,00	0,81823
(SABONETES, PAPELHIGIENICO, OLEOS, ARROZ, SABAOEMPO)	1981,00	0,80843
(SABONETES, CREMESDENTAIS, PAPELHIGIENICO, OLEOS, CAFES)	1963,00	0,80109
(SABONETES, PAPELHIGIENICO, ACUCARCRISTAL, OLEOS, ARROZ)	1963,00	0,80109

Percebe-se que a maioria das categorias combinadas em grupos de 5 são as categorias mais compradas, exceto açúcar cristal, arroz e papel higiênico. Algumas das 24 categorias mais compradas não fazem parte de nenhum dos 30 conjuntos de 5 categorias mais comprados. São elas: CERVEJASLATA, MARGARINAS/MANTEIGA, SUCOLIQUIDO, FRANGOSACOUGUE e CARNEBOVINA1. Provavelmente estas categorias são compradas mais vezes em compras pontuais. São categorias de produtos que são consumidos mais rapidamente que um papel higiênico, um arroz ou um creme dental. Muitas vezes também, consumidores entram nos supermercados com o objetivo de comprar somente carne ou frango para um churrasco ou algumas cervejas em lata para assistir um jogo de futebol. Isto faz com que a proporção de todas as compras em que estas categorias são compradas aumente em compras pontuais e diminua em compras maiores. Não é objetivo deste trabalho deduzir o porque as compras acontecem desta maneira, pois só quem tem condições de avaliar essas informações são profissionais com conhecimento de domínio. A avaliação do conhecimento

gerado, isto é, se é um conhecimento útil ou não a tomada de decisões deve ser realizada por alguém que tenha experiência no tipo de negócio em que a técnica está sendo utilizada.

Com esta análise foram geradas 14.858 regras. Foram buscadas então as regras mais interessantes. Após a poda através do *lift*, permaneceram 13.802 regras o que ainda é um número alto o suficiente para tornar a observação de regra por regra impraticável.

Devido ao grande número de regras geradas foi definido um objetivo mais específico para a escolha das regras. Foram selecionadas algumas categorias para se analisar como estas categorias interagem com outras categorias e entre si. São elas: PADARIAX, VERDURAS, FRUTAS, LEITESLIQUIDOS, REFRIG.PETACIMA600ML, BISCOITOS, PADARIA, QUEIJOS, CONGELADOS e IOGURTES.

A Tabela 3.7 mostra algumas regras encontradas relacionando as categorias mais compradas e outras categorias.

Tabela 3.7 – Regras entre as categorias mais compradas e outras categorias

Índice	Categoria 1	==>	Categoria 2	Suporte (%)	Confiança (%)	Lift
1	ACHOCOLATADOS	==>	BISCOITOS	1,82948	46,34071	2,913887
2	MACARRAOINSTANTANEO	==>	BISCOITOS	1,62217	45,78438	2,878906
3	MOLHOSDETOMATE	==>	CONGELADOS	1,46995	34,62463	2,539354
4	MASSAS	==>	CONGELADOS	1,66747	28,05741	2,057717
5	SUCOLIQUIDO	==>	FRUTAS	3,07866	40,70138	2,035708
6	SUCOLIQUIDO	==>	IOGURTES	2,52038	33,32074	2,712799
7	MACARRAOINSTANTANEO	==>	IOGURTES	1,14715	32,37733	2,635991
8	ACHOCOLATADOS	==>	LEITESLIQUIDOS	1,74746	44,26297	2,493514
9	MARGARINAS/MANTEIGA	==>	LEITESLIQUIDOS	2,64934	38,57856	2,173288
10	CAFES	==>	LEITESLIQUIDOS	2,58119	35,84382	2,019229
11	MARGARINAS/MANTEIGA	==>	PADARIA	2,46407	35,88068	2,357494
12	SUCOLIQUIDO	==>	PADARIA	2,55099	33,72538	2,215883
13	MASSAS	==>	QUEIJOS	2,05516	34,58079	2,246665
14	MOLHOSDETOMATE	==>	QUEIJOS	1,70297	40,11343	2,606113
15	CERVEJASLATA	==>	REFRIG.PETACIMA600ML	3,12722	35,12077	2,010574
16	MAIONESE	==>	REFRIG.PETACIMA600ML	0,85618	38,08314	2,180163
17	CARNEBOVINA1	==>	VERDURAS	4,01809	50,79447	2,117231
18	CALDOSTEMPERO	==>	VERDURAS	1,24305	57,37427	2,391492
19	ARROZ	==>	VERDURAS	2,54895	50,89220	2,121305

A lista abaixo mostra por extenso como é a interpretação destas regras. O índice da lista corresponde exatamente ao índice das regras na tabela.

1. Em 1,83% de todas as compras, foram compradas em conjunto as categorias ACHOCOLATADOS e BISCOITOS. Considerando agora todas as compras em que a categoria ACHOCOLATADOS foi levada, em 46,34% destas a categoria BISCOITOS também foi levada. Uma pessoa que compra algum produto da categoria ACHOCOLATADOS tem 2,91 chances a mais de levar algum produto

da categoria BISCOITOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

2. Em 1,62% de todas as compras, foram compradas em conjunto as categorias MACARRAOINSTANTANEO e BISCOITOS. Considerando agora todas as compras em que a categoria MACARRAOINSTANTANEO foi levada, em 45,78% destas a categoria BISCOITOS também foi levada. Uma pessoa que compra algum produto da categoria MACARRAOINSTANTANEO tem 2,88 chances a mais de levar algum produto da categoria BISCOITOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

3. Em 1,47% de todas as compras, foram compradas em conjunto as categorias MOLHOSDETOMATE e CONGELADOS. Considerando agora todas as compras em que a categoria MOLHOSDETOMATE foi levada, em 34,62% destas a categoria CONGELADOS também foi levada. Uma pessoa que compra algum produto da categoria MOLHOSDETOMATE tem 2,54 chances a mais de levar algum produto da categoria CONGELADOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

4. Em 1,67% de todas as compras, foram compradas em conjunto as categorias MASSAS e CONGELADOS. Considerando agora todas as compras em que a categoria MASSAS foi levada, em 28,06% destas a categoria CONGELADOS também foi levada. Uma pessoa que compra algum produto da categoria MASSAS tem 2,06 chances a mais de levar algum produto da categoria CONGELADOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

5. Em 3,09% de todas as compras, foram compradas em conjunto as categorias SUCOLIQUIDO e FRUTAS. Considerando agora todas as compras em que a categoria SUCOLIQUIDO foi levada, em 40,70% destas a categoria FRUTAS também foi levada. Uma pessoa que compra algum produto da categoria SUCOLIQUIDO tem 2,03 chances a mais de levar algum produto da categoria FRUTAS do que uma pessoa sobre a qual não se tem informação e vice-versa.

6. Em 2,52% de todas as compras, foram compradas em conjunto as categorias SUCOLIQUIDO e IOGURTES. Considerando agora todas as compras em que a

categoria SUCOLIQUIDO foi levada, em 33,32% destas a categoria IOGURTES também foi levada. Uma pessoa que compra algum produto da categoria SUCOLIQUIDO tem 2,71 chances a mais de levar algum produto da categoria IOGURTES do que uma pessoa sobre a qual não se tem informação e vice-versa.

7. Em 1,15% de todas as compras, foram compradas em conjunto as categorias MACARRAOINSTANTANEO e IOGURTES. Considerando agora todas as compras em que a categoria MACARRAOINSTANTANEO foi levada, em 32,38% destas a categoria IOGURTES também foi levada. Uma pessoa que compra algum produto da categoria MACARRAOINSTANTANEO tem 2,63 chances a mais de levar algum produto da categoria IOGURTES do que uma pessoa sobre a qual não se tem informação e vice-versa.

8. Em 1,75% de todas as compras, foram compradas em conjunto as categorias ACHOCOLATADOS e LEITESLIQUIDOS. Considerando agora todas as compras em que a categoria ACHOCOLATADOS foi levada, em 44,26% destas a categoria LEITESLIQUIDOS também foi levada. Uma pessoa que compra algum produto da categoria ACHOCOLATADOS tem 2,49 chances a mais de levar algum produto da categoria LEITESLIQUIDOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

9. Em 2,65% de todas as compras, foram compradas em conjunto as categorias MARGARINAS/MANTEIGA e LEITESLIQUIDOS. Considerando agora todas as compras em que a categoria MARGARINAS/MANTEIGA foi levada, em 38,58% destas a categoria LEITESLIQUIDOS também foi levada. Uma pessoa que compra algum produto da categoria MARGARINAS/MANTEIGA tem 2,17 chances a mais de levar algum produto da categoria LEITESLIQUIDOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

10. Em 2,58% de todas as compras, foram compradas em conjunto as categorias CAFES e LEITESLIQUIDOS. Considerando agora todas as compras em que a categoria CAFES foi levada, em 35,84% destas a categoria LEITESLIQUIDOS também foi levada. Uma pessoa que compra algum produto da categoria CAFES tem 2,02 chances a mais de levar algum produto da categoria LEITESLIQUIDOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

11. Em 2,46% de todas as compras, foram compradas em conjunto as categorias MARGARINAS/MANTEIGA e PADARIA. Considerando agora todas as compras em que a categoria MARGARINAS/MANTEIGA foi levada, em 35,88% destas a categoria PADARIA também foi levada. Uma pessoa que compra algum produto da categoria MARGARINAS/MANTEIGA tem 2,36 chances a mais de levar algum produto da categoria PADARIA do que uma pessoa sobre a qual não se tem informação e vice-versa.

12. Em 2,55% de todas as compras, foram compradas em conjunto as categorias SUCOLIQUIDO e PADARIA. Considerando agora todas as compras em que a categoria SUCOLIQUIDO foi levada, em 33,52% destas a categoria PADARIA também foi levada. Uma pessoa que compra algum produto da categoria SUCOLIQUIDO tem 2,21 chances a mais de levar algum produto da categoria PADARIA do que uma pessoa sobre a qual não se tem informação e vice-versa.

13. Em 2,05% de todas as compras, foram compradas em conjunto as categorias MASSAS e QUEIJOS. Considerando agora todas as compras em que a categoria MASSAS foi levada, em 34,58% destas a categoria QUEIJOS também foi levada. Uma pessoa que compra algum produto da categoria MASSAS tem 2,25 chances a mais de levar algum produto da categoria QUEIJOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

14. Em 1,70% de todas as compras, foram compradas em conjunto as categorias MOLHOSDETOMATE e QUEIJOS. Considerando agora todas as compras em que a categoria MOLHOSDETOMATE foi levada, em 40,11% destas a categoria QUEIJOS também foi levada. Uma pessoa que compra algum produto da categoria MOLHOSDETOMATE tem 2,61 chances a mais de levar algum produto da categoria QUEIJOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

15. Em 3,13% de todas as compras, foram compradas em conjunto as categorias CERVEJASLATA e REFRIG.PETACIMA600ML. Considerando agora todas as compras em que a categoria CERVEJASLATA foi levada, em 35,12% destas a categoria REFRIG.PETACIMA600ML também foi levada. Uma pessoa que compra algum produto da categoria CERVEJASLATA tem 2,01 chances a mais

de levar algum produto da categoria REFRIG.PETACIMA600ML do que uma pessoa sobre a qual não se tem informação e vice-versa.

16. Em 0,86% de todas as compras, foram compradas em conjunto as categorias MAIONESE e REFRIG.PETACIMA600ML. Considerando agora todas as compras em que a categoria MAIONESE foi levada, em 38,08% destas a categoria REFRIG.PETACIMA600ML também foi levada. Uma pessoa que compra algum produto da categoria MAIONESE tem 2,18 chances a mais de levar algum produto da categoria REFRIG.PETACIMA600ML do que uma pessoa sobre a qual não se tem informação e vice-versa.

17. Em 4,02% de todas as compras, foram compradas em conjunto as categorias CARNEBOVINA1 e VERDURAS. Considerando agora todas as compras em que a categoria CARNEBOVINA1 foi levada, em 50,79% destas a categoria VERDURAS também foi levada. Uma pessoa que compra algum produto da categoria CARNEBOVINA1 tem 2,12 chances a mais de levar algum produto da categoria VERDURAS do que uma pessoa sobre a qual não se tem informação e vice-versa.

18. Em 1,24% de todas as compras, foram compradas em conjunto as categorias CALDOSTEMPERO e VERDURAS. Considerando agora todas as compras em que a categoria CALDOSTEMPERO foi levada, em 57,37% destas a categoria VERDURAS também foi levada. Uma pessoa que compra algum produto da categoria CALDOSTEMPERO tem 2,39 chances a mais de levar algum produto da categoria VERDURAS do que uma pessoa sobre a qual não se tem informação e vice-versa.

19. Em 2,55% de todas as compras, foram compradas em conjunto as categorias ARROZ e VERDURAS. Considerando agora todas as compras em que a categoria ARROZ foi levada, em 50,89% destas a categoria VERDURAS também foi levada. Uma pessoa que compra algum produto da categoria ARROZ tem 2,12 chances a mais de levar algum produto da categoria VERDURAS do que uma pessoa sobre a qual não se tem informação e vice-versa.

Buscou-se também observar como as 10 categorias mais compradas interagem ente si. A Tabela 3.8 mostra algumas das regras que foram geradas entre estas categorias.

Tabela 3.8 – Regras entre as categorias mais compradas

Índice	Categoria 1	==>	Categoria 2	Suporte (%)	Confiança (%)	Lift
1	IOGURTES	==>	FRUTAS	5,01220	40,80670	2,040976
2	FRUTAS	==>	VERDURAS	11,48783	57,45719	2,394949
3	IOGURTES	==>	PADARIA	4,03972	32,88923	2,160944
4	PADARIA	==>	QUEIJOS	6,10508	40,11262	2,606060
5	CONGELADOS	==>	PADARIA	4,22173	30,96193	2,034314
6	IOGURTES	==>	CONGELADOS	3,47818	28,31750	2,076792
7	CONGELADOS	==>	QUEIJOS	5,45866	40,03352	2,600921
8	IOGURTES	==>	BISCOITOS	4,64002	37,77660	2,375379
9	IOGURTES	==>	QUEIJOS	4,22132	34,36773	2,232823

A lista abaixo mostra por extenso como é a interpretação destas regras. Novamente, o índice da lista corresponde exatamente ao índice das regras na tabela.

1. Em 5,01% de todas as compras, foram compradas em conjunto as categorias IOGURTES e FRUTAS. Considerando agora todas as compras em que a categoria IOGURTES foi levada, em 40,81% destas a categoria FRUTAS também foi levada. Uma pessoa que compra algum produto da categoria IOGURTES tem 2,04 chances a mais de levar algum produto da categoria FRUTAS do que uma pessoa sobre a qual não se tem informação e vice-versa.

2. Em 11,49% de todas as compras, foram compradas em conjunto as categorias FRUTAS e VERDURAS. Considerando agora todas as compras em que a categoria FRUTAS foi levada, em 57,46% destas a categoria VERDURAS também foi levada. Uma pessoa que compra algum produto da categoria FRUTAS tem 2,39 chances a mais de levar algum produto da categoria VERDURAS do que uma pessoa sobre a qual não se tem informação e vice-versa.

3. Em 4,04% de todas as compras, foram compradas em conjunto as categorias IOGURTES e PADARIA. Considerando agora todas as compras em que a categoria IOGURTES foi levada, em 32,89% destas a categoria PADARIA também foi levada. Uma pessoa que compra algum produto da categoria IOGURTES tem 2,16 chances a mais de levar algum produto da categoria PADARIA do que uma pessoa sobre a qual não se tem informação e vice-versa.

4. Em 6,10% de todas as compras, foram compradas em conjunto as categorias PADARIA e QUEIJOS. Considerando agora todas as compras em que a categoria PADARIA foi levada, em 40,11% destas a categoria QUEIJOS também foi levada. Uma pessoa que compra algum produto da categoria PADARIA tem 2,61

chances a mais de levar algum produto da categoria QUEIJOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

5. Em 4,22% de todas as compras, foram compradas em conjunto as categorias CONGELADOS e PADARIA. Considerando agora todas as compras em que a categoria CONGELADOS foi levada, em 30,96% destas a categoria PADARIA também foi levada. Uma pessoa que compra algum produto da categoria CONGELADOS tem 2,03 chances a mais de levar algum produto da categoria PADARIA do que uma pessoa sobre a qual não se tem informação e vice-versa.

6. Em 3,48% de todas as compras, foram compradas em conjunto as categorias IOGURTES e CONGELADOS. Considerando agora todas as compras em que a categoria IOGURTES foi levada, em 28,32% destas a categoria CONGELADOS também foi levada. Uma pessoa que compra algum produto da categoria IOGURTES tem 2,08 chances a mais de levar algum produto da categoria CONGELADOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

7. Em 5,46% de todas as compras, foram compradas em conjunto as categorias CONGELADOS e QUEIJOS. Considerando agora todas as compras em que a categoria CONGELADOS foi levada, em 40,03% destas a categoria QUEIJOS também foi levada. Uma pessoa que compra algum produto da categoria CONGELADOS tem 2,60 chances a mais de levar algum produto da categoria QUEIJOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

8. Em 4,64% de todas as compras, foram compradas em conjunto as categorias IOGURTES e BISCOITOS. Considerando agora todas as compras em que a categoria IOGURTES foi levada, em 37,78% destas a categoria BISCOITOS também foi levada. Uma pessoa que compra algum produto da categoria IOGURTES tem 2,37 chances a mais de levar algum produto da categoria BISCOITOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

9. Em 4,22% de todas as compras, foram compradas em conjunto as categorias IOGURTES e QUEIJOS. Considerando agora todas as compras em que a categoria IOGURTES foi levada, em 34,37% destas a categoria QUEIJOS também foi levada. Uma pessoa que compra algum produto da categoria

IOGURTES tem 2,23 chances a mais de levar algum produto da categoria QUEIJOS do que uma pessoa sobre a qual não se tem informação e vice-versa.

A Figura 3.4 mostra graficamente as relações existentes entre cada um das categorias mais vendidas. O gráfico faz um comparativo entre as regras. O tamanho da elipse indica o suporte relativo de cada regra e a cor indica a confiança relativa, sendo que quanto maior a elipse, maior é o suporte e quanto mais escura é a cor, maior é a confiança. Os eixos *x* e *y* representam respectivamente os antecedentes e conseqüentes das regras.

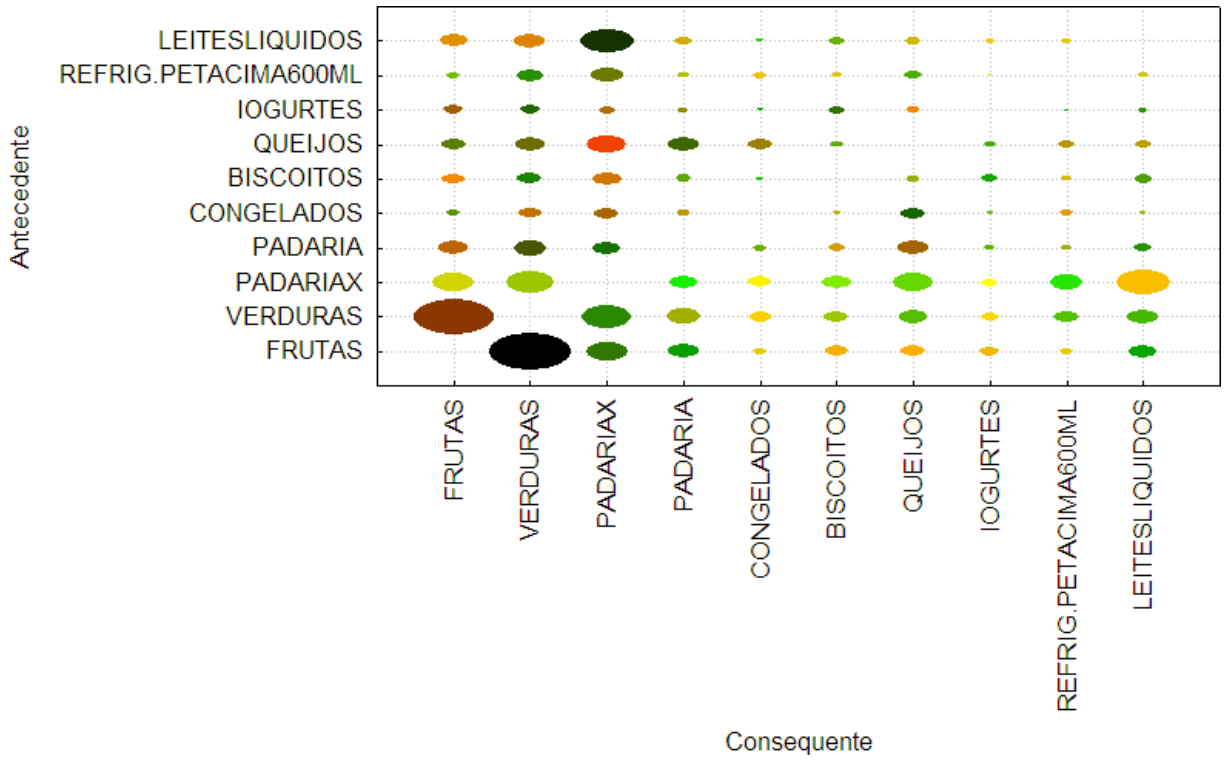

Figura 3.4 – Suporte e confiança relativos

Através do gráfico percebe-se que as regras com suporte mais altos em relação as outras envolve os conjuntos (VERDURAS; FRUTAS) e (LEITESLIQUIDOS; PADARIAX). Considerando estes conjuntos as regras que possuem maior confiança são: $FRUTAS \rightarrow VERDURAS$ e $LEITESLIQUIDOS \rightarrow PADARIAX$. Outras regras com confianças relativas mais altas são: $IOGURTES \rightarrow BISCOITOS$ e $CONGELADOS \rightarrow QUEIJOS$.

A Figura 3.5 mostra um gráfico em rede onde também é possível observar as relações entre as categorias mais vendidas. Neste gráfico o tamanho da elipse representa o suporte relativo de cada item e não de cada regra como no gráfico anterior. Quanto maior a elipse, maior é o suporte daquele item, isto é, maior é a percentagem das compras que aquele item foi

levado. A espessura da linha representa o suporte relativo de cada regra. Quanto maior a espessura, maior é o suporte relativo. Finalmente, a cor da linha indica o *lift* das regras. Quanto mais escura é a cor, maior é o *lift*.

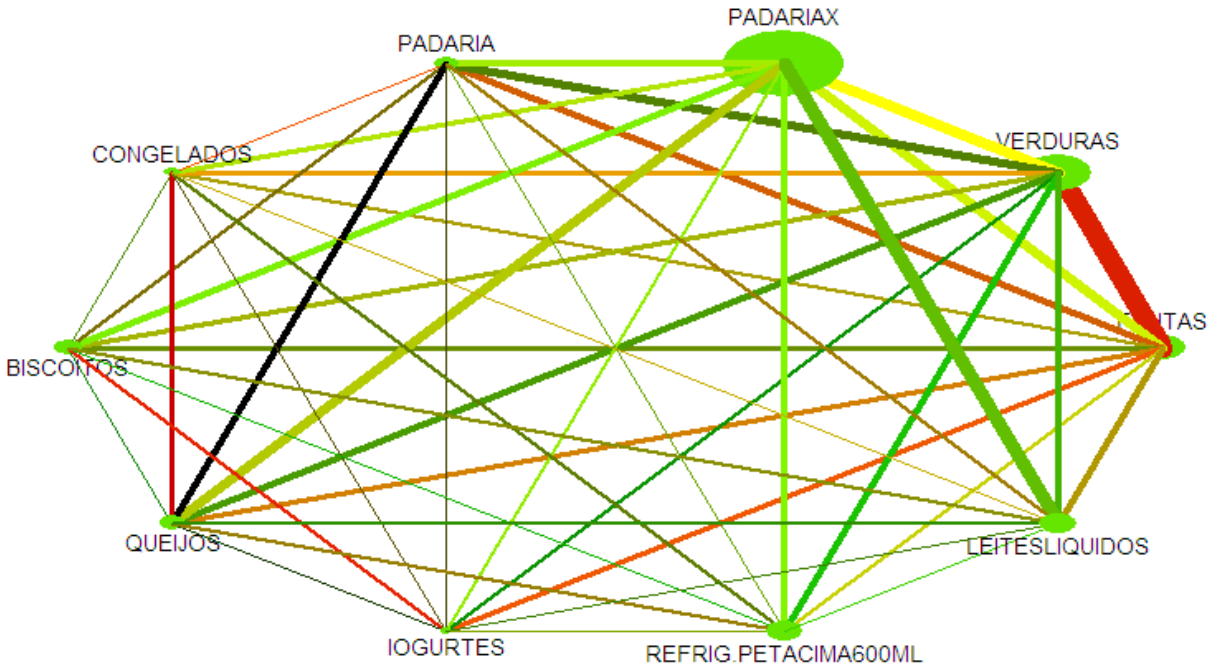

Figura 3.5 – Suporte e *lift* relativos

Percebe-se através do gráfico que a relação com o maior *lift* é entre as categorias PADARIA e QUEIJOS. A relação que possui maior suporte, isto é, as duas categorias que mais ocorrem em conjunto são: VERDURAS e FRUTAS. Os dois gráficos (Figura 3.4 e Figura 3.5) são também úteis quando o objetivo é saber qual é o produto que está mais fortemente relacionado com um outro determinado produto. Por exemplo, se o objetivo fosse encontrar a categoria que está mais fortemente relacionada com a categoria LEITELIQUIDOS, as categorias PADARIA e FRUTAS seriam facilmente identificadas.

Foi realizada uma outra análise com o objetivo de encontrar relações interessantes entre as 10 categorias mais vendidas e aquelas não tão vendidas. Aqui foram consideradas apenas as compras de ≤ R$ 40,00 (“sacolinha”). Existem 201.361 compras dentro desta especificação. Foi adotado um suporte igual a 0 e uma confiança de 10%. Foram geradas combinações de até no máximo 2 categorias. O tempo de mineração foi de 3 horas. Foram geradas 1607 regras. Considerou-se apenas as regras com *lift* igual a 2, o que resultou em um conjunto final de 97 regras. A Tabela 3.9 mostra algumas das regras mais interessantes.

Tabela 3.9 – Categorias mais freqüentes X categorias menos freqüentes

Índice	Categoria 1	==>	Categoria 2	Suporte (%)	Confiança (%)	Lift
1	DIETSOBREM	==>	FRUTAS	0,041716	33,20158	2,279797
2	DIETBISCOITOS	==>	FRUTAS	0,025824	32,91139	2,259871
3	DIETCEREAIS	==>	FRUTAS	0,109257	30,01364	2,060896
4	CEREALMATINAL	==>	IOGURTES	0,185736	30,40650	3,951648
5	CATCHUP	==>	CONGELADOS	0,132598	27,41273	2,858547
6	ACHOCOLATADOS	==>	LEITESLIQUIDOS	0,663982	34,799584	2,485115
7	DIETACHOCOLATADOS	==>	LEITESLIQUIDOS	0,014899	34,482759	2,462490
8	CATCHUP	==>	REFRIG.PETACIMA600ML	0,148490	30,69815	2,305894
9	BATATAPALHA	==>	REFRIG.PETACIMA600ML	0,298469	28,82494	2,165188
10	PROD.INF.SABONETES	==>	IOGURTES	0,037247	18,98734	2,467607
11	PROD.INF.CONDICIONADOR	==>	IOGURTES	0,010926	18,64407	2,422995

Percebe-se uma relação forte entre produtos dietéticos (DIETSOBREM, DIETBISCOITOS e DIETCEREAIS) e frutas. Pessoas que compram produtos de uma destas categorias têm mais de 2 chances a mais de levar frutas e vice-versa do que pessoas sobre as quais não se tem informação. Isto reflete um hábito saudável das pessoas, já que a preocupação com a saúde alimentar tem aumentado nos últimos anos. Na maioria das vezes os produtos dietéticos são produtos que provêem alto retorno para os supermercadistas. Este padrão poderia ser utilizado para incentivar a venda de produtos deste tipo, seja colocando-os ao lado das prateleiras e cestas de frutas ou de uma outra maneira.

A regra $CATCHUP \rightarrow CONGELADOS$ também se mostra como uma regra interessante. Provavelmente, uma promoção na categoria CATCHUP ou uma simples mudança no *layout*, poderia aumentar as vendas tanto de *catchups* como de hambúrgueres, por exemplo.

Nas regras 10 e 11 percebe-se que a compra destas categorias provavelmente estão sendo realizadas para crianças. Esta informação poderia ser utilizada, por exemplo, para o desenvolvimento de uma campanha de marketing personalizada para mães e pais jovens.

Possivelmente, as categorias mais levadas podem ser vistas como uma das principais razões pelas quais os consumidores freqüentam os supermercados. Uma análise deste tipo pode ser utilizada para alavancar as vendas destas categorias e de produtos vendidos com menor freqüência.

Através destas análises realizadas até aqui, chegou-se a uma importante conclusão: mesmo com todas as medidas existentes de avaliação do conhecimento, é extremamente difícil aplicar a técnica de mineração de dados para gerar conhecimento sem que exista um problema específico a ser resolvido. Isto acontece por causa do grande número de padrões gerados. Então o pesquisador optou por criar um *case* baseado em um problema real do mundo do varejo e mostrar através dele como a mineração de regras de associação pode ser útil a um varejista.

Case

O supermercado irá lançar uma batata palha própria para concorrer com as que já estão à venda. Algumas informações são necessárias para responder perguntas como, por exemplo:

- Como produtos como este se relacionam com os outros produtos?
- Como produtos como este se relacionam com os produtos mais vendidos?
- Quais são os produtos mais vendidos desta categoria e como estes se relacionam com os outros produtos?
- Existe algum fator de sazonalidade que tem influência na venda deste produto?
- Existe algum fator de sazonalidade que tem influência na venda dos produtos mais associados a este?

Com estas e outras informações disponíveis é possível expor o novo produto no local certo, na hora certa, com o preço certo, para as pessoas certas e desenvolver campanhas de marketing e promoções mais precisas a fim de promover a venda do produto lançado.

Primeiramente buscou-se descobrir como a categoria BATATAPALHA se relaciona com outras categorias em geral. As regras mais interessantes encontradas com a categoria em questão no antecedente estão na Tabela 3.10.

Tabela 3.10 – Regras com BATATAPALHA no antecedente

Antecedente	==>	Conseqüente	Suporte(%)	Confiança(%)	Lift
BATATAPALHA	==>	FRANGOSACOUGUE	0,15346	14,8201	3,0253
BATATAPALHA	==>	FRIOSAZEITONAS	0,04668	4,5084	4,1912
BATATAPALHA	==>	EXTRATOSDETOMATE	0,07747	7,4820	4,4130
BATATAPALHA	==>	CARNEBOVINA1	0,10727	10,3597	2,1210
BATATAPALHA	==>	MOLHOSDETOMATE	0,14452	13,9568	6,6313
BATATAPALHA	==>	MAIONESE	0,09883	9,5444	9,8355
BATATAPALHA	==>	ERVILHA/MILHO	0,21206	20,4796	13,7873
BATATAPALHA	==>	CREMEDELEITE	0,21007	20,2878	12,5427
BATATAPALHA	==>	CATCHUP	0,08542	8,2494	17,0545

A Tabela 3.11 mostra as regras com a categoria BATATAPALHA no conseqüente.

Tabela 3.11 – Regras com BATATAPALHA no conseqüente

Antecedente	==>	Conseqüente	Suporte(%)	Confiança(%)	Lift
FRANGOSACOUGUE	==>	BATATAPALHA	0,15346	3,1326	3,0253
FRIOSAZEITONAS	==>	BATATAPALHA	0,04668	4,3398	4,1912
EXTRATOSDETOMATE	==>	BATATAPALHA	0,07747	4,5694	4,4130
MOLHOSDETOMATE	==>	BATATAPALHA	0,14452	6,8664	6,6313
MAIONESE	==>	BATATAPALHA	0,09883	10,1842	9,8355
ERVILHA/MILHO	==>	BATATAPALHA	0,21206	14,2762	13,7873
CREMEDELEITE	==>	BATATAPALHA	0,21007	12,9874	12,5427
CATCHUP	==>	BATATAPALHA	0,08542	17,6591	17,0545

Uma regra com duas categorias pode ter duas formas: $A \rightarrow B$ ou $B \rightarrow A$. Percebe-se nas Tabelas 3.10 e 3.11 que o *lift* e o suporte não levam em consideração a implicação, isto é, as medidas são as mesmas para as duas formas possíveis. Já a confiança varia de acordo com a posição das categorias, considerando o sentido da implicação.

De acordo com a confiança, a compra de BATATAPALHA tem mais influência na compra de CREMEDELEITE do que o contrário. Já a compra de CATCHUP tem mais influência na compra de BATATAPALHA do que o contrário. Em algumas aplicações é interessante considerar o sentido da aplicação. Estas características podem ser observadas na Figura 3.6. A cor da elipse que representa a regra $CREMEDELEITE \rightarrow BATATAPALHA$ é mais clara do que a cor da elipse que representa a regra contrária. Da mesma forma que a cor da elipse que representa a regra $BATATAPALHA \rightarrow CATCHUP$ é mais clara que a cor da elipse que representa a regra oposta.

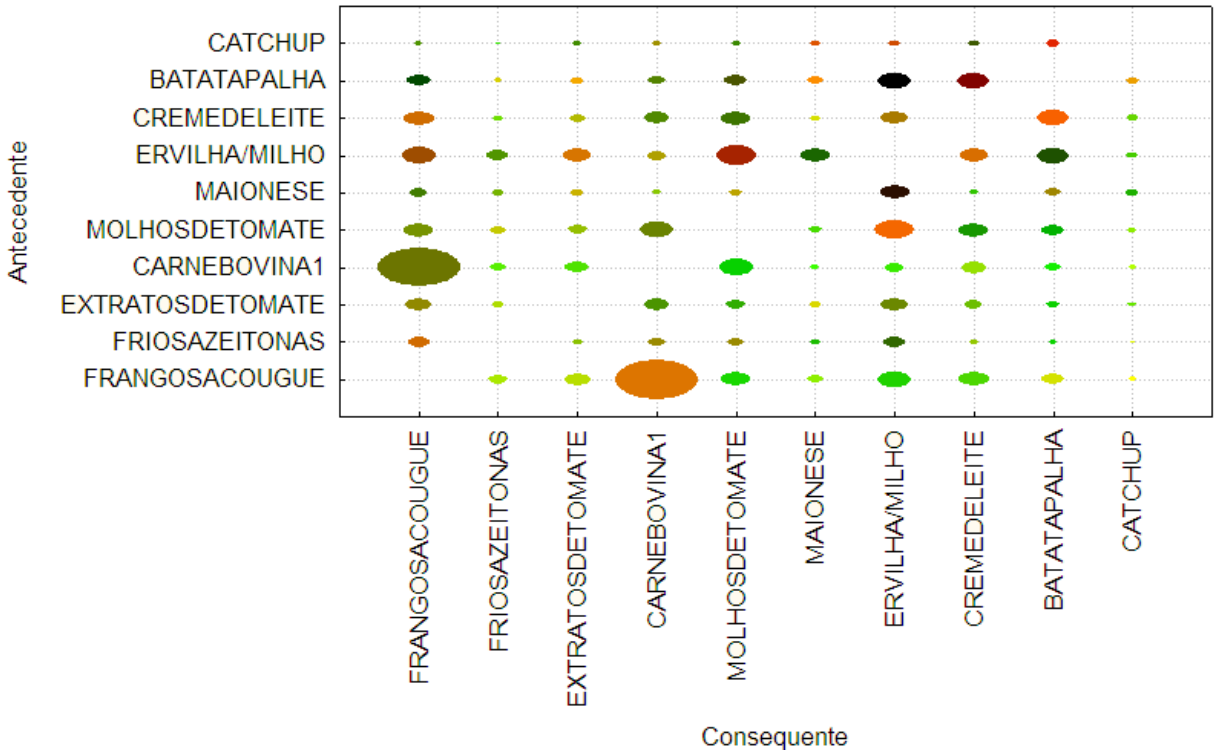

Figura 3.6 – Suporte e confiança relativos das regras com BATATAPALHA

De acordo com o *lift*, uma pessoa que compra BATATAPALHA tem 12,54 de chances a mais de comprar CREMEDELEITE assim como uma pessoa que compra CREMEDELEITE tem 12,54 chances a mais de comprar BATATAPALHA. Segundo o *lift*, as categorias que se relacionam mais fortemente com a categoria BATATAPALHA são: ERVILHA/MILHO, CREMEDELEITE e CATCHUP. Isto pode ser visto na Figura 3.7. As linhas mais escuras são aquelas que ligam a BATATAPALHA a estas 3 categorias, o que indica um valor de *lift* mais alto em relação as outras regras. Percebe-se ainda que há na tabela algumas categorias que em conjunto com a BATATAPALHA são os principais ingredientes de um *strogonoff*.

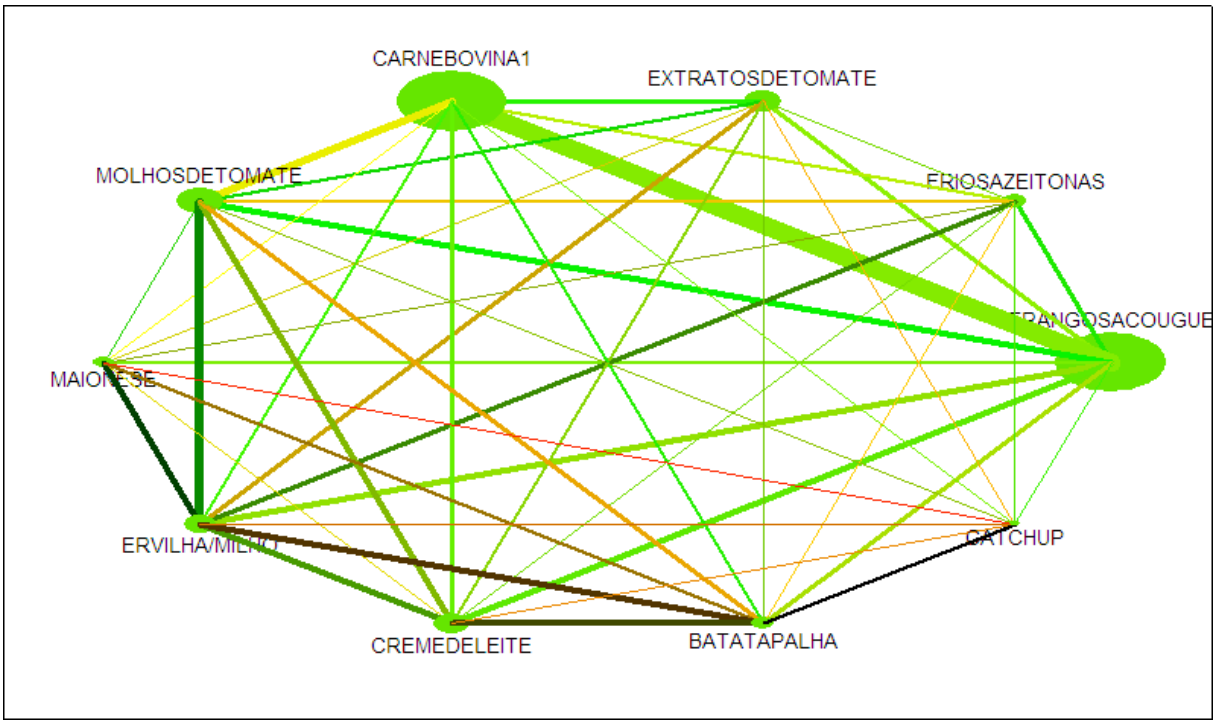

Figura 3.7 – Suporte e *lift* relativos das regras com BATATAPALHA

O gráfico da Figura 3.7 mostra ainda como as categorias mais relacionadas com a categoria BATATAPALHA se relacionam entre si. Percebe-se que, por exemplo, há uma forte relação entre: CREMEDELEITE e MOLHOSDETOMATE; e MAIONESE e ERVILHA/MILHO.

O próximo passo foi descobrir como a categoria BATATAPALHA se relaciona com as categorias mais vendidas. A Figura 3.8 mostra que entre as categorias mais vendidas, as que têm relações mais significativas com a categoria BATATAPALHA são: REFRIG.PETACIMA600ML e CONGELADOS.

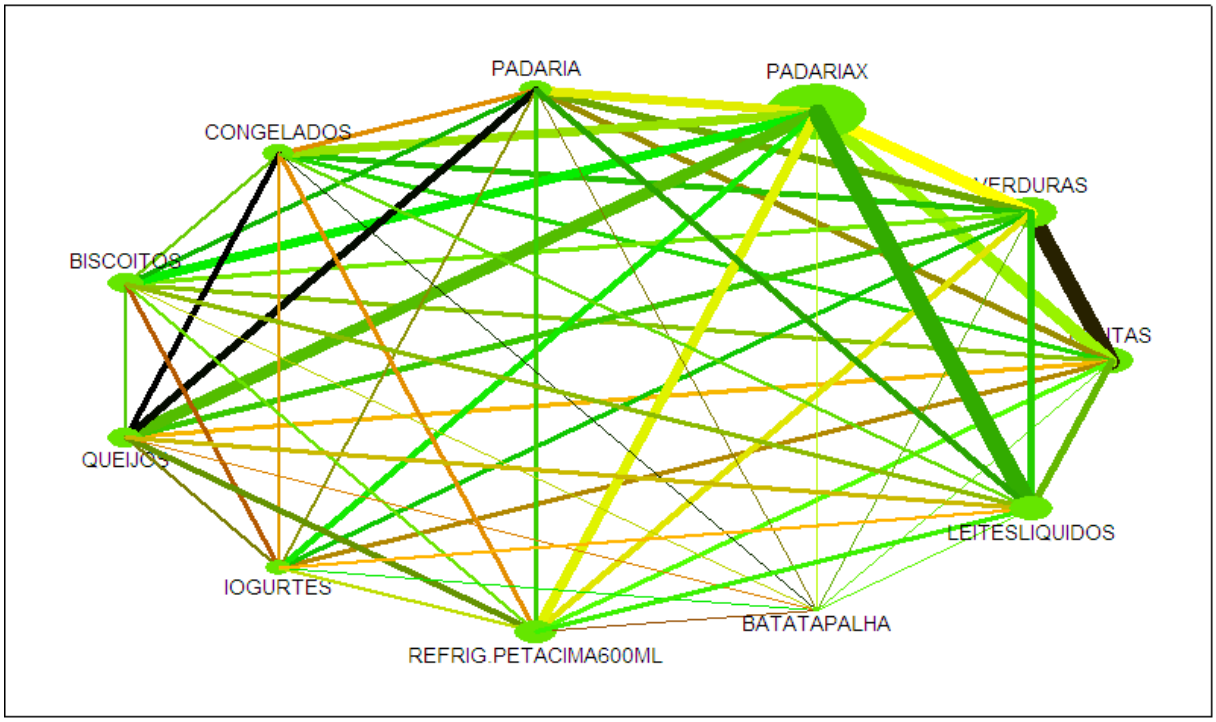

Figura 3.8 – Relações entre BATATAPALHA e categorias mais vendidas

A tabela 3.12 confirma este fato mostrando que o lift para estas relações são maiores que para as outras.

Tabela 3.12 – Relações entre a categoria BATATAPALHA e as mais vendidas

Antecedente	==>	Consequente	Suporte(%)	Confiança(%)	Lift
BATATAPALHA	==>	FRUTAS	0,149483	14,43645	0,991283
BATATAPALHA	==>	VERDURAS	0,275128	26,57074	1,464394
BATATAPALHA	==>	PADARIAX	0,323797	31,27098	0,892486
BATATAPALHA	==>	PADARIA	0,196165	18,94484	1,847964
BATATAPALHA	==>	CONGELADOS	0,238874	23,06954	2,405648
BATATAPALHA	==>	BISCOITOS	0,136074	13,14149	1,180752
BATATAPALHA	==>	QUEIJOS	0,192689	18,60911	1,771367
BATATAPALHA	==>	IOGURTES	0,080949	7,81775	1,015999
BATATAPALHA	==>	REFRIG.PETACIMA600ML	0,298469	28,82494	2,165188
BATATAPALHA	==>	LEITESLIQUIDOS	0,134088	12,94964	0,924762
BISCOITOS	==>	BATATAPALHA	0,136074	1,22261	1,180752
CONGELADOS	==>	BATATAPALHA	0,238874	2,49094	2,405648
IOGURTES	==>	BATATAPALHA	0,080949	1,05202	1,015999
LEITESLIQUIDOS	==>	BATATAPALHA	0,134088	0,95755	0,924762
PADARIA	==>	BATATAPALHA	0,196165	1,91348	1,847964
PADARIAX	==>	BATATAPALHA	0,323797	0,92413	0,892486
QUEIJOS	==>	BATATAPALHA	0,192689	1,83417	1,771367
REFRIG.PETACIMA600ML	==>	BATATAPALHA	0,298469	2,24195	2,165188
VERDURAS	==>	BATATAPALHA	0,275128	1,51631	1,464394
FRUTAS	==>	BATATAPALHA	0,149483	1,02643	0,991283

Na Tabela 3.12 existem alguns valores de confiança significativos como para a regra $BATATAPALHA \rightarrow PADARIAX$ que é de aproximadamente 31%. Isso quer dizer que em

31% das compras que ocorreu a categoria BATATAPALHA, ocorreu também a categoria PADARIAX. Como já foi comentado anteriormente, a confiança pode induzir a erros. A categoria PADARIAX já ocorre em 35% das compras a priori. Então o fato de ocorrer a categoria BATATAPALHA diminui as chances de ocorrer PADARIAX. Observa-se então o *lift* e é verificado que este possui um valor menor do que 1, o que indica que há uma correlação negativa entre estas duas categorias, evitando conclusões erradas.

Buscou-se também descobrir se existe algum efeito de sazonalidade que tem influência sobre a venda da categoria BATATAPALHA. Para isso, foram considerados os dias da semana (DOMINGO, SEGUNDA, TERÇA, QUARTA, QUINTA, SEXTA e SÁBADO), os períodos do dia (MANHÃ, TARDE e NOITE) e os tipos de dia (NORMAL, VÉSPERA DE FERIADO, FERIADO e PÓS-FERIADO) e foram geradas regras que relacionassem a ocorrência de BATATAPALHA à estes atributos.

O gráfico da Figura 3.9 mostra que há uma probabilidade maior de ocorrência de BATATAPALHA aos sábados e principalmente aos domingos. Já era esperado que as vendas desta categoria fosse maior aos sábados, pois este é o dia da semana em que ocorre uma maior quantidade de compras grandes. Domingo é o dia em que a quantidade total de compras é a menor delas, até porque neste dia o supermercado só funciona no período da manhã. Mesmo assim é no domingo que existe a maior probabilidade de vendas de BATATAPALHA. Se é domingo, há uma probabilidade duas vezes maior de ocorrência desta categoria em uma compra.

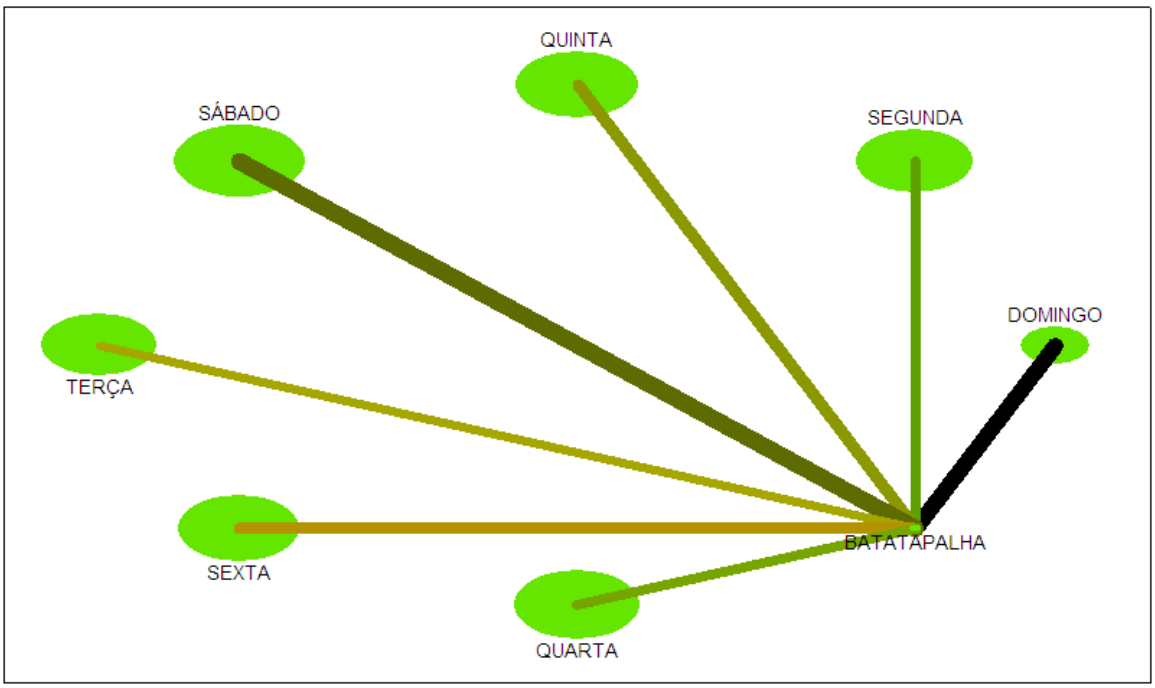

Figura 3.9 –BATATAPALHA X dias da semana

O gráfico da Figura 3.10 mostra que apesar de a diferença entre as probabilidades de venda de BATATAPALHA em cada período do dia não ser significativa e o número total de compras no período noturno ser menor, há uma maior probabilidade de venda de BATATAPALHA neste período.

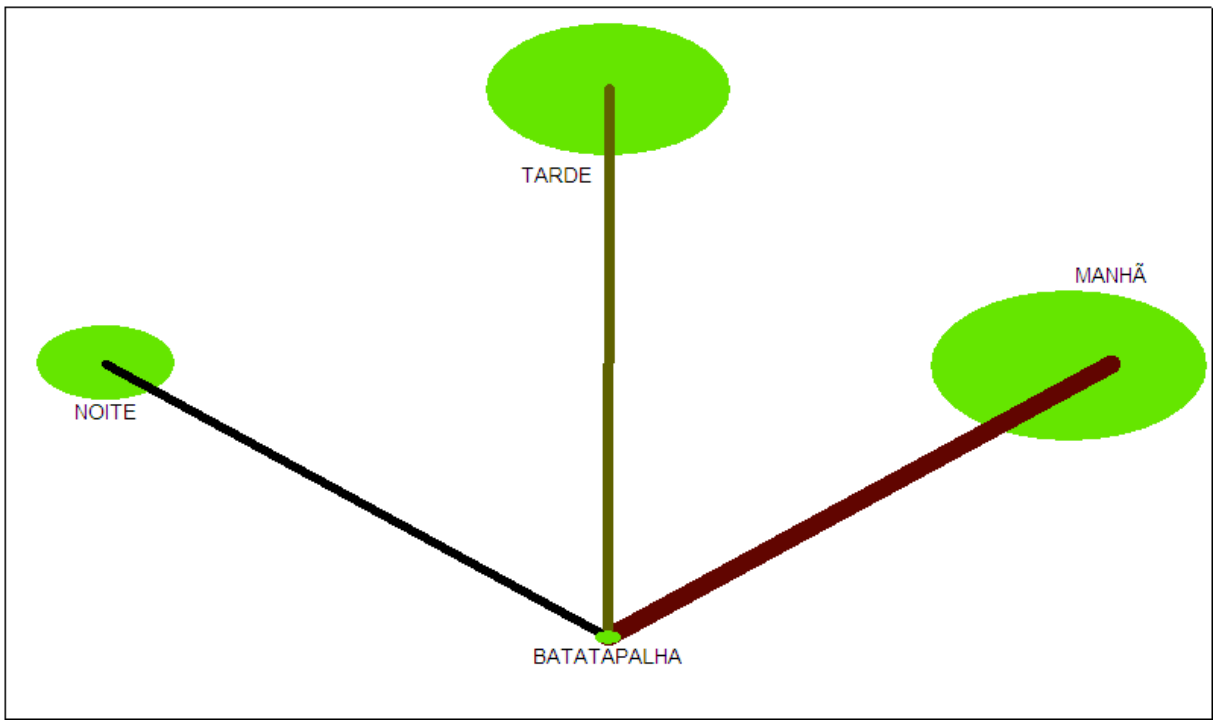

Figura 3.10 –BATATAPALHA X períodos do dia

A Figura 3.11 mostra as relações existentes entre a ocorrência de BATATAPALHA e o tipo de dia. Também não há uma diferença significativa entre as probabilidades de ocorrência desta categoria nos diferentes tipos de dia. Percebe-se através da espessura de cada linha que a batata palha é muito mais comprada em dias normais porque existe uma quantidade maior de dias normais do que feriados, vésperas e pós-feriados. Ainda assim, se é feriado, há uma probabilidade maior de ocorrência de BATATAPALHA em relação a qualquer outro tipo de dia. É importante informar que em alguns feriados o supermercado estudado funciona normalmente ou em meio período.

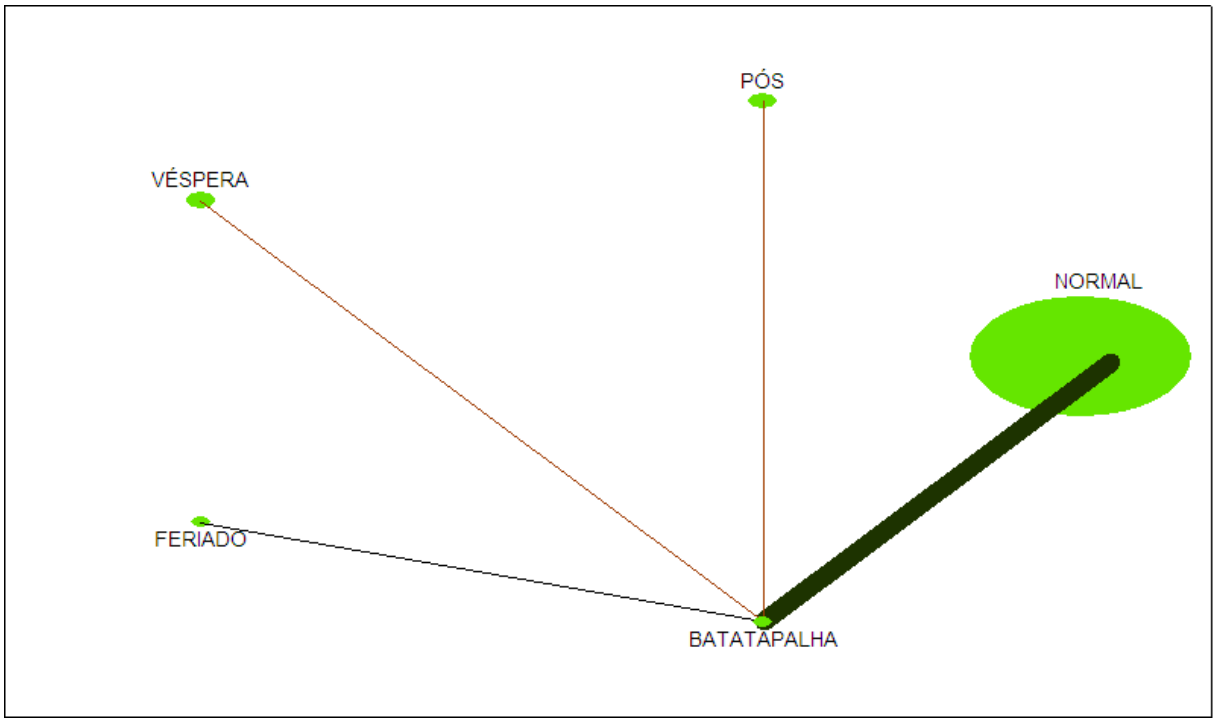

Figura 3.11 –BATATAPALHA X tipos de dia

Finalmente, foi realizada uma análise para descobrir se o fator dia da semana exerce alguma influência sobre a ocorrência de algumas das categorias que são correlacionadas com a categoria BATATAPALHA. As seguintes categorias foram analisadas: CREMEDELEITE, FRANGOSACOUGUE, REFRIG.PETACIMA600ML, MOLHOSDETOMATE e ERVILHA/MILHO. A Figura 3.12 mostra as relações existentes.

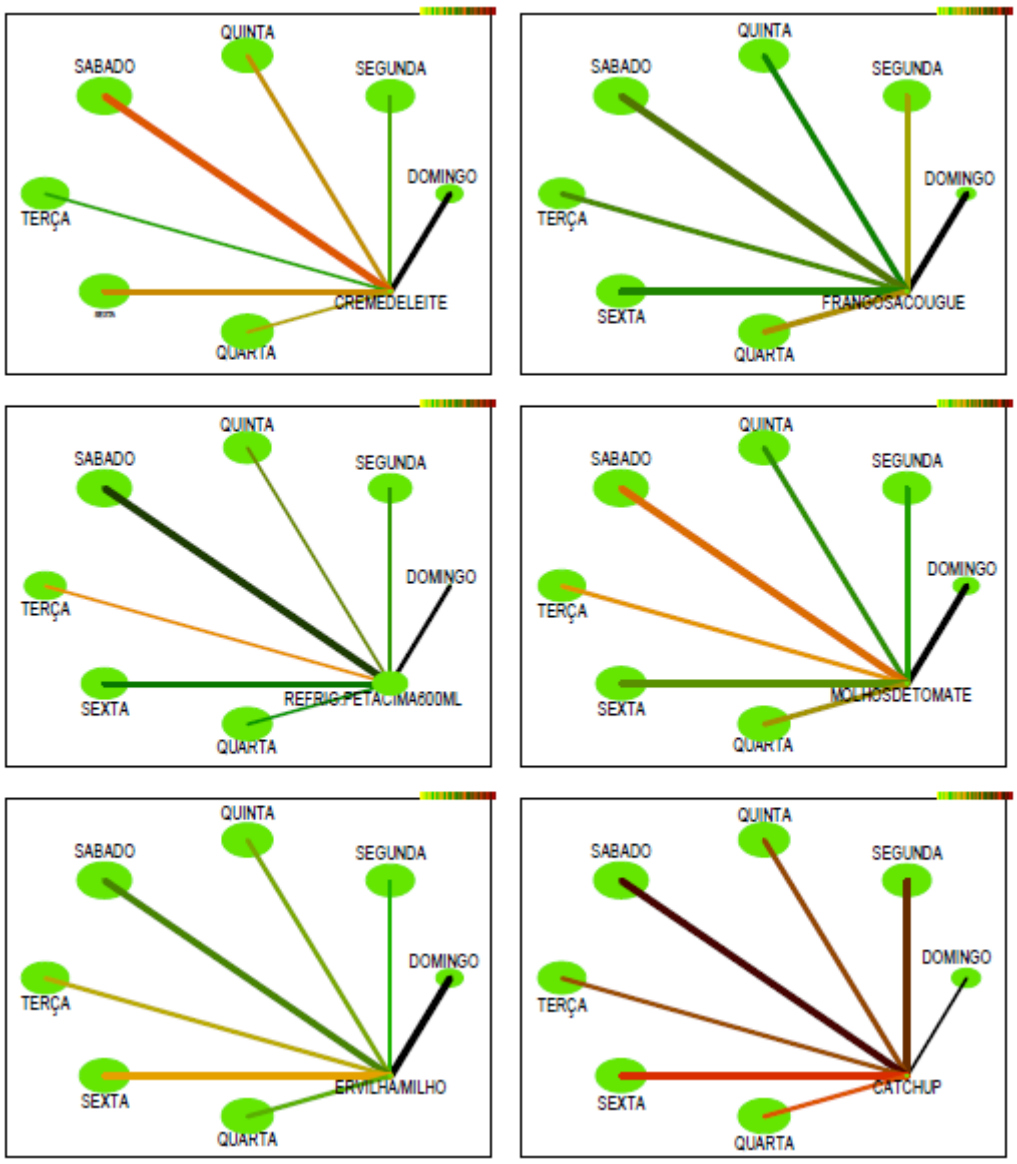

Figura 3.12 – Outras categorias X dias da semana

Percebe-se através dos gráficos que todas as categorias analisadas têm maior probabilidade de ocorrer aos sábados e principalmente aos domingos.

Além das informações consideradas neste trabalho, outras informações poderiam ser interessantes ao tomador de decisão como, por exemplo, o retorno que cada produto provê para o supermercadista e o preço de cada produto. Poderia ser descoberta uma relação entre o preço de um produto e o abandono do mesmo por parte dos clientes para responder a seguinte pergunta: A partir de que preço as pessoas deixam de escolher este produto e passam a escolher outro? Relacionar a marca de um produto a uma marca de outro produto também pode gerar informações de grande valor.

A grande quantidade de informações diferentes que pode ser gerada através das regras de associação pode ser encarada como um ponto positivo quando se têm um problema específico

a ser resolvido. Esse montante de informações aliado a fácil interpretação dos resultados que esta técnica fornece faz com que o sucesso de uma aplicação dependa mais do entendimento dos conceitos envolvidos. Entender perfeitamente o que é uma associação entre eventos e que tipo de informação a técnica de mineração de regras de associação pode gerar é fundamental. A partir do momento em que estes conceitos são assimilados, a probabilidade de sucesso de uma aplicação aumenta exponencialmente. Na Seção 3.4 é realizada uma discussão sobre a aplicabilidade da técnica de mineração de regras de associação.

3.4 Discussão

Nesta pesquisa foi aplicada a técnica de mineração de regras de associação na modelagem dos dados de um supermercado com o objetivo de gerar conhecimento. O pesquisador não possuía nenhum problema específico para ser resolvido com auxílio da mineração de dados. Isso dificulta a análise, pois geralmente é encontrado um grande número de padrões que podem ser utilizados no auxílio a diversos tipos de decisões. O pesquisador acredita que a mineração de regras de associação pode ser muito mais útil quando existe um problema específico que tem que ser resolvido ou existe uma decisão específica que tem que ser tomada. Isto fornece um foco à análise, diminuindo o campo de busca de informações e reduzindo o número de padrões gerados. O exemplo dado pelo case na Seção 3.3 ilustrou um problema específico e como as regras de associação podem prover informações que auxiliem na tomada de decisão.

No varejo, regras de associação podem ser aplicadas com o objetivo de responder as seguintes perguntas:

- Quais as interações ou conjunto de interações são os melhores indicadores de ações futuras dos clientes?
- Quais clientes são mais previsíveis?
- Como as respostas acima variam quando observa-se apenas os clientes mais ou menos rentáveis?
- As conversões recentes podem ser atribuídas a última campanha de marketing ou outros fatores em jogo?
- Com base no comportamento de compra e no perfil, é possível identificar os melhores clientes antes de estes realizarem uma compra?
- Como o comportamento de clientes na loja afeta ao comportamento de compra destes on-line?

- Qual é a relação entre o abandono de uma compra on-line e as compras subseqüentes, on-line ou na loja?
- Porque alguns clientes compram após um simples *download* de um papel de parede enquanto outros não compram nem após um alto investimento da empresa em uma campanha de marketing?
- O que torna um cliente fiel? Preços? Produtos? Serviços?
- Qual é o comportamento dos clientes mais leais?
- Quais segmentos existem no grupo dos clientes mais leais e como eles variam em termos de retorno para a loja?
- Quais produtos e promoções realmente aumentam as vendas para os clientes fiéis?
- Quais interações tornam clientes casuais em clientes fiéis?
- Existem promoções beneficiando apenas clientes fiéis, os quais já comprariam mesmo sem estas promoções?
- Quais são os primeiros sinais que mostram que um cliente irá deixar de ser fiel e que promoções provavelmente contribuirão para retê-lo?
- Quais campanhas de marketing são mais efetivas para atração de clientes novos e rentáveis?
- Qual é a melhor maneira de atingir os clientes mais rentáveis?
- Quais segmentos de clientes respondem e não respondem a diferentes tipos de mensagens?
- Quais ações indicam que um cliente está pronto para uma promoção específica?
- Quais canais de comunicação são mais efetivos para determinados segmentos e produtos?
- É mais eficiente gastar com promoções on-line ou com cartazes impressos?
- Os segmentos de clientes mais rentáveis estão sendo atingidos pelas campanhas?
- Quais são as campanhas mais e menos lucrativas?
- Quais produtos são vendidos em conjunto e com que freqüência?

- Quais são os produtos relacionados a itens em promoção que provêem maior retorno? Como essas relações variam de acordo com a loja, cidade e estado?
- Qual é o *layout* ideal para a loja e como ele deve variar de acordo com a localização da loja e dados demográficos dos clientes?
- Quando algum produto está em falta no estoque qual a reação dos clientes? Eles escolhem outro diferente na loja ou deixam o local sem finalizar a compra?
- Quais são os novos e emergentes padrões de compra que devem ser considerados na definição do *layout*?
- Como as estratégias de promoções devem variar de acordo com os dados demográficos dos clientes?
- Qual é o impacto de longo prazo causado por determinada promoção?
- Qual valor de retorno deveria ser esperado para a próxima semana se determinada promoção fosse lançada hoje?
- Quais são as promoções mais eficientes para atingir os clientes mais rentáveis?
- Como fatores sazonais, regionais e outros afetam o comportamento de compra?
- Como os padrões de compra dos clientes fiéis estão mudando com o tempo?
- Quais produtos são vendidos freqüentemente juntos em determinada região, mas em outras não? Por quê?
- Qual seqüência de eventos quase sempre faz com que um cliente realize uma compra on-line?
- O que leva alguém ao site da loja ou a própria loja e o que o mantém lá por mais tempo?

Percebe-se através das perguntas anteriores que a aplicabilidade das regras de associação ou de uma análise de cesto de compras pode ser relacionada também a funções específicas que profissionais ocupam, isto é, a que tipo de decisões que estas têm que tomar. Diferentes profissionais como profissionais de CRM, vendas, serviço de atendimento ao cliente, marketing, merchandising, analistas de negócio, analistas *web*, entre outros, podem ser beneficiados por uma aplicação de regras de associação.

Na opinião do pesquisador, a melhor forma de se utilizar as regras de associação é iniciar a partir de um problema específico e buscar as informações que serão interessantes ao

tomador de decisão através da mineração de dados. Para isso é necessário conhecer a técnica, isto é, saber que tipo de informações ela pode prover e também ter experiência no domínio de aplicação para saber quais informações buscar.

Um dos grandes problemas de ferramentas analíticas com aplicativos de mineração de dados e geração de relatórios está no fato de que uma pessoa gera o relatório e outra toma as decisões. Quem gera o relatório não sabe quais informações são importantes para o tomador de decisão e este não sabe como extrair informações realmente úteis ao processo de tomada de decisões. Muitas vezes se gasta tempo e dinheiro para a geração de relatórios inúteis. A técnica de regras de associação ainda tem a vantagem de ser de fácil interpretação. São necessários apenas alguns fundamentos de estatística básica para entender o que pode ser extraído como resultado de uma aplicação. O mais difícil é saber quais informações são interessantes para a tomada de alguma decisão.

A técnica de mineração de regras de associação técnica provê um resumo exato do que está acontecendo e de como as coisas se relacionam entre si. Basicamente, ela mede o grau de interação entre eventos ocorridos. Mesmo assim a maioria dos varejistas ainda não descobriu as vantagens que uma ferramenta deste tipo pode oferecer ao seu negócio.

4. Conclusões

4.1 Conclusões

O objetivo principal deste trabalho de dissertação foi realizar uma análise de cesto de compras através da mineração de regras de associação sobre os dados transacionais de um supermercado de forma a proporcionar maior conhecimento sobre o comportamento de compra dos seus consumidores. A existência de um banco de dados estruturado, organizado e livre de ruídos se mostrou de fundamental importância para a análise. Uma base de dados com estas características facilita o pré-processamento das informações e torna possível a geração de resultados de qualidade, que representem exatamente o comportamento dos clientes. Dependendo do tipo de análise, a existência de ruídos, dados faltantes ou erros de digitação, impossibilita a extração de conhecimento, como ocorreu com a análise do efeito das promoções.

O processo de descoberta de conhecimento em base de dados não é um processo genérico de descoberta. Não se aplica uma técnica de mineração de dados com o objetivo geral de gerar qualquer tipo de conhecimento. Sempre há um foco que especifica o tipo de conhecimento que se quer gerar. Este tipo de conhecimento é determinado pela escolha do tipo de tarefa (previsão, classificação, associação e *clustering*) e também pelo tipo de algoritmo escolhido.

Como já se sabia, o grande número de regras gerado pela técnica utilizada neste trabalho dificulta a análise fazendo com que a seleção das melhores regras seja uma tarefa impraticável, mesmo com todas as medidas de avaliação de conhecimento existentes. Mas quando é dado um foco mais específico à análise para resolver um problema em particular, a quantidade de padrões gerados nem sempre é um problema. Quando há uma decisão específica a ser tomada e o tomador de decisão sabe que tipo de informação a técnica de mineração de regras de associação pode gerar como resultado, a quantidade de padrões gerados por esta técnica pode ser entendida como uma vantagem, pois pode prover suporte a uma grande variedade de decisões.

A capacidade de processamento e armazenamento de dados é um limitador da técnica utilizada neste trabalho e de qualquer outro processo de extração de conhecimento com base em dados. Porém, para as regras de associação, já existem algoritmos eficientes que são capazes de minerar grandes quantidades de dados em algumas horas de processamento.

A técnica de mineração de regras de associação assim como qualquer outra técnica de mineração de dados, pode ser entendida como uma ferramenta de auxílio à tomada de decisões e não substitui o conhecimento e a experiência de um especialista do domínio de aplicação, apenas os enriquece. Para a utilização desta técnica, não é necessário um

profissional especialista em mineração de dados, o que pode aproximar o tomador de decisão da ferramenta analítica. A compreensão dos conceitos relacionados à associação entre eventos ocorridos é o bastante para que um tomador de decisão obtenha sucesso em uma aplicação.

O conhecimento útil gerado por um processo de mineração de dados nem sempre é um conhecimento novo. Muitos padrões gerados pelas regras de associação já são conhecidos e até mesmo óbvios, mas não são mensurados. A mensuração do conhecido ou do óbvio pode gerar um conhecimento mais preciso para a tomada de alguma decisão.

A técnica de mineração de regras de associação se mostrou capaz de gerar grande quantidade de conhecimento útil, seja ele novo ou não. No supermercado estudado, os principais tomadores de decisões são os próprios donos, mas decisões relativas a *layout*, promoções, preços e campanhas de marketing em geral, são tomadas pelos diretores da rede da qual este supermercado faz parte. Devido aos seus resultados, este trabalho está sendo indicado à esta rede.

4.2 Trabalhos Futuros

Recomenda-se para trabalhos futuros um estudo mais aprofundado das medidas objetivas de avaliação de conhecimento em regras de associação e a aplicação destas medidas em problemas específicos no varejo com o objetivo de atenuar a dependência da técnica em relação ao conhecimento do domínio.

Além das medidas objetivas de avaliação de conhecimento, existem outras que consideram questões monetárias como o retorno de produtos associados, o retorno total das compras nas quais estes produtos ocorrem, entre outras variáveis, o que pode gerar um conhecimento de maior valor.

Considerar dados de cartões de fidelidade como idade do comprador, sexo, estado civil, renda, entre outros, também pode agregar grande valor a análise e prover maior quantidade de conhecimento para personalização de serviços.

Devido ao grande número de algoritmos de mineração de regras de associação, recomenda-se ainda um estudo comparativo entre aqueles existentes para aplicações em varejo, onde as bases de dados são mais densas.

REFERÊNCIAS BIBLIOGRÁFICAS

AGGELIS, V. *Association rules model of e-banking services.* 5th International Conference on Data Mining, Text Mining and their Business Aplications, 2004. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.8937&rep=rep1&type=pdf. Acesso em 20/03/2009.

AGRAWAL, R.; IMIELINSKI, T; SWAMI, A. *Mining association rules between sets of items in large databases.* ACM SIGMOD International Conference on Management of Data, p. 207-216, 1993.

AGRAWAL, R.; SRIKANT, R. *Fast algorithms for mining association rules.* 20th VLDB Conference, p. 478-499, 1994.

BERTRAND, J. W. M.; FRANSOO, J. C. *Operations management research methodologies using quantitative modeling.* International Journal of Operations & Production Management, v. 22, n. 2, p. 241-264, 2002.

BRADLEY, P.; FAYYAD, U.; MANGASARIAN, O. *Data mining: Overview and optimization opportunities.* Technical Report MSR-TR-98-04, Microsoft Research Report, Redmond, WA, 1998.

BRIJS, T.; SWINNEN, G; VANHOOF, K; WETS, G. *Building an Association Rules Framework to Improve Product Assortment Decisions.* Data Mining and Knowledge Discovery, v. 8, p. 7-23, 2004.

BRIN, S.; SILVERSTEIN, C.; MOTWANI, R. *Beyond Market Baskets: Generalizing Association Rules to Dependence Rules.* Data Mining and Knowledge Discovery, v. 2, p. 39-68, 1998.

BRUHA, I.; FAMILI, A. *Postprocessing in machine learning and data mining.* ACM SIGKDD Explorations Newsletter, v. 2, p. 110-114, 2000.

CARVALHO, V. O. *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado.* Tese de Doutorado. Instituto de Ciências Matemáticas de Computação, Universidade de São Paulo (USP), 2007.

CHEN, Y.; TANG, K.; SHEN, R.; HU, Y. *Market basket analysis in a multiple store environment.* Decision Support Systems, v. 40, p. 339-354, 2005.

CHENG, J; KE, Y.; NG, W. *Effective elimination of redundant association rules.* Data Mining and Knowledge Discovery, v. 16, p. 221-249, 2008. DOI 10.1007/s10618-007-0084-8

FACTPOINT GROUP. *Leading Practices in Market Basket Analysis: How Top Retailers are Using Market Basket Analysis to Win Margin and Market Share.* Research, 2008.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From Data Mining to Knowledge Discovery in Databases.* Artificial Intelligence Magazine, v. 17, n. 3, p. 37-54, 1996a.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *Knowledge Discovery and Data Mining: Towards a Unifying Framework.* 2nd International Conference on Knowledge Discovery and Data Mining, p. 82-88, 1996b.

GARCÍA, M., N., M.; ROMÁN, I. R.; PEÑALVO, F., J., G.; BONILLA, M., T. *An association rule mining method for estimating the impact of project management policies on software quality, development time and effort.* Expert Systems with Applications, v. 34, p. 522-529, 2008.

GENG, L.; HAMILTON, H. J. *Interestingness Measures for Data Mining: A Survey.* ACM Computing Surveys, v. 38, n. 3, 2006.

GOETHALS, B.; ZAKI, M. J. *Advances in frequent itemset mining implementations: report on FIMI'03***.** ACM SIGKDD Explorations Newsletter, v. 6, n. 1, p. 109-117, 2004.

GREWAL, D.; LEVY, M. *Retailing research: Past, present, and future.* Journal of Retailing, v. 83, n. 4, p. 447-464, 2007.

GROTH, R. DATA MINING: building competitive advantage. Prentice Hall PTR, Upper Saddle River, New Jersey, 2000.

HAN, J.; CHENG, H.; XIN, D.; YAN, X. *Frequent pattern mining: current status and future directions.* Data Mining and Knowledge Discovery, v. 15, n. 1, p. 55-86, 2007.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques.* 2.ed. Morgan Kaufmann, San Francisco, CA, 2006.

HIPP, J.; GUNTZER U.; NAKHAEIZADEH, G. *Algorithms for Association Rule Mining - A General Survey and Comparison.* ACM SIGKDD Explorations, v. 2, n. 1, p. 58-64, 2000.

HIPP, J.; GUNTZER U.; NAKHAEIZADEH, G. *Data Mining of Association Rules and the Process of Knowledge Discovery in Databases.* Data Mining in E-Commerce, Medicine, and Knowledge Management, p. 15-36, 2002.

KARABATAK, M.; INCE, M. C. *An expert system for detection of breast cancer based on association rules and neural network.* Expert Systems with Applications, v. 36, p. 3465-3469, 2009. DOI: 10.1016/j.eswa.2008.02.064

KAZIENKO, P. *Mining Indirect Associations Rules for Web Recommendation.* International Journal of Applied Mathematics *and* Computer Science, *v. 19, n. 1, p. 165-186, 2009.* DOI: 10.2478/v10006-009-0015-5

LAKATOS, E. M.; MARCONI, M. A. *Fundamentos da metodologia científica*. Editora Atlas, São Paulo, 2001.

LAVRAC, N.; FLACH, P.; AND ZUPAN, B. *Rule evaluation measures: A unifying view.* 9th International Workshop on Inductive Logic Programming (ILP '99). Bled, Slovenia, Springer-Verlag, p. 174-185, 1999.

LENCA, P.; MEYER, P.; VAILLANT, B.; LALLICH, S. *A multicriteria decision aid for interestingness measure selection.* Tech. Rep. LUSSI-TR-2004-01-EN, LUSSI Department, GET/ENST, Bretagne, France, 2004.

LI, X. *Two Essays on "Mining Basket Data: Models and Applications in Marketing".* Tese de Doutorado. Faculty of School of Business of The George Washington University, 2008.

MANCHANDA, P.; ANSARI, A.; GUPTA, S. *The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions.* Marketing Science, v. 18, n. 2, p. 95-114, 1999.

MELANDA, E. A. *Pós-processamento de regras de associação.* Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP), 2004.

METWALLY, A.; AGRAWAL, D.; ABBADI, A. E. *Using Association Rules for Fraud Detection in Web Advertising Networks.* 31st VLDB Conference, p. 169-180, 2005.

OHSAKI, M.; ABE, H.; TSUMOTO, S.; YOKOI, H.; YAMAGUCHI, T. *Evaluation of rule interestingness measures in medical knowledge discovery in databases.* Artificial Intelligence in Medicine, v. 41, p. 177-196, 2007.

OHSAKI, M.; KITAGUCHI, S.; OKAMOTO, K.; YOKOI, H.; YAMAGUCHI, T. *Evaluation of rule interestingness measures with a clinical dataset on hepatitis.* 8th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2004), Pisa, Italy, p. 362-373, 2004.

PIATETSKY-SHAPIRO, G. *Discovery, Analysis and Presentation of Strong Rules*. Knowledge Discovery in Databases, p. 229-248, 1991.

PIATETSKY-SHAPIRO, G. *Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics".* Data Mining and Knowledge Discovery, v. 15, p. 99-105, 2007. DOI 10.1007/s10618-006-0058-2.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. *Mineração de Dados.* REZENDE, S. O. (Ed.), Sistemas Inteligentes: Fundamentos e Aplicações, Editora Manole, Barueri, São Paulo, cap. 12, p. 307-335, 2003.

RIBEIRO, M. X.; TRAINA, A. J. M.; TRAINA, C.; AZEVEDO-MARQUES, P. M. *An Association Rule-Based Method to Support Medical Image Diagnosis With Efficiency.* IEEE Transactions on Multimedia, v. 10, n. 2, 2008. DOI 10.1109/TMM.2007.911837

RUSSELL, G. J.; PETERSEN, A. *Analysis of Cross Category Dependence in Market Basket Selection.* Journal of Retailing, v. 76, n. 3, p. 367-392 , 2000.

SÁNCHEZ, D.; VILA, M. A.; CERDA, L.; SERRANO, J. M. *Association rules applied to credit card fraud detection*. Expert Systems with Applications, v. 36, p. 3630-3640, 2009.

TAN, P.; KUMAR, V.; SRIVASTAVA, J. *Selecting the right objective measure for association analysis.* Information Systems, v. 29, p. 293-3313, 2004.

VAILLANT, B.; LENCA, P.; LALLICH, S. *A clustering of interestingness measures.* 7th International Conference on Discovery Science (DS 2004). Padova, Italy, p. 290-297, 2004.

WANG, K.; ZHOU, S.; HAN, J. *Profit mining: From patterns to actions.* 8th Conference on Extending Database Technology (EDBT 2002). Prague, Czech Republic, p. 70-87, 2002.

WEISS, S. M.; INDURKHYA, N. *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann, San Francisco, CA, 1998.

WITTEN, I. H.; FRANK, E. *Data Mining: practical machine learning tools and techniques with java implementations.* 2.ed. Morgan Kaufmann, San Francisco, CA, 2005.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; McLACHLAN, G. J.; NG, A.; LIU, B.; PHILIP, S.; YU, P. S.; ZHOU, Z.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. *Top 10 algorithms in data mining.* Knowledge Information Systems, v. 14, p. 1-37, 2008. DOI 10.1007/s10115-007-0114-2

ZAKI, M.J. *Generating Non-Redundant Association Rules.* 6th International Conference on Knowledge Discovery and Data Mining, p. 34-43, 2000.

ZHANG, C.; ZHANG, S. *Association Rules Mining: Models and Algorithms*. Lecture Notes in Artificial Intelligence, v. 2307, Springer, 2002.

ZHENG, Z,; KOHAVI, R.; MASON, L. *Real World Performance of Association Rule Algorithms.* 7th International Conference on Knowledge Discovery and Data Mining, p. 401-406, 2001.