ESCOLA POLITÉCNICA
DE PERNAMBUCO

d.s.c. UPE
Departamento de Sistemas e Computação

Aplicação de Regras de Associação para Auxílio na Gestão de Vendas de uma Empresa Varejista Utilizando a ferramenta WEKA

Trabalho de Conclusão de Curso

Engenharia da Computação

André Luiz Vale de Araújo
Orientador: Prof. Meuser Valença

UNIVERSIDADE
DE PERNAMBUCO

André Luiz Vale de Araújo

Aplicação de Regras de Associação para Auxílio na Gestão de Vendas de uma Empresa Varejista Utilizando a ferramenta WEKA

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia da Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, Novembro de 2009

Dedico este trabalho científico aos meus pais: Alfredo Rodolfo Beuttenmüller de Araújo e Hulda Vale de Araújo.
Aos meus irmãos Rodolfo e César
Ao meu inesquecível amigo Jorge Augusto de Oliveira Costa.

Agradecimentos

À Universidade de Pernambuco por disponibilizar total estrutura humana e física para minha formação.

À minha família por sempre acreditar no meu esforço e por me apoiar nos momentos difíceis.

À toda gerência e equipe de administração de banco de dados da CSI – Comércio Soluções Inteligentes, LTDA.

Ao administrador de banco de dados Maurício Marques pela intensa colaboração durante o desenvolvimento do trabalho.

Aos amigos pelo apoio dado em todas as fases da minha vida.

Ao Professor Meuser Valença, pelas orientações durante a elaboração e desenvolvimento do projeto.

Aos amigos e colegas da Escola Politécnica de Pernambuco por todo apoio e amizade cultivada durante estes cinco anos de convivência.

Por fim, agradeço a Deus por sempre me dar força, saúde e paz de espírito, para que eu pudesse finalizar o estágio com máxima dedicação.

Resumo

Mineração de dados é um processo interativo entre homem e máquina, onde o objetivo é extrair informações úteis para diversas áreas de uma organização comercial ou científica. Informações estas que auxiliem na criação ou modificação de novas tarefas e/ou processos da mesma. Este trabalho visa realizar um estudo analítico, utilizando a regra de associação como técnica de mineração de dados, em uma base de dados de uma empresa do ramo varejista, onde não há tradição em extração de conhecimento em seus dados, concebendo a esta corporação uma visão melhor dos seus clientes e produtos auxiliando na gestão de vendas.

Abstract

Data mining is an interactive process between man and machine, where the goal is to extract useful information for various areas of a commercial or scientific. Information that help in these creation or modification of new tasks and processes of the same. This study aims at an analytical study using the rules association as a technique of data mining in a database of a company in the retail industry, where there is no tradition in mining knowledge in their data, conceiving this corporation a better view its customers and products helping manage sales.

Sumário

Índice de Figuras

Índice de Tabelas

Tabela de Símbolos e Siglas

(Dispostos por ordem de aparição no texto)

KDD – Knowledge Discovery in Databases

MD – Mineração de Dados

DW – Data Warehouse

BD – Banco de Dados

TI – Tecnologia da Informação

Capítulo 1

Introdução

Empresas e pessoas convivem cada vez mais com dados. Dados de todo tipo são armazenados: de pessoais a profissionais. A facilidade de armazenamento que os discos de múltiplos gigabytes oferecem faz com que decisões a serem tomadas sobre o que fazer com esses dados sejam quase sempre adiadas, basta comprar outro disco e continuar a fazer como antes [1].

A cada caminho ou decisão que uma pessoa ou uma empresa toma, um novo registro é armazenado, resultando num amontoado de registros que aparenta não ter fim. A quantidade e a diversidade dos dados registrados nos bancos de dados dificultam ainda mais o entendimento que se possa ter deles. À medida que mais dados são inseridos, menos as pessoas têm entendimento deles. No entanto, o que está escondido no meio desse monte é informação. Extrair essas informações é possível utilizando o processo de DataMining [1].

O estudo em questão propõe a utilização de uma ferramenta open source, chamada WEKA, para a realização da mineração de dados. Será analisada a seguinte técnica de mineração: regra de associação; utilizando a técnica Apriori.

O estudo será baseado em uma base de dados de uma empresa do ramo varejista, na qual não há exploração de informação nos registros armazenados. O propósito disso é obter um melhor conhecimento dos dados lá inseridos, buscando padrões que auxiliem na gestão de vendas para determinados produtos, utilizando técnicas de Data Mining (MD).

A aplicação do processo de mineração de dados tornará possível uma análise matemática para derivar padrões e tendências que existem nos dados. É possível saber, por exemplo, quais os produtos que são mais vendidos, em que época do ano, qual o perfil dos clientes que mais compram o produto levantado, etc. Estes padrões não podem ser levantados com a exploração de dados convencional, pois as relações podem ser bastante complexas ou por haver uma quantidade de dados muito grande [2].

1.1 Objetivos e Metas

O processo de KDD (*Knowledge Discovery in Databases*) consiste em uma nova geração de técnicas que buscam não só a extração de informação, mas a descoberta de conhecimento a partir das bases de dados. Ele é composto por diversas etapas entre as quais se destaca a Mineração de Dados. Este trabalho visou aplicar justamente o processo de Mineração de Dados na base de dados de uma empresa varejista em busca de padrões que auxiliem a gestão da empresa.

Objetivo Geral

Identificar padrões no banco de dados de uma empresa varejista, através da aplicação do processo de MD, para subsidiar o processo de gestão de vendas da empresa.

Objetivos específicos:

- Selecionar e aplicar uma metodologia que oriente o processo de Mineração de Dados;
- Selecionar e aplicar uma tarefa, método e algoritmo de Mineração de Dados que sejam capazes de permitir uma solução do problema;
- Selecionar e aplicar uma ferramenta que implemente o algoritmo escolhido.

1.2 Estratégia de ação

Para a execução de uma dada tarefa para a escolha da técnica de MD mais adequada o trabalho foi dividido nas etapas:

1. **Pesquisa**

   Para a sua execução é importante saber alguma coisa a respeito do domínio da aplicação que se pretende realizar. O embasamento teórico em torno das técnicas de KDD e MD serão extraídos do livro “Data Mining – Pratical Machine

Learning Tools and Techniques" de Ian H. Witten & Eibe Frank [1], além de artigos variados em torno do assunto em questão.

Para a ferramenta escolhida será utilizado o sistema WEKA (*Waikato Environment for Knowledge Analysis*) para o estudo das técnicas de Mineração de dados para auxiliar na tomada de decisões. O software WEKA é composto por um conjunto de implementações algorítmicas e de diversas técnicas de Mineração de Dados, implementado em Linguagem Java, que tem como principal característica a portabilidade e também é um software livre.

Também será feito um estudo em torno da base de dados em questão, remodelando-a envolvendo apenas os elementos que estejam ligados diretamente ao estudo.

Esta é a parte mais interessante do projeto, a que mais alavanca e auxilia o empresário a descobrir filões de mercado. O cérebro humano, comprovadamente, consegue fazer até oito comparações ao mesmo tempo. A função do Data Mining é justamente ampliar esta comparação para "infinito" e tornar isso visível ao olho humano [2].

2. Escolha dos algoritmos de MD

A tarefa de MD escolhida para o trabalho foi Regras de associação. O algoritmo utilizado foi o algoritmo apriori.

3. Análise

Análise do resultado será feita observando o resultado obtido com os testes realizados na técnica e algoritmo escolhidos e aplicados à base de dados, avaliando os resultados para se chegar a um padrão que seja interessante para a corporação.

4. Conclusão

Ao final da fase analítica, a conclusão se deu de forma a expressar o resultado conseguido. Foi mostrado em quais casos é melhor utilizar qual algoritmo ou técnica, destacando a importância de cada qual bem como comentando suas

vantagens e desvantagens. Dessa forma, a metodologia aqui apresentada permitiu que, com a utilização de Mineração de dados seja realizada uma busca eficiente em uma base de dados de uma empresa varejista a procura de padrões que tenham valor para a organização.

1.3 Resultados e impactos esperados

Neste trabalho foi detalhado um processo de MD, primeiramente diferenciando MD do processo de KDD, o que não raro é confundido na literatura. Depois foram apresentadas as abordagens amplas da MD, seguidas da distinção e conceituação das tarefas, métodos e algoritmos, e de que maneira cada elemento desses pode contribuir para o processo de descoberta de padrões.

Através de um estudo de caso foi possível explorar na prática e de maneira plena e efetiva todas as etapas de um processo de MD em uma base de dados de uma empresa varejista. Com o estudo de caso foram elaboradas soluções práticas para situações adversas, tais como limpeza, transformação de dados, discretização de valores contínuos, dentre outras.

Com relação ao estudo de caso, este trabalho cumpriu com o objetivo relacionado à identificação de padrões na base de dados estudada.

Capítulo 2

Processo de KDD

O processo de KDD *(Knowledge Discovery in Databases)* ou ainda, descoberta de conhecimento em bases de dados trata de um processo não trivial, cíclico e estruturado em fases, consistindo de interações entre homem e máquina, onde o objetivo maior é identificar padrões úteis e compreensíveis em dados com o fim de se extrair conhecimentos deles [1] [2].

O processo é cíclico, pois conforme necessário é possível retornar um passo anterior. Esta necessidade dá-se em função de uma melhor análise nos dados ou em uma nova fase de alguma hipótese a ser testada [3]. À medida que se for tendo um melhor entendimento da base de dados em que o trabalho é desenvolvido, novas necessidades podem surgir e o retrocesso no trabalho de KDD é iminente.

O processo é interativo entre homem e máquina. Esta interação ocorre pela necessidade de envolvimento de pessoas que não são necessariamente da área de TI, mas o conhecimento destas para o negócio é essencial para o entendimento de forma concisa da aplicação que se deseja fazer. Estas pessoas interagem com outras da área de TI, que são analistas de dados [4]. Juntas, estas podem levantar os requisitos necessários para a aplicação da mineração de dados. Antes de iniciar alguma tarefa do processo de KDD, é necessário fazer alguns levantamentos como pessoas e áreas envolvidas, inventário de bases e dados disponíveis, existência de DataWarehouses, interesses em relação ao negócio por parte do cliente, avaliar a quantidade dos dados disponíveis e identificar e documentar o conhecimento previamente conhecido.

Bancos de dados no mundo real estão altamente pré-dispostos à armazenagem de dados incoerentes, inconsistentes, com grande quantidade de valores nulos e na maioria dos casos estão armazenados em bases de dados gigantes, com até milhões de registros. O grande desafio do processo de KDD é preparar um banco de dados com esta configuração para o processo de mineração de dados. Isto significa melhorar a qualidade dos dados que lá estão inseridos para

que em conseqüência possa extrair resultados com qualidade. A preparação do BD para este tipo tarefa envolve as fases de limpeza, integração, seleção e transformação dos dados, mineração dos dados, avaliação dos padrões e apresentação e assimilação do conhecimento [3]. A figura 1, extraída de [5] mostra uma representação gráfica de como estas etapas estão dispostas, em ordem, facilitando o entendimento do processo.

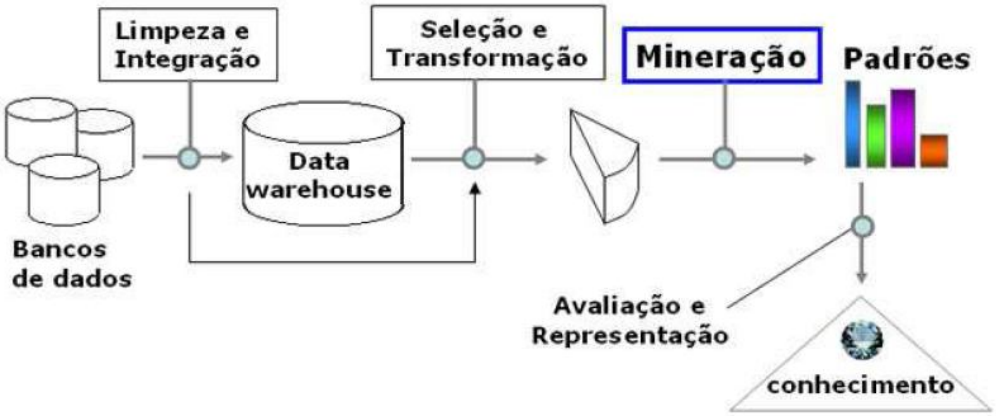

Figura 1. Etapas do processo de KDD

2.1 Limpeza dos dados

Como descrito antes, no mundo real bases de dados estão suscetíveis a erros quando dados são inseridos. Erros como ausência de dados, dados fora dos padrões e inconsistentes são bastante comuns. O trabalho de limpeza dos dados dá-se no preenchimento dos dados ausentes, padronização dos dados fora de padrões e correção dos inconsistentes. Alguns métodos de limpeza podem ser aplicados no início da etapa de KDD, porém alguns podem ser aplicados em etapas posteriores [3] [4].

2.1.1 Valores ausentes

Ao se aplicar mineração de dados pode ser necessário trabalhar com valores que por algum motivo não estão dispostos nas bases e para isto há alguns métodos que propõem soluções para este caso [3]:

1. Ignorar o registro – Técnica pouco funcional, pois apenas na condição da tupla conter vários valores ausentes é que se torna funcional [3].
2. Valor constante para preencher dados – Não muito recomendado. Dependendo da variância do atributo ou de sua importância para o processo pode mascarar resultados [3].
3. Valor médio d mesma classe a qual a tupla pertença – Utilizado quando o valor do atributo é numérico e quando seu significado é passível de atribuição a um valor médio. Um estudo de classificação de um produto pode determinar um valor médio para um determinado valor ausente encontrado na mesma classe [3].
4. Valores estatísticos para preencher dados – Podem-se utilizar técnicas de regressão ou ferramentas de inferências, tais como um formalismo baysiano ou indução por árvores de decisão [3].

Os métodos de um a quatro inferem sobre o dado. Há possibilidade do dado inserido não ser o correto. O método quatro é o mais confiável, pois utiliza de modelos matemáticos considerando mais informações sobre os dados para predizer valores ausentes. Dessa forma a utilização de outros atributos para estimar um valor ausente, dá uma confiabilidade maior na preservação do relacionamento entre o atributo estimado e os utilizados no processo de estimação [3].

2.1.2 Valores fora de padrão

São erros encontrados em atributos que possam sofrer desvios acentuados em seus valores ou cadastrados de forma incorreta, fugindo, assim, do padrão dos outros registros. Atributos do tipo lucro, rendimento ou faturamento, por exemplo, podem estar sujeitos a este tipo de erro. A seguir algumas técnicas são explicadas para correção desses valores [3]:

1. Binning – Este método consiste em ordenar os valores e agrupar eles. Após agrupamento aplicar uma técnica escolha de uma medida para ajuste dos valores em cada grupo, como média aritmética, mediana, valor limite. Após isso substituir os valores pelos calculados em cada grupo [3]:
2. Agrupamento – Valores fora do padrão podem ser detectados quando dados similares são dispostos em clusters ou grupos. Os valores fora

destes grupos podem ser considerados fora do padrão e, assim, excluídos. Porém, para algumas técnicas de mineração valores fora do padrão podem ser interessantes. Se o interesse do cliente é conhecer fraudes em cartões de crédito, por exemplo, os dados fora do padrão irão ser essenciais para detecção da infração [3] [4].

3. Inspeção humana e computador – Às vezes o próprio programador pode fazer algumas medições, identificar os valores e excluí-los [4].
4. Regressão – Dados podem ser ajustados por funções de regressão. Estas funções ajustam duas variáveis, num gráfico, desde que uma possa ser predita por outra [3].

2.1.3 Dados inconsistentes

Há várias formas de ocorrer inconsistências num banco de dados. Algumas delas podem ser retiradas manualmente através de referências externas. Estes erros são mais comuns quando um usuário da entrada incorretamente num dado. Outras formas de retirar dados inconsistentes de um BD é utilizar ferramentas de engenharia de conhecimento. Inconsistências podem ser causadas por integrações de dados. Dados de bancos de dados diferentes, por exemplo, ou ainda de bases diferentes. Há a possibilidade também de ocorrer redundância de dados. Dados com nomes parecidos pertencentes a um mesmo atributo [2].

2.2 Integração de dados

Dados podem ser integrados de várias fontes diversas como: banco de dados, arquivos textos, flat file, etc. Este processo assemelha-se ao de construção de um DW, onde uma nova base de dados será construída de forma consistente, de acordo com os registros reunidos através das outras fontes de dados. Existem três pontos relevantes a serem tratados em relação a esta etapa [3]:

1. Integração de sistemas internos – No mundo real pode acontecer o caso de entidades com mesmo valor semântico estarem incluídas em diversos esquemas com nomes e atributos diferentes. Isto trata de um típico caso

de identificação de entidades. Em bases operacionais e em datawarehouses este problema é minimizado por conterem metadados, que ajudam a evitar este problema [3].

2. Dados redundantes - Um atributo pode ser redundante se o mesmo for derivado de outra tabela. Inconsistências em atributos ou dimensões podem ser causa de redundância em conjuntos de dados. Uma forma de tratar este problema é a utilização de análise de correlação. Esta técnica consiste em verificar o quanto dois atributos são correlatos. Outra forma de se ter atributos redundantes é tendo duas tuplas idênticas cadastradas na base de dados [3].
3. Detecção e resolução de valores conflitantes - Um enorme desafio na integração dos dados está na diferença de valores que os dados podem apresentar nas diversas fontes de dados que os mesmos provêm. Em bases de dados do mundo real, dados de muitas bases diferentes, de várias tabelas podem diferir em seus valores dependendo de que região se encontre. Por exemplo, determinados produtos podem ter seus preços variados, por conta de taxa de imposto atribuído sobre ele, de acordo com a cidade, o estado ou país em que se encontra a base [3].

Há outros fatores que inferem na redundância e na inconsistência dos dados, dificultando ainda mais o processo de integração que são formatos de armazenamento em banco de dados relacionais, arquivos de texto, campos fixos e variáveis. Estes formatos dificultam a recuperação dos dados para integração [3].

Se cuidados com as formatações dos dados e verificações de dados redundantes e inconsistentes forem tomados a integração dos dados poderá ser realizada de forma muito mais agradável, dando consistência ao processo e maior agilidade nos processos seguintes [3].

2.3 Seleção dos dados

Nesta etapa é interessante a participação de pessoas ligadas ao negócio em que as tarefas de mineração de dados serão aplicadas [3]. O método exige que a essa

altura do processo de KDD sejam escolhidos os atributos relevantes à tarefa de MD selecionada [4]. Outra forma de descrever a seleção dos dados é que se trata de responder às perguntas que o cliente quer saber, em forma de atributos selecionados e métodos de MD aplicados. Por exemplo, o cliente quer saber se um determinado produto sai em seqüência de algum outro. Selecionam-se então os atributos relacionados com as transações comerciais do estabelecimento e aplica-se uma técnica de regra de seqüência, que define que diante da saída de um determinado produto, algum tempo depois outro produto, associado a este, será vendido também. Um exemplo desta técnica seria compras de rádios automotivos. Algum tempo após a venda de um rádio, autos-falantes podem ser vendidos também. Isso pode ajudar na estratégia da empresa a combinar a venda destes produtos. Esta técnica bem como outras serão melhores abordadas no capítulo dedicado à mineração de dados.

2.4 Transformação dos dados

Além da importância da integração dos dados, outra etapa bastante relevante é a de transformação dos dados. Nesta etapa é enriquecedor alterar algumas formas dos dados ou dos atributos mudando sua semântica e adaptando sua nova forma à aplicação que irá realizar a mineração de dados. As principais regras de transformação são [4]:

1. Agregação – Agrega e sumariza os dados [3]. Por exemplo, numa tabela há atributos de vendas diárias. Estes atributos agora serão sumarizados e agregados em vendas semanais, mensais e anuais.
2. Aplainamento – Retira dados ruidosos. Utilizam técnicas de agrupamento, bining e regressão [3].
3. Generalização – Dados podem ser alterados de forma a constituir em um novo contexto mais abstrato. Por exemplo, atributo idade pode agora ser alterado para faixa etária [4].
4. Construção de atributos – Atributos novos são construídos de acordo com informações existentes [4].
5. Redução de dados – Dividido em:

a. Agregações,
b. Redução dimensional – Eliminação de atributos irrelevantes à técnica a ser utilizada.
c. Compressão dos dados – Utilização de alguma forma de codificação para reduzir o conjunto de dados.
d. Redução numérica – instâncias, por exemplo.

2.5 Mineração de Dados

Esta é a etapa mais importante de todo o processo de KDD, merecendo um capítulo especial dedicado a sua fundamentação, suas técnicas e funcionalidades. Será descrita em maiores detalhes no capítulo 3.

2.6 Avaliação dos padrões

Após a técnica de MD ser aplicada, resultados serão mostrados e cabe ao analista de dados verificar se os mesmos são compatíveis com os objetivos inicialmente planejados. Os padrões obtidos com MD podem, ou não ser bons para o estudo em questão. Nesta etapa devem-se analisar precisamente quais padrões podem ser utilizados. As técnicas de MD podem, por exemplo, não determinar padrão algum. As bases do mundo real estão suscetíveis a esta problemática. Porém o processo só terá validade se as analises forem feitas e os padrões utilizados mediante sua expressividade estatística [3].

2.7 Apresentação e assimilação do conhecimento

Consiste basicamente nas seguintes etapas:

1. Apresentar as descobertas obtidas.

2. Determinar a melhor forma de utilizar tais informações na tomada de decisão
3. Definir as vantagens e desvantagens do projeto
4. Reavaliar o projeto
5. Criar novos projetos

2.8 Considerações Finais

Após toda apresentação do processo de KDD, é possível fazer algumas reflexões a respeito do mesmo. No mundo real, podemos perceber que bases de dados, em muitos casos, são inconsistentes, redundantes e em muitos casos não normalizadas. Trabalhar para acabar com estes problemas é uma tarefa bastante árdua. Além disso, como as bases são construídas em sua plenitude para prover subsídios lógicos e físicos para aplicações a que lhe foi destinada, elas não foram feitas nem pensadas para aplicações de MD. Logo, em quase todas as bases em que um cliente deseje realizar uma determinada tarefa de MD, será necessário a seqüência de KDD desde o processo inicial, pois é a partir das técnicas de limpeza, integração, seleção e transformação dos dados que a mineração de dados será concluída com sucesso, ou não. Quanto mais precisa e consistente for a base, melhores serão os resultados obtidos.

Capítulo 3

Mineração de Dados

Neste capítulo será explicado de forma mais detalhada a etapa de mineração de dados no processo de KDD. Por ser uma fase onde serão escolhidas técnicas para extração de conhecimento e, sendo assim, tornando-se a mais importante, julgou-se necessário uma explicação com maior granularidade de suas tarefas e técnicas. Por ser um trabalho de tentar identificar padrões compreensivos que possam ajudar uma empresa na gestão de vendas e pela configuração da base de dados ser limitada, será dado um enfoque maior nas tarefas análise de regras de associação.

Como destacamos anteriormente, a mineração de dados é um processo de exploração de grandes bases de dados com o propósito de extrair informações que possam ser transformadas em conhecimento útil para uma determinada corporação, seja ela comercial ou científica, para que a mesma tenha capacidade de fazer novas atividades baseadas nas informações obtidas [3].

A mineração de dados é classificada como uma mistura de áreas da matemática, inteligência computacional e banco de dados, incluindo estatística, Inteligência artificial e banco de dados. Inicialmente pesquisadores da área estatística não deram muito crédito à área, porém, como sua aplicação prática foi bastante aceita entre organizações comerciais e científicas das outras áreas envolvidas, o campo tem tido um crescimento acentuado e diversos congressos científicos e ferramentas comerciais surgiram fazendo com que sua importância aumentasse [4].

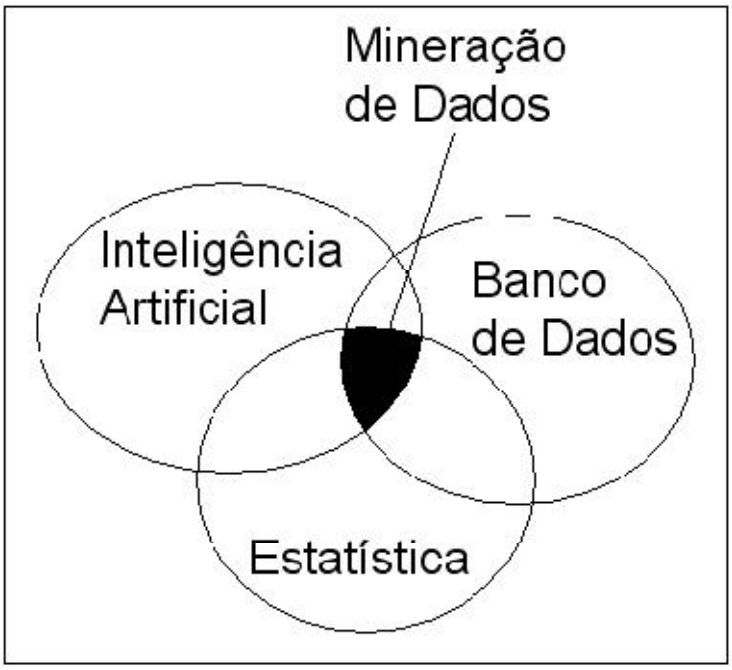

Figura 2. Mineração de Dados é uma interseção de recursos de várias áreas [4]

O objetivo maior em extrair conhecimento de bases de dados é encontrar e descrever padrões significativos em registros como ferramenta para ajudar a explicar determinados fenômenos nos dados e fazer previsões nestes. Os dados devem revestir a forma de um conjunto de exemplos – Exemplos de clientes que trocaram de preferência de alguma loja ou situações em que possa descrever algum tipo de produto. As saídas têm a forma de previsões sobre novos exemplos – Prever se um cliente vai mudar de comportamento em relação à sua lealdade com a loja, por exemplo, ou prever que tipo de produto irá ser mais propício à venda em determinadas circunstâncias. O desempenho também é bastante útil para fornecer uma representação explícita do conhecimento que é adquirido, refletindo duas condições de aprendizagem importantes: a aquisição de conhecimento e a capacidade de usá-lo [2].

Muitas técnicas de aprendizagem procuram por descrições compreensíveis do que foi aprendido. Descrições estas que podem se tornar bastante complexas e geralmente são expressas num conjunto de regras. Como estas descrições podem ser compreendidas por pessoas, elas servem para explicar o que foi aprendido e explicar uma base para novas previsões. Experiências mostram que em aplicações

de aprendizado de máquina para MD, estruturas de conhecimento explícito que é adquirido, descrições compreensíveis, são tão importantes, e às vezes muito mais importantes, que a capacidade de executar bem em novos exemplos [2].

3.1 Tarefas e Técnicas de MD

Em primeiro lugar, é essencial diferir tarefas de MD, de técnicas de MD. Tarefas de mineração de dados estão relacionadas às perguntas feitas na etapa de seleção dos dados, no processo de KDD, descrita no sub-tópico 2.3 do capítulo 2, ou seja, nesta fase a preocupação é relacionar as perguntas feitas com o que se tem em mãos e, assim, determinar, pela tarefa a ser realizada, que técnica usar. Por exemplo, um gasto exagerado de cartão de crédito de um cliente, fora do padrão dos seus gastos usuais. Neste exemplo a tarefa associada a esse argumento será análise de desvios e as técnicas para este tipo de tarefa podem ser árvores de decisão ou redes neurais [5] [6]. Deste modo a técnica de MD trata da especificação de métodos que serão executados para descoberta do conhecimento a que se tem interesse.

O próximo passo agora é entender como saber que técnica escolher dada uma determinada tarefa. Como já foi dito, a tarefa depende do objetivo a ser alcançado. Conhecendo estes, é possível, através de alguns aspectos, que chamaremos de inibidores e motivadores, chegar a uma conclusão sobre qual técnica aplicar ao problema em questão. A tabela 1, extraída de [7], mostra algumas técnicas e seus aspectos inibidores e motivadores que ajudam na escolha.

Tabela 1. Aspectos inibidores e motivadores para técnicas de mineração de dados

Técnica	Aspectos Motivadores	Aspectos Inibidores
Redes Neurais	• versatilidade • bons resultados em domínios complicados • aplicam-se a dados categóricos e numéricos	• conversão dos dados de entrada e saída • resultados não explicativos • podem convergir a uma solução inferior a esperada
Árvores de Decisão	• geram resultados compreensíveis • executam uma classificação sem muito processamento • aplicam-se a dados categóricos e numéricos • indicam campos mais importantes para predição/classificação	• muitas classes ocasionam propensão a erro • processamento intensivo • problemas com regiões não-retangulares
Regras Associativas	• geram resultados compreensíveis • trabalham com dados de maneira uniforme	• dificuldade de uso de SQL • escolha difícil de um bom suporte e da confiança • número possível de regras • necessidade do uso de dados intervalares
Algoritmos Genéticos	• geram resultados compreensíveis • aplicação imediata dos resultados • aplicam-se a vários tipos de dados • aplicáveis para otimização • integram-se bem com redes neurais	• dificuldade em codificar muitos problemas • não há garantia de otimização • processamento intensivo
Descoberta Automática de Agrupamento	• suporte a mineração de dados não dirigida • aplica-se a dados categóricos, numéricos e textuais • facilidade de aplicação da técnica	• dificuldade de definir os parâmetros iniciais • pode ser difícil interpretar os agrupamentos resultantes

No próximo sub-tópico serão explicadas as principais tarefas e algumas de suas técnicas que podem ser aplicadas.

3.1.1 Classificação e Predição

Segundo [5], Classificação é processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. Este modelo é baseado na análise prévia de um conjunto de dados de amostragem ou dados de treinamento, contendo objetos corretamente classificados. Como exemplo suponha uma empresa credora que deseja do risco

a prover crédito a seus solicitantes. A empresa pode determinar uma regra: clientes da faixa econômica C, com idade superior a 50 representam um alto risco para a corporação. Já clientes da faixa econômica B, com idade entre 40 e 55 anos representam um risco médio. Clientes da faixa econômica B, com idade entre 30 e 50 anos representam um risco baixo. A predição é um processo de inferir um comportamento futuro baseado em várias informações. Por exemplo, baseado em informações sobre a formação acadêmica de uma determinada pessoa, associada com informações sobre seu trabalho atual e a área de atividade, pode-se predizer qual será seu salário em um determinado espaço de tempo [3]. As técnicas aplicadas para estas tarefas podem ser árvores de decisão e redes neurais.

3.1.2 Análise de *Clusters*

Nesta tarefa o objetivo é identificar classes de objetos ainda não classificados. Diferentemente do processo de classificação e predição, onde as classes dos objetos são conhecidas. O método consiste em identificar agrupamentos de objetos e estes agrupamentos identificarem uma classe [5]. Por exemplo, podemos considerar um censo nacional para formar grupos de domicílios, utilizando atributos como escolaridade, número de filhos, faixa etária, profissão e sexo. Observa-se que não existem classes definidas e poderemos ter grupos domiciliares idênticos numa região geográfica diferente, porém, com valores dos atributos diferentes [3]. Uma boa técnica para esta tarefa é clusterização [6].

A figura 3, extraída de [3] mostra claramente uma visualização de possíveis agrupamentos dado um conjunto de dados.

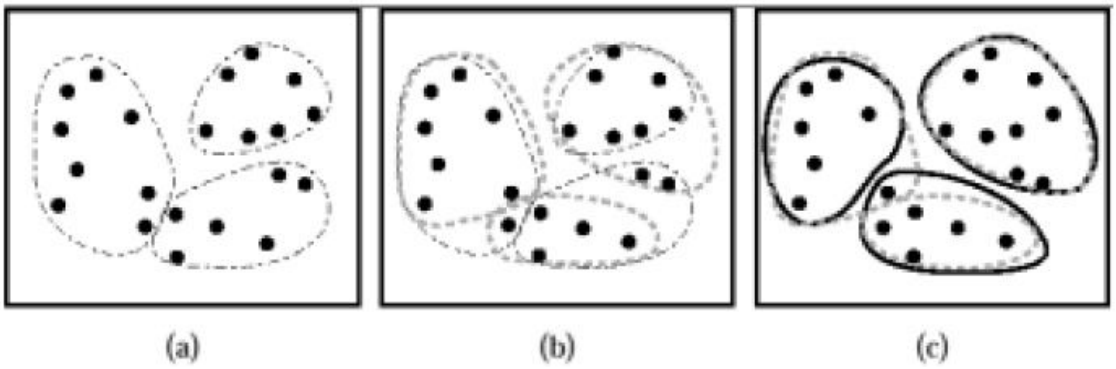

Figura 3. Possíveis agrupamentos, dado um conjunto de dados.

3.1.3 Análise de desvios

Desvios em dados são valores não condizentes com o padrão dos outros que estão na mesma categoria em um banco de dados. A análise de desvios consiste em identificar estes valores para auxílio na tomada de decisão. A tomada de decisão mais comum é excluí-lo da base, já que ele é considerado discrepante em relação aos outros, sendo tratado, assim, como uma exceção. Porém, para algumas aplicações essas exceções são bastante interessantes. Por exemplo, podemos citar uma empresa de cartões de crédito que deseja descobrir fraudes no uso dos seus cartões por uso dos seus clientes. Valores discrepantes podem dar uma boa dica para identificar essas infrações [5]. As técnicas utilizadas para isso podem ser, por exemplo, Árvores de decisão ou redes neurais [6].

3.1.4 Análise de regras de associação

Segundo [8], a análise de regras de associação é o estudo de atributos ou características que ocorrem em conjunto. Também conhecida como Análise de Cesta de Compras, a descoberta de associações tem como objetivo identificar itens que ocorram de forma simultânea em transações de banco de dados. A descoberta de itens “casados” pode ser bastante útil para a estratégia de uma empresa interessada em impulsionar as vendas induzindo seus clientes a comprarem mais tais itens [5].

3.1.4.1 Técnica Apriori

Formular determinadas perguntas sobre determinados produtos em relação às vendas dos mesmos pode ser uma estratégia interessante para o departamento de marketing de uma determinada empresa. Perguntas como: quais produtos saem juntos numa mesma transação e quantas vezes isso acontece em um determinado tempo? As respostas a essas indagações podem revelar amostras bastante agradáveis para este setor. Como resultado disso, a empresa pode planejar melhor os cartazes de promoções dos produtos, fazer combinações de produtos em uma prateleira, dispondo eles de forma mais acessível aos clientes, para que desta forma eles sintam-se mais à vontade para comprá-los [5]. Esta estratégia chama-se cesta de compras e a figura 4, extraída de [3] mostra uma ilustração da relação entre empresário e produto [5].

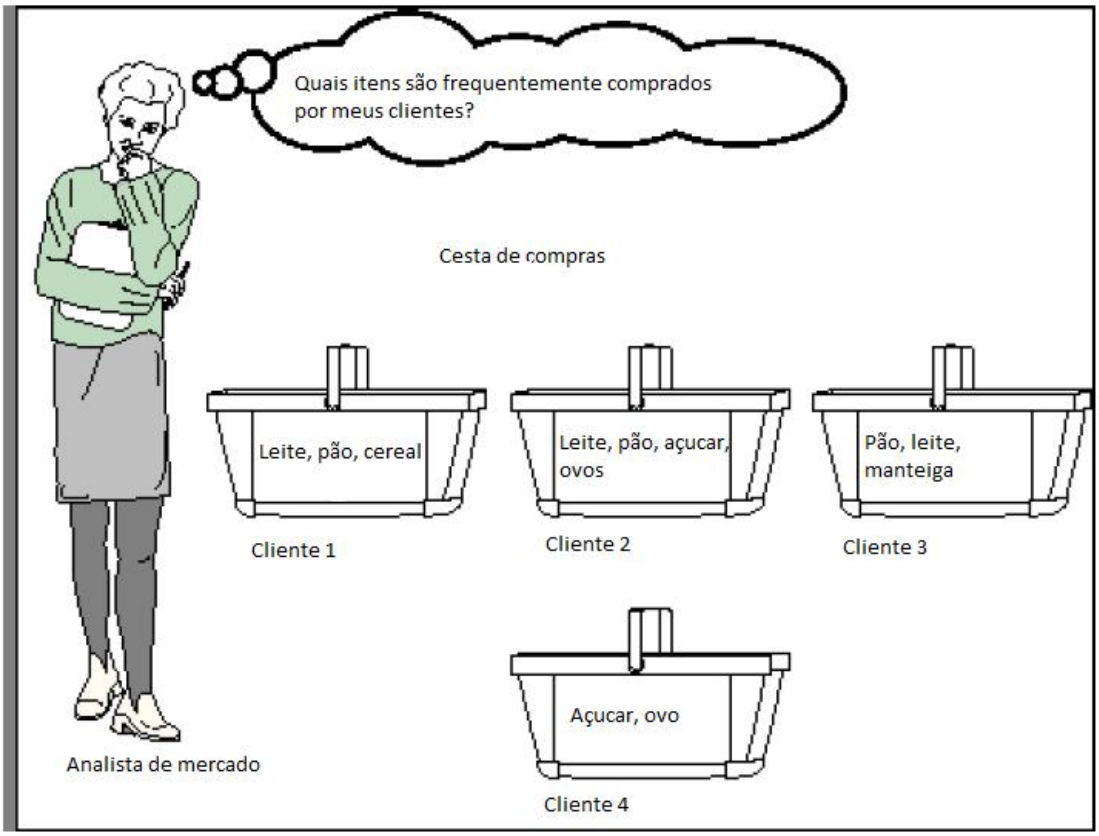

Figura 4. Cesta de compras.

Supondo um caso prático, podemos ilustrar uma pequena base de dados de produtos de uma empresa fictícia. Nesta base a sua configuração está disposta da seguinte forma: atributo Artigo e atributo Número que o representa. A figura 5, extraída de [5] mostra uma associação desses artigos com seus respectivos números [5].

Artigo (item)	número que o representa
Pão	1
Leite	2
Açúcar	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Figura 5. Representação de cada artigo da empresa fictícia.

A cada grupo de itens que é vendido pela empresa dá-se o nome de *itemset.* Um *itemset* com k elementos é chamado de *k-itemset.* A partir desta definição pode-se pensar em regras para decidir se um determinado produto pode ser chamado de freqüente. Por exemplo, um gerente pode decidir que se um *itemset* sai em pelo menos 50% das transações comerciais, este produto é freqüente. A figura 6, extraída de [5] mostra esses *itemsets* em transações comerciais. Usando o exemplo da figura 6, o *itemset* {1, 3} pode ser considerado freqüente, pois aparece em quatro, das seis transações. A porcentagem que representa a quantidade de aparições de um itemset numa transação é chamada de *suporte* [5].

TID	Itens comprados
101	{1,3,5}
102	{2,1,3,7,5}
103	{4,9,2,1}
104	{5,2,1,3,9}
105	{1,8,6,4,3,5}
106	{9,2,8}

Figura 6. Banco de transações.

O atributo TID identifica o número da transação.

Uma regra de associação é uma expressão A -> B, onde A e B são itemsets. Um exemplo seria {pão, leite} -> {café}. Isto é uma regra de associação. O significado desta transação é que clientes que compram pão e leite tendem a comprar café também. Porém, a recíproca desta relação não necessariamente é verdadeira [5].

A toda associação A -> B, determinamos um grau de confiança, denotado por conF(A->B). Esta confiança significa a porcentagem das transações que suportam B dentre todas as transações que suportam A, ou seja:

conF(A->B) = Número de transações que suportam (AUB)/ Número de transações que suportam (A) [5]

Por exemplo, o grau de confiança da regra {cerveja} -> {manteiga}, com relação ao banco de dados da figura 6 é um, ou seja, 100%.

Este resultado é bem impactante e interessante. Podemos refletir alguns aspectos sobre isso. Primeiro, seria mesmo uma regra “boa” considerar um grau de confiança alto? Se percebermos, o numero de transações em que cerveja e manteiga saíram juntas, é um. Porém, em todas as vezes que cerveja saiu, manteiga saiu junto. O bom senso nos diz que considerar apenas o grau de confiança, mesmo que alto, em situações assim, torna a regra pouco confiável, logo “ruim”. Então, para conseguirmos uma

regra de associação boa, ou interessante, uma boa sugestão seria considerar o grau de confiança e o suporte e ambos serem relativamente altos [5].

A toda regra de associação, associamos um suporte, definido por sup(A->B), sendo o suporte do itemset (AUB). Por exemplo, se calcularmos o suporte da regra {cerveja} -> {manteiga} no banco de dados da figura 6, o resultado é 0.1666%. [5]

Uma regra de associação r interessante seria se conF(r) >= a e sup(r) >= b, onde a e b são, respectivamente um grau mínimo de confiança e um grau mínimo de suporte, definidos pelo usuário [5]. Para este caso, a = 0.8 e b = 0.1 torna a regra interessante [5].

O problema da mineração de regras de associação em banco de dados é o seguinte: dadas as transações, um mínimo de confiança a e um mínimo de suporte b, encontrar todas as regras de associação interessantes com relação às transações, a e b. A principal técnica para identificar regras de associações é o algoritmo apriori [5].

3.1.4.2 Algoritmo Apriori

Trata-se de um algoritmo que resolve o problema da mineração de itemsets freqüentes, isto é, dados um banco de transações *D* e um nível mpinimo de suporte *B*, o algoritmo encontra todos os itemsets freqüentes com relação a *D* e *B* [5].

O algoritmo é composto por três fases: geração, poda e validação. As duas primeiras fases são realizadas na memória principal e não necessitam que o banco de dados seja utilizado. A memória secundária só é exigida quando o conjunto de itemsets é grande demais. Apenas na terceira fase, cálculo de suporte dos candidatos, é que o banco de dados é utilizado [5].

Fase de geração dos candidatos – Nesta fase são gerados os itemsets cadidatos (não necessariamente frequentes) de tamanho *k* a partir de um conjunto L_{k-1}. A intenção é gerar somente itemsets que tenham alguma chance de serem freqüentes. Como o algoritmo Apriori é feito de forma

interativa, ou seja, os itemsets freqüentes de tamanho *k* são calculados a partir de itemsets freqüentes de tamanho *k-1*, que já foram calculados num passo anterior. Diante disso, todos os itemsets de tamanho *k* - 1 contidos nos candidatos de tamanho *k* deverão ser freqüentes, portanto, deverão pertencer ao conjunto L_{k-1}. Assim, o conjunto C'_k de itemsets candidatos de tamanho *k* é construído juntando-se pares de itemsets de tamanho *k* - 1 que tenham *k* - 2 elementos comum. Desta maneira pode-se ter certeza de obter um itemset de tamanho *k* onde pelo menos dois de seus subconjuntos de tamanho *k* - 1 são frequentes.

Um exemplo prático pode facilitar o entendimento da construção de um itemset candidato. Considerando o banco de transações ilustrado na figura 6 e supondo que num passo dois de uma iteração, pôde ser obtido o seguinte conjunto de itemsets freqüentes de tamanho 2:

$L_2 = \{\{1,3\}, \{1,5\}, \{1,4\}, \{2,3\}, \{3,4\}, \{2,4\}\}$

Então o conjunto dos pré candidatos C'_3 da iteração seguinte será:

$C'_3 = \{\{1,3,5\}, \{1,3,4\}, \{1,4,5\}, \{2,3,4\}\}$

Fase de poda dos cadidatos – Sabe-se que se um itemset C'_k possuir um subconjunto de itens de tamanho *k-1* que não estiver em L_{k-1} ele poderá ser descartado, pois não terá a menor chance de ser freqüente. Assim, nesta fase é calculado o conjunto $C_k = C'_k$ – {conjunto de itemsets podados}.

Por exemplo, considerando a situação apresentada anteriormente. Neste caso, $C_3 = C'_3$ – {{1,4,5}, {1,3,5}} = {{1,3,4}, {2,3,4}}. O itemset {1,4,5} foi podado pois não terá a menor chance de ser freqüente: ele contém o 2-itemset {4,5} que não é freqüente, pois não aparece em L_2 . O mesmo vale para o itemset {1,3,5}. O 2-itemset {3,5} não parece em L_2.

Fase do cálculo do suporte – Nesta fase é calculado o suporte de cada um dos itemsets do conjunto C_k. Isto pode ser feito varrendo-se uma única vez o banco de dados. Para cada transação do banco de dados, verifica-se quais são os candidatos suportados e para estes candidatos incrementa-se de uma unidade o contador do suporte.

Os itemsets de tamanho 1 são computados considerando-se todos os conjuntos unitários possíveis, de um único item. Em seguida, varre-se uma vez o banco de dados para calcular o suporte de cada um destes conjuntos unitários, eliminando-se aqueles que não possuem suporte superior ou igual ao mínimo exigido pelo usuário.

3.2 Abordagens de Mineração de Dados

As abordagens descrevem a maneira como o usuário vai preparar o ambiente para a obtenção das tarefas e técnicas de MD. Existem, basicamente, duas abordagens a serem feitas: top-down e bottom-up. Na primeira, o usuário parte do princípio de que já tem uma idéia do que buscar. Não por poucas vezes, ele acerta saber o que tem em mãos, e apenas quer refutar ou confirmar. Na segunda hipótese, também chamada de busca de conhecimento, um processo de exploração nos dados é iniciado, a fim de se descobrir algo ainda desconhecido [3].

Nas duas hipóteses, o usuário pode optar por uma abordagem de forma supervisionada ou não-supervisionada. Na primeira, é necessário um ambiente a ser montado para aplicação da técnica a ser usada, para que se possa utilizá-lo em novas amostras de dados [7]. Na segunda não existe essa necessidade, pois não há necessidade de treinamento. Logo, também não será necessário nenhum ambiente programado [3].

3.3 Considerações finais

Conforme descrito, para utilização do processo de mineração de dados é preciso deixar bem claro qual a funcionalidade e quais resultados se deseja chegar. Para escolha da funcionalidade, pode ser necessária a participação de um profissional que entenda muito bem do negócio em que as técnicas de MD estão sendo utilizadas. Este usuário não precisamente pertence a um grupo ligado a TI, mas sim, pertencente a um grupo que entenda bem da missão da empresa para que possa analisar e interpretar os resultados corretamente. Diversas técnicas podem ser

utilizadas para obtenção dos resultados, porém cada uma tem suas peculiaridades e cabe ao analista de negócio e de dados escolherem a mais adequada. Indicadores inibidores e motivadores, mostrados anteriormente, ajudam neste processo de escolha.

No próximo capítulo será descrito um estudo de caso, no qual o processo de KDD foi aplicado. Na fase de mineração de dados será discutida a funcionalidade escolhida, bem como os resultados a serem obtidos. Durante esta fase também será descrita a tarefa de MD utilizada e a técnica apropriada para tal.

Capítulo 4

Estudo de caso

Neste capítulo será descrito o processo de KDD na prática, durante a tentativa de extração de conhecimento de uma base de dados de uma empresa varejista. Todo o processo foi baseado nos princípios apresentados até então, da limpeza dos dados à apresentação dos resultados. Na etapa de mineração de dados foi utilizada a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) para utilização das técnicas e apresentação dos resultados.

O projeto é tentar extrair padrões interessantes, utilizando técnicas de mineração de dados em uma base de dados de uma empresa do ramo varejista. Padrões estes que auxiliem a gestão de vendas da empresa. Diante disso, foram traçadas algumas tarefas possíveis a serem executadas. Algumas tarefas foram levantadas, como:

- Conhecer perfis de clientes que freqüentam e compram na loja.
- Conhecer clientes que compraram na loja pela primeira vez.
- Tentar conhecer os produtos que estão associados em transações, para uma análise de cesta de mercados.

4.1 Aquisição da base de dados

O ponto inicial do projeto é a aquisição da base de dados. A iniciativa partiu do autor deste trabalho de conclusão e não da corporação em questão. Foram feitas reuniões com a célula de banco de dados e com a gerência da empresa que presta serviço à corporação que teve sua base analisada.

Após a escolha da base de dados, foi dado início ao processo de backup do banco para que o mesmo pudesse ser portado para outra máquina, para que pudesse ser construída uma base de teste e interagir a mesma com a ferramenta de extração de conhecimento.

4.1.1 Dificuldades apresentadas

Foram feitas algumas tentativas de backup's sem sucesso. Por a base de dados ser bastante grande, 20GB, foi necessário uma análise dos objetos necessários à aplicação das tarefas levantadas antes do processo de bakcup.

4.2 Caracterização da tarefa de MD

Nesta etapa foi feito estudo para identificação do problema alvo que é extrair informações na base que auxiliem a gestão de vendas numa empresa varejista, diante das tarefas já levantadas. Durante este processo foi feito um levantamento dos objetos que tivessem relação com a área de estudo em questão.

Os objetos levantados revelaram um problema crítico a duas tarefas: os dados dos clientes. A empresa não tinha dado algum sobre os clientes que compravam em suas lojas. Logo, o escopo do projeto foi reduzido às tarefas de associações de produtos em transações de vendas.

4.2.1 Dificuldades apresentadas

A fase de levantamento dos objetos relacionados às tarefas de MD aplicadas foi bastante problemática. A começar por não existir um modelo de dados que mostrasse as relações entre suas 746 tabelas, dificultando enormemente o processo de entendimento do negócio por parte do analista de dados.

Com ajuda da célula de banco de dados da empresa concedente da base de dados, foi disponibilizado um mini-modelo mostrando 61 tabelas e suas dependências. Dessas, pode-se perceber algumas tabelas que se relacionavam com vendas e transações. A identificação das outras tabelas foi mediante ajuda intensa da célula de banco de dados.

4.3 Criação da base de testes

Após a caracterização das tarefas a serem aplicadas durante o processo de mineração de dados, iniciou-se o processo de montagem de uma base de testes, para preparar a base para o processo de KDD, preservando, assim, a base original.

Como o escopo do projeto foi reduzido, por conta da falta de informações referentes aos clientes das lojas, nesta etapa foram consideradas apenas tabelas de transações e produtos. Dessa forma o ambiente de simulação ficou pronto para o processo de KDD.

4.4 Processo de KDD

Depois de feito todo o levantamento inicial, a caracterização das tarefas e a criação da base de testes, o processo de KDD pôde, enfim, ser implementado.

4.4.1 Limpeza dos dados

Este processo foi revisto por todas as fases descritas na fundamentação teórica, Capítulo 2, como a segue:

4.4.1.1 Valores ausentes

Nesta etapa foi verificada que em tabelas de transações de vendas, como se tratam de produtos já existentes, inseridos na tabela via aplicação, foi observado que tais valores não existiam.

4.4.1.2 Valores fora de padrão

Para a tarefa de associação de produtos, não foram considerados valores preço, ou qualquer outro tipo de valor que pudesse estar fora dos padrões, fazendo com que esta regra fosse descartada da análise.

4.4.1.3 Dados inconsistentes

Esta é uma etapa que por ser bastante crítica, merece uma atenção especial. Durante esta fase foram identificadas inconsistências de bastante impacto na análise: códigos diferentes para o mesmo produto, nomes de produtos diferentes referenciando o mesmo produto, falha da modelagem da tabela de produto.

Como uma transação utiliza a tabela de produtos para colocar estes dados na tabela de transação, julgou-se necessário um tratamento na tabela de transação com relação a estas inconsistências. Todas as inconsistências foram tratadas via programação SQL, utilizando o SGBD Oracle 10g.

Para os códigos diferentes para o mesmo produto, foram contados quantos produtos de mesmo nome se repetiam nas transações, com códigos diferentes. Consultando a tabela de produtos, foi identificado o código que referenciava o produto nas transações era um tipo de código que muda de acordo com o tempo. Logo, havia vários produtos iguais com códigos diferentes. A correção foi feita através de atualização de todos os produtos nesta situação pelo código do produto extraído da tabela de produtos.

Para os nomes de produtos diferentes referenciando o mesmo produto foi necessária uma análise profunda no significado dos produtos presentes nas transações. Em algumas transações produtos como “Camiseta Hering” era encontrado. Em outras, “Camiseta Hrng”. Tratar inconsistências deste tipo geralmente custa muito tempo ou exige-se a utilização de ferramentas de software. Como todas as ferramentas encontradas em pesquisas eram pagas, foi decidido utilizar, mais uma vez, programação SQL para sanar este problema.

Esta solução foi trabalhada em conjunto para sanar este problema e o da falha de modelagem nos produtos. A falha consistia em não haver categorias para os produtos, fazendo com que estas fossem cadastradas juntamente com o nome dos produtos no campo “Descricao_Produto”.

Diante da grande quantidade de produtos, 6859 peças, a solução para esta problemática foi trabalhar com os nomes genéricos, pois não

haveria tempo hábil para fazer todas as verificações exatas produto a produto. Dessa forma, “Camiseta Hering” ou “Camiseta Hrng” agora, chama-se “Camiseta”. Deixando a análise menos rígida, mas ainda atendendo aos propósitos do trabalho. Todos os produtos com esta falha foram atualizados da mesma forma com a qual o exemplo de “Camiseta” mostra.

4.4.2 Integração dos dados

Como o tratamento foi feito em uma tabela de transação, os dados de outras lojas e outros SGBD’s ou fontes diversas foram integrados via aplicação, eliminando, assim a possibilidade de haver problemas com integração. Com relação a dados conflitantes, não foi considerado, no trabalho, valores que possam inferir sobre a heterogeneidade semântica dos dados, como salário, taxas de impostos, etc.

4.4.3 Seleção dos dados

Tendo os dados limpos, o próximo passo foi a seleção dos dados que foram necessários para o processo de mineração de dados. O processo de associação de produtos em transações leva em conta apenas o número da transação e os produtos nela compostos. Diante disso dados como valor da venda, código do produto, taxas, e outras informações características com finalidade de informação sobre a transação foram excluídos.

Como a empresa analisada funciona com varejo variado, há muitas sessões de produtos e, conseqüentemente, muitos produtos. Foram contabilizados 6.859 produtos. Destes, observou-se que o os produtos que vendem muito, considera-se mais que quinhentos itens vendidos, formam um conjunto de aproximadamente cem itens. Foram listados, então, os produtos para se trabalhar nas associações. Foi observado na consulta, que quase todos os cem tipos de produtos são das classes vestimenta ou cama, mesa e banho. Este foi o grupo de produtos que o trabalho focou.

O raciocínio para isso é que se alguns produtos saem muito, a probabilidade deles estarem associados é maior que um produto que sai pouco.

Diante dessa premissa e da constatação dos produtos mais vendidos serem de vestimenta ou cama, mesa e banho, foi iniciado o trabalho de tentar associações entre eles. A figura 7 mostra a disposição dos produtos mais vendidos *versus* a quantidade de produtos.

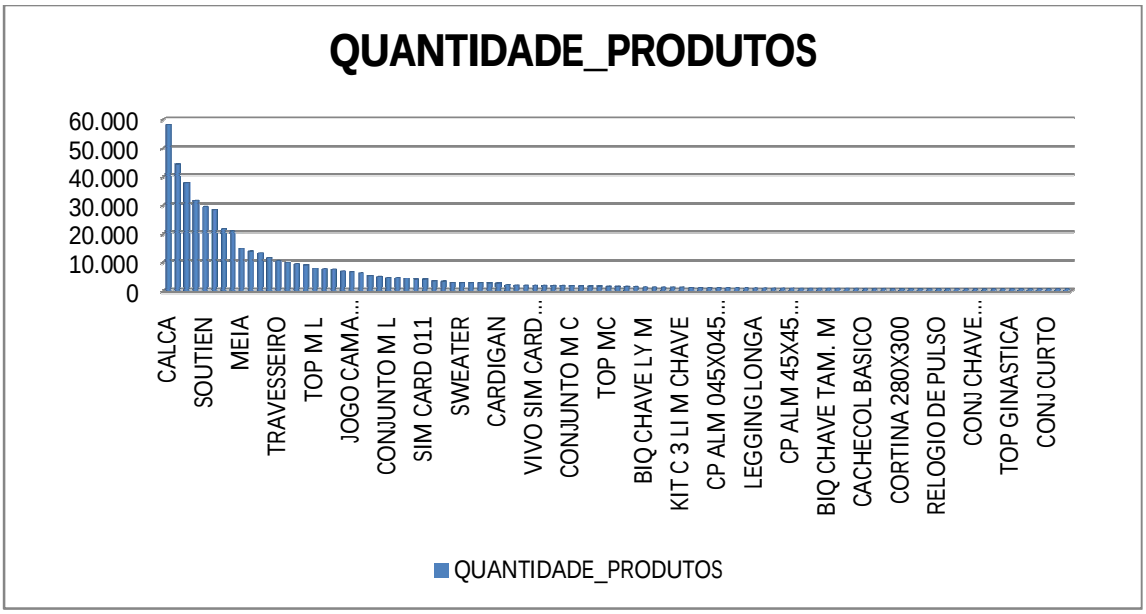

Figura 7. Alguns dos produtos mais vendidos

Também nessa fase foi determinado o período de análise dos dados. O período foi de 01/02/2007 à 01/06/2207. A razão para isto foi que a empresa que disponibilizou os dados liberou apenas por este período. Durante este período foram contabilizadas 128.197 transações.

4.4.4 Transformação dos dados

Após selecionar os dados que foram utilizados nas práticas de mineração de dados, foi necessário fazer alterações nas estruturas das tabelas, com os dados selecionados, e nos tipos dos dados.

Algumas etapas do processo de transformação dos dados foram tratadas nas etapas anteriores do processo de KDD, como generalização, por exemplo. A normalização foi necessária, pois a ferramenta de mineração de dados lê os dados de uma forma diferente ao que está cadastrado no banco de dados.

Então foi preciso determinar valores em caractere, varchar, valorados em “s” ou “n”, para identificar se um determinado produto saiu ou não numa transação. Houve necessidade também de se criar novos atributos. Na base de dados da empresa uma transação é identificada unicamente por quatro atributos, a seguir: “numero_loja”, refere-se ao número que identifica uma loja da empresa; “cod_componente”, refere-se ao código do PDV que executou a venda; “data_transacao”, refere-se à data em que a transação foi realizada; “nsu”, identifica um cupom de venda. Uma transação completa pode ter vários registros, como mostra a figura 8.

NUMERO_LOJA	COD_COMPONENTE	DATA_TRANSACAO	NSU	DESC_COMPLETA
0052	7	01022007	83	LENCOL SOLTEIRO
0052	7	01022007	83	JOGO CAMA SOLTEI
0052	7	01022007	83	JOGO CAMA SOLTEI
0052	7	01022007	83	FRONHA
0052	7	01022007	83	FRONHA
0012	11	01022007	57	BIQUINI
0052	7	01022007	85	TOALHA BANHO
0012	2	01022007	81	SAIA
0012	2	01022007	81	CAMISETA
0012	2	01022007	81	CAMISETA

Figura 8. Transações de venda na base da empresa varejista

No caso da figura 9, o número de loja de valor 52, o código de componente de valor 7, a data de transação de 01/02/2007 e o nsu de 83 contabilizam cinco registros e tratam de uma única transação, onde cinco produtos foram vendidos.

A ferramenta de mineração de dados, WEKA, lê apenas os números das transações, aqui identificados pelo número de loja, o código de componente, a data de transação e o nsu, e os produtos que saíram nesta transação, em um único registro. Então, a transformação foi feita via programação SQL, usando SGBD Oracle 10g, gerando a seguinte tabela de dados:

NUMERO_LOJA	COD_COMPONENTE	DATA_TRANSACAO	NSU	CAMISA	CALCA	CAMISETA	CALCINHA	BLUSA	BOLSA	BLUSAO	SAIA	TENIS	BERMUDA	MEIA_CALCA	MEIA	SAPATO	BIQUINI	SUNGA
0006	10	13022007	28	N	N	N	N	N	N	N	S	N	N	N	N	N	N	N
0006	10	13022007	32	N	N	N	N	N	N	N	S	N	N	N	N	N	N	N
0006	10	13022007	36	N	N	N	N	N	N	N	S	N	N	N	N	N	N	N
0006	10	13022007	51	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
0006	10	13022007	64	N	N	N	N	S	N	N	S	N	N	N	N	N	N	N
0006	10	13022007	68	N	N	N	N	S	N	N	N	N	N	N	N	N	N	N
0006	10	13022007	74	N	N	N	N	N	N	N	N	N	N	N	N	N	S	N
0006	10	13022007	79	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
0006	10	13022007	85	N	N	N	N	N	N	N	N	N	S	N	N	N	N	N

Figura 9. Transações transformadas para leitura no WEKA

Neste exemplo, pode-se perceber que há apenas uma transação por registro. Os produtos estão em colunas agora e para cada produto um valor “s” ou “n” é associado, identificando se o produto saiu ou não na transação.

Desta forma, os dados que tinham sido selecionados sofreram uma nova restrição, em sua quantidade. Como os produtos são dispostos em colunas foram considerados trinta produtos, entre os antes selecionados, para se trabalhar.

Na etapa de redução dos dados não foi necessário a construção de um data warehouse, porém outras métricas de redução foram utilizadas, como redução de dimensão que acabou sendo feita por outras partes do processo de KDD, na retirada de dados redundantes, por exemplo e na compreensão dos dados, para reduzir o conjunto de dados. A decisão de trabalhar com dados de vestimenta e cama, mesa e banho é outro exemplo de redução.

4.4.5 Mineração de dados

Esta é a etapa mais importante de todo o processo de KDD. Para se chegar nela, é essencial dar uma atenção especial às outras fases, pois esta é uma fase crítica do processo. Se o sistema entrar com varáveis inconsistentes, falhas ou ausentes, provavelmente todo o conhecimento extraído não será verdadeiro.

Como definido antes, o escopo inicial deveria fazer tarefas de mineração para:

- Conhecer perfis de clientes que freqüentam e compram na loja.
- Conhecer clientes que compraram na loja pela primeira vez.
- Tentar conhecer os produtos que estão associados em transações, para uma análise de cesta de mercados.

No entanto, ficou constatado que a base de dados não continha nenhuma informação dos seus clientes, fazendo com que o escopo fosse reduzido e as tarefas concentradas em associações para cesta de mercados.

Com as tarefas definidas e a base totalmente preparada, as técnicas de mineração de dados puderam ser escolhidas. No tópico 4.4.5.2, a escolha da tarefa de regras de associação será melhor abordada. A ferramenta utilizada para realizar esta tarefa foi a WEKA, descrita no sub-tópico a seguir.

4.4.5.1 Ferramenta utilizada - WEKA

Weka é uma ferramenta, de código aberto, desenvolvida em JAVA, que usa algoritmos de aprendizagem de máquina para tarefas de Mineração de Dados. Os algoritmos podem ser aplicados diretamente ao conjunto de dados ou chamados por um código JAVA. WEKA possui ferramentas para pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização [9].

O WEKA possui uma GUI que provê um ponto inicial para o usuário poder explorar as suas ferramentas e aplicações.

A GUI do WEKA é composta de quatro botões, como mostra a ilustração, retirada de [9].

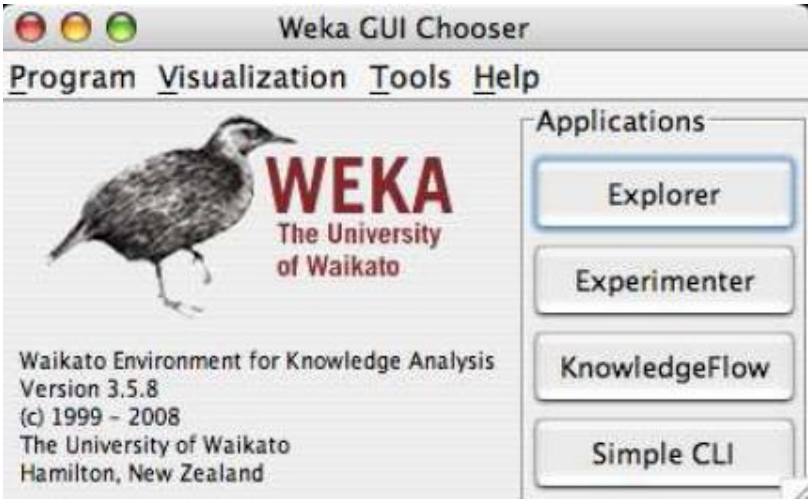

Figura 10. GUI WEKA

Os botões são usados para iniciar as seguintes aplicações:

- Explorer – Um ambiente para exploração dos dados no WEKA (durante a explicação da ferramenta maiores detalhes serão inseridos).
- Experimenter – Um ambiente de testes estatísticos e experimentos de desempenho entre os esquemas.
- KnowledgeFlow – Um ambiente que suporta praticamente as mesmas funcionalidades do ambiente Explorer, com a diferença de ter uma interface drag-and-drop.
- SimpleCLI – Uma interface de linha de comando que fornece um ambiente para executar os comandos WEKA.

Para este trabalho foi utilizado apenas o botão Explorer.

No ambiente Explorer iremos apenas dar uma breve explicação em cada aplicação de mineração de dados para não fugirmos do escopo do projeto, que é extrair padrões utilizando regras de associação.

No ambiente Explorer é possível ver uma aba onde o usuário pode escolher quais aplicações quer utilizar. A figura 11, extraída de [9], mostra uma ilustração deste ambiente.

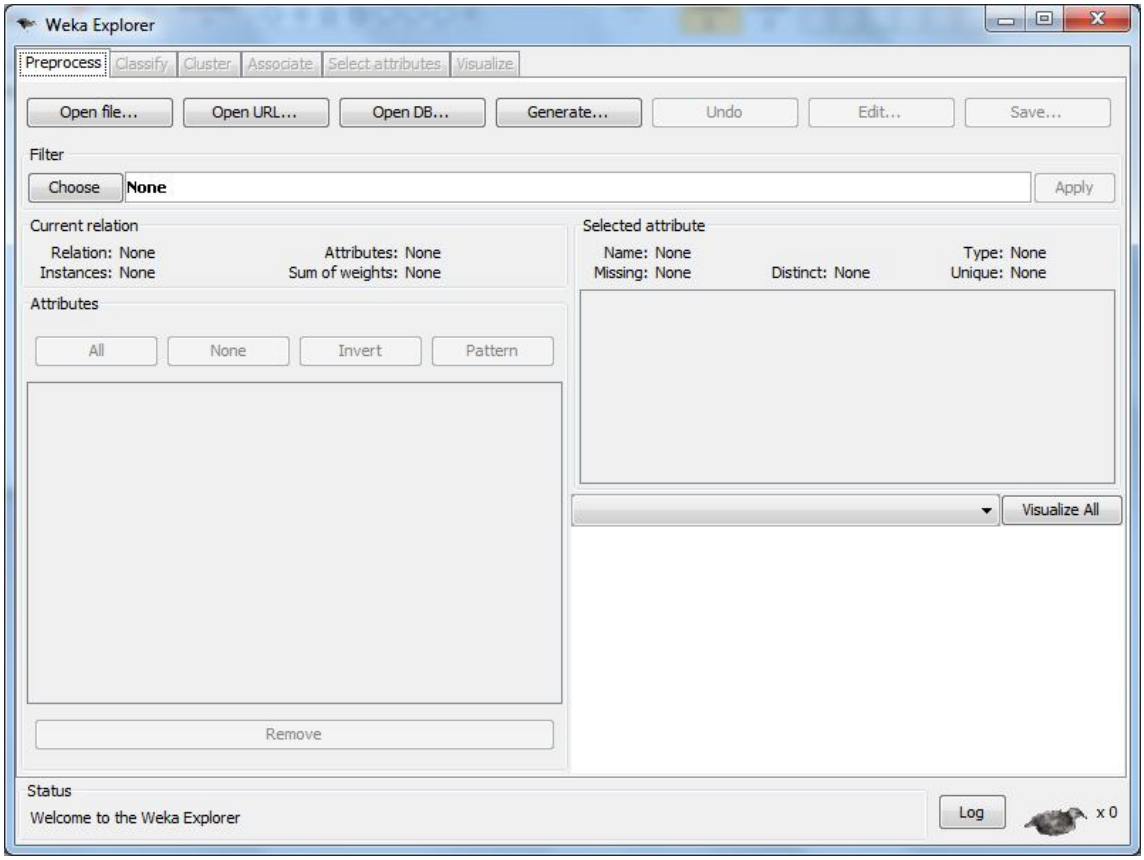

Figura 11. Ambiente Explorer

As abas de tarefas de mineração de dados são as seguintes [9]:

- Preprocess – Escolhe e modifica os dados.
- Classify – Treina e testa esquemas de aprendizado.
- Cluster – Agrupa os dados.
- Associate – Aplica regras de associação para os dados. Foi esta funcionalidade que o trabalho utilizou nos experimentos.
- Select Attributes – Seleciona e mostra os atributos mais relevantes dos dados.
- Visualize – Mostra uma interface 2D dos dados.

Na aba de pré-processamento dos dados é possível visualizar abas para habilitação de leitura dos dados. Dentre elas, o trabalho focou na aba “Open DB”. A funcionalidade desta aba é permitir que o WEKA acesse diretamente o banco de dados para obtenção dos dados a serem minerados [9]. A figura 12 mostra uma conexão do WEKA com o banco de dados.

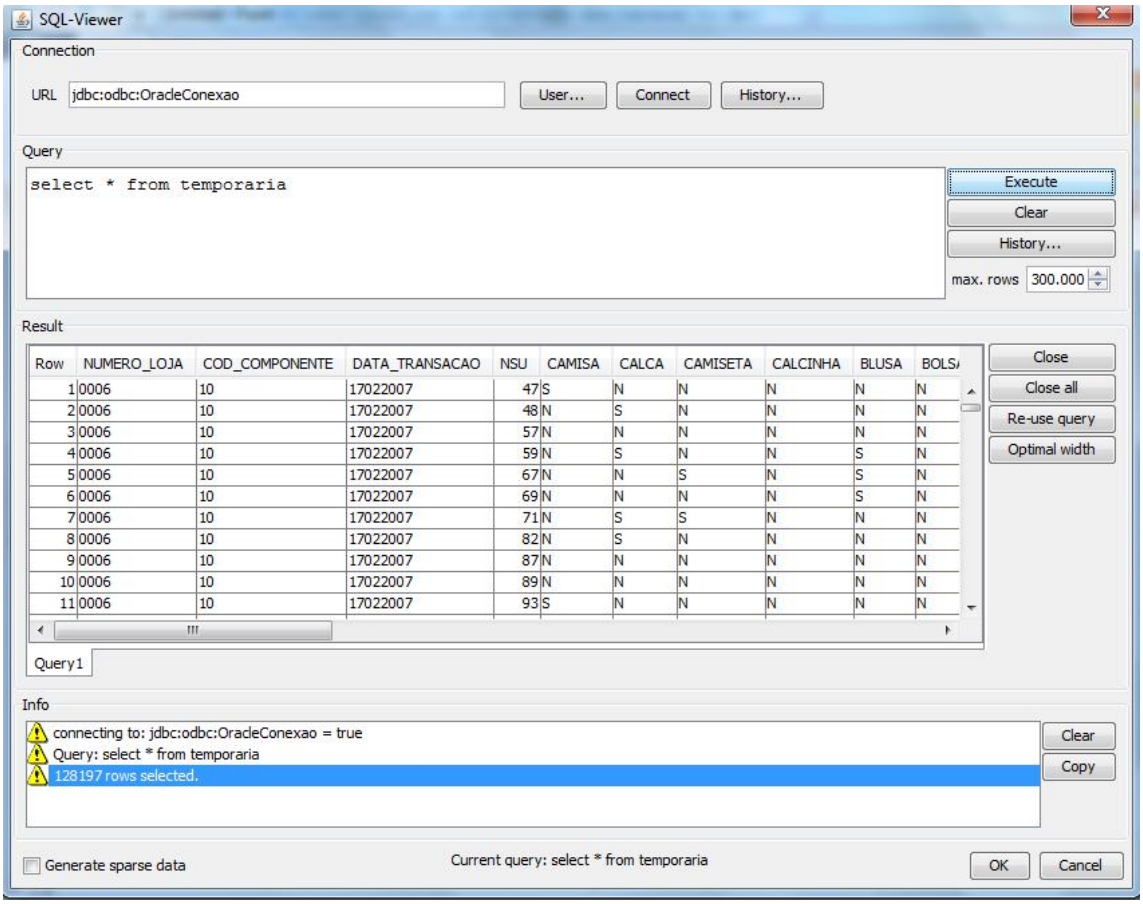

Figura 12. Conexão Oracle com WEKA

Foi necessário criar uma variável de conexão ODBC para a conexão ser bem sucedida. Após a utilização da variável e conexão com o usuário, podem-se fazer consultas diretamente do banco no ambiente do WEKA. A partir daí é que os dados foram enviados para o ambiente de pré-processamento do WEKA para que alguma técnica de MD fosse utilizada.

4.4.5.2 Técnica utilizada

Como descrito antes, a tarefa realizada foi associar produtos em transações para fazer uma análise de cesta de mercado. Para esta tarefa, foram utilizadas uma regras de associação, com o algoritmo apriori, utilizando a ferramenta WEKA.

O processo de aplicação de regras de associação no WEKA começa pela conexão ao banco de dados como mostrado na figura 13. Após isso os dados são transferidos para o ambiente Explorer para pré-processamento como mostra a figura 13.

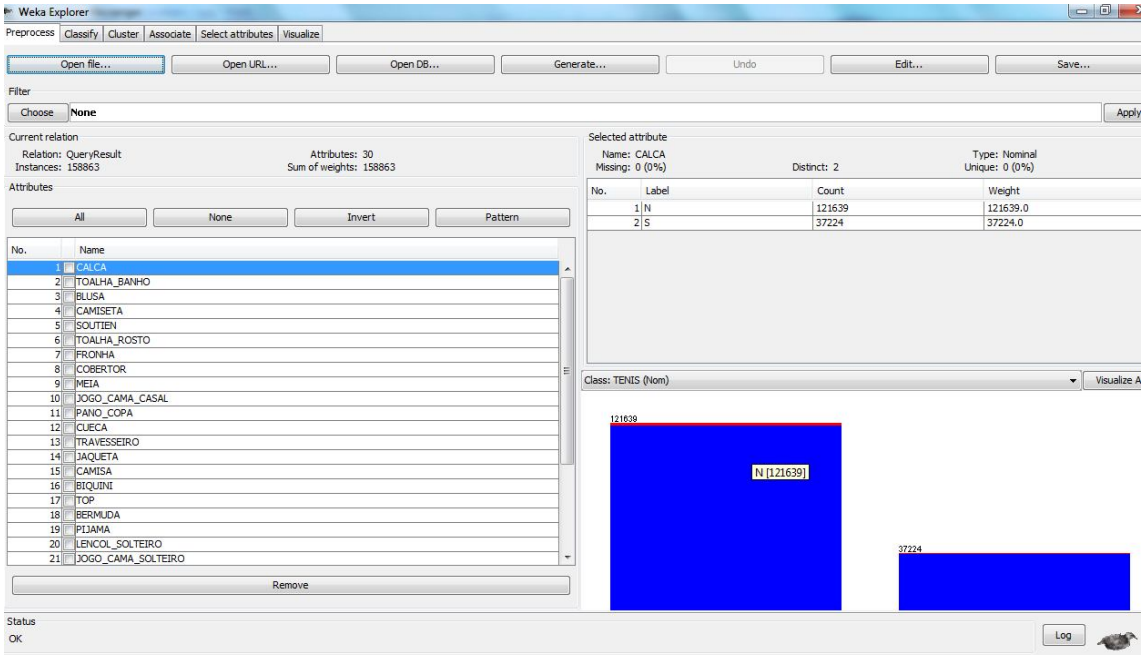

Figura 13. Dados prontos para pré-processamento no ambiente Explorer do WEKA

O próximo passo será escolher os atributos que se deseja trabalhar e se necessário, aplicar algum filtro para adequar os dados às técnicas a serem utilizadas. No caso do trabalho em questão este processo de adequação dos dados já foi feito via programação SQL, visto que o autor já conhecia como os dados são lidos no WEKA. Os atributos a serem

selecionados para aplicação do algoritmo também pode ser escolhido via programação SQL e, assim, o programa já transfere os dados para o ambiente Explorer de forma reduzida.

O próximo passo é escolher a aba “Associate” e utilizar a regra de associação, como mostra a figura 14.

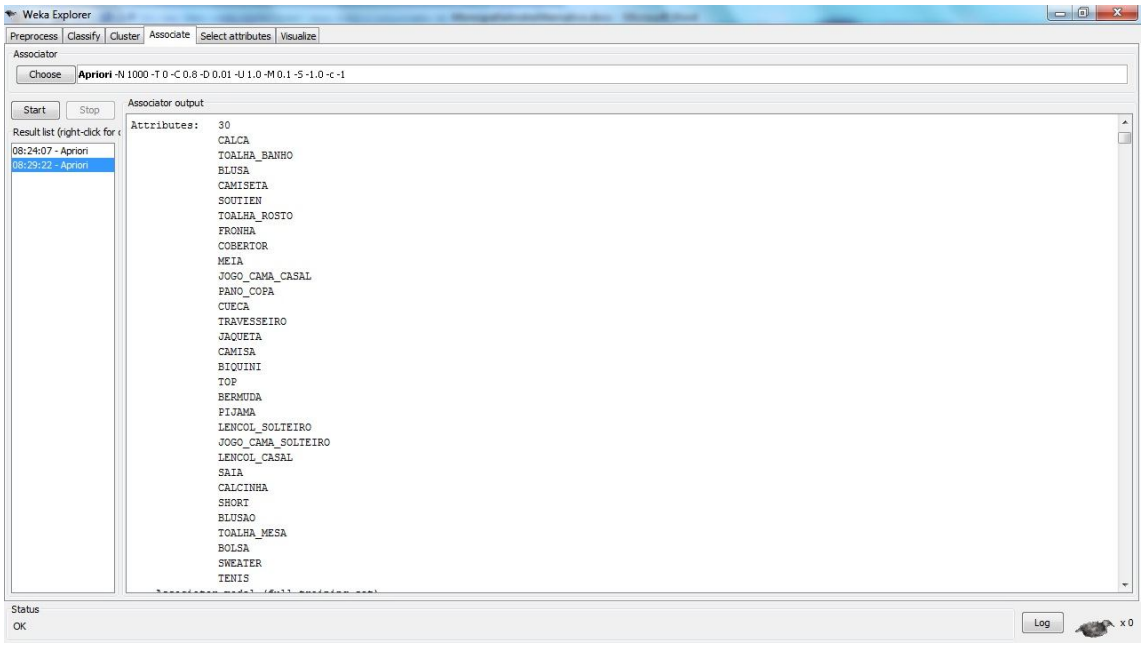

Figura 14. Ambiente de Regras de Associação

Neste ambiente, o usuário pode escolher os algoritmos de regras de associação via botão “Choose”. O trabalho em questão focou no algoritmo Apriori. Como descrito no capítulo 3, este algoritmo utiliza métricas de associação usando o mínimo suporte e a confiança para gerar uma regra interessante para o problema. Dessa forma, nos experimentos foram feitos testes alterando estes valores. Os experimentos serão descritos com maiores detalhes no sub-tópico seguinte.

4.4.6 Avaliação dos resultados e assimilação do conhecimento

Foram realizados sete experimentos, cada um com vários testes alterando as métricas suporte mínimo, que determina a porcentagem mínima exigida em que os produtos estejam contidos nas transações e a confiança, que determina a porcentagem das transações que contendo um produto ou mais de um, contenha outro associado. Os experimentos foram os seguintes:

- Aplicação da regra de associação a todos os produtos mais vendidos.
- Aplicação da regra de associação a todos os produtos de vestimentas.
- Aplicação da regra de associação a todos os produtos de vestimenta masculina.
- Aplicação da regra de associação a todos os produtos de vestimenta feminina.
- Aplicação da regra de associação a todos os produtos de cama, mesa e banho.
- Aplicação da regra de associação a todos os produtos de cama e mesa.
- Aplicação da regra de associação a todos os produtos de banho.

```
Apriori
=======

Minimum support: 0.4 (42 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 60

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

  1. TOALHA_ROSTO=S 58 ==> TOALHA_BANHO=S 57    conf:(0.98)
  2. TOALHA_ROSTO=S BIQUINI=N 47 ==> TOALHA_BANHO=S 46    conf:(0.98)
  3. TOALHA_ROSTO=S 58 ==> BIQUINI=N 47    conf:(0.81)
  4. TOALHA_BANHO=S TOALHA_ROSTO=S 57 ==> BIQUINI=N 46    conf:(0.81)
  5. TOALHA_ROSTO=S 58 ==> TOALHA_BANHO=S BIQUINI=N 46    conf:(0.79)
  6. TOALHA_BANHO=S BIQUINI=N 60 ==> TOALHA_ROSTO=S 46    conf:(0.77)
```

Figura 15. Resultados obtidos com o WEKA

Antes de começar a explicar os experimentos, é necessário entender como o WEKA disponibiliza os resultados. Com a figura 16 é possível fazer esta análise. Alguns pontos de interesse da listagem da figura 15 são:

- O resultado mostra alguns dos valores default para os parâmetros do algoritmo.
- Logo abaixo os tamanhos dos conjuntos de itemsets com suporte mínimo são mostrados.
- Logo abaixo as melhores regras de associação são mostradas ordenadas por confiança. Os valores depois dos antecedentes e conseqüentes das regras são o número de instâncias ou ocorrências para as quais a associação ocorre.
- É interessante observar que o algoritmo usa tantos positivos “s” como negativos “n”, encontrando muitas regras significativas onde a associação é “Se A não foi comprado então B também não foi”.

Foram feitos vários testes para os experimentos. Sempre variando o mínimo suporte e a confiança. Para o primeiro experimento, todos os produtos que são mais vendidos foram considerados e para o primeiro teste foi considerado o mínimo suporte com 10% e a confiança com 80%. Isto significa que o WEKA procurou, em todos os produtos selecionados, aqueles que estavam presentes em, no mínimo, 10% das transações (suporte) e que estivessem juntos em, pelo menos, 80% das vezes em que saíram nas transações (confiança). Na figura 16 os resultados obtidos são mostrados.

```
Minimum support: 0.9 (23670 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 2

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17

Size of set of large itemsets L(2): 21

Size of set of large itemsets L(3): 5

Best rules found:

 1. PANO_COPA=N 24384 ==> TOALHA_MESA=N 24058    conf:(0.99)
 2. JOGO_CAMA_SOLTEIRO=N 24369 ==> TOALHA_MESA=N 23915    conf:(0.98)
 3. LENCOL_SOLTEIRO=N 24384 ==> TOALHA_MESA=N 23911    conf:(0.98)
 4. LENCOL_CASAL=N 24274 ==> TOALHA_MESA=N 23794    conf:(0.98)
 5. BOLSA=N 25396 ==> TOALHA_MESA=N 24877    conf:(0.98)
 6. PIJAMA=N 24416 ==> TOALHA_MESA=N 23912    conf:(0.98)
 7. SHORT=N 25178 ==> TOALHA_MESA=N 24658    conf:(0.98)
 8. TENIS=N 25322 ==> TOALHA_MESA=N 24796    conf:(0.98)
 9. CALCINHA=N 24279 ==> TOALHA_MESA=N 23771    conf:(0.98)
```

Figura 16. Resultados obtidos com todos os produtos

Diante dos resultados obtidos, foram observados alguns pontos: foram solicitadas ao programa mil regras de associação. Algumas regras foram encontradas podem dizer alguma coisa sobre o comportamento dos clientes que compram na loja. Várias regras de produtos com vestimenta e cama, mesa e banho foram encontradas. Na figura 17 podem-se ver algumas delas. Por exemplo, as regras CALCA=S CUECA=S ==> TOALHA_MESA=N, CALCA=S CUECA=S SHORT=N ==> TOALHA_MESA=N ou CALCA=S MEIA=S ==> TOALHA_MESA=N, podem indicar que os clientes que compram produtos de vestimenta não costumam comprar produtos de cama, mesa e banho. Por outro lado, regras do tipo TOALHA_ROSTO=S ==> TOALHA_BANHO=S, sugere que clientes que compram toalhas de rosto também levam toalha de banho.

Durante os outros testes foi alterado o mínimo suporte para que o algoritmo procurasse entre os produtos mais freqüentes. Gradualmente subindo o valor do mínimo suporte e procurando manter a confiança em no máximo 50%, pois abaixo disso as regras geradas, em todas as combinações testadas foram negativas. Logo, foi assumido que um valor de confiança abaixo de 50%

não seria uma boa regra. Todos os testes seguintes geraram regras semelhantes às do primeiro teste, mesmo variando o suporte mínimo e a confiança. O fato interessante observado é que os produtos de classes diferentes não se misturavam, ou seja, um cliente que compra um artigo de vestimenta, não costuma comprar artigos de cama, mesa ou banho. Esta foi a conclusão para estes testes.

No segundo experimento foram considerados todos os produtos relacionados apenas com vestimenta. Alguns resultados interessantes foram observados. Regras como CAMISETA=S MEIA=S ==> CALCA=S ou CAMISETA=S MEIA=S ==> CALCA=S ou CALCA=S CAMISETA=S CUECA=S ==> MEIA=S ou MEIA=S ==> SHORT=N sugerem que camisetas e meias estão sempre associadas às calças ou cuecas. Meias também são associadas a estes itens. Por exemplo, quando camiseta e meia saem, há uma tendência de calças saírem. Regras como CUECA=S ==> CALCA=S podem sugerir que o produto calça sai quando cueca também sai. Porém, muitas regras com os outros artigos foram negativas, do tipo, SAIA=N BOLSA=N ==> CALCA=S podem sugerir que, sabendo que o artigo calça se relaciona com muita freqüência com camiseta, meia, e cueca, os produtos saia e bolsa são comprados separadamente. Muitas regras deste tipo, com os outros produtos que não sejam calça, camiseta, meia ou cueca, foram encontradas. Isto sugere que as melhores associações nesta loja, para artigos de vestimenta, são feitas com produtos camiseta, calça, meia ou cueca.

No teste seguinte foi alterado o suporte para 30% e a confiança para 70%. Porém, foi observado que as regras geradas são semelhantes ao teste anterior.

No terceiro teste alterou-se o suporte para 40% e a confiança para 50%. Algumas poucas regras com produtos diferentes de camiseta, calça, meia ou cueca foram encontradas. Por exemplo, SWEATER=N ==> SOUTIEN=S conf:(0.61) pode sugerir que quando sweater não sai, soutien sai.

Outros testes foram executados, porém com regras muito semelhantes às encontradas anteriormente. O que se observou foi que quanto mais o suporte mínimo é aumentado, menos regras ele consegue criar. O que é bem

esperado, pois o suporte trata da porcentagem em que estes produtos aparecem nas transações. Quanto mais exigente, mais óbvias as regras serão. Como o interessante é encontrar padrões escondidos a relação com suporte mínimo de 30% e confiança de 70% foi a melhor entre todas testadas.

Tendo esta última relação sido classificada como a melhor entre todas, outros dois experimentos foram testados com relação aos artigos de vestimenta: vestimenta feminina e vestimenta masculina

Na primeira tentativa, observaram-se muitas regras negativas entre produtos femininos. Regras como CALCA=S BLUSA=S TOP=N ==> SHORT=N, ou ainda CALCA=S SAIA=N ==> SHORT=N, ou SOUTIEN=N BIQUINI=N SAIA=N ==> CALCA=S, ou ainda SOUTIEN=S ==> SHORT=N, BLUSA=S ==> CALCA=S, indicam que novamente calça é um artigo bastante associado com outros. Porém, observa-se que muitos produtos como saia, blusa e soutien, por exemplo, não fracamente associados. Poucas regras foram encontradas com estes itens. A mais interessante percebida foi BLUSA=S ==> CALCA=S. Outras regras não tão interessantes foram encontradas, como: SOUTIEN=S ==> SWEATER=N, SOUTIEN=S ==> SAIA=N ou ainda BLUSA=S ==> TOP=N.

Alterando-se a confiança e o suporte novamente o comportamento mostrou-e levemente melhor. Regras como CALCA=S SHORT=N ==> SOUTIEN=S puderam ser encontradas. No entanto, calça continua sendo um artigo muito forte e soutien, apesar de pela primeira vez conseguir ser associado a outro produto, foi com uma confiança muito baixa, 40% com suporte de 10%, indicando uma fragilidade na mesma.

O outro experimento de vestimenta foi com artigos masculinos. Considerando valores baixos de métricas de mínimo suporte 30% e confiança 70%, observaram-se muitas regras positivas.

Em todos os testes realizados, os produtos calca, camiseta, cueca e meia se associam com freqüência. Então, nos testes seguintes, os valores de mínimo suporte e confiança foram aumentados para 40% e 80%, respectivamente.

A figura 17 mostra os resultados.

```
Best rules found:

  1. CAMISETA=S MEIA=S CUECA=S 43 ==> CALCA=S 43    conf:(1)
  2. CALCA=S CAMISETA=S CUECA=S 45 ==> MEIA=S 43    conf:(0.96)
  3. MEIA=S CUECA=S 52 ==> CALCA=S 49    conf:(0.94)
  4. CAMISETA=S MEIA=S 50 ==> CALCA=S 47    conf:(0.94)
  5. CAMISETA=S CUECA=S 48 ==> CALCA=S 45    conf:(0.94)
  6. CALCA=S CAMISETA=S MEIA=S 47 ==> CUECA=S 43    conf:(0.91)
  7. CAMISA=N BLUSAO=N 54 ==> TENIS=N 49    conf:(0.91)
  8. CUECA=S 61 ==> CALCA=S 55    conf:(0.9)
  9. CAMISETA=S CUECA=S 48 ==> MEIA=S 43    conf:(0.9)
 10. CAMISETA=S CUECA=S 48 ==> CALCA=S MEIA=S 43    conf:(0.9)
 11. CALCA=S CUECA=S 55 ==> MEIA=S 49    conf:(0.89)
 12. MEIA=S 64 ==> CALCA=S 57    conf:(0.89)
 13. CAMISETA=S 71 ==> CALCA=S 63    conf:(0.89)
 14. CALCA=S MEIA=S CUECA=S 49 ==> CAMISETA=S 43    conf:(0.88)
 15. CAMISETA=S MEIA=S 50 ==> CUECA=S 43    conf:(0.86)
 16. CAMISETA=S MEIA=S 50 ==> CALCA=S CUECA=S 43    conf:(0.86)
 17. CAMISETA=S TENIS=N 57 ==> CALCA=S 49    conf:(0.86)
 18. CALCA=S MEIA=S 57 ==> CUECA=S 49    conf:(0.86)
 19. BLUSAO=N 75 ==> TENIS=N 64    conf:(0.85)
 20. CUECA=S 61 ==> MEIA=S 52    conf:(0.85)
```

Figura 17. Resultados obtidos vestimenta masculina

Mesmo com valores mais altos sendo avaliados, as principais regras positivas foram mantidas, até que ao escolher mínimo suporte 60% e confiança 80%, nenhuma regra foi encontrada. Sendo assim, o melhor resultado encontrado foi o da figura 17.

Uma conclusão que se pode ter desse experimento em comparação com o de vestimenta masculina é que alguns produtos que tanto podem ser masculinos, como femininos, como por exemplo, calça, camisa, camiseta, se associam com mais freqüência quando analisados com produtos como cueca, meia. Isso pode indicar que estes produtos do vestuário masculino são melhores associados que produtos do vestuário feminino. Isso pode indicar uma preferência de compras de artigos de vestimenta.

O próximo experimento tratou de avaliar os produtos de cama, mesa e banho. Os testes começaram com valores baixos das métricas. Mínimo suporte 30% e confiança 50%. Ainda assim, algumas regras interessantes foram encontradas. Por exemplo, JOGO_CAMA_CASAL=S ==> TOALHA_BANHO=S

é uma associação de jogo de cama com toalhas. Com esses valores também foram encontradas regras como JOGO_CAMA_SOLTEIRO=N ==> TOALHA_BANHO=S e TOALHA_ROSTO=S ==> PIJAMA=N, que podem indicar que os clientes que compram artigos de cama, não compram artigos de banho.

Em outro teste foi aumentado o valor do mínimo suporte e da confiança para deixar a regra mais exigente. Os valores foram 50% e 70%, para mínimo suporte e confiança, respectivamente. Foram os melhores resultados encontrados e na figura 18 é possível visualizá-los.

```
Best rules found:

 1. TOALHA_ROSTO=S 58 ==> TOALHA_BANHO=S 57    conf:(0.98)
 2. PANO_COPA=N LENCOL_SOLTEIRO=N 56 ==> TOALHA_MESA=N 55    conf:(0.98)
 3. PANO_COPA=N JOGO_CAMA_SOLTEIRO=N 56 ==> TOALHA_MESA=N 55    conf:(0.98)
 4. TOALHA_BANHO=S PANO_COPA=N 55 ==> TOALHA_MESA=N 54    conf:(0.98)
 5. PANO_COPA=N TRAVESSEIRO=N 55 ==> TOALHA_MESA=N 54    conf:(0.98)
 6. TRAVESSEIRO=N JOGO_CAMA_SOLTEIRO=N 55 ==> TOALHA_MESA=N 54    conf:(0.98)
 7. PANO_COPA=N 78 ==> TOALHA_MESA=N 76    conf:(0.97)
 8. PANO_COPA=N LENCOL_CASAL=N 59 ==> TOALHA_MESA=N 57    conf:(0.97)
 9. PANO_COPA=N PIJAMA=N 56 ==> TOALHA_MESA=N 54    conf:(0.96)
10. COBERTOR=N 62 ==> TOALHA_MESA=N 58    conf:(0.94)
11. JOGO_CAMA_SOLTEIRO=N 71 ==> TOALHA_MESA=N 66    conf:(0.93)
12. TRAVESSEIRO=N 76 ==> TOALHA_MESA=N 70    conf:(0.92)
13. PIJAMA=N 72 ==> TOALHA_MESA=N 66    conf:(0.92)
14. TRAVESSEIRO=N LENCOL_SOLTEIRO=N 58 ==> TOALHA_MESA=N 53    conf:(0.91)
15. FRONHA=N 65 ==> TOALHA_MESA=N 58    conf:(0.89)
16. LENCOL_SOLTEIRO=N 74 ==> TOALHA_MESA=N 65    conf:(0.88)
17. TOALHA_BANHO=S 80 ==> TOALHA_MESA=N 70    conf:(0.88)
18. LENCOL_SOLTEIRO=N LENCOL_CASAL=N 64 ==> TOALHA_MESA=N 56    conf:(0.88)
19. LENCOL_CASAL=N 79 ==> TOALHA_MESA=N 69    conf:(0.87)
20. LENCOL_SOLTEIRO=N 74 ==> LENCOL_CASAL=N 64    conf:(0.86)
21. FRONHA=N 65 ==> TRAVESSEIRO=N 56    conf:(0.86)
22. FRONHA=N 65 ==> LENCOL_SOLTEIRO=N 56    conf:(0.86)
23. LENCOL_SOLTEIRO=N TOALHA_MESA=N 65 ==> LENCOL_CASAL=N 56    conf:(0.86)
24. LENCOL_SOLTEIRO=N TOALHA_MESA=N 65 ==> PANO_COPA=N 55    conf:(0.85)
25. JOGO_CAMA_SOLTEIRO=N TOALHA_MESA=N 66 ==> PANO_COPA=N 55    conf:(0.83)
26. FRONHA=N 65 ==> LENCOL_CASAL=N 54    conf:(0.83)
27. LENCOL_CASAL=N TOALHA_MESA=N 69 ==> PANO_COPA=N 57    conf:(0.83)
```

Figura 18. Resultados obtidos com cama, mesa e banho

As regras um e quatro podem dizer algo. A regra um pode insinuar uma associação direta entre toalha de banho e toalha de rosto. A regra quatro pode

ratificar o que se levantou quando testes menos exigentes foram executados: clientes que compram artigos de cama, não compram artigos de banho. Este pode ser considerado um teste conclusivo para estes artigos, já que as regras são semelhantes mesmo variando as métricas. Os resultados no teste mais significativo são parecidos com outros menos exigentes.

Considerando a hipótese de que produtos de cama não saem junto com produtos de banho, foram feitos mais dois testes para saber então que produtos de cama e que produtos de banho podem ser associados.

O primeiro foi com produtos de cama. Foram feitos testes com valores iguais aos do teste anterior e não foi encontrada nenhuma regra interessante. Uma hipótese poderia ser que estes produtos são vendidos separadamente, idéia não tão absurda, já que o número médio de produtos por transações, na loja, é baixo, aproximadamente dois.

O segundo teste foi com artigos de banho. Como apenas toalhas de rosto e toalhas de banho fazem parte dos produtos de banho, foi inserido um artigo de vestimenta, biquíni, e algumas regras interessantes puderam ser relatadas, como a figura 19 mostra.

```
Best rules found:

  1. TOALHA_ROSTO=S 58 ==> TOALHA_BANHO=S 57    conf:(0.98)
  2. TOALHA_ROSTO=S BIQUINI=N 47 ==> TOALHA_BANHO=S 46    conf:(0.98)
  3. TOALHA_ROSTO=S 58 ==> BIQUINI=N 47    conf:(0.81)
  4. TOALHA_BANHO=S TOALHA_ROSTO=S 57 ==> BIQUINI=N 46    conf:(0.81)
  5. TOALHA_ROSTO=S 58 ==> TOALHA_BANHO=S BIQUINI=N 46    conf:(0.79)
  6. TOALHA_BANHO=S BIQUINI=N 60 ==> TOALHA_ROSTO=S 46    conf:(0.77)
```

Figura 19. Resultados obtidos com artigos de banho

Como observado antes, toalhas de banho e rosto sempre se associam. Biquíni não foi associado a nenhum produto, ratificando o que já tinha sido levantado, produtos de vestimenta não costumam ser vendidos juntos com produtos de cama, mesa e banho.

4.5 Considerações finais

A conclusão das análises dos resultados obtidos, considerando todos os testes e possibilidades de associação é que a empresa vende muitos produtos de cama, mesa e banho e vestuário masculino e feminino. Destes produtos foi observado que não é uma compra freqüente produtos das duas classes. Na classe de produtos de banho, toalhas de banho e rosto se relacionam bem e podem ser um indicativo forte que eles podem ser associados em conjunto. Na classe de vestuário, os testes indicam que os produtos masculinos são mais freqüentes, principalmente calça, meia, camiseta e cueca. Estes produtos talvez também possam ser associados em conjunto.

Uma grande vantagem de utilizar este método de mineração é este jogo de combinações entre os produtos que o usuário não precisa fazer. A ferramenta e o algoritmo fazem a maior parte do trabalho e o trabalho mais árduo. Cabe ao analista de negócio saber interpretar as regras para que isto possa servir para o departamento de marketing começar a trabalhar na possibilidade de organizar e juntar estes produtos melhores avaliados para impulsionar a venda.

A desvantagem pode ser encontrada na dificuldade de chegar ao resultado final, já que é necessária uma grande experimentação com os valores nas métricas para se chegar a um resultado plausível. Outra dificuldade pode ser encontrada caso o usuário queira fazer alguma tarefa que a ferramenta WEKA não faça. Neste caso, será necessário modificar o código do sistema para fazer tais alterações. Isto nem sempre é fácil e geralmente tem um custo alto. Cabe ao usuário decidir a que ponto ele pode arcar com este custo para utilizar a ferramenta ou construir uma própria.

Capítulo 5

Conclusão

Ao final do estudo, foi possível realizar uma análise com os resultados conseguidos. Isso permitiu mostrar que com a utilização de Mineração de dados pode-se realizar buscas eficientes em uma base de dados a procura de padrões que tenham valor para uma organização.

Dessa forma o objetivo proposto de aplicar uma técnica e um algoritmo selecionado foi atingido e foi mostrado, também, que uma empresa pode obter mais conhecimento sobre seus próprios dados. Assim, foi possível fornecer subsídios para que empresários possam planejar melhor suas estratégias e decisões.

5.1 Dificuldades encontradas e trabalhos futuros

Das várias dificuldades encontradas durante o processo de obtenção da base de dados e definição das tarefas de mineração de dados a serem abordadas, algumas dificuldades devem ser citadas com maiores detalhes por inferir diretamente no resultado das análises feitas.

Foi constatada da ausência de informações sobre os clientes da loja. Isso dificultou a elaboração das tarefas de mineração e conseqüentemente em suas técnicas. Diante disso, julgou-se necessário reduzir o escopo do projeto a tarefas e técnicas que não envolvessem clientes de forma direta, desconsiderando o fator humano, importantíssimo para o entendimento do comportamento dos clientes que compram nas lojas da empresa.

Falhas na modelagem dos dados em relação aos produtos, como ausência de categorias de produtos, inferiram diretamente na análise dos resultados da técnica de mineração aplicada. Diante da situação foi entendido como melhor opção,

generalizar as descrições dos produtos para que a mineração de dados pudesse ser aplicada. Dessa forma a análise não pôde refletir com total precisão quais produtos estavam sendo considerados, especificamente. Apenas a descrição geral foi trabalhada.

Diante dessas situações, o desenvolvimento do projeto ficou limitado a trabalhar com produtos generalizados e desprezando o fator humano. Como trabalhos futuros a sugestão para esta empresa é organizar melhor os dados de produtos, criando categorias e subcategorias para que dessa forma uma análise mais profunda e detalhada possa ser feita.

A consideração de informações sobre os clientes que compram nas suas lojas é outro fator que a empresa pode começar a coletar. Isto enriquece bastante a análise de comportamento dos clientes, ajudando a empresa a entender melhor como os mesmos escolhem os produtos que querem comprar. Dessa forma o departamento de marketing pode atuar com mais certeza sobre como e pelo que os consumidores das lojas se interessam, e assim, tomar as decisões sobre qual melhor estratégia utilizar para impulsionar as vendas.

Com relação ao projeto feito, outras técnicas podem ser aplicadas como regras de seqüências, agrupamentos, classificação e previsão. Para aplicação de regras seqüências é preciso alterar o formato de como os dados estão dispostos na base de dados. Isso por que o WEKA tem essa limitação. Para alguns algoritmos, o WEKA exige que os dados estejam dispostos de forma diferente. Para a aplicação de agrupamentos e classificação seria necessário o conhecimento das especificações dos produtos e, novamente, esbarramos na limitação da base. Para previsões, seria necessário considerar as datas das vendas. Estes atributos são conhecidos e estão dispostos, tornando viável a execução desta tarefa.

Bibliografia

[1] WITTEN, I. H.; FRANK, E. **Data Mining - Pratical Machine Learning Tools and Taechniques**. 2ª Edição. ed. San Francisco: Elsevier, v. I, 2005.

[2] LOUZADA - NETO, F.; DINIZ, C. A. R. Data Mining. **Data Mining - Uma introdução**, São Paulo, 2000.

[3] HAN, J.; KAMBER, M. **Data Mining - Concepts and Techniques**. [S.l.]: Morgan Kaufmann Publishers, 2001.

[4] CORTES, S. D. C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de Dados. **Mineração de Dados - Funcionalidades, Técnicas e Abordagens**, Rio de Janeiro, Maio 2002. 1-5.

[5] SILVA, M. P. D. S. Mineração de Dados - Conceitos, Aplicações e Experimentos com WEKA, Mossoró.

[6] AMO, S. D. Técnicas de Mineração de Dados, Uberlância.

[7] CARDOSO, T. E. Desenvolvimento de um Algoritmo Híbrido para Mineração de Dados na Área de Vendas, Santa Cruz do Sul, Dezembro 2007.

[8] SIMM, M. R. . D. V. Trabalho de Graduação. **Uma Overview do Processo de Descoberta de Conhecimento em Base de Dados, com Ênfase em Técnincas de Mineração de Dados**, Santa Cruz do Sul, 2000.

[9] LAROSE, D. T. **Discovering Knowledge in Data, An Introduction to Data**. [S.l.]: John Wiley & Sons, 2005.

[10] WEKA Project. Disponivel em: <http://www.cs.waikato.ac.nz/~ml/weka/index.html>. Acesso em: 29 out. 2009.