UNIVERSIDADE FEDERAL DO RIO GRANDE I

ESCOLA DE ADMINISTRAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

GESID - Grupo de Estudos em Sistemas de Informação e de Apoio à Decisão

AVALIAÇÃO DE FERRAMENTAS DE MINERAÇÃO DE DADOS COMO FONTE DE DADOS RELEVANTES PARA A TOMADA DE DECISÃO: APLICAÇÃO NA REDE UNIDÃO DE SUPERMERCADOS, SÃO LEOPOLDO-RS

Lóren Pinto Ferreira Gonçalves

Dissertação apresentada ao Programa de Pós-Graduação em Administração (PPGA/EA/UFRGS) como requisito parcial para a obtenção do grau de Mestre em Administração.

Orientador: Prof. Henrique M. R. de Freitas

Porto Alegre, junho de 2001

Agradecimentos

Gostaria de agradecer a todos aqueles que contribuíram, de alguma maneira, para a realização deste trabalho.

Em especial ao Everaldo, meu marido, que com sua compreensão, amor e amizade, sempre me deu apoio e incentivo para que eu continuasse, mesmo nas horas mais difíceis.

À minha família, pela confiança, força e incentivo e, principalmente por ter me dado a estrutura necessária para chegar até aqui.

Aos amigos Ricardo, Virgínia e Vitória que por diversas vezes me acolheram no aconchego do seu lar para que eu pudesse vir com nossa turma a Porto Alegre, muito obrigado pela sua hospitalidade e amizade.

A meus amigos e colegas Cláudio Albano, Léu Carate, Ricardo Bernardes e Ramão Dornelles pela amizade, pelos diversos intercâmbios de idéias e materiais e pelas muitas horas que convivemos juntos, durante estes três anos.

Ao professor Henrique, meu orientador, pelos ensinamentos, pela dedicação e até pelos merecidos puxões de orelha.

À Edimara, pelas diversas dicas e revisões do trabalho, pelo seu empenho e dedicação.

À Universidade da Região da Campanha (Urcamp), pela iniciativa em desenvolver um Mestrado Distribuído e pela ajuda financeira concedida.

À empresa Comercial Unida de Cereais (Rede Unidão), em especial ao Gustavo e à Iara, que abriram as portas da empresa e forneceram seus dados para a realização deste trabalho.

Ao pessoal das empresas desenvolvedoras das ferramentas utilizadas, em especial à Ana Pavan, da Hycones IT, pela licença de utilização do software Aira Data Mining® e pelo suporte fornecido, e ao Rob Gerritsen, Presidente da Exclusive Ore, pela contribuição através dos vários e-mails trocados.

Sem a colaboração e o apoio de vocês este trabalho não teria se realizado.

Muito obrigado!

SUMÁRIO

Lista de figuras

Lista de quadros

Resumo

Esta pesquisa tem como tema a avaliação de ferramentas de mineração de dados disponíveis no mercado, de acordo com um site de descoberta do conhecimento, chamado Kdnuggets (http://www.kdnuggets.com). A escolha deste tema justifica-se pelo fato de tratar-se de uma nova tecnologia de informação que vem disponibilizando diversas ferramentas com grandes promessas e altos investimentos, mas que, por outro lado, ainda não é amplamente utilizada pelos tomadores de decisão das organizações. Uma das promessas desta tecnologia é vasculhar grandes bases de dados em busca de informações relevantes e desconhecidas e que não poderiam ser obtidas através de sistemas chamados convencionais.

Neste contexto, realizar uma avaliação de algumas destas ferramentas pode auxiliar a estes decisores quanto à veracidade daquilo que é prometido sem ter de investir antes de estar seguro do cumprimento de tais promessas.

O foco da pesquisa é avaliar sistemas que permitem a realização da análise de cesta de supermercado (market basket analysis) utilizando bases de dados reais de uma rede de supermercados.

Os seus objetivos são: avaliar ferramentas de mineração de dados como fonte de informações relevantes para a tomada de decisão; identificar, através da revisão de literatura, as promessas da tecnologia e verificar se tais promessas são cumpridas pelas ferramentas; identificar e caracterizar ferramentas de mineração de dados disponíveis no mercado e comparar os tipos de resultados gerados pelas diferentes ferramentas e relatar problemas encontrados durante a aplicação destas ferramentas.

O desenvolvimento do trabalho segue o método estudo de caso múltiplo: os dados foram coletados a partir da aplicação das ferramentas às bases de dados e da entrevista com tomadores de decisão da empresa. Foram seguidos procedimentos já utilizados de avaliação de sistemas para a realização desta pesquisa.

A partir da análise dos dados coletados, pôde-se conhecer alguns problemas apresentados pelas ferramentas e concluiu-se que as ferramentas, que foram utilizadas neste trabalho, não estão prontas para serem disponibilizadas no mercado.

Abstract

This research has as subject the evaluation of data mining tools available in the market, in agreement with a knowledge discovery site, called Kdnuggets (http://www.kdnuggets.com). The choice of this subject is justified by the fact that data mining is a new information technology that is releasing several tools with great promises and high investments, however that, on the other hand, is not still used thoroughly by the organizations' decision makers. One of the promises of this technology is to search great databases in order to obtain important and unknown information which could not be obtained through conventional systems.

In this context, accomplishing an evaluation of some of these tools can aid these people concerning the truthfulness of what is promised without making them invest before being sure about the veracity of such promises.

The focus of the research is to evaluate systems that allow the accomplishment of the market basket analysis using real databases of a supermarket chain.

Its objectives are: to evaluate data mining tools as source of important information for decision making, to identify, through literature revision, the promises of the technology and to verify if such promises are achieved by the tools, to identify and characterize data mining tools available in the market and to compare the type of the results generated by the different tools and to relate some problems found during the tool´s application.

The development of the work follows the multiple case study method, where the data were collected from the application of the tool upon the databases and from interviews with the organization's decision makers. System evaluation methods that have already been used were applied to accomplish this research.

Starting from the analysis of the collected data, some problems in the tools could be found and the conclusion was that the data mining tools, that were used in this research, are not ready for being in the market.

Capítulo 1

Tema e justificativa

Mineração de dados, ou *data mining*, é definida como uma etapa na descoberta do conhecimento em bancos de dados que consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis. Para isto são utilizadas técnicas que envolvem métodos matemáticos, algoritmos e heurísticas[1] que visam a descobrir padrões e regularidades entre os dados pesquisados (Brusso, 1998).

Em um mundo globalizado, sem fronteiras geográficas, onde as empresas competem mundialmente, a informação torna-se um fator crucial na busca pela competitividade. Almeida (1995) afirma que o fato de uma empresa dispor de certas informações possibilita-lhe aumentar o valor agregado de seu produto ou reduzir seus custos em relação àquelas que não possuem o mesmo tipo de informação. Para Freitas e Lesca (1992), as informações e o conhecimento compõem um recurso estratégico essencial para o sucesso da adaptação da empresa em um ambiente de concorrência. Oliveira (1997) garante que toda empresa tem informações que proporcionam a sustentação para as suas decisões, entretanto apenas algumas conseguem otimizar o seu processo decisório e aquelas que estão neste estágio evolutivo seguramente possuem vantagem empresarial.

As ferramentas de mineração de dados, por definição, devem trabalhar com grandes bases de dados e retornar, como resultado, conhecimento novo e relevante (Niederman, 1997); porém devemos ser céticos quanto a esta afirmação, pois este tipo de ferramenta irá criar inúmeras relações e equações, o que pode tornar impossível o processamento destes dados. Conforme Santos & Becker (1990), este também foi um dos aspectos que gerou

[1] Heurísticas são processos ou regras de pesquisa e busca de soluções, conduzidos por processos de associações de idéias em geral incompletas, pela complexidade que os problemas tratados envolvem procurando simular ou substituir os processos de inferência dedutiva do raciocínio humano (Torres, 1995)

ceticismo por parte de usuários potenciais de Inteligência Artificial quando começava-se a falar em Redes Neurais e na segunda geração de Sistemas Inteligentes.

A grande promessa da mineração de dados resume-se na afirmação de que ela 'vasculha' grandes bases de dados em busca de padrões escondidos, que extrai informações desconhecidas e relevantes e as utiliza para tomar decisões críticas de negócios. Outra promessa em relação a esta tecnologia de informação diz respeito à forma como elas exploram as interrelações entre os dados. Segundo Figueira (1998), as diversas ferramentas de análise disponíveis dispõem de um método baseado na verificação, isto é, o usuário constrói hipóteses sobre interrelações específicas e então verifica ou refuta estas hipóteses, através do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos, e em refinar a análise, baseado nos resultados de consultas potencialmente complexas ao banco de dados. Já o processo de mineração de dados, para o autor, ficaria responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude dos resultados.

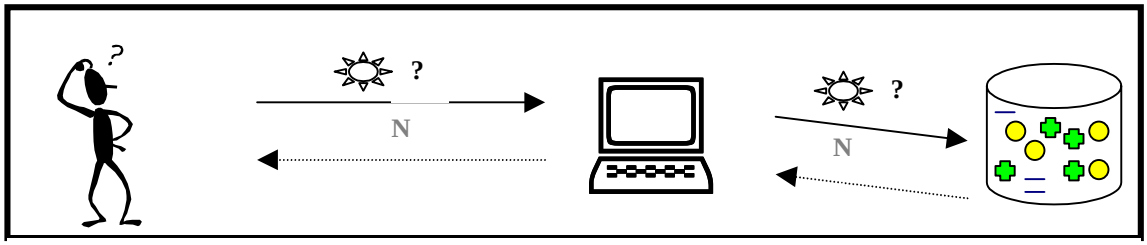

Figura 1 - Busca por informações em sistemas convencionais

(Fonte: Figueira, 1998)

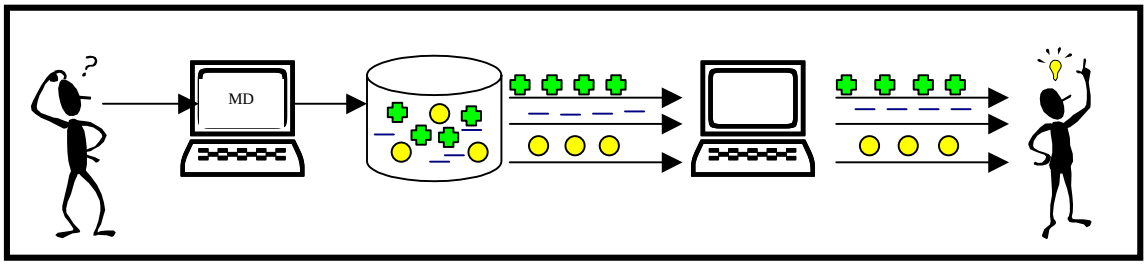

Figura 2 - Busca por informações em sistemas de mineração de dados (Fonte: Figueira, 1998)

Conforme Figueira (1998), a cada ano, companhias acumulam mais e mais dados em seus bancos de dados. Estes dados muitas vezes são mantidos mesmo depois de esgotados seus prazos legais de existência, como no caso de notas fiscais. Com o passar do tempo, este volume de dados passa a armazenar internamente o histórico das atividades da

organização. Como conseqüência, estes bancos de dados passam a conter verdadeiros ‘tesouros’ de informação sobre vários dos procedimentos dessas companhias. Toda esta informação pode ser usada para melhorar seus procedimentos, permitindo que a empresa detecte tendências e características disfarçadas, e reaja rapidamente a um evento que ainda pode estar por vir. No entanto, apesar do enorme valor desses dados, a maioria das organizações é incapaz de aproveitar totalmente o que está armazenado em seus arquivos. Esta informação está implícita, escondida sob uma montanha de dados, e, segundo o autor, não pode ser descoberta utilizando-se sistemas de gerenciamento de banco de dados convencionais. Para ele a quantidade de informação armazenada está explodindo, e ultrapassa a habilidade técnica e a capacidade humana na sua interpretação. Simon (1957 *apud* Freitas, 1993) explicou este fato dizendo que os indivíduos têm uma capacidade limitada de aquisição e análise de informações e que essa capacidade é rapidamente saturada. Por isso diversas ferramentas têm sido usadas para examinar os dados que possuem, no entanto, a maioria dos analistas tem reconhecido que existem padrões, relacionamentos e regras escondidos nestes dados que não podem ser encontrados utilizando estes métodos tradicionais. Para Newing (1996) a resposta é usar softwares de mineração de dados que utilizam algoritmos matemáticos avançados para examinar grandes volumes de dados detalhados.

A necessidade de transformar a ‘montanha’ de dados armazenados em informações significativas é óbvia, entretanto, a sua análise ainda é demorada, dispendiosa, pouco automatizada e sujeita a erros, mal entendidos e falta de precisão (Newing, 1996). A automatização dos processos de análise de dados, com a utilização de softwares ligados diretamente à massa de informações, tornou-se uma necessidade (Figueira, 1998). Este motivo deve ser o responsável pelo crescimento do mercado de tecnologias de informação.

O mercado de mineração de dados tem crescido consideravelmente nos últimos anos. Conforme Krivda (1996 *apud* Niederman, 1997), em 1996 havia mais de cinqüenta produtos de mineração de dados oferecidos no mercado. Fayad (1996 *apud* Niederman, 1997), chama a atenção para o fato de que existem poucas ferramentas de mineração de dados bem desenvolvidas; o autor salienta que a maioria delas não foi testada em uma variedade de ambientes, que a maioria não é robusta quanto à falta de dados e ao surgimento de erros, e que não está claro o quanto elas podem ser utilizadas por outras

pessoas – que não sejam os seus desenvolvedores. Daí a importância da avaliação de ferramentas disponíveis no mercado utilizando-se bases de dados reais.

Devido ao número de tarefas e de ferramentas de mineração de dados existentes serão tratadas, neste trabalho, somente as ferramentas de mineração de dados que utilizam a tarefa associação ou análise de cesta de supermercado (*market basket analysis*).

Para desenvolver este documento, trata-se no capítulo dois dos objetivos desta pesquisa, no capítulo três do referencial teórico pertinente, no capítulo quatro do método de pesquisa utilizado, no capítulo cinco dos resultados, falando sobre ferramentas de mineração de dados, supermercados e da aplicação das ferramentas às bases de dados, e, por fim, no capítulo seis, são tratadas as considerações finais do trabalho, ou seja, as conclusões, limites, contribuições e perspectivas futuras de pesquisa.

Capítulo 2

Objetivos

Sabendo-se do número de ferramentas de mineração de dados disponíveis no mercado e da falta de casos demonstrando os resultados da sua utilização, este estudo tem os objetivos explicitados a seguir.

2.2.1 Objetivo Geral

Avaliar algumas ferramentas de mineração de dados como fonte de dados relevantes para a tomada de decisão, buscando verificar o cumprimento, por parte das ferramentas, das promessas desta tecnologia.

2.2.2 Objetivos Específicos

- Identificar, através da revisão de literatura, as promessas das ferramentas de mineração de dados e verificar se tais promessas são cumpridas pelas ferramentas.
- Identificar e caracterizar algumas ferramentas de mineração de dados disponíveis no mercado.
- Comparar os tipos de resultados gerados pelas diferentes ferramentas.
- Relatar problemas encontrados durante a aplicação destas ferramentas.

Capítulo 3

A utilização de técnicas de mineração de dados como fonte de informação para a tomada de decisão

Há algum tempo se fala em mineração de dados ou redes neurais, como sendo uma segunda geração dos sistemas especialistas (Santos e Becker, 1990).

Mineração de dados é uma etapa na descoberta do conhecimento em bancos de dados que promete analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis. Para isto são utilizadas técnicas que envolvem métodos matemáticos, algoritmos e heurísticas para descobrir padrões e regularidades entre os dados pesquisados (Brusso, 1998).

Harrison (1998) definiu um padrão como sendo uma afirmação sobre uma distribuição probabilística e pode ser demonstrado, principalmente na forma de regras, fórmulas e funções.

As informações obtidas através da mineração de dados, se forem relevantes, poderão ser utilizadas na tomada de decisão das empresas.

3.1 Dados, informação e conhecimento

Suliman Jr. e Souza (1997) afirmam que os conceitos de dados, informação e conhecimento podem variar. Para os autores existe uma hierarquia de complexidade em que os dados constituem a parte mais simples dessa hierarquia e o conhecimento constitui a parte mais complexa.

Basicamente, se é atribuído algum significado especial a um dado, este se transforma em informação. Se os especialistas no domínio do problema elaboram uma

norma (regra), a interpretação do confronto entre esta informação e essa regra constitui um conhecimento (Suliman Jr. e Souza, 1997). Para Freitas *et. al.* (1997) a informação é considerada como um dado dotado de relevância e propósito, para cuja conversão é necessário conhecimento.

3.2 Administração e Decisão

Newman, Summer e Warren (*apud* Albertin, 1998) citam uma concepção de administração como um processo de aplicação de princípios e de funções para o alcance de objetivos. Na abordagem dada por Dale (*apud* Albertin, 1998) para a administração, são apresentadas cinco funções essenciais: planejamento, organização, pessoal, direção e controle.

Precisamos tomar decisões a todo momento. Estas decisões vão desde as mais simples até as mais complexas, ou seja, aquelas que necessitam de um tratamento mais aprofundado, as quais, se não forem bem estruturadas, podem trazer conseqüências desastrosas. O mesmo acontece nas empresas. É preciso que haja disponibilidade de informações de qualidade que auxiliem aos executivos no momento da tomada de decisão. As tecnologias da informação começam, então, a aparecer no intuito de propiciar o suporte necessário aos tomadores de decisão das empresas que buscam um diferencial em relação aos seus concorrentes.

Nem todas as informações apresentam importância para uma tomada de decisão. Umas são mais importantes, mais relevantes do que outras. Suliman Jr. e Souza (1997) definem relevância como grau de importância que uma informação possui para a tomada de decisão.

A racionalidade, segundo Freitas *et. al*. (1997), se ocupa da seleção de alternativas que mais se encaixem em algum sistema de valores e são, até certo ponto, uma aceitação do razoável. Alter (1996) também nos fala sobre a escolha de uma alternativa satisfatória ao invés de uma alternativa ótima. Segundo este autor esta idéia é consistente com a teoria da limitação da racionalidade, segundo a qual as pessoas decidem num período de tempo limitado, baseadas em informações limitadas a com uma habilidade limitada para processar estas informações.

3.3 Banco de dados

Freedman (1995) define banco de dados e base de dados como qualquer área eletrônica de dados, ou seja, qualquer coleção de dados armazenada eletronicamente.

Date (1991) diz que um sistema de banco de dados é basicamente um sistema de manutenção de registros por computador – ou seja, um sistema cujo objetivo global é manter as informações e torna-las disponíveis quando solicitadas. Trata-se de qualquer informação considerada como significativa ao indivíduo ou à organização servida pelo sistema – em outras palavras, que seja necessária ao processo de tomada de decisão daquele indivíduo/organização.

Um dos propósitos dos bancos de dados é recuperar informação de maneira eficiente. A informação recuperada, em alguns casos, pode não ser a mesma contida no banco de dados, mas uma informação inferida desse mesmo banco. Existem duas técnicas de inferência destacadas por Holsheimer e Kersten (*apud* Ávila, 1998):

- dedução - técnica de inferir informações que são conseqüências lógicas das informações contidas na base de dados;
- indução - técnica de inferir informações que são generalizações das informações contidas na base de dados.

Para Ávila (1998) a diferença mais importante entre dedução e indução é que a dedução resulta em descrições absolutamente corretas em relação ao mundo real descrito pela base de dados, enquanto que indução resulta em descrições que são suportadas pela base de dados, porém não são necessariamente verdadeiras sobre o mundo real. O autor afirma que a capacidade em inferência de informações em bases de dados está cada vez mais distante da capacidade humana devido à crescente evolução no tamanho das bases de dados; isto explica por que a maioria dos Sistemas Gerenciadores de Bancos de Dados (SGBD) auxiliam na dedução de informações, mas nenhum deles suporta a indução de informações.

3.4 A descoberta do conhecimento em bases de dados (DCBD) e mineração de dados

A pesquisa em descoberta do conhecimento em bases de dados tem crescido e atraído esforços, baseada na disseminação da tecnologia de bancos de dados e na premissa de que as grandes coleções de dados hoje existentes podem ser fontes de conhecimento útil, que está implicitamente representado e pode ser extraído (Feldens, 1997).

O alto desenvolvimento das tecnologias de bancos de dados fez com que hoje a capacidade de coletar e armazenar informações supere os recursos para efetivamente utilizar a informação armazenada (Feldens e Castilho, 1997). As tecnologias de coleta e armazenamento de informações têm evoluído muito nas últimas décadas, por outro lado o processamento desta informação tornou-se cada vez mais difícil de ser realizado, podendo-se dizer que as capacidades de coleta e armazenamento de dados atualmente superam os recursos para efetivamente utilizar a informação.

A descoberta de conhecimento em bases de dados é o processo de extração de conhecimento novo, útil e interessante a partir de bases de dados. Esta é uma das áreas que investem no desenvolvimento de tecnologias mais eficientes para a recuperação de informações, ao projetar, implementar e validar sistemas para a extração de conhecimento útil a partir de bases de dados (Fayyad, 1996, *apud* Feldens e Castilho, 1997)

O processo DCBD tem natureza iterativa e interativa. É dito iterativo (Feldens, 1997) por ser composto de uma série de etapas seqüenciais, podendo haver retorno a etapas anteriores, isto é, as descobertas realizadas (ou a falta delas) podem levar a novas hipóteses de descoberta. Nesse caso, o usuário pode decidir pela retomada dos processos de mineração, ou uma nova seleção de atributos, por exemplo, para validar hipóteses que surgiram durante o processo, por isso é dito interativo.

O processo DCBD é cooperativo (Brusso, 1998) entre humanos e computadores. Os humanos projetam as bases de dados, descrevem problemas e definem objetivos. Os computadores processam os dados, procuram por padrões que coincidem com as metas estabelecidas.

As etapas do processo DCBD são mostradas na Figura 3.

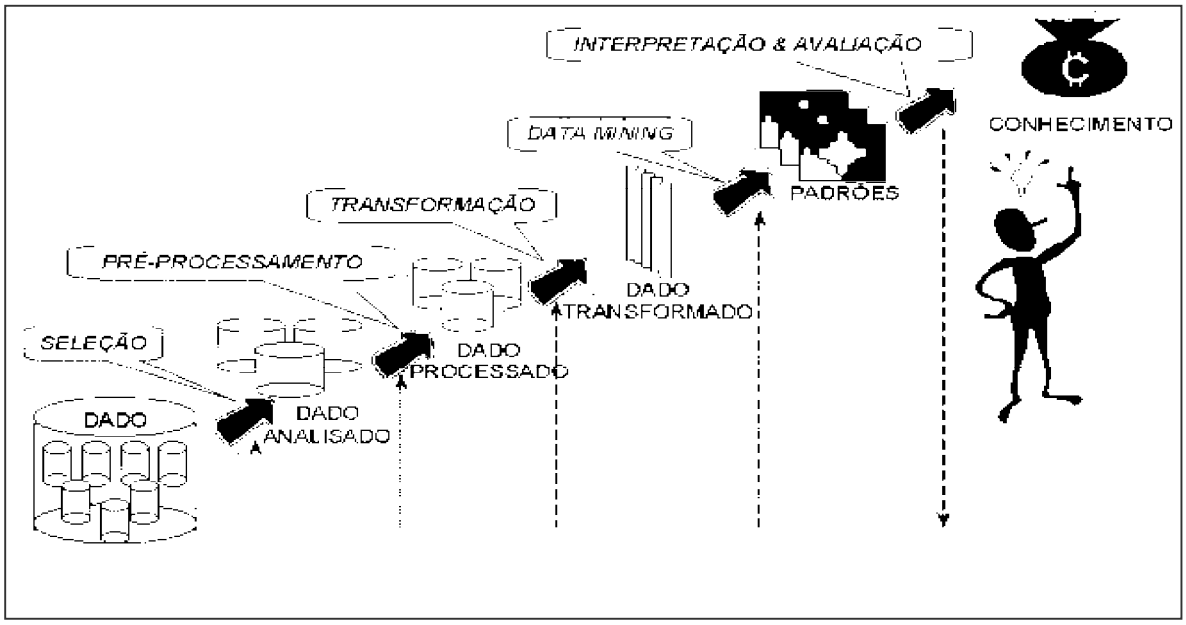

Figura 3 - Etapas do processo DCBD (Fonte: Fayyad, 1996, apud Pilla *et. al,* 1998)

As etapas que fazem parte do processo de DCBD são (Figueira, 1998):

- seleção - agrupamento organizado de uma massa de dados;
- pré-processamento - limpeza dos dados visando adequá-los aos algoritmos (integração de dados heterogêneos, eliminação de incompletude dos dados, repetição de registros, problemas de tipagem, etc.);
- transformação - armazena adequadamente os dados visando facilitar o uso das técnicas de mineração de dados;
- mineração de dados (data mining) - pode ser considerada o núcleo da descoberta de conhecimento em bases de dados, consistindo na aplicação de algoritmos de extração de padrões a partir de dados;
    - interpretação e avaliação - interpretação das descobertas, realizadas pela mineração de dados, feita pelos analistas de mineração.

Devido à importância da etapa de mineração, o termo 'mineração de dados' tem sido utilizado para identificar todo o processo, como um sinônimo para a descoberta de conhecimento em bases de dados.

A expressão mineração de dados é comumente usada por estatísticos, analistas de dados e pela comunidade *MIS* (*Management Information Systems*), gerentes de sistemas de

informações, enquanto DCBD é mais usada pelos pesquisadores de inteligência artificial (Suliman Jr. e Souza, 1997).

As fases do processo de descoberta de conhecimento em banco de dados, segundo Suliman Jr. e Souza (1997), são identificadas como:

a) desenvolver a compreensão do domínio da aplicação, o conhecimento anterior relevante e os objetivos do usuário final;
b) criar um conjunto-alvo de dados em que a prospecção deverá ser efetuada;
c) realizar a redução e projeção de dados, reduzindo o número efetivo de variáveis consideradas ou encontrar representações não variáveis para os referidos dados;
d) escolher as tarefas de mineração de dados: decidindo se o objetivo do processo DCBD é a classificação, regressão, clusterização ou outro;
e) escolher os algoritmos de mineração de dados, selecionando métodos para uso na busca de padrões nos dados;
f) mineração de dados;
g) interpretação dos padrões obtidos;
h) consolidação do conhecimento.

Mineração de dados tem sido definida como a extração não trivial da informação importante, implícita, previamente desconhecida, de dados. Ela usa o aprendizado de máquina, técnicas estatísticas e de visualização para descobrir e apresentar o conhecimento em uma forma facilmente compreensível pelos humanos.

Berry e Linoff (1997) definem mineração de dados como a exploração e análise, por meio automático ou semi-automático, de grandes quantidades de dados no intuito de descobrir padrões e regras.

Mineração de dados tem sido descrita como a interseção entre a inteligência artificial, aprendizagem da máquina e tecnologias de bancos de dados. Muitas vezes a meta é construir automaticamente um modelo de software que prediga um valor de saída dado um conjunto de valores de entrada. Uma variedade de técnicas podem ser usadas, e cada uma tem sua estrutura própria.

Para Moxon (1996) uma diferença significativa entre mineração de dados e outras ferramentas de análise é a abordagem que elas utilizam na exploração de relacionamentos entre os dados. Muitas ferramentas de análise disponíveis suportam uma abordagem

baseada na verificação, na qual o usuário cria hipóteses sobre específicos relacionamentos entre os dados e então utiliza as ferramentas para verificar ou refutar tais hipóteses. Esta abordagem confia na intuição do analista para propor hipóteses interessantes e refina a análise baseada nos resultados das consultas potencialmente complexas feitas à base de dados. Já a mineração de dados fica responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude aos resultados (Figueira, 1998).

3.4.1 Por que utilizar mineração de dados?

Vivemos em um mundo em que um dos fatores mais fortes de competitividade para qualquer empresa, em qualquer ramo de negócios, é o uso da informação e da tecnologia da informação (Torres, 1995).

As empresas têm construído sistemas para colecionar dados. O próximo desafio é a interpretação destes dados (Newing, 1996)

Kotler (1998) afirma que na sociedade da informação, de hoje, o desenvolvimento de informações confiáveis pode proporcionar à empresa um salto sobre suas concorrentes; para ele, a administração deve desenvolver e administrar informações para conhecer as mudanças de desejos dos consumidores, dos canais de distribuição, e as novas iniciativas dos concorrentes, etc.

Os pequenos varejistas usam o conhecimento do cliente para inspirar sua fidelidade (Harrison, 1998). Uma empresa pequena constrói seus relacionamentos com os clientes atendendo as suas necessidades, lembrando suas preferências e aprendendo através das interações passadas como servi-los melhor no futuro (Berry e Linoff, 1997).

Como uma grande empresa pode realizar tais procedimentos, se as interações existentes geralmente ocorrem com funcionários diferentes? Então, como a empresa poderá atender suas necessidades, lembrar-se de suas preferências e aprender com as interações passadas? O que pode substituir a intuição da pessoa que conhece os clientes por nome, fisionomia e voz, e lembra-se dos seus hábitos e preferências? Através da aplicação inteligente da tecnologia da informação, mesmo a maior empresa pode vir a ficar próxima dos seus clientes. Conforme Harrison (1998) o *data warehouse* fornece a memória para a empresa, mas ele salienta que memória sem inteligência tem pouco uso.

> *“A inteligência nos permite vasculhar nossa memória observando padrões, inventando regras, tendo novas idéias para fazer previsões sobre o futuro”(Harrison, 1998, p. 154).*

Segundo Feldens *et. al.* (1997) os sistemas de mineração de dados são capazes de aprender e apoiar a realização de descobertas a partir de bases de dados. Estes sistemas podem auxiliar o processo, analisando volumes muito grandes de dados, evidenciando relacionamentos difíceis de se perceber, muitas vezes revelando situações inesperadas, trazendo à tona problemas com a qualidade de serviço/produto ou das próprias informações, possíveis de erros nas bases de dados e até fraudes.

Com o rápido crescimento da informatização, da automação dos processos, e por conseqüência, da quantidade de informações armazenadas, o desenvolvimento de ferramentas eficientes de mineração de dados se tornou um desafio importante em diversas áreas de pesquisa, como em bancos de dados, estatística, inteligência artificial e aprendizado de máquina, entre outros (Figueira, 1998).

Um exemplo da utilização de mineração de dados: a rede americana Wal-Mart descobriu que as pessoas que vão às suas lojas às quintas-feiras para comprar fraldas Huggies tendem a adquirir dezenove itens adicionais. Assim, toda quinta-feira a Wal-Mart altera a disposição dos produtos de suas lojas para assegurar que os compradores de Huggies encontrem os tais dezenove produtos (Menconi, 1998).

3.4.2 Ciclo Virtuoso de mineração de dados

Berry e Linoff (1997) afirmam que o ciclo virtuoso de mineração de dados reconhece que mineração de dados é um passo num processo que requer ganho de conhecimento através do entendimento crescente dos consumidores, mercados, produtos e competidores para os processos internos. Este é um processo contínuo que traz resultados a toda hora.

O ciclo virtuoso de mineração de dados é composto por quatro estágios:

- identificação do problema do negócio;
- utilização de técnicas de mineração de dados para transformar dados em informações;
- ação a partir da informação;

- medição dos resultados.

O ciclo virtuoso de mineração de dados, conforme Harrison (1998) está demonstrado na Figura 4.

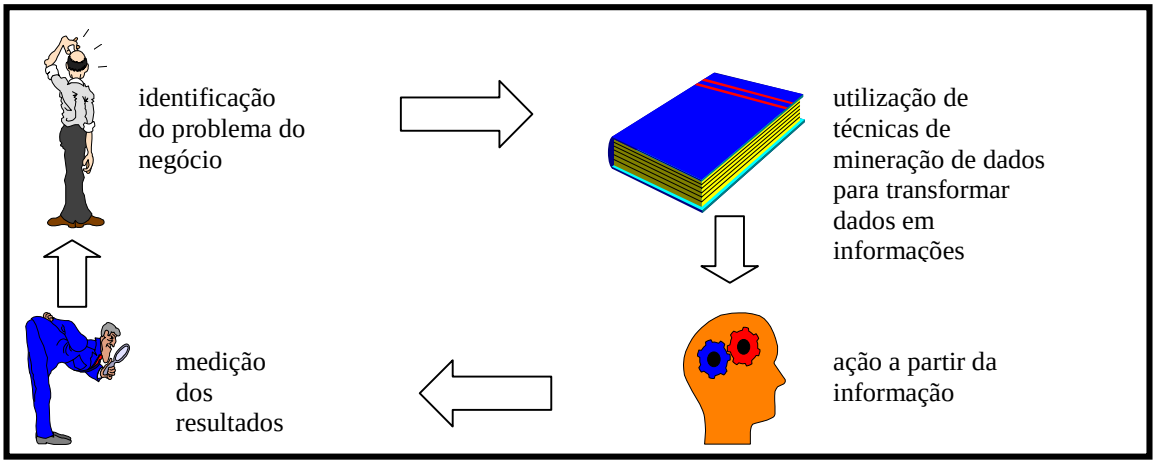

Figura 4 - Ciclo virtuoso de mineração de dados. Fonte: Harrison (1998)

3.4.3 Metodologia de mineração de dados

Segundo Berry e Linoff (1997) há três variações de metodologia para mineração de dados. Os autores apresentam cada uma das variações passo a passo:

a) Teste de hipóteses

1º. criar hipóteses;
2º. definir os dados necessários para testar as hipóteses;
3º. alocar os dados;
4º. preparar os dados para análise;
5º. desenhar os modelos computacionais e as questões dos bancos de dados para confrontar as hipóteses com os dados;
6º. avaliar os resultados dos modelos e das questões;
7º. agir baseado nos resultados da mineração de dados;
8º. medir os efeitos das ações tomadas;

9º. reiniciar o processo de mineração de dados tirando vantagem dos novos dados gerados a partir das ações tomadas.

b) Descoberta de conhecimento direto

1º. identificar fontes de dados preclassificados;
2º. preparar os dados para análise;
3º. selecionar técnicas apropriadas de descoberta de conhecimento baseado em características dos dados e na meta da mineração de dados;
4º. dividir os dados em conjuntos de formação, teste e avaliação;
5º. usar o conjunto de dados de formação para construir um modelo computacional;
6º. afinar o modelo aplicando-o ao conjunto de dados de teste;
7º. avaliar a acuracidade do modelo aplicando-o ao conjunto de dados de avaliação;
8º. agir baseado nos resultados da mineração de dados;
9º. medir o efeito das ações tomadas;
10º. reiniciar o processo de mineração de dados tirando vantagem dos novos dados gerados através das ações tomadas.

c) Descoberta de Conhecimento Indireto

1º. identificar fontes de dados disponíveis;
2º. preparar os dados para análise;
3º. selecionar técnicas apropriadas de descoberta de conhecimento indireto baseado em características dos dados e na meta da mineração de dados;
4º. usar a técnica selecionada para descobrir estruturas escondidas nos dados;
5º. identificar alvos potenciais para a descoberta de conhecimento indireto;
6º. gerar novas hipóteses a serem testadas.

3.4.4 Tarefas de mineração de dados

A mineração de dados pode desempenhar uma série limitada de tarefas, e apenas sob certas circunstâncias (Berry e Linoff, 1997).

Harrison (1998) identificou seis tarefas de mineração de dados. São elas: classificação, estimativa, previsão, associação ou análise de cesta de supermercado, segmentação ou *clustering* e descrição.

a) Classificação

Segundo Berry e Linoff (1997) a classificação é a tarefa mais comum de mineração de dados. Ela consiste em examinar os aspectos de um objeto e ligá-lo a uma das classes predefinidas. Os objetos a serem classificados são geralmente representados por registros nos bancos de dados e a ação da classificação consiste em atualizar cada registro preenchendo um campo com o código da classe. Conforme os autores, a classificação trata com valores discretos.

A tarefa de classificação é caracterizada por uma boa definição das classes, adquirida em um conjunto de exemplos pré-classificados (Harrison, 1998).

Entre os algoritmos de classificação, dois são largamente utilizados: árvores de decisão e redes neurais.

Alguns exemplos de classificação, segundo Harrison (1998):

- atribuir palavras-chave a artigos jornalísticos;
- classificar pedidos de créditos como de baixo, médio e alto risco;
- determinar que número telefônico corresponde ao fax;
- esclarecer pedidos de seguro fraudulentos.

b) Estimativa

Harrison (1998) afirma que a estimação trabalha com resultados contínuos. Tendo algum dado de entrada, nós usamos a estimativa para estipular um valor a uma variável contínua desconhecida, tal como renda, altura ou limite de cartão de crédito.

Alguns exemplos de estimativa, segundo Harrison (1998):

- estimar o número de filhos em uma família;
- estimar a renda total de uma família;
- estimar o valor em tempo de vida de um cliente;

- estimar a probabilidade com que alguém responderá ao pedido de transferência de saldo.

c) Previsão

Previsão é o mesmo que classificação ou estimativa, exceto pelo fato de que os registros são classificados de acordo com algum comportamento futuro previsto ou valor futuro estimado (Harrison, 1998).

Na previsão, a única forma de checar a acuracidade da classificação é esperar e ver (Berry e Linoff, 1997).

Alguns exemplos de previsão, conforme Harrison (1998):

- previsão da quantia de dinheiro que um cliente utilizará caso seja oferecido a ele um certo limite de crédito;
- previsão de quais clientes sairão nos próximos seis meses;
- previsão de quais assinantes de telefone usariam um serviço extra, como segmentação por telefone, CPA, ou redirecionamento de ligação.

d) Associação ou análise da cesta de supermercado

Conforme Berry e Linoff (1997) a tarefa da associação é determinar quais são os itens que são vendidos em conjunto, por exemplo. Assim, a associação pode ser usada para identificar oportunidades de vendas de pacotes de produtos ou de produtos e serviços.

Devido ao fato de a tarefa análise de cesta de supermercado ser o objeto de estudo deste trabalho, ela será detalhada no Capítulo 7.

e) Segmentação ou *Clustering*

Segmentação ou *clustering* é a tarefa de segmentar uma população heterogênea em um número maior de subgrupos homogêneos ou clusters. No *clustering* não há classes predefinidas (Berry e Linoff, 1997).

Na segmentação os registros são agrupados de acordo com a semelhança. Depende de quem está analisando determinar qual o significado que cada um dos segmentos resultantes terá (Harrison, 1998).

A segmentação é freqüentemente um dos primeiros passos na análise de mineração de dados (FAQ, 1996).

f) Descrição

Conforme Berry e Linoff (1997), às vezes o objetivo da mineração de dados é simplesmente descrever o que está acontecendo em uma base de dados complicada no intuito de aumentar nosso entendimento sobre as pessoas, produtos ou processos que produziram os dados.

Para Harrison (1998), a divergência de gênero na política americana é um exemplo de como uma simples descrição “o número de mulheres que apóiam os democratas é maior do que o de homens” pode provocar grande interesse e estudos por parte de jornalistas, sociólogos, economistas e cientistas políticos, sem contar os próprios candidatos.

3.4.5 Técnicas de mineração de dados

Há muitas técnicas diferentes de mineração de dados. A técnica a ser usada é determinada pelo tipo de informação que você está tentando determinar através dos dados. As técnicas devem ser aplicadas nas áreas corretas e nos dados corretos.

Berry e Linoff (1997) salientam que nenhuma técnica resolve todos os problemas de mineração de dados. A familiaridade com uma variedade de técnicas é necessária para encontrar o melhor caminho para resolver estes problemas.

Harrison (1998) identificou as seguintes técnicas de mineração de dados: análise de seleção estatística, MBR, algoritmos genéticos, detecção de agrupamentos, análise de vínculos, árvores de decisão e indução de regras e redes neurais artificiais.

a) Análise de seleção estatística

É uma forma de agrupamento usada para encontrar grupos de itens que tendem a ocorrer em conjunto em uma transação ou seleção estatística. Como técnica de agupamento, é útil quando desejamos saber quais itens ocorrem ao mesmo tempo ou em uma seqüência particular. A informação resultante pode ser usada para vários objetivos, como planejar a arrumação de lojas, criar “pacotes” de produtos, entre outros (Harrison, 1998).

b) Raciocínio baseado em memória (MBR)

O MBR (Memory Based Reasoning) ou raciocínio baseado em memória é uma técnica que usa exemplos conhecidos como modelo para fazer previsões sobre exemplos desconhecidos (Harrison, 1998).

Berry e Linoff (1997) dizem que o MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão.

Segundo Harrison (1998), uma das maiores vantagens do MBR é a habilidade de ser executado em qualquer fonte de dados, mesmo sem modificações. Para ele, os dois elementos-chave do MBR são a função de distância usada para encontrar os vizinhos mais próximos para fazer uma previsão. O autor identifica outra vantagem principal do MBR: sua habilidade de aprender sobre novas classificações simplesmente introduzindo novos exemplos no banco de dados.

c) Algoritmos Genéticos

Os algoritmos genéticos aplicam mecanismos de seleção genéticos e naturais para uma busca usada para encontrar conjuntos de parâmetros ótimos que descreve uma função preditiva. É usado para mineração de dados direta (Berry e Linoff, 1997).

Harrison (1998) diz que os algoritmos genéticos usam os operadores de seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções.

d) Detecção de agrupamentos

Harrison (1998) define esta técnica como a construção de modelos que encontram registros de dados semelhantes e diz que essas reuniões por semelhança são chamadas grupos (clusters).

Segundo Berry e Linoff (1997), a detecção de agrupamentos é uma mineração de dados indireta, uma vez que a meta é encontrar similaridades não conhecidas previamente.

Reunir dados é uma boa maneira de começar qualquer análise. Agrupar por semelhança pode fornecer o ponto de partida para saber o que há nos dados e descobrir como usá-los melhor (Harrison, 1998).

e) Análise de vínculos

Harrison (1998) afirma que a análise de vínculos segue as relações entre registros para desenvolver modelos baseados em padrões nas relações.

O mesmo autor salienta que como técnica de mineração de dados, a análise de vínculos não é muito compatível com a tecnologia de bancos de dados relacionais. A maior área onde é aplicada, segundo ele, é a área policial, onde pistas são ligadas entre si para solucionar crimes.

Conforme Berry e Linoff (1997), as poucas ferramentas disponíveis enfocam com maior freqüência a visualização dos vínculos ao invés de analisar os padrões.

f) Árvores de decisão e indução de regras

As árvores de decisão são usadas para a mineração de dados direta (Harrison, 1998), particularmente para a classificação. As árvores de decisão são um modelo poderoso produzido por uma classe de técnicas que inclui árvores de regressão e de classificação e indução qui-quadrado automática (Berry e Linoff, 1997).

Harrison (1998) identificou como uma das principais vantagens das árvores de decisão a facilidade de explicação de seu modelo, devido a sua forma de regras explícitas.

g) Redes Neurais

As redes neurais são provavelmente a técnica de mineração de dados mais comum, talvez sinônimo de mineração de dados para algumas pessoas (Harrison, 1998). Elas têm sua origem em pesquisas neurológicas (Almeida, 1995), e seu modelo de base é o cérebro humano. Como no cérebro humano, as redes neurais possuem neurônios interconectados de modo que os dados os percorram. Esses neurônios transmitem informação através de sinapses ou conexões.

O conceito-chave das redes neurais é a utilização de dados na criação de bases de conhecimentos. As redes neurais, ao contrário dos sistemas especialistas não precisam de um especialista para a criação da sua base de conhecimentos. Não trabalham com regras, sua aquisição é feita automaticamente a partir de exemplos coletados em bancos de dados (Almeida, 1995).

Ao examinar repetidamente milhares de registros de dados, o software pode desenvolver um modelo estatístico poderoso descrevendo os relacionamentos e os padrões de dados importantes - nada que um pesquisador humano tenha tempo (ou capacidade visual) de fazer de maneira rigorosa e consistente (Kotler, 1998).

Nas redes neurais não há uma codificação de programas a fim de introduzir o conhecimento sobre um problema. Por um processo iterativo (processo de aprendizado) as redes neurais lêem os exemplos fornecidos sobre um problema e criam assim um modelo de resolução. Elas são bem adaptadas a dois tipos de tarefas: reconhecimento de formas e generalização (Almeida, 1995).

Harrison (1998) diz que não há uma técnica que resolva todos os problemas de mineração de dados. A familiaridade com as técnicas é necessária para proporcionar a melhor abordagem de acordo com os problemas apresentados. O autor afirma que a escolha das técnicas de mineração de dados dependerá da tarefa específica a ser executada e dos dados disponíveis para análise.

A técnica a ser escolhida dependerá do tipo de dados que temos e do tipo de informação que estamos tentando determinar (FAQ, 1996).

3.4.6 Aplicações

A mineração de dados tem se mostrado importante em um grande número de aplicações, incluindo segmentação de mercado, detecção de fraude em cartões de crédito, análises financeiras e de investimentos, detecção e predição de erros em grandes negócios, análise de informações, ferramentas inteligentes e limpeza em bases de dados (Greenfeld, 1996).

Feldens (1997) apresentou algumas das aplicações atuais para a mineração de dados, além das citadas acima, são elas: marketing e melhoria do processo industrial.

3.4.7 Problemas encontrados em mineração de dados

Podem surgir problemas quando se utiliza bases de dados em tarefas de aprendizado. Alguns desses problemas são (Ávila, 1998):

- informação incompleta - nem todas as variáveis podem estar presentes em uma determinada base de dados;
- ruído - podem existir atributos baseados em medidas ou julgamentos subjetivos, o que facilita a atribuição de valores errados a esses atributos;
- valores desconhecidos - existem valores desconhecidos de atributos;
- relações tamanho das bases de dados - as bases de dados utilizadas em aprendizado de máquina são diferentes das bases de dados do mundo real, pois uma base de dados em aprendizado de máquina com milhares de exemplos é considerada grande, porém, uma base de dados do mundo real pode possuir um número maior de objetos e de informações por objeto. Em mineração de dados quanto mais informação disponível, maior será a chance de se construir descrições que reflitam informações verdadeiras, porém, o espaço de busca por descrições também torna-se muito maior. Conforme Santos e Becker (1990), um dos aspectos que gerou ceticismo por parte dos usuários potenciais de Inteligência Artificial foi o enorme número de equações e relações necessárias para minerar grandes bases de dados;

Além dos problemas verificados por Ávila (1998), Pressman (1995) considera a execução incorreta da fase de testes no período de desenvolvimento dos sistemas como um dos maiores problemas das ferramentas de mineração de dados disponíveis no mercado. O

autor afirma que a atividade de teste exige que o desenvolvedor descarte noções preconcebidas da ‘corretitude’ do sistema que ele acabou de desenvolver e supere um conflito de interesses que ocorre quando erros são descobertos.

Beizer (*apud* Pressman, 1995) diz que há um mito segundo o qual, se fôssemos realmente bons para programar, não haveria *bugs* a ser procurados. Se pudéssemos realmente nos concentrar, se todos usassem programação estruturada, projeto *top-down*, tabelas de decisão, se os programas fossem escritos em uma única e fácil linguagem de programação, então não haveria *bugs*. Assim segue o mito. Existem *bugs*, diz o mito, porque somos ruins naquilo que fazemos; e, se somos ruins nisso, devemos sentir-nos culpados por isso. Por conseguinte, a atividade de teste e o projeto de casos de teste são uma admissão de falha, o que promove uma boa dose de culpa. E o tédio de testar é apenas uma punição pelos erros cometidos.

#####

Neste capítulo, foi abordada a teoria inerente à mineração de dados como fonte de informação para a tomada de decisão. Os itens abordados são a base para o estudo, constituindo elementos importantes e merecedores de consideração.

Esta revisão de literatura nos permitiu verificar as promessas da tecnologia de Mineração de dados, que não são poucas, porém não encontramos casos concretos que confirmem tais promessas.

As promessas encontradas através da revisão de literatura são mostradas no quadro abaixo:

Quadro 1 – Promessas da mineração de dados encontradas na revisão de literatura

Promessa	Descrição	Fonte
P1	Analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis	Brusso, 1998
P2	Trabalhar com grandes bases de dados	Niederman, 1997
P3	Retornar conhecimento novo e relevante	Niederman, 1997
P4	A mineração de dados é responsável pela geração de hipóteses	Figueira, 1998
P5	Encontrar padrões que não são encontrados por sistemas ditos tradicionais	Figueira, 1998 Moxon, 1996
P6	Os sistemas são capazes de aprender e apoiar à realização de descobertas a partir dos dados	Feldens *et. al.*, 1997

Através dos conhecimentos teóricos obtidos, conseguimos também, entender o processo de mineração de dados e identificamos dentre as três metodologias existentes, aquela que poderia ser utilizada (descoberta de conhecimento indireto) e identificamos, também a tarefa adequada à base de dados disponível (análise de cesta de supermercado), a utilização dos mesmos será detalhada nos Capítulos 4 e 5.

Capítulo 4

Método de pesquisa

Este estudo é de natureza exploratória, pois tem como objetivo a obtenção de informação sobre possibilidades práticas de realização de pesquisa em situação de vida real. Este tipo de estudo tem como objetivo principal a descoberta de idéias e intuições e pode ser utilizado para buscar familiarizar-se com o fenômeno ou conseguir nova compreensão deste (Selltiz *et. al.*, 1974).

O método utilizado é qualitativo, pois não empregou-se instrumental estatístico como base do processo de análise do problema (Richardson, 1985). O presente trabalho não pretendeu numerar ou medir unidades de categorias homogêneas.

O método caracteriza-se como um estudo de caso múltiplo: o caso de quatro ferramentas de mineração de dados as quais são aplicadas às bases de dados de uma rede de supermercados.

Boyd & Stach (*apud* Yin, 1990) salientam que no estudo de caso é dada ênfase à descrição e ao entendimento do relacionamento dos fatores de cada situação, não importando os números envolvidos. Este tipo de pesquisa, como o experimento, não representa uma amostragem, e o objetivo do observador é difundir e generalizar teorias (generalização analítica) e não enumerar freqüências (generalização estatística).

Segundo Yin (1990), quando se tem perguntas do tipo ‘como’ ou ‘porque’ um programa funcionou (ou não) os métodos de pesquisa que podem ser utilizados são estudo de caso ou experimento.

O estudo de caso é um estudo profundo de um ou poucos objetos de maneira a permitir o seu conhecimento amplo e detalhado (Gil, 1999).

Yin (*apud* Oliveira, 1999) definiu este método de pesquisa como uma forma de se fazer pesquisa social empírica ao investigar-se um fenômeno atual dentro de seu contexto

de vida real, sendo que os limites entre o fenômeno e o contexto não estão claramente definidos e na situação em que múltiplas fontes de evidências são usadas.

Para esta pesquisa foram usados dados primários - oriundos das entrevistas com os tomadores de decisão, e secundários – aqueles contidos nas bases de dados da empresa.

Abaixo é apresentado o desenho da pesquisa, procurando-se demonstrar todas as etapas do trabalho.

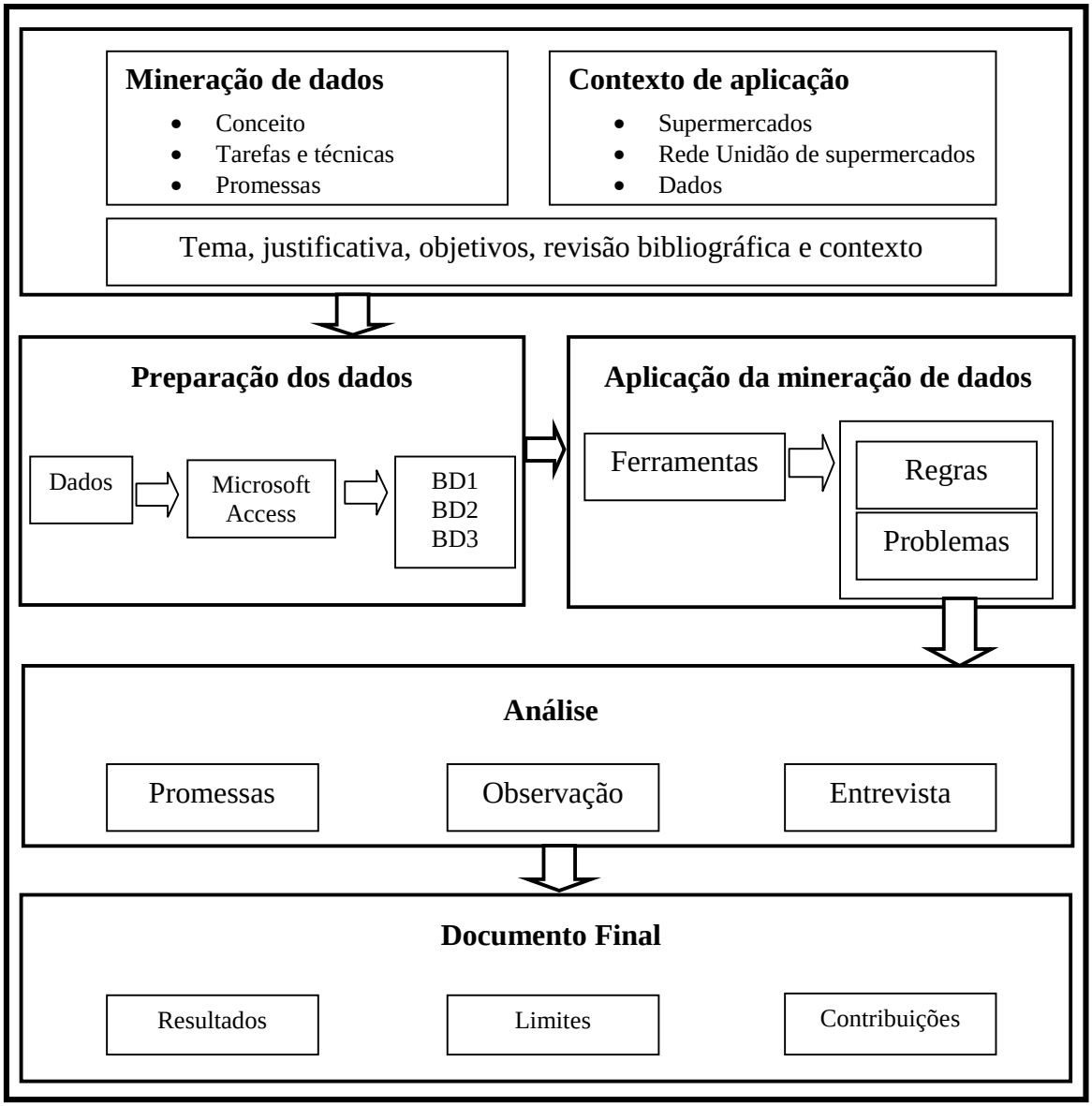

Figura 5 – Desenho da pesquisa

4.1. A escolha do contexto de aplicação

O contexto de aplicação da pesquisa poderia ser qualquer área da administração, porém foi escolhido o setor de supermercados devido à riqueza de seus dados e ao fato deste setor viver em forte competição. A empresa escolhida foi selecionada por conveniência, em função de: (1) disponibilidade de sua(s) base(s) de dados e (2) porte da empresa, considerando o ranking da AGAS (Associação Gaúcha de Supermercados), onde a empresa encontrava-se na quinta posição em 1998, conforme o anexo A.

A empresa em questão foi contactada via e-mail e posteriormente foram realizadas reuniões na sede da mesma, na cidade de São Leopoldo-RS. O contexto está melhor detalhado no Capítulo 5.

4.2 Coleta de dados

Em estudos de caso, as principais técnicas empregadas para a coleta de dados são: literatura, documentos de arquivo, entrevistas, observação – participativa ou não, experiências e mesmo artefatos (Yin, 1990). Campomar (1991) identifica alguns documentos de arquivo que servem como instrumento de coleta de dados neste tipo de pesquisa. São eles: relatórios, memorandos, cartas, mensagens, artigos e outros.

Neste trabalho, os instrumentos de coleta de dados utilizados foram:

a) literatura, para o conhecimento das promessas da tecnologia;
b) arquivos contendo os dados da movimentação de quatro lojas, durante dois meses;
c) observação participativa durante a aplicação das ferramentas;
d) entrevistas com os tomadores de decisão da empresa.

Com base na literatura existente, foi elaborado um conjunto de promessas da tecnologia de mineração de dados e buscou-se a aquisição do conhecimento necessário sobre mineração de dados, suas técnicas e ferramentas existentes.

As bases de dados utilizadas são aquelas que guardam a movimentação diária realizada pelos clientes, pois, segundo Menconi (1998), para minerar dados, é necessário arquivar toda e qualquer transação realizada pelo cliente.

a) Obtenção dos dados

Os dados começaram a ser armazenados no mês de maio e os arquivos foram enviados em discos zip[2] pelo correio (via Sedex). No início do mês de junho foi enviado o disco contendo os dados de maio e no início do mês de julho os do mês de junho, e assim sucessivamente. Em cada mês foram gerados quatro arquivos, um de cada uma das quatro lojas.

Quadro 2 - Arquivos recebidos contendo os dados das transações da empresa

Nome do Arquivo	Formato	Tamanho Kb	N° de registros	Mês	Loja
Movgeral01-mai	DBF	65.143	758.018	maio	Feitoria
Movgeral01-jun	DBF	68.333	795.138	junho	Feitoria
Movgeral01-jul	DBF	62.870	731.569	julho	Feitoria
Movgeral10-mai	DBF	26.403	307.231	maio	Canela
Movgeral10-jun	DBF	26.292	305.937	junho	Canela
Movgeral10-jul	DBF	27.562	320.714	julho	Canela
Movgeral11mai	DBF	24.037	279.691	maio	Sapiranga
Movgeral11-jun	DBF	25.135	292.472	junho	Sapiranga
Movgeral11-jul	DBF	20.646	240.237	julho	Sapiranga
Movgeral-28-mai	DBF	24.745	287.935	maio	Viamópolis
Movgeral28-jun	DBF	25.820	300.437	junho	Viamópolis
Movgeral28-jul	DBF	26.049	303.109	julho	Viamópolis

Devido ao grande volume de dados, ficou acertado que se trabalharia somente com dados de dois meses (maio e junho).

b) Preparação dos dados (pré-processamento)

Como os dados começaram a ser armazenados para a realização deste trabalho, não houve muito esforço para prepará-los para a mineração.

[2] Mídia utilizada para armazenar grandes volumes de dados

Como o tamanho da base de dados é um fator referenciado nas promessas da tecnologias encontradas através da bibliografia, foram utilizadas três bases de dados de tamanhos diferentes para a aplicação das mesmas, assim, se poderia verificar se o volume de dados influenciava na performance dos sistemas. Estas bases estão descritas no quadro abaixo.

Quadro 3 - Descrição das três bases de dados utilizadas

Base	Nº de registros	Nº de Entidades	Média Registros/entidade	Composição da base
B1	10.000	1.801	5,55	10.000 primeiros registros do movimento de maio/00 na loja de Viamão
B2	287.935	38.842	7.41	Movimento completo da loja de Viamão no mês de maio/00
B3	3.326.859	148.115	22,46	Movimento das quatro lojas nos meses de maio/00 e junho/00

Filtrou-se o menor arquivo para que fosse gerada a base de dados com 10.000 registros (B1); a empresa selecionou a base de dados da loja de Viamão, com o movimento do mês de maio para a base de dados de tamanho médio, com 287.935 registros (B2) e uniu-se os oito arquivos para criar a base de dados com mais de 3.000.000 de registros (B3).

Foram realizadas algumas filtragens e classificações no Microsoft Access, no intuito de verificar a existência de dados incompletos e registros repetidos nas bases de dados, o que poderia gerar problemas na hora da mineração. Não foram encontradas inconsistências nos dados.

As bases de dados encontravam-se no formato vertical, usando múltiplas linhas para armazenar cada entidade (venda). As linhas para uma entidade particular são ligadas através do campo chamado cupom, o qual identifica o número do cupom fiscal da venda.

As bases de dados utilizadas têm a seguinte estrutura:

Quadro 4 - estrutura das bases de dados

Nome do Campo	Descrição	Tipo	Tamanho	Casas Decimais
Código	Código EAN do Produto	String	13	-
Descrição	Descrição do Produto	String	30	-
Qtd	Quantidade Vendida	Numérico	10	3
Valor	Preço do Produto	Numérico	11	2
PDV	Ponto de Venda	String	3	-
Cupom	Número do Cupom Fiscal	Numérico	6	-
Data	Data da Venda	String	10	-
Diasemana	Dia da Semana	Numérico	1	-
Loja	Identificação da Loja do Unidão	String	3	-

4.3 Obtenção e aplicação das ferramentas de mineração de dados

As ferramentas obtidas foram seis: Aira Data Mining®, Xaffinity®, SuperQuery Discovery Edition®, SuperQuery Office Edition®, PolyAnalyst® e CBA®. Todas foram escolhidas por conveniência, em função de: (1) realização da técnica de análise de cesta de supermercado, (2) rodar em plataforma Windows e, (3) sua disponibilização por parte de seus proprietários, sem ônus ao pesquisador. Devido ao fato dos objetos de análise deste estudo serem ferramentas de mineração de dados, as mesmas são tratadas com maior detalhamento no Capítulo 5.

Para a aplicação das ferramentas de mineração de dados foi utilizada a metodologia que Berry e Linoff (1997) chamaram de descoberta de conhecimento indireto, visto que, primeiramente, foram identificadas as bases de dados disponíveis, depois os dados foram preparados para a análise, foi escolhida a tarefa apropriada, a técnica foi aplicada e no final novas hipóteses foram geradas e testadas.

A aplicação destas ferramentas ocorreu fora das instalações da empresa.

O passo posterior foi a avaliação das ferramentas.

4.4 Análise

A análise da pesquisa foi realizada através de: observação do pesquisador, como usuário das ferramentas, ou seja, a pessoa que utilizou as ferramentas e realizou as minerações; e a avaliação dos resultados obtidos pelas ferramentas (informações geradas) efetuando-se uma entrevista com os tomadores de decisão da empresa.

4.4.1 Avaliação das ferramentas

Para a avaliação de Sistemas de Apoios a Decisão Sprague & Carlson (1982 *apud* Freitas, 1993) sugerem quatro medidas:

- de produtividade – o impacto das decisões, o tempo, o custo da tomada de decisão e de sua implantação, os resultados decorrentes;
- de processo, o impacto sobre a tomada de decisão, as alternativas, as análises, os participantes, o tempo empregado em cada fase, os dados necessários;
- de percepção, o impacto sobre os decisores, o controle exercido, a utilidade percebida, a facilidade de utilização, a compreensão do problema, a convicção na solução;
- e de produto – para avaliar o mérito técnico, o tempo de resposta, a disponibilidade, o prazo de reparação de uma pane ou falha, o custo de desenvolvimento, o custo de implantação e de manutenção, o custo de formação e o custo de aquisição dos dados.

Avaliou-se o pontencial dos sistemas na geração de informações relevantes para a tomada de decisão da organização em estudo. Para realizar esta avaliação, utilizou-se, em primeiro lugar, um instrumento de avaliação de Sistemas de Informação desenvolvido por Freitas, Ballaz & Moscarola (1994 *apud* Stumpf, 1998). O modelo de avaliação apresentado por estes autores, baseia-se em dois pilares básicos: a utilidade e a facilidade. A utilidade é ligada ao usuário enquanto a facilidade é ligada ao sistema. Sendo assim, analisamos a opinião do usuário verificando as seguintes afirmações (A1 e A2):

A1: O sistema é útil ao decisor (contribui para o processo de decisão);

A2: O sistema é fácil de utilizar (apresenta boa ergonomia, segundo o usuário).

Com relação aos dois pilares básicos, que são objeto de avaliação do sistema, utilidade e facilidade, Davis (1989 *apud* Stumpf, 1998) os define da seguinte forma: utilidade percebida é o grau atribuído, pelo usuário, ao uso efetivo de um determinado sistema, considerando a capacidade de aumentar seu desempenho no trabalho. Este conceito parte da definição da palavra útil: algo capaz de ser usado com vantagens. Ainda, na visão do autor, a utilidade percebida em um sistema de informação acontece quando o usuário acredita na existência de um relacionamento positivo entre o uso efetivo do sistema e o seu desempenho ao utilizá-lo.

A facilidade de uso refere-se ao grau de credibilidade depositado pelo usuário ao usar um determinado sistema. O uso efetivo do sistema o liberaria de determinados esforços. Sendo assim, a definição da palavra fácil está relacionada à liberdade de grandes esforços pelo usuário. Um esforço caracteriza-se por um recurso que um determinado usuário despende para realizar diversas atividades pelos quais é responsável.

Formas de avaliar a percepção dos usuários do sistema quanto a sua qualidade, impacto causado e benefícios alcançados a partir do seu uso também foram consideradas. Esta avaliação foi fundamentada em trabalhos desenvolvidos por Vandenbosch & Higgins (1996 *apud* Stumpf, 1998). Além destes aspectos, os autores desenvolveram entrevistas semi-estruturadas, as quais serviram de base, para avaliar o uso efetivo do instrumento.

4.4.2 Entrevista

O número de respondentes por parte da empresa foi estipulado pela direção da mesma, de acordo com a função exercida pela pessoa. Foram entrevistados os diretores geral e de sistemas e tecnologia.

As entrevistas foram conduzidas no mês de março de 2001.

Algumas variáveis referentes à percepção de utilidade e facilidade para o usuário foram definidas de acordo com os modelos definidos por Freitas (1993) e Stumpf (1998), permitindo, assim, a elaboração de um referencial para a realização da entrevista com o usuário dos sistemas e com os tomadores de decisão da empresa.

As informações obtidas possibilitaram a verificação das afirmações (A1 e A2) apresentadas neste capítulo. Do mesmo modo, para investigarmos as percepções dos

usuários em relação à qualidade, ao impacto e aos benefícios proporcionados pelo sistema, adotamos o modelo de entrevistas também desenvolvido por Stumpf (1998) e adaptado pela pesquisadora usando as medidas da qualidade segundo Pressman (1995) que inclui nas medidas mais amplamente utilizadas de qualidade de sistemas além da useabilidade (facilidade), a corretitude. Gilb (*apud* Pressman, 1995) definiu corretitude como o grau em que o sistema executa a função que é dele exigida. O sistema deve operar corretamente; caso contrário, oferecerá pouco valor aos seus usuários. Sua medida mais comum é o defeito, ou seja, a falta verificada em conformidade aos requisitos. Os defeitos são registrados pelos usuários do sistema depois que este foi liberado para uso geral.

Foi necessário incluir, também, as medidas de interesse objetivas e subjetivas, definidas por Brusso (2000), por tratarem-se de parâmetros de qualidade de regras de associação que, neste caso, são as informações geradas pelas ferramentas utilizadas.

Para cada uma das variáveis, dentro de cada um dos pilares utilidade, facilidade, qualidade, impacto e benefícios foram elaboradas questões para serem respondidas pelos usuários, as quais serviram para a avaliação. São elas:

Quadro 5 - Instrumento de coleta de dados

Pilar	Variável	Questões	Respondente	Referencial
Utilidade	Rapidez na realização de tarefas	Usar as informações geradas em seu trabalho permite executar suas tarefas com mais rapidez?	Tomadores de decisão	Davis (*apud* Stumpf, 1998)
	Desempenho e produtividade no trabalho	Usar as informações geradas no seu trabalho aumenta seu desempenho?	Tomadores de decisão	
		Usar as informações geradas em seu trabalho aumenta sua produtividade?	Tomadores de decisão	
	Eficácia	Usar as informações geradas em seu trabalho aumenta sua eficácia?	Tomadores de decisão	
	Facilidade no trabalho	Usar as informações geradas torna seu trabalho mais fácil de ser executado?	Tomadores de decisão	
		O sistema proporciona a você informações de que precisa para fazer seu trabalho?	Tomadores de decisão	Vandenbosch & Higgins (*apud* Stumpf, 1998)
	Utilidade	Você considera as informações geradas pelo sistema úteis ao seu trabalho?	Tomadores de decisão	
		Que informações geradas você considera inúteis?	Tomadores de decisão	
	Tempo de resposta	Como foi o tempo gasto pelo sistema para gerar as informações?	Usuário	Freitas (1993)
	Aprendizagem	As informações geradas lhe proporcionaram algum tipo de aprendizagem?	Tomadores de decisão	
	Autonomia e independência	Você conseguiu utilizar o sistema sem necessitar de auxílio/interferência de seus desenvolvedores	Usuário	
Facilidade	Aprendizado	Aprender a operar o sistema é fácil para você?	Usuário	Davis (*apud* Stumpf, 1998)
	Domínio (controle)	É fácil para você conseguir que o sistema faça aquilo que você quer que ele faça?	Usuário	
	Interação	Sua interação com o sistema é clara e compreensível	Usuário	
	Flexibilidade	Você considera o sistema flexível para interagir com ele?	Usuário	
	Habilidade	É fácil para você tornar-se um perito em usar o sistema?	Usuário	
	Facilidade de uso	Você considera o sistema fácil de usar?	Usuário	
	Funcionalidade	O sistema atendeu a sua função?	Tomadores de decisão	Freitas (1993)
	Impacto da apresentação gráfica	Era possível visualizar as informações em forma de gráficos?	Usuário	
	Qualidade das imagens gráficas	Como era a qualidade das imagens gráficas geradas pelo sistema?	Tomadores de decisão	

Qualidade	Precisão e confiabilidade	Como é a qualidade do sistema quanto à precisão e confiabilidade das informações?	Tomadores de decisão	Vandenbosch & Higgins (*apud* Stumpf, 1998)
	Oportunidade	Como é a qualidade do sistema quanto à oportunidade das informações?	Tomadores de decisão	
	Dificuldade para obter informações	É difícil para você interpretar as informações contidas no sistema?	Tomadores de decisão	
	Facilidade de interpretação	É fácil para você obter informações significativas no sistema?	Tomadores de decisão	
	Fonte de in-formação	Você considera o sistema eficiente como uma fonte de informação?	Tomadores de decisão	
	Corretitude	O sistema apresentou defeitos? Quais?	Tomadores de decisão e usuário	Pressman (1995)
	Medidas subjetivas de interesse	Em sua opinião as regras geradas têm alguma utilidade?	Tomadores de decisão	Brusso (2000)
		Em sua opinião as regras geradas eram inesperadas?	Tomadores de decisão	
	Medidas objetivas de interesse	Era possível ajustar o grau de confiança da mineração?	Usuário	
		Era possível ajustar o suporte mínimo da mineração?	Usuário	
Impacto e benefícios	Contribuição	Em sua opinião qual a contribuição do sistema para você?	Tomadores de decisão	Vandenbosch & Higgins (*apud* Stumpf, 1998)
	Impacto	Teve algum impacto na tomada de decisão?	Tomadores de decisão	
	Entendimento do negócio	O sistema proporciona a você um melhor entendimento do negócio?	Tomadores de decisão	
	Discussões dentro da organização	O sistema mudou a natureza das reuniões ou discussões na organização? De que forma?	Tomadores de decisão	
	Competitividade	Em sua opinião o sistema torna a organização mais competitiva? De que forma?	Tomadores de decisão	

####

Cada um dos pilares do instrumento de coleta de dados gerou um quadro onde foram colocadas as ferramentas, as variáveis e as suas respectivas avaliações.

Para o preenchimento destes quadros foram utilizadas escalas de presença e de satisfação/eficiência, com os valores ***nenhum***, ***pouco***, ***médio*** e ***muito***; dependendo da questão relacionada à variável em estudo.

Na escala de presença foram usados os valores ***sim*** (significando a presença de um fator) e ***não*** (significando a ausência de um fator).

Na escala de satisfação/eficiência foram usados os seguintes valores: ***nenhum*** (significando o valor mais baixo, ou seja, insatisfação total), ***pouco*** (significando uma

insatisfação alta), ***médio*** (significando uma insatisfação regular), ***muito*** (significando satisfação).

As variáveis e as questões estão mostradas no quadro 5.

A aplicação das ferramentas às bases de dados e as suas avaliações encontram-se no próximo capítulo.

Capítulo 5

Resultados

Para uma melhor explanação dos resultados desta pesquisa torna-se necessária a contextualização da mesma.

Por isso, num primeiro momento tratou-se do mercado de mineração de dados e das ferramentas utilizadas para a realização deste trabalho.

Depois é tratado o contexto de aplicação, ou seja, a rede de supermercados que cedeu suas bases de dados.

E, finalmente, é demonstrada a aplicação das ferramentas às bases de dados e os seus resultados.

5.1 O mercado de mineração de dados

Existem diversas ferramentas de mineração de dados disponíveis no mercado. Empresas renomadas e empresas desconhecidas investiram no desenvolvimento de ferramentas que utilizam esta tecnologia. Abaixo, são listadas as principais ferramentas, respectivos desenvolvedores e tipo de distribuição, segundo o catálogo organizado pelo Kdnuggets[3], separados por tarefa.

[3] http://www.kdnuggets.com

Quadro 6 – Algumas ferramentas de mineração de dados disponíveis no mercado

Ferramenta	Desenvolvedor	Distribuição
Tarefa: Associação		
Aira Data Mining®	Hycones IT	Comercial
Apriori®	Christian Borgelt	Livre
ARMiner®	Umass	Livre
CBA®	Bing Liu	Livre
Clementine®	SPSS	Comercial
Data Mining Suite®	Information Discovery Inc.	Comercial
Intelligent Miner®	IBM	Comercial
KDD Explorer Suite®	SRA	Comercial
Magnum Opus®	GI Weeb & Associates	Comercial
Mineset®	SGI	Comercial
Nuggets Suite®	Data Mining Technologies Inc.	Comercial
PolyAnalist®	Megaputer	Comercial
SuperQuery Discovery Edition®	Azmy	Comercial
SuperQuery Office Edition®	Azmy	Comercial
Wizsoft®	Wizsoft	Comercial
Xaffinity®	Exclusive Ore	Comercial
Xpertrule Miner®	Attar	Comercial
Tarefa: Classificação		
Answer Tree®	SPSS	Comercial
Clementine®	SPSS	Comercial
Darwin®	Oracle	Comercial
Knowledge Studio®	Angoss	Comercial
Market Miner®	Market Miner Inc.	Comercial
Model 1 Suite®	Unica	Comercial
PolyAnalyst®	Megaputer	Comercial
Previa Classpad®	Elseware	Comercial
Rock Convex Hull Program®	Ton Fawcett	Livre
Tarefa: Árvore de decisão		
AC2®	Alice Soft	Comercial
C4.5®	Ross Quinlan	Livre
C5.0®	Rulequest	Comercial
EC4.5®	Pisa KDD Laboratory	Livre
IND®	Nasa	Livre
Jam®	Columbia University	Livre
LMDT®	Carla E. Brodley	Livre
MLC++®	SGI	Livre
ODBC Mine®	Intelligent Systems Research	Livre
PC4.5®	New York University	
XpertRule Miner®	Attar	
Tarefa: Redes Neurais		
BrainMaker®	California Scientific Software	Comercial
DBProphet®	Trajecta	Comercial
Neural Connection 2®	SPSS	Comercial
Neural Innovation Proforma®	Neural Innovation	Comercial
Neural Net Toolbox®	Matlab	Comercial
Neural Solutions®	Neuro Solutions	Comercial
Neural Ware®	Neural Ware	Comercial
SNSS®	Informatik	Comercial
Statistica Neural Networks®	Statsoft	Comercial

Tarefa: Análise estatística		
Auto Fit®	Lava	Livre
Cubist®	Rulequest	Comercial
Data Desk®	Data Description Inc.	Comercial
DM Statware®	DM Stat-1	Comercial
Evolutionary Regression®	Evolutionary Software	Comercial
JMP®	SAS	Comercial
MARS®	Salford Systems	Comercial
Matlab®	Mathworks	Comercial
Modstat®	Robert C. Knodt	Comercial
Previa Studio®	Elseware	Comercial
R®	Cran	Livre
SAS®	SAS	Comercial
Simstat for Windows®	Mycomputer	Comercial
S-Plus®	Mathsoft	Comercial
SPSS Base®	SPSS	Comercial
Statistica®	Statsoft	Comercial
Statware Statit(R) ®	Statware	Comercial
Tarefa: Estimativa®		
TCBM®	TCBM	Comercial
WizWhy®	Wizsoft	Comercial
Xlisp Stat®	Statlib	Livre
Tarefa: Clustering		
Autoclass C®	Nasa	Livre
C Viz Cluster Visualization®	IBM	Comercial
Clustan Graphics®	Clustan	Comercial
Darwin Suite®	Oracle	Comercial
Data Mining Suite®	Information Discovery Inc.	Comercial
Ecobweb®	Yoram Reich	Livre
Intelligent Miner for Data®	IBM	Comercial
Mclust/Emclust®	University of Washington	Livre
Mine Set®	SGI	Comercial
Mixture Modeling®	David Dowe	Livre
PolyAnalist®	Megaputer	Comercial
Snob®	David Dowe	Livre
SO Mine®	Eudaptics	Comercial

Conforme Fayad (*apud* Niederman, 1997) existem muitas ferramentas disponíveis no mercado, porém, ainda não se sabe o quanto as ferramentas de mineração de dados podem ser utilizadas por outras pessoas que não sejam seus desenvolvedores.

Para a liberação da utilização das ferramentas neste trabalho, foi feita uma visita aos sites dos desenvolvedores de todas as ferramentas de mineração de dados, que realizam a tarefa associação, citadas no site Kdnuggets[4]. Através do site do desenvolvedor, foi realizado contato via e-mail solicitando a licença para a utilização da ferramenta. Cinco desenvolvedores liberaram suas ferramentas para a aplicação nesta pesquisa.

As ferramentas liberadas foram as caracterizadas no quadro abaixo:

[4] http://www.kdnuggets.com

Quadro 7 - Caracterização das ferramentas utilizadas

Ferramentas / Item	Aira Data Mining®	Xaffinity®	Poly Analyst®	Super Query Discovery Edition®	Super Query Office Edition®	CBA®
Desenvolvedor	Hycones IT	Exclusive Ore	Mega Puter	Azmy	Azmy	National University of Singapure
Ano	1996	2000	1994	*	*	*
Idioma	Português e Inglês	Inglês	Inglês	Inglês	Inglês	Inglês
Sistema Operacional	Windows	Windows	Windows	Windows	Windows	Windows
Configuração Mínima	Pentium, 32MB Ram, 20 MB HD	Pentium 266, 64 MB Ram	Pentium 266, 32 MB Ram, 20 MB HD	Pentium 266, 32 MB Ram, 20 MB HD	Pentium 266, 32 MB Ram, 20 MB HD	Pentium , 32 MB Ram, 5 MB HD
Preço (U$)	5.000	5.000	*	449,95	149,95	Livre[5]
Bases de Dados	Oracle, Sybase, Informix, SQL Server, Paradox e Dbase	SQL Server, MS Access, Oracle, Red Brick e White Cross	MS Excel, ODBC, Oracle, IBM Visual Warehouse, arquivs .csv	SQL Server, MS Access, Oracle e dBase	MS Excel e MS Access	Texto
Regras Geradas	Associativas e hierárquicas	Associativas e seqüenciais	Associativas	Associativas	Associativas	Associativas e hierárquicas
Formato dos Resutados	Regras em MS-Word e HTML	Regras em linguagem natural e em colunas	Lista de regras, histograma, gráficos 3D e 2D	Lista de regras e gráfico com os valores mais freqüentes	Lista de regras e gráfico com os valores mais freqüentes	Lista de regras

** Não foram informados valores*

Após a obtenção destas seis ferramentas verificou-se a possibilidade de aplicação das mesmas aos dados da rede de supermercados que disponibilizou a base de dados das vendas aos seus clientes.

A verificação e a aplicação estão relatadas no item 5.3.4.

Nos itens seguintes será realizada a contextualização desta pesquisa.

5.2 Os Supermercados

Segundo Rojo (1998), os supermercados são lojas com o método de auto-serviço no varejo de alimentos. Os produtos oferecidos pelos supermercados incluem uma ampla

[5] A ferramenta CBA é um software freeware

variedade de produtos como: hortifrutigranjeiros, mercearia, frios e laticínios, carnes frescas e não-alimentos básicos (perfumaria e limpeza).

De acordo com a ABRAS (Associação Brasileira de Supermercados) não há outro setor da atividade econômica no País que tenha crescido tanto, do zero ao estágio atual, em prazo tão curto (vide quadros 8 e 9). Em mais de quatro décadas, o auto-serviço brasileiro impôs-se como a forma mais moderna, econômica e racional de se adquirir uma infinidade de produtos.

As vendas deste setor vêm crescendo a cada ano, como podemos verificar no quadro abaixo.

Quadro 8 - Evolução das vendas de supermercados (Rojo, 1998)

Ano	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**
Vendas (Bilhões de dólares)	28,7	25,7	26,9	28,1	37,5	43,7	46,5	46,6

Os supermercados conquistaram a condição de maiores abastecedores de alimentos e artigos de higiene e limpeza, longe de quaisquer subsídios ou favores oficiais, graças a uma vocação para o crescimento impulsionada por investimentos contínuos e crença no desenvolvimento do Brasil (ABRAS, 1997). O setor é o responsável pela distribuição de mais de 82% dos gêneros de primeira necessidade.

Os números gerais do setor, no ano de 1997, são mostrados no quadro 9.

Quadro 9 - Números gerais do setor (Agas, 1998)

Item	**Valor**
Faturamento em 1997	US$ 46,6 bilhões
Percentual sobre o PIB	6,2%
Número de lojas	47.847
Número de empregados	655.000

Uma análise do setor desenvolvida pela AGAS (Rojo, 1998) reforça a percepção de que, após a estabilização da economia, ocorreu um acirramento da concorrência, o volume

de vendas cresceu, enquanto as margens de lucro foram pressionadas para baixo, levando as empresas a perseguir vantagens competitivas por meio de serviços melhores e da busca incessante da eficiência administrativa. A AGAS, segundo Rojo (1998), afirma que o segmento supermercadista está sendo obrigado a repensar seu negócio, necessitando de muitos ajustes em busca de um novo modelo de operação baseado no controle mais eficiente dos negócios e na satisfação dos clientes.

Um dos grandes desafios a serem enfrentados pelo setor de supermercados no Brasil será sem dúvida o de como aglutinar e dominar a informação. Diante de uma concorrência, tanto interna como internacional que se prenuncia cada vez mais acirrada e competente, conhecer o próprio negócio e ter acesso a diferentes fontes de informação serão condições vitais para a sobrevivência no mercado.

O varejo, juntamente com os bancos é um dos ramos de negócios que mais investiu em novas tecnologias de informação nos últimos tempos (Marcovitch, 1996).

5.2.1 Automação nos supermercados brasileiros

Desde a década de 80, o varejo brasileiro tem incorporado novas tecnologias de forma crescente. Num primeiro plano buscou-se a melhoria da eficiência interna. A partir da venda de um produto ao consumidor final, dava-se baixa no estoque e acionava-se o setor de compras quando os níveis de estoque baixavam, para que fosse providenciada a reposição. A idéia era controlar e melhorar a eficiência interna da empresa, principalmente através de um giro mais rápido de estoques. Em seguida, implantou-se a leitora óptica, o código de barras e as máquinas de preenchimento de cheques. Estas tinham como intuito aumentar a velocidade de passagem do cliente pelo '*check-out*' e, portanto, reduzir filas. Trata-se de uma ação com dois benefícios claros, um do lado do cliente - oferecer maior rapidez e menor espera – e outro do lado do varejo - permitir o atendimento de maior número de clientes com o mesmo número de *check-outs* (Marcovitch, 1996).

Um importante marco neste aspecto aconteceu em 1987, com a implantação e padronização do código de barras, realizada pela Associação Brasileira de Automação Comercial (ABAC), hoje EAN Brasil.

A partir dos anos 90 a automação dos supermercados ganha impulso, com o fim da reserva de mercado de informática. Nasce uma nova era para os supermercados, que agora podem oferecer aos seus consumidores o que há de melhor no mundo em tecnologia.

Na segunda etapa deste processo, na metade da década de 90, a inovação tecnológica passou a ser utilizada como uma forma de agregar valor ao cliente, através de informação, serviços e facilidades (Gonçalves e Gonçalves Filho, *apud* Marcovitch, 1996).

Com a estabilidade monetária, o setor de supermercados é obrigado a se profissionalizar, porque as margens de comercialização diminuem. Algumas empresas não se adaptam aos novos tempos e fecham, outras vencem os desafios e crescem ainda mais. Inicia-se um período de fusões e aquisições no setor. A abertura econômica, ampliada pelo governo de Fernando Henrique Cardoso, torna o mercado brasileiro mais atraente às grandes redes de varejo (ABRAS, 1998).

A tecnologia de informação surgiu como uma ferramenta de redução de custos e agilizadora do processo de troca de informações (Gonçalves e Gonçalves Filho, *apud* Marcovitch, 1996).

De acordo com o Super Censo da Abras (Albuquerque, 2000), 41,4% das lojas, em 2000, tinham leitor óptico na frente de caixa. Os campeões em automação eram em primeiro lugar o Espírito Santo, com 84,2% e em segundo o Rio Grande do Sul, com 52,2%.

A figura 6 mostra a evolução do número de lojas automatizadas, no Brasil, até 1997.

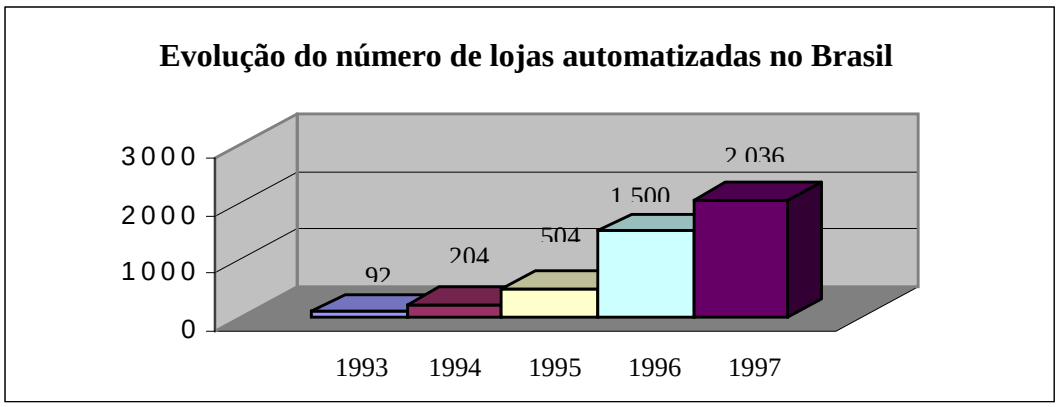

Figura 6 - Número de lojas automatizadas.

Fonte: AGAS, 1998

As novas tecnologias deixam de ser apenas uma forma de melhorar a eficiência interna ao ponto de venda e passam a intermediar as relações do varejo com fornecedores e clientes finais (Marcovitch, 1996).

5.2.2 A Comercial Unida de Cereais

A Comercial Unida de Cereais[6] Ltda. foi fundada em 1962, em São Leopoldo-RS. A princípio era um armazém de secos e molhados denominado Santos Dumont. Da pequena casa comercial surgiu a Rede de Supermercados Unidão.

Atualmente, a Rede Unidão é uma das principais redes gaúchas de supermercados. A mesma encontrava-se em 1998 em quinto lugar no *ranking*[7] da AGAS (Associação Gaúcha de Supermercados), conforme Anexo A.

A empresa atua no empacotamento, beneficiamento, comércio de carga para terceiros, importação e exportação de produtos alimentícios, bebidas, brinquedos, vestuário e calçados. Também conta com uma central de hortifrutigranjeiros e um setor próprio de publicidade e propaganda.

O objetivo da empresa é sempre oferecer produtos de qualidade, das melhores marcas e menores preços, regra perseguida durante os mais de 35 anos de existência da empresa. A empresa procura, também, prestar um bom atendimento e disponibilizar uma grande diversidade de produtos.

Os segmentos de negócio da empresa são:

a) Atacado - vende mercadorias para pequenos estabelecimentos comerciais (bares, minimercados e mercearias) de todo o estado do Rio Grande do Sul e de Santa Catarina;
b) Varejo - oferece através de suas lojas, produtos alimentícios e bazar em geral, diretamente aos seus consumidores;
c) Transporte - realiza transporte de carga, para clientes de expressão nacional.

---

[6] A Sede da Comercial Unida de Cereais está localizada na Av. Thomas Edison, n° 2598, na cidade de São Leopoldo-RS.

[7] Disponível via WWW [http://www.agas.com.br/]

Atualmente a empresa possui aproximadamente 1700 funcionários nos setores de atacado, varejo (supermercado) e transporte.

A Rede Unidão de Supermercados é composta por vinte lojas espalhadas pelo Vale dos Sinos, Região das Hortências e em Porto Alegre.

Os dados referentes ao ano de 1999 estão demonstrados no quadro abaixo:

Quadro 10 - dados sobre as lojas da Rede Unidão. Fonte: Leite (2000)

Loja	Localidade	Área (m²)	Venda média/mês (R$)
01 - Feitoria	São Leopoldo	1.150	1.150.000
03 - Campo Bom	Campo Bom	591	410.000
04 - Scharlau	São Leopoldo	1.006	540.000
05 - São Chico	São Francisco de Paula	1.140	460.000
07 - Canudos (*)	Novo Hamburgo	451	220.000
08 - Mini Preço	São Leopoldo	116	85.000
09 - Pernambuco	Novo Hamburgo	516	370.000
10 - Canela	Canela	1.535	715.000
11 - Sapiranga	Sapiranga	986	510.000
12 - Campo Bom Sesi (*)	Campo Bom	621	320.000
13 - Sapiranga Sesi (*)	Sapiranga	812	300.000
16 - Portão	Portão	974	470.000
18 - Aliança (*)	Novo Hamburgo	438	195.000
19 - Rio dos Sinos	São Leopoldo	1.190	430.000
22 - Campo Bom Shop	Campo Bom	675	390.000
23 - Ivoti (*)	Ivoti	500	270.000
24 - Gramado	Gramado	1.582	480.000
27 - Restinga	Porto Alegre	1.250	405.000
28 - Viamópolis	Viamão	744	480.000
29 - Viamão Cent.	Viamão	520	415.000

Obs.: Os números mostrados à frente do nome da loja indicam o número de identificação utilizado pela empresa.

(*) lojas não automatizadas

As lojas que fazem parte desta pesquisa são: a filial 01 - Feitoria, a filial 10 - Canela, a filial 11 - Sapiranga e a filial 28 - Viamópolis. Estas lojas foram escolhidas pela

própria empresa, que as considerou como representativas de núcleos diferenciados de mercado e com potencial de crescimento.

a) Automação das lojas

A automação na Rede Unidão teve início em outubro de 1994. Esta automação teve o objetivo de registar somente a venda tributada, ou seja não existia um controle de vendas por cupom, não havia identificação de clientes, etc.

Atualmente, das 20 (vinte) lojas da rede, somente 05 (cinco) não são automatizadas (vide quadro 10).

No momento da primeira reunião com os responsáveis pela empresa não havia dados armazenados para o desenvolvimento do trabalho, visto que o sistema existente não armazenava os dados das transações. Após o fechamento do acordo para o desenvolvimento do trabalho foram necessárias alterações nos sistemas para que os mesmos armazenassem os dados sobre as vendas aos clientes.

Atualmente, os sistemas armazenam os dados sobre as vendas ao consumidor. Os dados armazenados são: código EAN do produto, descrição do produto, quantidade vendida, preço de venda, número do ponto de venda, número do cupom fiscal, data da venda, dia da semana e identificação (número) da loja.

A empresa não possui informação alguma sobre o comportamento de seus clientes, pois não armazenava os dados das transações e também não realizava qualquer tipo de análise.

5.3 Aplicação de mineração de dados em bases de dados de supermercados

A tarefa de mineração de dados aplicada foi a de associação, devido ao fato de as bases de dados disponíveis tratarem de dados referentes às vendas de uma rede de supermercados.

A descoberta de regras de associação tem por objetivo encontrar relacionamentos ou padrões freqüentes em conjuntos de dados, ou seja, determinar os produtos que são comprados em conjunto em uma cesta ou carrinho de supermercado. A associação pode ser

utilizada para identificar oportunidades de vendas de pacotes de produtos ou de produtos e serviços (Berry & Linoff, 1997).

Um exemplo do resultado de tal mineração seria a declaração de que 80% das transações nas quais foram adquiridas fraldas também foi comprado leite. As cadeias de varejo usam esta técnica para planejar a disposição dos produtos nas prateleiras das lojas ou em um catálogo, de modo que os itens geralmente adquiridos na mesma compra sejam vistos próximos entre si (Harrison, 1998).

O interesse nesta busca de informações ocorre devido, principalmente, ao progresso feito na tecnologia de código de barras, que tornou possível para as organizações de varejo coletar e armazenar grandes quantidades de dados referentes às vendas efetuadas, conhecidos como ‘dados da cesta’. Um registro destes dados tipicamente consiste da data da transação e dos itens comprados. Organizações de sucesso atualmente vêem tais bancos de dados como importantes peças da sua infra-estrutura de marketing, pois permitem que o processo de marketing seja dirigido, além de auxiliar em programas e estratégias customizadas como reorganização do layout das lojas e projeto de catálogos. Como exemplo de uma regra que poderia ser encontrada em um banco de dados de um supermercado seria o fato de que 90% dos clientes que compram o produto A, também adquirem, na mesma ocasião, o produto B (análise do comportamento do consumidor no comércio varejista).

Os dados usados por um algoritmo de associação são formados por entidades e atributos. A entidade pode ser uma cesta de supermercado e os atributos todos os itens comprados na mesma compra.

Os dados usados por um algoritmo de associação precisam estar em um dos seguintes formatos: a) horizontal, onde há uma linha para cada entidade e há colunas para cada atributo. Para a análise de cesta de supermercado com o formato horizontal, há uma linha para cada cesta e colunas para cada produto. Um problema significante do formato horizontal é o grande número de colunas que ele pode alcançar. Para a análise de cesta de supermercado, onde o número de produtos pode exceder a 100.000, os produtos similares precisam ser agrupados para reduzir o número de colunas para uma quantidade razoável. Outro problema com este formato é que o esquema é dependente dos dados. Quando um novo produto é adicionado à cesta, ou quando os produtos são classificados de maneira diferente, então o esquema precisa ser mudado para adicionar ou reorganizar as colunas; e

b) o formato vertical, o qual é mais usado pelos produtos de mineração de dados, elimina estes problemas usando múltiplas linhas para armazenar uma entidade, usando uma linha para cada atributo. As linhas para uma entidade particular (cesta) são ligadas por uma identificação comum. Este tipo de representação é mais normalizado no senso relacional e funciona muito melhor quando uma entidade pode ter grande variedade em termos de número de atributos. Por exemplo, algumas pessoas compram somente dois itens enquanto outras enchem vários carrinhos com centenas de itens (Brand e Gerritsen, 2000).

Devido ao fato dos dados já existirem em um formato alguns produtos, tais como o Intelligent Miner da IBM, suportam a operação de conversão do formato horizontal para vertical.

Os algoritmos de associação podem operar somente com dados discretos. Se você usar atributos contínuos, o dado deve ser classificado em faixas, por exemplo 0 - 20.000, 20.001 - 40.000 e > 40.000), transformando cada faixa em um atributo.

Outra característica comum dos geradores de regras de associação é uma hierarquia de itens, uma hierarquia de produto que pode ser usada para agrupar itens similares. Na realidade, as lojas vendem itens específicos (cerveja tal de lata, cerveja tal de garrafa, cerveja de outra marca, etc.). No entanto, às vezes pode ser interessante gerar regras com itens genéricos.

A análise de uma cesta de supermercado nos fala de um consumidor, mas a análise de todas as compras feitas pelos consumidores fornece muito mais informações (Berry e Linoff, 1997). Consumidores não são iguais, cada um compra uma combinação de produtos diferentes, em diferentes quantidades, em horários e dias diferentes durante a semana. A análise de cesta de supermercado utiliza a informação sobre o que os consumidores adquirem para nos dar a compreensão de quem são eles e por quê eles fazem certas compras. Esta análise dá o discernimento para a comercialização, nos dizendo quais produtos tendem a ser comprados em conjunto e quais são mais adequados a promoções. Esta informação pode sugerir novos *layouts* para as lojas, determinar quais produtos podem ser vendidos em pacotes promocionais, entre outros. Suas raízes estão em analisar transações feitas nos pontos de vendas.

Esta análise também é realizada como um ponto de partida quando os dados das transações estão disponíveis e não se sabe que padrões específicos devem ser procurados. Este é um exemplo de mineração de dados indireta, porém a análise de cesta de

supermercado pode ser utilizada para mineração de dados direta e indireta (Berry & Linoff, 1997).

As técnicas que estão por baixo da análise de cesta de supermercado são oriundas da probabilidade e da estatística.

5.3.1 Regras de associação

O apelo da análise de cesta de supermercado vem da clareza e utilidade dos seus resultados, os quais são apresentados em forma de regras de associação.

Uma regra tem duas partes: a condição e o resultado, e geralmente é representada da seguinte forma: ‘Se condição então resultado’.

Segundo Berry & Linoff (1997), os três tipos de regras produzidas por esta análise são: as regras úteis, as triviais e as inexplicáveis.

As regras úteis contêm alta qualidade, informação para ação. Uma vez que o padrão é encontrado, não há dificuldade em justificá-lo. A conhecida regra sobre fraldas e cervejas nas quintas-feiras sugere que ao anoitecer de quinta-feira, jovens casais se preparam para o final de semana estocando fraldas para os bebês e cerveja para o pai (o qual irá assistir futebol na sexta-feira tomando cervejas). Mais importante do que sugerir as causas é o fato de que os gerentes agora podem agir. Localizando a estante de fraldas perto do corredor onde encontram-se as cervejas, eles podem aumentar as vendas de outros produtos. Devido à facilidade de entendimento da regra outras ações podem ser tomadas, tais como: colocar outros produtos para bebês e/ou para homens entre as cervejas e as fraldas, colocar outros produtos, tais como salgadinhos e aperitivos perto dos dois produtos (Berry e Linoff, 1997).

Na prática, as regras 'acionáveis' possuem somente um item como resultado. Então uma regra como: "Se fralda e quinta-feira então cerveja" é mais útil do que a seguinte regra: "Se quinta-feria então cerveja e fralda".

Os resultados inexplicáveis não sugerem um curso de ação.

As regras triviais já são conhecidas por qualquer pessoa que conheça o negócio, elas reproduzem um conhecimento comum sobre o negócio. Comumente, os resultados triviais simplesmente medem ações prévias, tais como campanhas de marketing, mas não oferecem nenhuma direção para ações futuras.

Segundo Brusso (2000), os algoritmos de mineração de dados incorporam alguma medida para representar o quanto interessante é um padrão. Essas medidas são utilizadas para decidir o que deve ser mantido ou o que deve ser descartado. Um dos problemas centrais no campo da descoberta do conhecimento é o desenvolvimento de boas medidas de interesse, uma vez que deveriam ser apresentados aos usuários somente os padrões que de fato fossem originais e interessantes.

O autor classifica as medidas de interesse em subjetivas e objetivas. As subjetivas são aquelas que não dependem apenas da regra descoberta e dos dados utilizados, mas também do usuário que a examina, ou seja, uma regra pode ser interessante para uma pessoa e não para a outra. As duas principais razões pelas quais um padrão pode ser considerado interessante do ponto de vista subjetivo do usuário são a utilidade – um padrão é interessante se o usuário pode fazer algo a partir dele, isto é reagir em seu proveito - e a inesperabilidade – um padrão pode ser considerado surpreendente, se ele for capaz de contradizer as expectativas do usuário, o que depende de suas convicções, ou seja, o que ele imagina que esteja armazenado nos dados. As medidas objetivas de interesse são aquelas que avaliam o grau de interesse de um padrão em termos de sua estrutura e dos dados utilizados no processo de descoberta. Tais medidas podem ser utilizadas como filtros para selecionar padrões potencialmente interessantes entre os muitos descobertos por um algoritmo de mineração devolvendo um conjunto menor ao usuário. Como principais exemplos de medidas objetivas de interesse podemos citar os graus de confiança e suporte mínimo das regras de associação.

a) Confiança

Confiança é a relação entre o número de transações que suportam a regra e o número de transações que a parte condicional da regra suporta. Outra maneira de explicá-la é: a relação entre o número de transações entre todos os itens e o número de transações somente com os itens que fazem parte do ‘se’ (‘if’) (Berry & Linoff, 1997). A confiança mede o quanto um item depende do outro (Brand e Gerritsen, 2000).

O que a confiança realmente diz é o seguinte: uma regra ‘se B e C então A’ com uma confiança de 0,33 é equivalente a dizer que, quando B e C aparecem na transação, há uma chance de 33% de que A também apareça.

Quanto maior a confiança da regra, melhor ela é.

b) Suporte mínimo

Existem técnicas para reduzir o número de itens e combinações de itens a ser considerados em cada passo para reduzir o tempo de processamento. A mais comum é chamada de suporte mínimo e ela garante que a regra abrange um número mínimo de transações na base de dados. O suporte mínimo mede a freqüência da ocorrência de itens em conjunto, como uma porcentagem do total de transações (Brand e Gerritsen, 2000). Por exemplo, se há 1.000.000 transações na base e o suporte mínimo é de 1%, então somente as regras suportadas por, no mínimo, 10.000 transações são interessantes. Em outras palavras, o suporte mínimo elimina itens que não aparecem em um número suficiente de transações (Berry & Linoff, 1997).

A análise de cesta de supermercado começa com transações contendo uma ou mais ofertas de produtos ou serviços e alguma informação pouco desenvolvida sobre a transação.

Os dados usados para análise de cesta de supermercado são tipicamente os dados detalhados das transações capturadas nos pontos de vendas.

A análise de cesta de supermercado tem utilidade para a área de varejo, tal como supermercados, lojas de conveniências, drogarias e redes de lancherias, onde muitas das compras são efetuadas à vista. Transações à vista são anônimas, significando que a loja não tem conhecimento algum sobre os clientes, porque não há informação alguma o identificando na transação. Para transações anônimas, geralmente a única informação conhecida sobre a compra é a data e o tempo, a localização da loja, o caixa, os itens comprados, e o troco. Segundo Berry e Linoff (1997) para a análise de cesta de supermercado mesmo estes dados limitados fornecem resultados interessantes e que podem gerar ação.

5.3.2 Aplicação das ferramentas nas bases de dados

Todos os sistemas foram executados no mesmo computador. Este computador tem a seguinte configuração:

- Processador K6 550 MHz
- 128 MB de memória RAM
- 10 GB de disco rígido
- Monitor SVGA

Cabe salientar que esta configuração está além de todos os requisitos mínimos exigidos pelos sistemas avaliados, conforme o quadro 7, visto no item 5.1.1 .

Primeiramente aplicou-se as ferramentas com a base de dados menor, B1, depois utilizou-se a base de dados B2 e por último a base de dados maior, B3.

Abaixo serão descritas indivualmente as aplicações de cada uma das ferramentas avaliadas.

a) Aira Data Mining®[8]

Foram utilizadas as versões 3.1.1.55, 3.1.1.67 e 3.5.13.256 do Aira Data Mining®.

O primeiro passo foi criar um projeto, e para isto foi necessário seguir a um assistente que conduz à realização desta tarefa. Primeiramente, o assistente solicitou o caminho completo onde o projeto seria salvo. O segundo passo foi informar a fonte dos dados com os quais iríamos trabalhar. O Aira Data Mining® trabalha com bases de dados (dBase, Paradox, Oracle, SQL Server, Access, Sybase, entre outros), fonte de dados ODBC, Planilhas do Excel e arquivos SPSS.

Após informar a fonte dos dados foi necessário ligar os dados ao projeto, informando o nome da base de dados que seria utilizada.

Depois de realizados estes passos, a ferramenta criou uma nova mineração - deixou uma mineração esperando para ser iniciada, pois no Aira Data Mining® pode-se ter mais de uma mineração por projeto, e com uma mineração criada a ferramenta ficou pronta para começar a trabalhar. Para que a mineração (processamento) possa ser iniciada é preciso configurá-la, ou seja, informar alguns parâmetros que são necessários para que a ferramenta possa iniciar o processamento. Nesta configuração devemos informar qual o campo será o índice, qual(is) campo(s) será(ão) o lado SE, qual será o lado ENTÃO da

[8] Maiores informações sobre a ferramenta podem ser obtidas no endereço http://www.hycones.com.br

regra e qual será o suporte mínimo utilizado. A figura 7 mostra a tela de configuração da mineração.

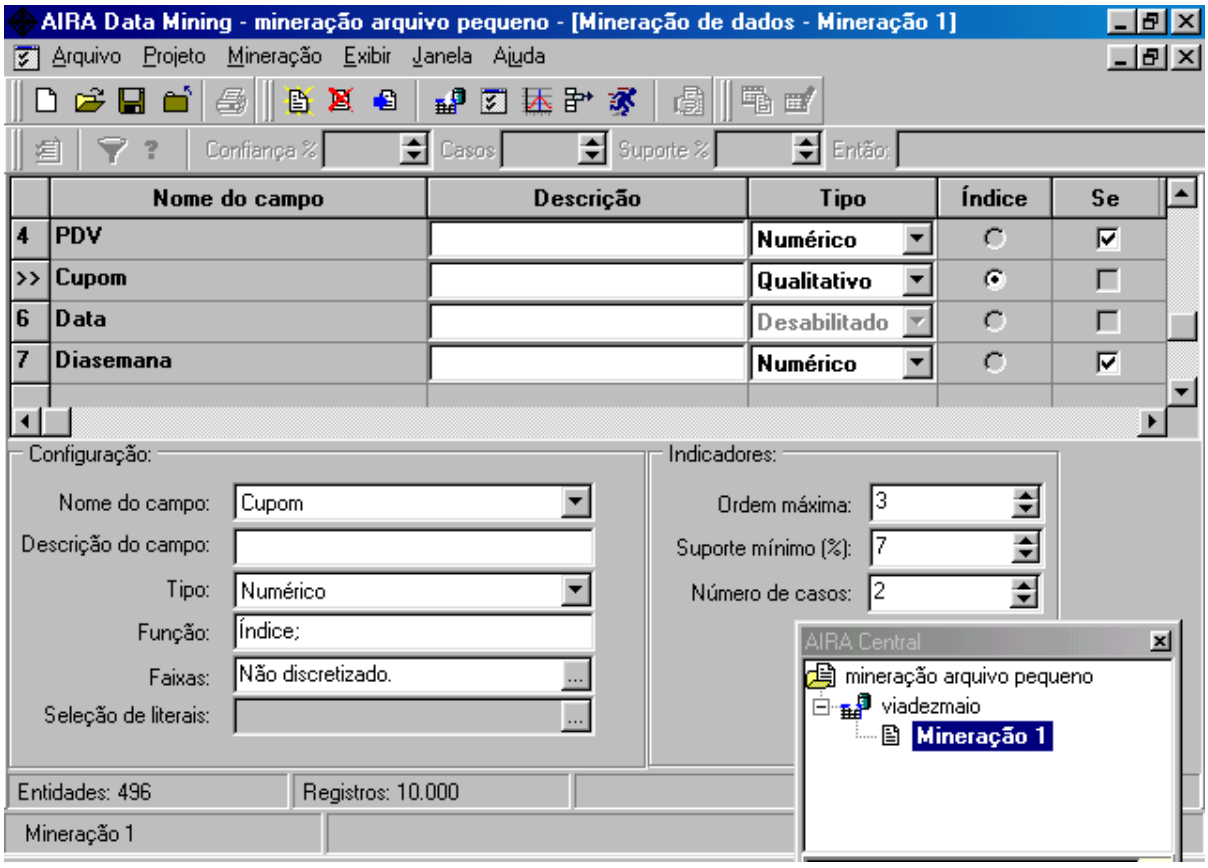

Figura 7 – Tela de configuração da mineração na Ferramenta Aira Data Mining®

A configuração utilizada nas três aplicações foi a mostrada no quadro abaixo:

Quadro 11 - Configuração de mineração utilizada na ferramenta Aira Data Mining®

Parâmetro	**Campo**
Índice	Cupom Fiscal
Se	Descrição, Data da venda, Dia da Semana, PDV, Quantidade, Valor, Loja[9]
Então	Descrição
Suporte Mínimo	7%

Os resultados gerados por esta ferramenta são mostrados em forma de regras associativas. O formato utilizado pelo Aira Data Mining® para mostrar os resultados pode ser visto na figura 8.

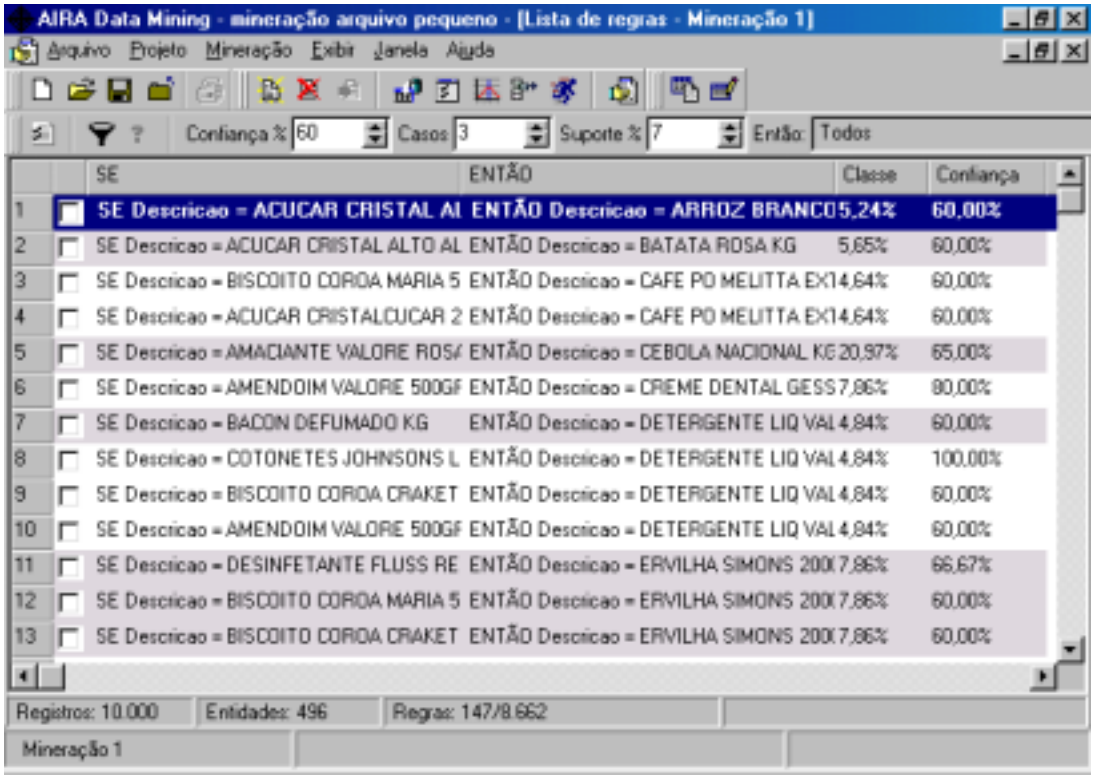

Figura 8 – Formato como as regras são apresentadas na ferramenta Aira Data Mining®

[9] O campo loja foi utilizado somente na aplicação à base de dados maior, pois as outras bases de dados possuem dados de uma única loja

Na visualização das regras podem ser utilizados filtros, que seleciona os itens que se deseja visualizar, conforme a figura 9.

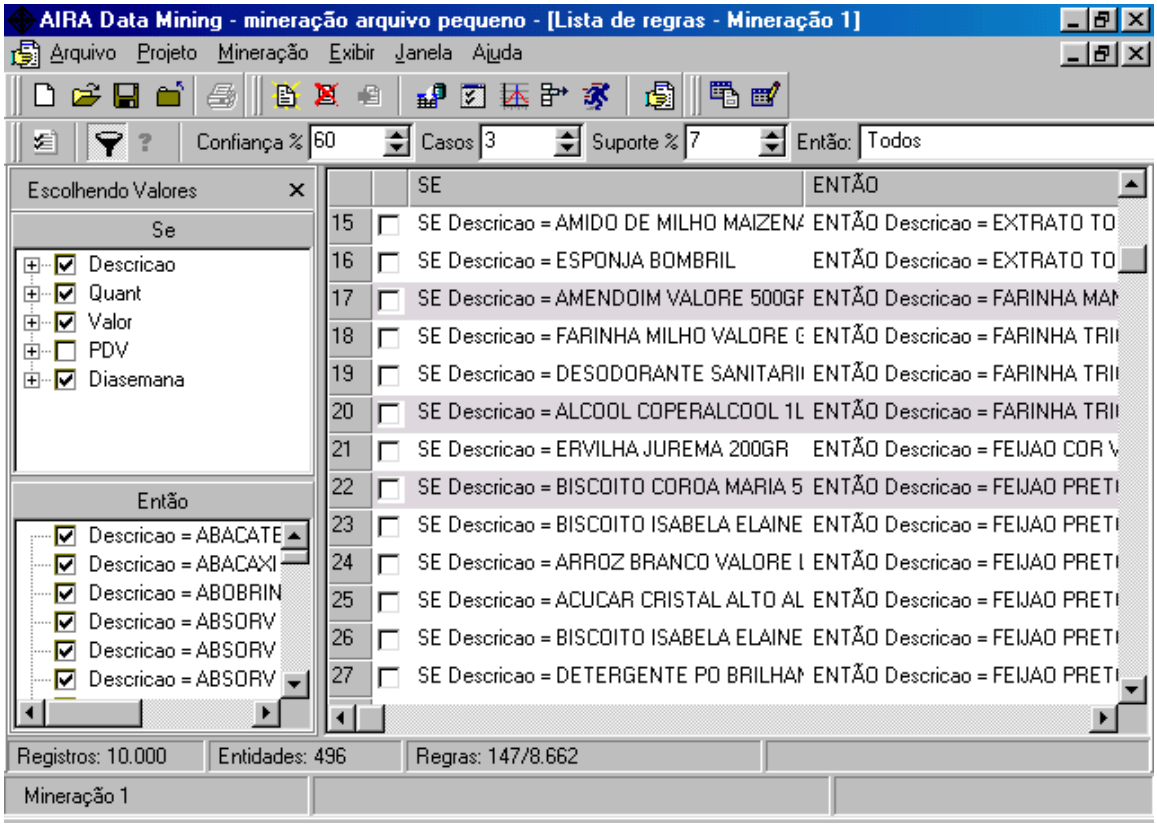

Figura 9 – Utilização de filtros na visualização de regras na ferramenta Aira Data Mining®

❖ Aplicação da ferramenta Aira Data Mining® à base de dados menor

O primeiro arquivo utilizado foi o menor, com 10.000 registros.

A ferramenta lê a base de dados e informa os seguintes dados iniciais: número de entidades = 496 e número de registros = 10.000, o que dá uma média de 20 itens por cupom.

Nas versões 3.1.1.55 e 3.1.1.67 a criação da nova mineração durou 10s. Após a ferramenta passou para a etapa de discretização dos valores, nesta etapa foram gastos 2s. O próximo passo foi a geração das literais, a qual levou 8s e passou para a inicialização da mineração. Nesta última etapa a ferramenta parou de responder, ou seja, trancou o processamento nas duas versões.

Com a versão 3.5.13.256 conseguiu-se realizar a mineração da base de dados B1. O tempo gasto para esta mineração foi de 2 minutos e 4 segundos e a ferramenta retornou 8.662 regras.

O cabeçalho das regras geradas pode ser visto no quadro 12.

Quadro 12 – Cabeçalho da lista de regras gerada pela ferramenta Aira Data Mining®

SE	Valor	Então	Valor	Classe %	Confiança %	Nº de Casos	Suporte
Descrição	Açúcar cristal alto alegre 5kg	Descrição	Arroz branco valore lt1 5kg	5,24	60	3	11,54
Descrição	Amaciante valore rosa	Descrição	Cebola nacional kg	20,97	65	13	12,50

As regras, para serem impressas, devem ser exportadas para outro software, pois a ferramenta não liberou a opção de impressão de regras. A forma como as regras são impressas pode ser vista no Anexo B.

❖ Aplicação da ferramenta Aira Data Mining® à base de dados média

Na segunda aplicação foi utilizada a mesma configuração da primeira (quadro 11).

Quando utilizou-se as versões 3.1.1.55 e 3.1.1.67, ao terminar de responder às perguntas do asssistente, ao passar para a etapa de criação de uma nova mineração, a ferramenta trancou e o sistema não respondeu mais. A única coisa que pôde ser feita foi sair do Aira Data Mining®, através interrupção feita pelas teclas CTRL, ALT e DEL pressionadas simultaneamente.

Na versão 3.5.13.256, a mineração durou 3 horas 4 minutos e 2 segundos. Os dados mostrados inicialmente foram: número de entidades = 53.360 e número de registros = 287.935.

Com esta base de dados, a ferramenta informa que foram geradas 104 regras, porém não conseguimos visualizar estas regras. Ao entrar na lista de regras a mesma aparece vazia.

❖ Aplicação da ferramenta Aira Data Mining® à base de dados maior

Utilizando as versões 3.1.1.55 e 3.1.1.67, quando se tentou aplicar a ferramenta à base de dados maior surgiu o mesmo problema da aplicação à base de dados média, ou seja, a ferramenta não conseguiu nem ligar os dados ao projeto nem criar a primeira mineração.

Na versão 3.5.13.256 conseguiu-se ligar a base de dados ao projeto. A leitura da base gastou 15 minutos. Os dados mostrados inicialmente foram: número de entidades = 148.115 e número de registros = 3.326.859. Quando a ferramenta foi começar a mineração ela trancou. Ocorreu o mesmo erro das outras versões.

❖ Outras aplicações

Diante dos problemas que ocorreram nas aplicações anteriores, foram realizadas novas aplicações; uma delas foi utilizando o arquivo menor e especificando uma configuração bem mais simples para a mineração. Esta configuração foi a seguinte:

Quadro 13 - Segunda configuração de mineração usada na ferramenta Aira Data Mining®

Parâmetro	**Campo**
Índice	Cupom Fiscal
Se	Dia da Semana, Quantidade, Valor
Então	Descrição
Suporte Mínimo	7%

Quando a ferramenta foi acionada para iniciar a mineração, ocorreu o mesmo problema da primeira aplicação, ou seja o sistema trancou e foi necessária a execução da interrupção utilizada na mineração anterior.

Como não se pôde utilizar a ferramenta, entrou-se em contato com os seus desenvolvedores e a ferramenta foi aplicada nas instalações da empresa desenvolvedora do sistema, utilizando a melhor máquina existente (de maior configuração[10]) e com seus desenvolvedores verificando erros e gerando soluções simultaneamente ao processamento

dos dados. Salienta-se que estes resultados foram obtidos através de artifícios de programação por parte dos desenvolvedores e não pela utilização normal da ferramenta, por parte de um usuário comum, o que se encaixa num dos problemas de mineração de dados que fala sobre até que ponto as ferramentas podem ser usadas por pessoas que não sejam seus desenvolvedores. O tempo gasto para esta mineração foi de onze horas e cinqüenta e três minutos.

Algumas regras foram geradas, mas a maioria delas não tem importância nenhuma, pois se tratam de regras sem sentido, que não dizem nada. Alguns exemplos destas regras podem ser vistos no quadro 14.

Quadro 14 – Exemplos de regras sem sentido, geradas pelo Aira Data Mining®

SE	Valor	ENTÃO	Valor	Confiança %	Nº de Casos	Suporte %
Valor	> 0,73 e PDV=10 e Loja=011	Diasemana	2	100	61	29,90
Valor	> 0,73 e loja= 011 e loja = 028	Diasemana	2	97,78	44	21,57

Outras regras geradas podem ter alguma importância, pois podem informar padrões de compras que estavam escondidos na base de dados. Exemplos destras regras são mostrados no quadro 15.

Quadro 15 – Regras que podem ter importância, geradas pela ferramenta Aira Data Mining®

SE	Valor	ENTÃO	Valor	Confiança %	Nº de Casos	Suporte %
Descrição	Bergamota poncan kg	Descrição	Maça gala especial kg	69,49	41	3,06
Descrição	Alvejante clorisol 2l	Descrição	Refresco tang abacaxi	60,00	3	25,00

b) SuperQuery Discover Edition[11]

Quando se inicia a utilizar esta ferramenta, a primeira coisa que deve ser feita é a leitura da base de dados. Através de um assistente, responde-se às perguntas que são realizadas até que todos os dados sejam lidos. Após a leitura dos dados, devemos informar

[10] A configuração da máquina utilizada foi Pentium III 750Mhz, 256Kb memória RAM, 40Gb HD.
[11] Maiores informações sobre a ferramenta podem ser obtidas no endereço http://www.azmy.com

a ação desejada; dentre as opções estão: selecionar colunas, classificar colunas em faixas (discretização), criar colunas virtuais, criar tabela sumário e descobrir fatos. Como se trabalharia com dados contínuos, foi escolhida a segunda opção em primeiro lugar. Após esta etapa, foi possível partir para a descoberta do conhecimento através da quinta opção.

O conhecimento descoberto é mostrado através de regras que podem mostrar fatos ou exceções. Na configuração, deve-se informar se o que se deseja descobrir são fatos, exceções ou ambos. Estas regras podem ser visualizadas e impressas na forma de tabela.

Deve-se configurar a mineração, informando os seguintes parâmetros: suporte mínimo, colunas que podem estar do lado 'SE' e colunas que podem estar do lado 'ENTÃO' da regra.

A configuração utilizada foi a demonstrada no quadro abaixo:

Quadro 16 – Configuração de mineração utilizada na ferramenta Super Query Discovery Edition

Parâmetro	**Campo**
Índice	Cupom Fiscal
Se	Descrição, PDV, Data da Venda, Dia da Semana, Quantidade, Valor
Então	Descrição, PDV, Data da Venda, Dia da Semana, Quantidade, Valor
Suporte Mínimo	7%
Regras	Fatos e exceções

Outro resultado obtido por esta ferramenta é a visualização dos sete valores mais freqüentes para cada campo da base de dados, o qual é mostrado em forma de gráfico de barras. Este gráfico só pode ser impresso, ou seja, não pode ser importado para editores de texto, planilhas eletrônicas ou outras ferramentas computadorizadas; assim, tais resultados serão mostrados em forma de quadros.

❖ Aplicação da ferramenta Super Query Discovery Edition à base de dados menor

Na primeira vez que foi executada a ferramenta utilizando a base de dados menor, não foram encontradas regras e nem exceções.

Os sete valores mais freqüentes para cada campo estão mostrados nos quadros abaixo:

Quadro 17 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Produto da base de dados B1

Produtos mais freqüentes		
Produto	**Nº de registros**	**% em relação à base**
Pão francês 50 gr c/5	435	4,35
Óleo de soja soya 900ml	133	1,33
Leite C nutrilat sc 1lt	129	1,29
Segunda moída kg	119	1,19
Segunda c/ osso kg	104	1,04
Queijo lanche kg	99	0,99
Coxa c/s coxa frango congelada	99	0,99

Através deste quadro pode-se verificar os sete produtos vendidos com maior freqüência. Este resultado não foi considerado inesperado pelos tomadores de decisão da empresa.

Quadro 18 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Quantidade da base de dados B1

Quantidades mais freqüentes		
Quantidade	**Nº de registros**	**% em relação à base**
1	272	2,72
2	208	2,08
3	111	1,11
0,16	75	0,75
0,15	74	0,74
0,17	73	0,73
0,14	72	0,72

Através deste quadro pode-se verificar que as vendas mais freqüentes são de itens em poucas quantidades. Esta informação também foi considerada já conhecida pelos tomadores de decisão da empresa.

Quadro 19 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Valor da base de dados B1

Valores mais freqüentes		
Valor	**Nº de registros**	**% em relação à base**
0,7	528	5,28
0,79	300	3,00
0,99	253	2,53
0,69	212	2,12
0,89	167	1,67
0,48	149	1,49

Este resultado mostra que nas vendas mais freqüentes são vendidos produtos de menor valor, o que já era conhecido pelos tomadores de decisão da empresa.

Quadro 20 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo PDV da base de dados B1

Pontos de Venda (PDV) mais freqüentes		
PDV	**Nº de registros**	**% em relação à base**
7	1659	16,59
6	1342	13,42
4	1305	13,05
3	1298	12,98
2	1164	11,64
5	1002	10,02

Este resultado foi considerado irrelevante, pois nem sempre todos os PDV encontram-se abertos, isto depende do movimento da loja, portanto, a freqüência não mostra o desempenho de cada PDV.

Quadro 21 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Cupom da base de dados B1

Cupons mais freqüentes		
Cupom	**Nº de registros**	**% em relação à base**
35719	65	0,65
97654	42	0,42
123144	39	0,39
35697	30	0,30
35700	25	0,25
123148	23	0,23

Este resultado também foi considerado irrelevante porque mostra os cupons fiscais em que foram vendidos os maiores números de itens.

Quadro 22 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Data da base de dados B1

Datas mais freqüentes		
Data	**Nº de registros**	**% em relação à base**
02/05/00	9325	93,25
03/05/00	675	6,75

Como esta base de dados possui somente os 10.000 primeiros registros do arquivo que contém a movimentação mensal de uma loja, os dados referem-se somente à movimentação de dois dias: 02/5/00 e 03/5/00, sendo que, deste último, foram usadas somente 675 transações para completar o número desejado. Portanto, este resultado torna-se irrelevante.

Quadro 23 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Dia da semana da base de dados B1

Dias da semana mais freqüentes		
Dia da semana	**Nº de registros**	**% em relação à base**
3 (terça-feira)	9325	93,25
4 (quarta-feira)	675	6,75

Este resultado neste arquivo não tem relevância pelo motivo explicado no item acima.

❖ Aplicação da ferramenta Super Query Discovery Edition à base de dados média

Quando ligamos esta base à ferramenta foi efetuada a sua leitura, a qual demorou 9 minutos. A etapa de discretização durou 2 minutos e, por fim, a descoberta dos fatos e exceções durou 30 minutos.

As regras encontradas foram:

Quadro 24 - Lista de regras encontradas pela ferramenta Super Query Discovery Edition na base de dados B1

Tipo	%	Se	Valor do se	Então	Valor do então
All	100	Codigo	0000032402735	Descricao	Pao frances 50gr c/5
All	100	Descricao	Pao frances 50gr c/5	Codigo	0000032402735
Some	30	QTD	1,00	Diasemana	7,00
Some	30	RValor	Low	Diasemana	7,00
All	100	Data	06/05/2000	Diasemana	7,00
Some	30	Loja	28	Diasemana	7,00
All	100	Data	12/05/2000	Diasemana	6,00
All	100	Data	13/05/2000	Diasemana	7,00
All	100	Data	20/05/2000	Diasemana	7,00
All	100	Data	27/05/2000	Diasemana	7,00
All	100	Data	31/05/2000	Diasemana	4,00

Verificamos a falta de nexo nas regras encontradas, onde somente um produto foi citado relacionando seu código e sua descrição, o que, num primeiro momento, poderia sugerir problemas (inconsistências) na base de dados, tais como produtos diferentes com o mesmo código e/ou códigos iguais para produtos diferentes. Porém, verificamos a base de dados e este tipo de problema não ocorre. Não foram encontradas exceções.

Os sete valores mais freqüentes de cada campo estão mostrados nos quadros abaixo:

Quadro 25 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Produto da base de dados B2

Produtos mais freqüentes		
Produto	**Nº de registros**	**% em relação à base**
Pão francês 50 gr c/5	14276	4,96
Segunda c/ osso kg	4469	1,55
Leite branc anube l vida integr	3773	1,31
Segunda moída kg	3547	1,23
Queijo lanche kg	2971	1,03
Óleo soja primor 900 ml	2901	1,01
Frango frangosul resfriado kg	2783	0,97

Este resultado mostrou os sete produtos vendidos com maior freqüência. Os tomadores de decisão já conheciam este resultado.

Quadro 26 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Quantidade da base de dados B2

Quantidades mais freqüentes		
Quantidade	**Nº de registros**	**% em relação à base**
2	8182	2,84
1	6252	2,17
3	3323	1,15
0,15	2262	0,79
0,16	2219	0,77
0,17	2185	0,76

Este resultado mostrou as quantidades vendidas com mais freqüência. Nota-se que estes valores são baixos. E os valores menores do que um indicam a venda de mercadorias por quilograma. Os tomadores de decisão da empresa disseram que este resultado já era esperado.

Quadro 27 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Valor da base de dados B2

Valores mais freqüentes		
Valor	**Nº de registros**	**% em relação à base**
0,65	13311	4,62
0,79	7687	2,67
0,99	6373	2,21
0,49	6056	2,10
0,69	5903	2,05
0,39	5661	1,97

Neste quadro pode-se verificar os valores mais freqüentes, ou seja os preços dos produtos que aparecem com maior freqüência na base de dados. Percebe-se que todos são valores abaixo de R$ 1,00 (um real). Este resultado também era esperado pelos tomadores de decisão.

Quadro 28 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo PDV da base de dados B2

Pontos de Venda (PDV) mais freqüentes		
PDV	**Nº de registros**	**% em relação à base**
7	49224	17,10
6	48497	16,84
4	48254	16,76
5	44209	15,35
8	40757	14,15
9	23773	8,26

Como foi explicado no quadro referente ao campo PDV da aplicação anterior, este resultado foi considerado irrelevante porque os valores mostrados não mostram o desempenho de cada PDV.

Quadro 29 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Cupom da base de dados B2

Cupons mais freqüentes		
Cupom	**Nº de registros**	**% em relação à base**
103931	206	0,07
100323	186	0,06
98150	155	0,05
103761	152	0,05
47940	151	0,05
45407	148	0,05

Este resultado foi considerado irrelevante porque mostra, apenas, os números dos sete cumpons fiscais que contiveram o maior número de itens vendidos.

Quadro 30 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Data da base de dados B2

Datas mais freqüentes		
Data	**Nº de registros**	**% em relação à base**
06/05/00	24580	8,54
13/05/00	22773	7,91
20/05/00	20119	6,99
27/05/00	19655	6,83
31/05/00	12724	4,42
12/05/00	12062	4,19

Este resultado mostrou a data onde foram vendidos os maiores números de itens. Os tomadores de decisão da empresa consideram este resultado irrelevante porque pensam não poder agir baseados neste tipo de informação. Eles julgam que o dia da semana (mostrado no quadro abaixo), ao invés da data seja uma informação mais interessante.

Quadro 31 - Resultado da ferramenta Super Query Discovery Edition em relação ao campo Dia da semana da base de dados B2

Dias da semana mais freqüentes		
Dia da semana	**Nº de registros**	**% em relação à base**
7	87127	30,26
4	52892	18,37
6	46069	16,00
3	40013	13,90
2	31854	11,06
5	26544	9,22

Este quadro mostra os dias da semana e o número de itens vendidos em cada um desses dias. Este resultado é considerado interessante pelos tomadores de decisão da empresa, porém eles salientam que ele pode ser obtido através do sistema convencional utilizado por eles.

❖ **Aplicação da ferramenta Super Query Discovery Edition à base de dados maior**

Ao ligar-se a base de dados maior à ferramenta, a mesma leu a base de dados, mostrou o número de registros existentes e passou para a etapa da discretização dos dados. Neste processo não foi obtida resposta, o sistema trancou e foi preciso sair da aplicação. Novas tentativas foram realizadas, porém o problema persistiu em todas elas.

c) SuperQuery Office Edition[12]

Aplicou-se esta ferramenta às três bases de dados e os resultados obtidos foram exatamente iguais ao da ferramenta SuperQuery Discovery Edition. É preciso salientar que o desenvolvedor trata-as como ferramentas distintas, inclusive com preços bem diferentes; porém, quando foram utilizadas, não apresentaram diferença alguma nem em relação a

[12] Maiores informações sobre a ferramenta podem ser obtidas no endereço http://www. Azmy.com

interfaces nem em relação à performance e resultados. Inclusive, na tela de apresentação das duas, aparece a palavra ‘office’.

d) XAffinity®[13]

A versão da ferramenta XAffinity® utilizada foi a 2.1.35.

O primeiro passo é criar um projeto e, para isto, segue-se um assistente que conduz à realização desta tarefa. Primeiramente o assistente solicita o caminho completo onde o projeto será salvo. O segundo passo é informar a fonte dos dados com os quais se irá trabalhar. Esta versão do Xaffinity® trabalha com bases de dados Access, Oracle 8i, Red Brick, SQL Server 6.5, SQL Server 7 e White Cross.

O tipo de base de dados utilizado em todas as minerações foi Access.

Após informar a fonte dos dados, foi preciso ligar os dados ao projeto, informando o nome da base de dados que seria utilizada.

Depois de realizados estes passos, a ferramenta criou o novo projeto e passou para o módulo de configuração da mineração, onde foram informados alguns parâmetros que são necessários para que a ferramenta pudesse iniciar o processamento. Nesta configuração, foi preciso informar o nome do projeto, o gerente do projeto, o nome da análise, a descrição desta análise, o algoritmo a ser utilizado, a tabela a ser utilizada, o campo chave da transação e o campo chave dos itens. Há outros itens que podem ser preenchidos, porém eles são opcionais e desnecessários para o tipo de mineração utilizada. A figura 10 mostra a tela de configuração do projeto.

[13] Maiores informações sobre a ferramenta no endereço www.exclusiveore.com

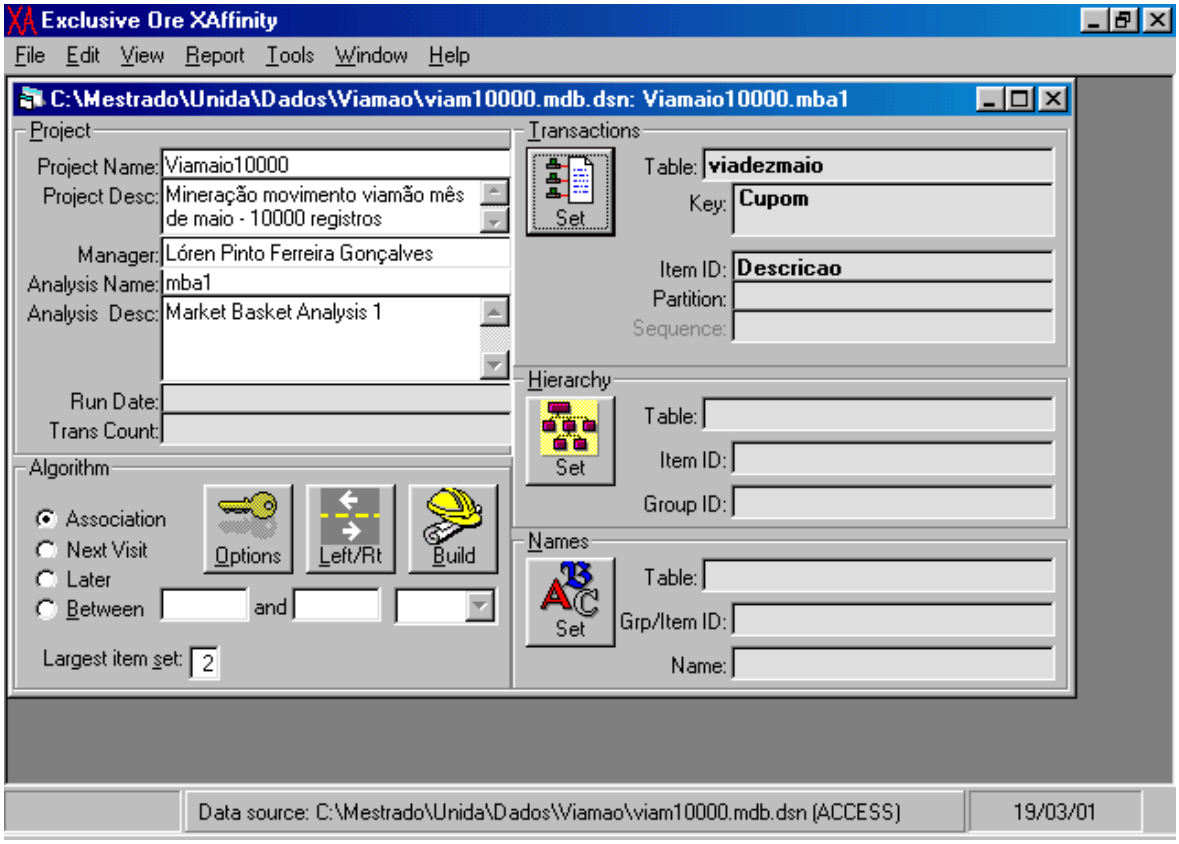

Figura 10 – Tela de configuração de projeto na ferramenta XAffinity®

Para a configuração da mineração informou-se os valores do suporte e confiança, conforme figura 11.

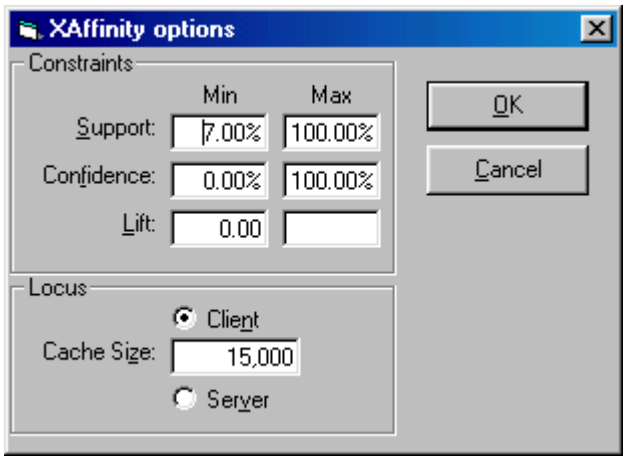

Figura 11 – Configuração da mineração na ferramenta XAffinity®

A configuração utilizada em todas as minerações foi a mostrada no quadro 32.

Quadro 32 – Configuração de projeto e mineração utilizada no Xaffinity®

Parâmetro	Valor
Nome do projeto	*
Gerente do projeto	Lóren Pinto Ferreira Gonçalves
Nome da análise	*
Descrição da análise	Market basket analysis do arquivo *
Algoritmo	Associação
Tabela	*
Chave de transação	Cupom
Chave de Item	Descrição
Confiança	0% - 100%
Suporte Mínimo	7% - 100%

** Estes campos foram preenchidos de acordo com o arquivo utilizado e serão descritos nas descrições das aplicações dos respectivos arquivos.*

Os resultados gerados por esta ferramenta são mostrados em forma de regras associativas. As regras geradas pelo Xaffinity® são mostradas inicalmente na forma de tabela, como mostrado na figura 12.

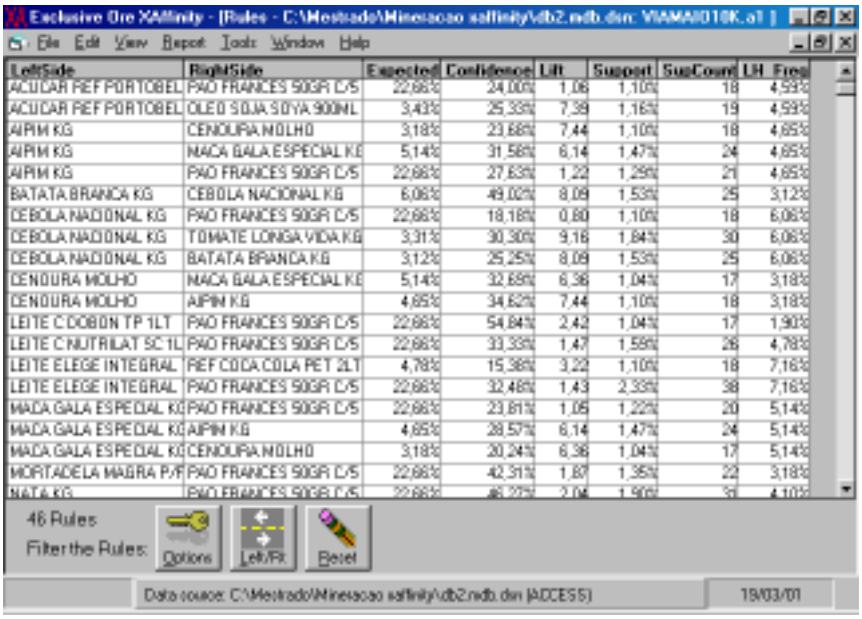

Figura 12 – Regras geradas pelo Xaffinity®

Pode-se escolher a forma de visualizar e imprimir as regras geradas. Na figura 13 é mostrado o formato em colunas.

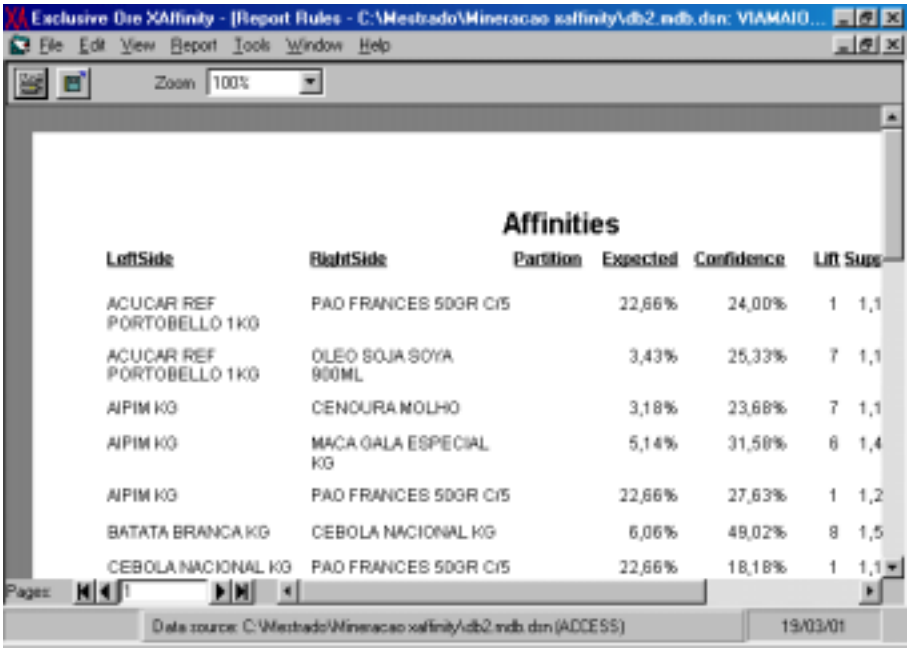

Figura 13 – Regras geradas pelo Xaffinity® mostradas em colunas

E na figura 14 é mostrado outro formato de visualização/impressão das regras, o formato chamado linguagem natural.

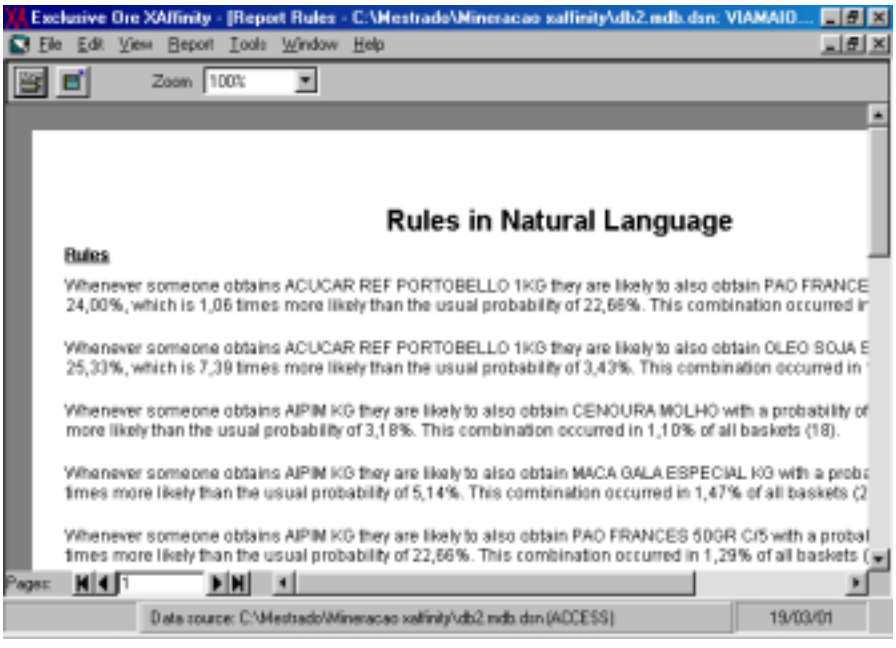

Figura 14 – Regras geradas pelo Xaffinity® mostradas em linguagem natural

Para facilitar a visualização das janelas que contém as regras e a configuração do projeto pode-se escolher entre três formas de visualização na tela: horizontal, vertical e em cascata.

A figura 15 mostra a visualização vertical.

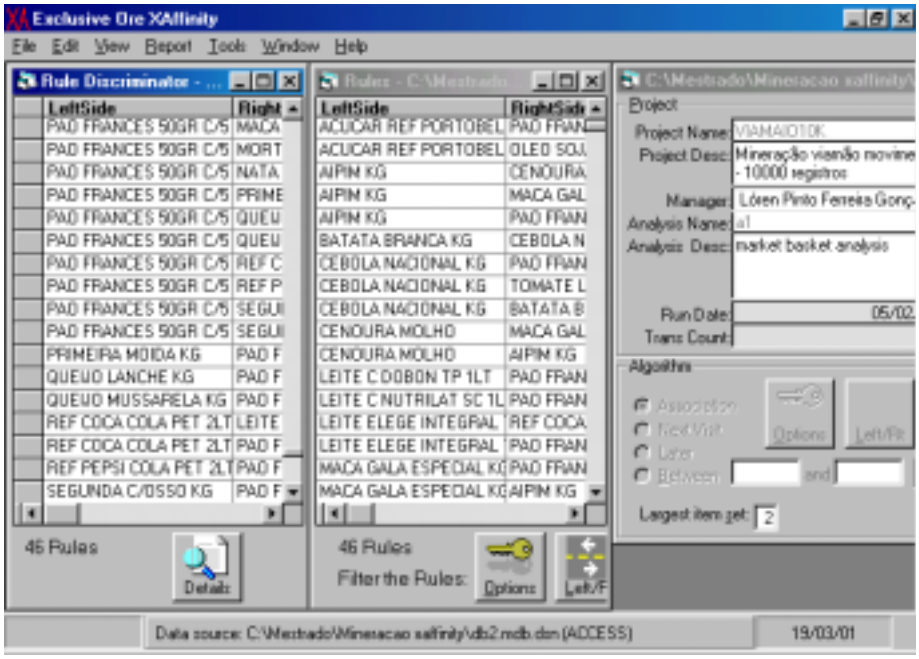

Figura 15 – Visualização vertical de regras e projeto no Xaffinity®

❖ Aplicação da ferramenta Xaffinity® à base de dados menor

A ferramenta leu a base de dados e, após a leitura, foram preenchidos os valores dos parâmetros de configuração de projeto e mineração, conforme o quadro 32. Os valores específicos desta aplicação são os mostrados no quadro 33.

Quadro 33 – Configuração de projeto utilizada na aplicação do Xaffinity® ao arquivo menor

Parâmetro	Valor
Nome do projeto	Viamao10000
Nome da análise	Mba10000
Tabela	Viamaio10000

A mineração durou 5 minutos.

A ferramenta não gerou regras. Devido a este fato, diminuímos o suporte mínimo para 1%. Nesta nova mineração a ferramenta retornou 46 regras de associação.

O cabeçalho destas regras pode ser visto no quadro 34 e a impressão das regras pode ser vista no Anexo C.

Quadro 34 – Cabeçalho da lista de regras gerada pela ferramenta Xaffinity®

SE	Valor	Então	Valor	Esperado	Confiança %	Nº de Casos	Suporte
Descrição	Açúcar ref portobello 1kg	Descrição	Pão francês 50 gr c/5	22,66	24	18	1,10
Descrição	Aipim kg	Descrição	Pão francês 50 gr c/5	22,66	18,18	21	1,29

❖ Aplicação da ferramenta Xaffinity® à base de dados média

Na segunda aplicação, foi utilizada a aplicação básica mostrada no quadro 32 e os valores referentes ao arquivo utilizado estão mostrados no quadro 35.

Quadro 35 – Configuração de projeto e mineração utilizada na aplicação do Xaffinity® ao arquivo médio

Parâmetro	Valor
Nome do projeto	Viamaob2
Nome da análise	Mbab2
Tabela	Viamaio
Suporte mínimo	1%

A mineração durou 24 minutos.

Com esta base de dados, a ferramenta gerou 80 regras de associação. A impressão de algumas destas regras pode ser vista no Anexo E.

❖ Aplicação da ferramenta Xaffinity® à base de dados maior

Na terceira aplicação, também foi utilizada a configuração básica mostrada no quadro 32 e os valores referentes ao arquivo utilizado estão mostrados no quadro 36.

Quadro 36 – Configuração de projeto e mineração utilizada na aplicação do Xaffinity® ao arquivo maior

Parâmetro	Valor
Nome do projeto	Viamaob3
Nome da análise	Mbab3
Tabela	Total
Suporte mínimo	1%

A mineração durou nove horas e cinqüenta minutos.

Com esta base de dados, a ferramenta gerou 4.804 regras de associação. A impressão de algumas destas regras pode ser vista no Anexo F.

❖ **Problemas que ocorreram durante a aplicação da ferramenta**

A ferramenta não reconhece o nosso padrão de datas (no formato dd/mm/aa) e se perde quando verifica a licença de utilização. Para resolver este problema o fabricante enviou arquivos que renovam a licença a cada entrada no sistema.

Durante a mineração, alguns problemas de sintaxe ocorrem, os quais não permitem que a mineração seja continuada. Para resolver este problema, o usuário passava por e-mail as mensagens dos problemas e o fabricante enviava alguns arquivos que deveriam ser rodados para a solução dos mesmos.

e) PolyAnalyst®[14]

Este sistema não pôde ser utilizado devido à incompatibilidade com o formato dos arquivos utilizados, pois estes estão dispostos no formato vertical, onde os dados de uma venda são identificados por um campo chave, que é repetido em cada linha. As linhas identificam os dados referentes aos itens vendidos. Já o formato utilizado por esta ferramenta é o formato horizontal, onde uma única linha representa uma venda, os itens vendidos são marcados nas colunas. Neste formato, é necessário colocar uma coluna para cada produto vendido pelos supermercados, o que seria inviável, devido à grande variedade de produtos existentes neste tipo de base de dados. A ferramenta não suporta a conversão dos dados do formato horizontal para o vertical.

f) CBA®[15]

Não foi possível utilizar esta ferramenta devido à necessidade de informarmos todos os valores possíveis para todos os campos da base de dados. Por tratar-se de bases de dados de supermercados, os quais possuem uma variedade enorme de produtos distintos, torna-se impraticável a entrada destes dados no início da mineração.

5.3.3 Avaliação das ferramentas

Abaixo será realizada a avaliação das ferramentas quanto aos seguintes itens: utilidade, facilidade, qualidade e impacto e benefícios. Essa avaliação foi realizada através do instrumento mostrado no quadro 5, do item 4.3, que foi criado baseado nos estudos de Freitas (1993) e Stumpf (1998).

Os resultados serão compostos por quadros e comentários referentes a cada item analisado.

Para melhorar o *layout* dos quadros, utilizaremos os nomes das ferramentas da seguinte forma: Aira para a ferramenta Aira Data Mining®, XA para a ferramenta Xaffinity®, SQDE para a ferramenta SuperQuery Discovery Edition e SQOE para a ferramenta SuperQuery Office Edition.

As respostas se referem às entrevistas realizadas com os gerentes da empresa e com o usuário dos sistemas.

a) Utilidade

Para a avaliação da utilidade foram considerados os parâmetros rapidez, desempenho e produtividade, eficácia, facilidade no trabalho, utilidade, tempo de resposta, aprendizagem, autonomia e independência.

---

[14] Maiores informações sobre a ferramenta no endereço www.megaputer.com
[15] Maiores informações sobre a ferramenta no endereço www.comp.nus.edu.sg/~liub

Quadro 37 - Avaliação da Utilidade da ferramenta

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
U1. Rapidez	Não	Não	Não	Não
U2. Desempenho e produtividade	Não	Não	Não	Não
U3. Eficácia	Nenhuma	Nenhuma	Nenhuma	Nenhuma
U4. Facilidade no trabalho	Nenhuma	Nenhuma	Nenhuma	Nenhuma
U5. Utilidade	Pouca	Pouca	Nenhuma	Nenhuma
U6. Tempo de resposta	Médio[16]	Médio	Pouco	Pouco
U7. Aprendizagem	Nenhuma	Nenhuma	Nenhuma	Nenhuma
U8. Autonomia e independência	Nenhuma	Nenhuma	Média	Média

Percebe-se, através do quadro acima, que as quatro ferramentas não foram consideradas muito úteis ao desenvolvimento do trabalho dos diretores da empresa. O que justifica esta posição é o fato de as ferramentas não funcionarem da forma como era esperado e porque seus resultados não foram retornados num formato adequado. Os diretores explicaram esta situação da seguinte forma:

> *"As ferramentas Aira Data Mining® e Xaffinity® seriam úteis se houvesse um trabalho maior sobre as informações geradas. Se fossem retornadas ordenadas pelo grau de confiança ou pelo número de casos, por exemplo, poderíamos saber as regras mais significativas; ou se pudéssemos selecionar, após olharmos o número de regras geradas, alguns produtos e verificarmos as regras geradas somente com estes produtos. Já nas ferramentas Super Query Discovery Edition® e Super Query Office Edition®, os valores retornados não têm utilidade porque um sistema convencional pode gerar as mesmas informações que estas ferramentas retornaram. Não precisaríamos investir em uma nova tecnologia."*

[16] Considerou-se o tempo de resposta da última versão utilizada, pois a cada nova versão este tempo diminuiu.

Quadro 38 – Avaliação da facilidade no trabalho

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
U4.1 Uso das informações torna o trabalho mais fácil	Não	Não	Não	Não
U4.2 O sistema proporciona informações necessárias ao trabalho	Não	Não	Não	Não

É importante salientar que os respondentes afirmam que as informações geradas pelos sistemas Aira Data Mining® e Xaffinity® não proporcionam informações necessárias ao trabalho desenvolvido por eles devido à forma como tais informações são apresentadas, porque seria preciso muito tempo de análise e seleção de regras que talvez tivessem alguma importância. Eles reclamam do volume de regras geradas e da falta de meios para navegar (selecionar, classificar, filtrar,etc.) entre as mesmas.

b) Facilidade

Para a avaliação da facilidade foram considerados os parâmetros aprendizado, domínio, interação, flexibilidade, facilidade no uso, funcionalidade, impacto da apresentação gráfica e qualidade da apresentação gráfica.

Quadro 39 – Avaliação da facilidade

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
F1. Aprendizado de operação	Muito Fácil	Muito Fácil	Muito Fácil	Muito Fácil
F2. Domínio	Muito Fácil	Muito Fácil	Muito Fácil	Muito Fácil
F3. Interação	Média	Média	Muita	Muita
F4. Flexibilidade	Média	Média	Média	Média
F5. Habilidade	Muito Fácil	Muito Fácil	Muito Fácil	Muito Fácil
F6. Facilidade no uso	Muito Fácil	Média	Muito Fácil	Muito Fácil
F7. Funcionalidade	Pouca	Pouca	Nenhuma	Nenhuma
F8. Impacto da apresentação gráfica	Nenhuma	Nenhuma	Nenhuma	Nenhuma
F9. Qualidade da apresentação gráfica	Nenhuma	Nenhuma	Nenhuma	Nenhuma

Quanto à facilidade, percebeu-se que os sistemas não apresentaram problemas, ou seja, quanto à interface entre os sistemas e os usuário é amigável. Os sistemas são fáceis de operar, utilizam padrão semelhante ao do Windows, o que facilita muito a sua utilização por pessoas que não conhecem os sistemas a muito tempo.

As quatro ferramentas utilizadas não apresentam informações através de recursos gráficos. Esta foi uma das reclamações dos usuários, pois, no ponto de vista deles, se estes recursos fossem utilizados, talvez a interpretação das informações se tornasse mais fácil.

Um problema apontado pelos usuários foi a funcionalidade dos sistemas.

A funcionalidade (F7) pode ser verificada através do cumprimento das promessas da teconologia por parte das ferramentas, conforme é mostrado no quadro abaixo.

Quadro 40 – Funcionalidade dos sistemas

Promessas	**Ferramentas**			
	Aira	**XA**	**SQDE**	**SQOE**
Análise de grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis	Não	Não	Não	Não
Trabalha com grandes bases de dados	Não	Não	Não	Não
Retorna conhecimento novo e relevante	Pouco	Pouco	Nenhum	Nenhum
A mineração de dados é responsável pela geração de hipóteses	Pouco	Pouco	Nenhum	Nenhum
Encontra padrões que não são encontrados por sistemas ditos	Sim	Sim	Não	Não
É capaz de aprender e apoiar a realização de descobertas a partir dos dados	Pouco	Pouco	Nenhum	Nenhum

Conclui-se, através das respostas, que as quatro ferramentas não cumprem as promessas encontradas na bibliografia. Alguns depoimentos:

Diretor Geral:

> *"Para mim, o maior problema encontrado foi com relação ao tratamento de grandes bases de dados. Na bibliografia pertinente esta é a grande promessa da tecnologia, porém, trabalhamos somente com dois meses de quatro lojas, o que para nós é uma base de dados bem pequena; mesmo assim, as quatro ferramentas 'travaram' devido ao volume de dados. No nosso entendimento deveríamos trabalhar com, no mínimo, dois anos para realmente verificarmos padrões de comportamento dos nossos consumidores sem cairmos em questões sazonais"*.

Diretora de Sistemas e Tecnologia:

> *"Os padrões encontrados pelas ferramentas Aira Data Mining® e Xaffinity® parecem trazer alguma informação importante para a empresa. É necessário analisá-las primeiro. Porém os resultados gerados pelas ferramentas Super Query Discovery Edition® e Super Query Office Edition® não trazem nada de novo para a empresa; portanto, não justificam um investimento em uma nova tecnologia"*.

Desta forma, verificamos que as ferramentas não cumpriram as promessas da tecnologia de mineração de dados e que para os tomadores de decisão da empresa as mesmas não seriam úteis para auxiliá-los na tomada de decisão e, portanto, não se justificaria o investimento em tal tecnologia.

c) Qualidade

Para a avaliação da qualidade, foram considerados os parâmetros precisão e confiabilidade, oportunidade, dificuldade para obter informação, facilidade de interpretação, fonte de informação, corretitude, medidas subjetivas de interesse e medidas objetivas de interesse.

Quadro 41 – Avaliação da qualidade

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
Q1. Precisão e confiabilidade	Média	Média	Nenhuma	Nenhuma
Q2. Oportunidade	Médio	Médio	Nenhuma	Nenhuma
Q3. Dificuldade para obter informação	Muita	Muita	Muita	Muita
Q4. Facilidade de interpretação	Nenhuma	Nenhuma	Nenhuma	Nenhuma
Q5. Fonte de Informação	Médio	Médio	Nenhuma	Nenhuma
Q6. Corretitude	Nenhuma	Nenhuma	Nenhuma	Nenhuma
Q7. Medidas subjetivas de interesse	Pouca	Pouca	Pouca	Pouca
Q8. Medidas objetivas de interesse	Sim	Sim	Sim	Sim

Com relação ao aspecto precisão e confiabilidade, os usuários ficaram preocupados, pois não se conseguiu obter resultados significativos sem a intervenção dos desenvolvedores das ferramentas Aira Data Mining® e Xaffinity®. Esta preocupação pode ser verificada no depoimento do Diretor Geral:

> *"Como poderíamos confiar em informações que foram geradas através da intervenção do desenvolvedor da ferramenta, no sentido de programar novas rotinas conforme os problemas vão surgindo. Como poderíamos saber se estas informações foram realmente geradas pelo algoritmo do sistema, que estava processando há um bom tempo ou pela pessoa naquele momento? Como poderíamos investir na compra de um sistema que não gera as informações necessárias sem que alguém introduza algum 'programinha milagroso' durante o processamento?".*

No caso das ferramentas Super Query® Discovery e Office Edition, esta intervenção não foi necessária, porém alguns valores apresentados não correspondiam à realidade da base de dados, conforme depoimento do usuário dos sistemas:

> *"Quando estávamos testando as ferramentas, no sentido de aprender a utilizá-las, aplicamos estas à base de dados da loja chamada Feitoria (a maior das quatro lojas disponíveis). Quando as ferramentas Super Query Discovery Edition® e Super Query Office Edition® retornaram os valores mais freqüentes, estranhou-se o fato de a venda de sábado ser a menor de todas. Então os dados foram abertos utilizando o Banco de Dados da Microsoft, o Access, e verificou-se o referido valor. Para nossa surpresa o valor era bem maior do que aquele mostrado pelas duas ferramentas.*

> *Através da descoberta deste fato foi preciso confirmar os valores de todas as outras aplicações, porém o problema ocorreu somente quando trabalhávamos com grandes volumes de dados".*

Como medidas subjetivas de interesse (Q7) foram avaliadas a utilidade e a inesperabilidade das regras geradas pelas ferramentas.

Quadro 42 – Avaliação de medidas subjetivas de interesse

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
Q7.1. Utilidade das regras geradas	Pouca	Pouca	Nenhuma	Nenhuma
Q7.2. Inesperabilidade das regras geradas	Média	Média	Nenhuma	Nenhuma

Os diretores da empresa consideraram as ferramentas fracas quanto às medidas subjetivas de interesse. A opinião dos mesmos está expressa nos depoimentos abaixo:

Diretor Geral:

> *"As regras geradas pelas ferramentas Aira Data Mining® e Xaffinity® não são muito úteis para nossa empresa por causa do formato em que as mesmas são apresentadas. Muitas regras são retornadas, o que torna difícil a verificação de regras inesperadas, pois a grande maioria delas são regras sem nexo ou óbvias. Quanto às outras duas ferramentas, não vejo utilidade alguma nos resultados gerados ".*

Diretora de Sistemas e Tecnologia:

> *"...talvez, se pudéssemos melhorar a forma de apresentação dos resultados gerados pelas duas primeiras ferramentas, as informações se tornassem mais úteis para nós; porém, da forma como os resultados nos são devolvidos, nos obriga a gastar muito tempo e utilizarmos outros recursos para captarmos estas informações. As duas últimas ferramentas não retornam informações desconhecidas, ou até mesmo relevantes, porque*

nosso sistema é capaz de retornar o tipo de informação gerada pelas mesmas."

Como medidas objetivas de interesse (Q8), foram avaliadas as possibilidades de ajuste, nas ferramentas, do grau de confiança e de suporte mínimo.

Quadro 43 – Avaliação das medidas objetivas de interesse

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
Q8.1. Ajuste do grau de confiança	Sim	Sim	Sim	Sim
Q8.2. Ajuste do suporte mínimo	Sim	Sim	Sim	Sim

Quanto à possibilidade de ajuste de medidas objetivas de interesse, as quatro ferramentas possuem recursos para os devidos ajustes.

d) Impacto e benefícios

Para a avaliação de impacto e benefícios, foram considerados os parâmetros contribuição, impacto, entendimento do negócio, discussões dentro da organização e competitividade.

Quadro 44 - Avaliação do impacto e benefícios

Parâmetro	Ferramentas			
	Aira	**XA**	**SQDE**	**SQOE**
IB1. Contribuição	Pouca	Pouca	Nenhuma	Nenhuma
IB2. Impacto	Nenhum	Nenhum	Nenhum	Nenhum
IB3. Entendimento do negócio	Pouco	Pouco	Nenhum	Nenhum
IB4. Discussões dentro da organização	Nenhuma	Nenhuma	Nenhuma	Nenhuma
IB5. Competitivade	Pouca	Pouca	Nenhuma	Nenhuma

Através do quadro acima, pode-se concluir que os sistemas não causaram impacto sobre a tomada de decisão da empresa, ou seja, não contribuíram para melhorar o processo decisório dentro da organização. O diretor geral da empresa salientou:

> *"As ferramentas de mineração de dados não mudaram em nada nossa tomada de decisão e nem geraram informações realmente úteis para auxiliá-la".*

#####

Algumas semelhanças foram encontradas nos resultados gerados pelas ferramentas. O tipo de regras geradas pelas ferramentas Aira Data Mining® e Xaffinity® (sob as condições já explicadas) foram semelhantes. Inclusive algumas destas regras eram iguais; alguns exemplos destas regras encontram-se no quadro abaixo.

Quadro 45 - Regras semelhantes encontradas nas ferramentas Aira Data Mining® e XAffinity®

Se	Então	Confiança	Suporte	Nº De Casos	Ferramenta
Açúcar ref portobello 1 kg	Leite bancanube l vida integ	27,4	2,09	3095	Xaffinity®
Açúcar ref portobello 1 kg	Leite bancanube l vida integ	27,4	2,09	3095	Aira Data Mining®
Banana Catarina/prata kg	Pão francês 50gr c/5	53,66	5,96	8829	Xaffinity®
Banana Catarina/prata kg	Pão francês 50gr c/5	53,66	5,96	8829	Aira Data Mining®

Concluiu-se que as ferramentas Super Query Discovery Edition® e Super Query Office Edition® são iguais, quanto ao módulo que trata da associação.

Conforme a análise proposta por Alter (1980 *apud* Freitas 1993), tentou-se nesta pesquisa responder aos seguintes pontos:

- O que os sistemas de mineração de dados estudados faziam e qual era a sua configuração técnica?
- Eles cumpriam as promessas da tecnologia de mineração de dados?
- Quem foram os usuários dos sistemas?

- Quais foram os problemas encontrados nos sistemas?
- Qual foi o impacto causado pela sua utilização?
- A avaliação: por que os sistemas são ou não um sucesso?
- Quais as contribuições e a aprendizagem trazidas pelos sistemas?

Estas questões encontram-se, juntamente com as respostas, no próximo capítulo, onde são apresentadas as conclusões do trabalho.

Capítulo 6

Considerações finais

Neste capítulo, são apresentadas as conclusões obtidas na pesquisa, seus limites, suas contribuições e é dada a sugestão à continuidade do tema proposto.

6.1 Conclusões

As empresas, em geral, estão passando por um momento de grande concorrência. Nos supermercados a competição ainda é mais acirrada. Desta forma, a necessidade de informações relevantes e desconhecidas torna-se imprescindível para obter vantagem competitiva.

Neste contexto, as tecnologias de informação tornam-se poderosas armas na ‘guerra pela competitividade’. Por isso, muitas novas tecnologias têm surgido nos últimos anos, algumas com grandes promessas. A mineração de dados é uma dessas tecnologias. Devido aos altos investimentos necessários à implantação destas tecnologias, é importante que se realize uma avaliação prévia do cumprimento das suas promessas.

De acordo com a análise proposta por Alter (1980 *apud* Freitas 1993), tentou-se nesta pesquisa responder às questões desenvolvidas a seguir.

“O que os sistemas de mineração de dados estudados faziam e qual era a sua configuração técnica?” Neste estudo foi possível perceber quais eram as promessas da tecnologia de mineração de dados e o que algumas das ferramentas existentes no mercado realmente faziam.

Outro ponto importante refere-se aos equipamentos necessários à utilização de cada uma destas ferramentas, pois uma das restrições citadas na bibliografia quanto ao uso desta

tecnologia tratava da necessidade de equipamentos caríssimos e de grande porte, quando do surgimento das primeiras ferramentas de mineração de dados.

"O problema: como o sistema poderá apoiar o usuário, eles cumpriam as promessas da tecnologia de mineração de dados?" Analisando os resultados obtidos e através da entrevista com os usuários potenciais, concluímos que as ferramentas em questão não auxiliaram aos tomadores de decisão da empresa, conforme o prometido pela tecnologia.

"Quem foram os usuários dos sistemas?" Os usuários deste tipo de sistemas foram as pessoas responsáveis pela tomada de decisão da empresa, ou seja, os diretores.

"Quais foram os problemas encontrados nos sistemas?" Os sistemas apresentaram vários problemas desde aqueles mais simples, como incompatibilidade de formato de datas (americano com o brasileiro), o que trancava a licença de utilização do sistema a cada vez que se tentava rodá-lo novamente, até problemas relacionados ao retorno de informações com valores errados.

Os maiores problemas relatados pelos usuários foram a falta de confiabilidade dos resultados gerados - devido aos fatos de os desenvolvedores precisarem intervir para a sua geração e à existência de valores incorretos que foram retornados por duas das ferramentas utilizadas; e a impossibilidade de trabalhar com grandes bases de dados, o que torna difícil obtermos padrões de comportamento do consumidor que realmente reflitam o que acontece no dia a dia dos negócios.

"Qual foi o impacto causado pela sua utilização?" Não foi possível obter o impacto desejado, devido aos diversos problemas ocorridos na utilização das ferramentas.

"A avaliação: por que os sistemas são ou não um sucesso?" Concluiu-se que as ferramentas não são um sucesso por não cumprirem suas promessas e por não trazerem retornos significativos à empresa.

"Quais as contribuições e a aprendizagem trazidas pelos sistemas?" Também não foram importantes na visão dos usuários potenciais da tecnologia na empresa.

Foi verificado por esta pesquisa que as ferramentas não apresentaram problemas com relação à facilidade de utilização. Todas as ferramentas utilizadas eram bastante amigáveis. Mas pecavam principalmente no aspectos referentes à qualidade e utilidade.

> "*O valor real de um sistema de informação não será dado por seu 'hardware' nem por seu 'software', mas pelo uso que dele fizer o sistema de*

> *decisão da empresa: tomando melhores decisões, descobrindo e aproveitando novas oportunidades de negócio, descobrindo e evitando futuros problemas" (Pereira & Perlingeiro, 1986).*

O que se pode concluir com a realização deste trabalho é que as ferramentas de mineração de dados utilizadas nesta pesquisa ainda não estão prontas para a utilização no ambiente empresarial. Muitos testes e melhorias deveriam ser realizados por parte dos desenvolvedores destas ferramentas, no intuito de melhorá-las no que se refere à utilidade para a gestão, o que ainda não está acontecendo.

Neste sentido, seria importante a realização de trabalhos em conjunto entre desenvolvedores e empresas interessadas na tecnologia, para que fosse possível o desenvolvimento e teste das ferramentas em ambiente e situações reais de negócios. Fayad (*apud* Niederman, 1997) já levantava a dúvida quanto à utilização de ferramentas de mineração de dados por pessoas que não fossem os seus desenvolvedores. O que se pode verificar neste trabalho é que este problema persiste e isto já é suficiente para impossibilitar estas ferramentas de encontrarem-se disponíveis no mercado.

6.2 Limites da pesquisa

A principal limitação desta pesquisa refere-se ao fato de ter trabalhado somente com ferramentas de mineração de dados que realizassem a tarefa ***associação***. Esta tarefa foi escolhida de acordo com o tipo de dados disponíveis e por causa do curto espaço de tempo, que não permitiria a análise de outras tarefas.

A base de dados de uma rede de supermercados também é um fator que pode ter sido limitador da pesquisa.

Outro limite diz respeito às ferramentas utilizadas que foram obtidas por conveniência, considerando além da tarefa associação, que fossem ferramentas para o ambiente Windows e liberadas para utilização sem ônus ao pesquisador.

6.3 Contribuições da pesquisa

As contriuições potenciais desta pesquisa são apresentadas nos parágrafos abaixo.

Para as empresas o conhecimento das promessas de uma nova tecnologia de informação que já está disponível no mercado, mas que ainda não é amplamente utilizada e a verificação do cumprimento destas promessas por parte de algumas ferramentas.

Para o pesquisador as contribuições deste trabalho para a sua vida profissional foram muitas, foram adquiridos conhecimentos sobre: a) uma nova tecnologia de informação; b) novas ferramentas disponíveis no mercado; c) avaliação de sistemas e d) aplicação de conhecimentos teóricos junto à realidade de uma empresa.

Para a área de Sistemas de Informação a pesquisa contribuiu mostrando as promessas da tecnologia de mineração de dados, os resultados obtidos através da aplicação de algumas ferramentas às bases de dados de uma rede de supermercados, e, principalmente, os problemas encontrados nestas ferramentas.

6.4 Sugestões para pesquisas futuras

Como continuidade para esta pesquisa, sugere-se a avaliação de outras ferramentas disponíveis no mercado, que utilizem outras tarefas de mineração de dados, ou até mesmo a avaliação de ferramentas, incluindo aquelas que utilizam a associação, em bases de dados de diferentes empresas, ou seja, com dados bem diferentes.

Outro trabalho a ser desenvolvido pode ser o teste de ferramentas que estão em desenvolvimento, a fim de detectar erros e corrigi-los antes da finalização da ferramenta.

Poderia ser realizado um teste de ferramentas utilizando uma base de dados real e uma outra criada aleatoriamente, no intuito de verificar se as ferramentas conseguem distinguir os dados reais dos aleatórios. Poderia ser verificada, também, a estabilidade das ferramentas.

Referências bibliográficas

ABRAS (Associação Brasileira de Supermercados) [25 de novembro de 1999] Disponível na World Wide Web <http://www.abrasnet.com.br>.

AGAS (Associação Gaúcha de Supermercados) [25 de novembro de 1999] Disponível na World Wide Web <http://www.agas.com.br>

ALBERTIN, Luiz. **Administração de informática: funções e fatores críticos de sucesso**. São Paulo: Atlas, 1998.

ALBUQUERQUE, Eliete. **Pequenos Supermercados têm espaço**. Zero Hora. 24 de setembro de 2000. Caderno Economia.

ALTER, S. **Information systems: a managerial perspective**. Menlo Park. CA: Benjamin e Cummings, 2ª ed. 1996.

ALMEIDA, Fernando C. **Desvendando o uso de redes neurais em problemas de administração de empresas**. RAE, São Paulo, v. 35, n. 1, p. 46-55, Jan./Fev. 1995.

ÁVILA, Bráulio Coelho. **Data mining**. Escola Regional de Informática - Regional Sul. Curitiba, 1998.

BERRY, Michael J. A., LINOFF, Gordon. **Data mining techniques: for marketing, sales and customer support**. USA: Wiley Computer Publishing, 1997.

BRAND, Estelle, GERRITSEN, Rob. **Association and sequencing**. 2000 [09 de setembro de 2000] Disponível na World Wide Web <http://www.dbmsmag.com/9807m03.html>

BRUSSO, Marcos José. **O paralelismo na mineração de regras de associação**. Porto Alegre: UFRGS, 1998 (Trabalho Individual I, Programa de Pós-Graduação em Computação).

_____, **Access miner: uma proposta para a extração de regras de associação aplicada à mineração de uso web.** 2000. [10 de setembro de 2000] Disponível na World Wide Web <http://www.inf.ufrgs.br/~brusso>.

CAMPOMAR, Marcos C. **Do uso de "estudo de caso" em pesquisas para dissertações e teses em administração**. RAE, São Paulo, v. 26, n. 3, p. 95-97, Jul./Set. 1991.

DATE, C. J. **Introdução a sistemas de bancos de dados**. Rio de Janeiro: Campus, 1991.

FAQ (Frequently Asked Questions). Data mining, 1996. [20 de dezembro de 1998] Disponível na World Wide Web <http://www.rpi.edu/faq.html>.

FELDENS, Miguel Artur. **Knowledge discovery in databases**. 1997. [20 de dezembro de 1998] Disponível na World Wide Web <http://www.ufrgs.br/~feldens>

_____, MORAES, Rodrigo Leal, PAVAN, Altino, CASTILHO, José Mauro Volkmer. **Mineração de dados na gestão hospitalar**. Porto Alegre: UFRGS, 1997. [20 de dezembro de 1998] disponível na World Wide Web <http://www.inf.ufrgs.br/~feldens/datamining.html>.

_____, CASTILHO, José Mauro V. **Engenharia da descoberta de conhecimento em bases de dados: estudo e aplicação na área de saúde**. Porto Alegre: UFRGS, 1997. [20 de dezembro de 1998] disponível na World Wide Web <http://www.inf.ufrgs.br/~feldens>.

FIGUEIRA, Rafael. **Mineração de dados e bancos de dados orientados a objetos**. Rio de Janeiro: UFRJ, 1998 (Dissertação, Mestrado em Ciência da Computação).

FREEDMAN, Alan. **Dicionário de informática: o guia ilustrado completo**. São Paulo: Makron Books, 1995.

FREITAS, Henrique. **A informação como ferramenta gerencial: um telessistema de informação em marketing para o apoio à decisão**. Porto Alegre: Ortiz, 1993.

_____, LESCA, Humbert. **A inovação e a informação: ser competitivo na era do conhecimento... também no Brasil**. Análise, Porto Alegre, v.3,nº 2, 1992.

_____, BECKER, João Luiz, KLADIS, Constantin Metaxa, HOPPEN, Norberto. **Informação e decisão: sistemas de poio e seu impacto**. Porto Alegre: Ortiz, 1997.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 1999.

GREENFELD, Norton. **Mineração de dados**. Unix Review, p. 9-14, may. 1996.

HARRISON, Thomas H. **Intranet *data warehouse***. São Paulo: Bekerley Brasil, 1998.

KOTLER, Philip. **Administração de marketing: análise, planejamento, implementação e controle**. 5ª ed. São Paulo: Atlas, 1998.

LEITE, Iara. [iara@unidão.com.br]. Informações. E-mail to Loren Pinto Ferreira Gonçalves. 05 de janeiro de 2000.

MARCOVITCH, Jacques (Org.). **Tecnologia da informação e estratégia empresarial**. São Paulo: FEA/USP, 1996.

MENCONI, Darlene. **A mineração de informações**. Info Exame. São Paulo, Ano 12, nº 144, p. 98-93, mar. 1998.

MOXON, Bruce. **Defining Data Mining**. DBMS, Data Warehouse Supplement, August, 1996.

NEWING, Rod. **Mineração de dados**. Management Accounting. p. 34-35. oct. 1996.

NIEDERMAN, Fred. **Data mining: a research framework**. Baltimore: Information Systems Research Center, 1997.

OLIVEIRA, Djalma de Pinho Rebouças. **Sistemas de informações gerenciais: estratégicas, táticas operacionais**. 4ª ed. São Paulo: Atlas, 1997

OLIVEIRA, Mirian. **Um método para obtenção de indicadores visando a tomada de decisão na etapa de concepção do processo construtivo: a percepção dos principais intervenientes**. Porto Alegre: UFRGS, Dissertação de Doutorado, PPGA/EA, 1999.

PEREIRA, Rogério C., PERLINGEIRO, Jayme E. **APX – avaliação e planejamento de sistemas de informação**. São Paulo: Edgard Blücher, 1986.

PILLA, A. D., CAPRETZ. **O processo KDD**. 1998. [08 de março de 2000] Disponível na World Wide Web <http://www.igce.unesp.br/igce/grad/computacao/~cintiab/datamine>

PRESSMAN, Roger S. **Engenharia de Software**. São Paulo: Makron Books, 1995.

RICHARDSON, Roberto Jarry. **Pesquisa social: métodos e técnicas**. São Paulo: Atlas, 1985.

ROJO, Francisco J. G. **Supermercados no Brasil: qualidade total, marketing de serviços, comportamento do consumidor**. São Paulo: Atlas, 1998.

SANTOS, Eduardo Ribas, BECKER, João Luiz. **BIACS – base de dados inteligente para aquisição de conhecimentos de sistemas.** Série Documentos para Estudo: PPGA, nº 13, 1990.

SELLTIZ, JAHODA, DEUTSCH, COOK. **Métodos de pesquisa nas relações sociais**. São Paulo: Editora Pedagógica e Universitária Ltda, 1974.

STUMPF, Evandro Carlos. **Concepção e desenvolvimento de um painel de controladoria em uma organização do setor de autopeças utilizando a tecnologia da informação.** Porto Alegre: UFRGS, Dissertação de Mestrado, PPGA/EA, 1998.

SULIMAN JR., Alberto, SOUZA, Jano Moreira. **Prospecção de Conhecimento em Bancos de Dados**. Developers Magazine, Rio de Janeiro, Ano1, Nº 6, p. 38-39,fev., 1997.

TORRES, Norberto. **Competitividade empresarial com a tecnologia da informação**. São Paulo: Makron Books, 1995.

YIN, Robert K. **Case study research: design and methods**. Second edition. Vol. 5. Sage Publications, 1994.