Universidade Federal de Pelotas
Instituto de Física e Matemática
Curso de Bacharelado em Informática

UTILIZAÇÃO DE TÉCNICAS *DATA MINING* NA BUSCA DE CONHECIMENTO NA *WEB*

Rodrigo Gonçalves Santos

Pelotas - RS
2000

Rodrigo Gonçalves Santos

UTILIZAÇÃO DE TÉCNICAS *DATA MINING* NA BUSCA DE CONHECIMENTO NA *WEB*

Monografia apresentada ao Curso de Bacharelado em Informática do Instituto de Física e Matemática da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Informática.
Ênfase: Sistemas de Computação
Orientadora: Flávia de Azambuja Carvalho
Universidade Federal de Pelotas

Monografia defendida e aprovada, em 28 de dezembro de 2001, pela banca examinadora constituída pelos professores:

Flávia de Azambuja Carvalho
Orientadora

Anderson Priebe Ferrugem

Fábio Zschornack

Dedico este trabalho a meus pais, pelo amor, pelo carinho, pela educação pela dedicação, enfim por tudo que me faz sentir verdadeiramente orgulhoso de ser seu filho

AGRADECIMENTOS

A Deus, por tudo que tenho. Aos meus pais, por me darem sempre total apoio em todos os momentos de minha vida. A minha orientadora, Professora Flávia de Azambuja Carvalho, pela dedicação e atenção. Aos meus amigos pela cooperação. Aos meus colegas de trabalho que sempre colaboraram com minhas atividades acadêmicas. A todos aqueles familiares, amigos e colegas, que sempre me ajudaram, torceram e acreditaram em mim.

SUMÁRIO

1 INTRODUÇÃO 13
2 DCBD – DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS 17
2.1 Introdução ao Processo de Descoberta em BD 17
2.2 A Mineração de Dados no Sistema de DCBD 18
2.2.1 Pré-Processamento 20
2.2.2 Mineração de Dados 22
2.2.3 Pós-Processamento 22
2.3. O Sistema de DCBD 23
2.4 Técnicas de Descoberta de Conhecimento 25
2.4.1 Correlação 25
2.4.2 Dependência de Atributos 25
2.4.3 Aprendizado Supervisionado 26
2.4.4 Aprendizado não Supervisionado - Clusterização 27
2.4.5 Detecção de Seqüências - Dimensão Temporal 27
2.4.6 Detecção de Desvio 27
2.4.7 Regressão 28
2.5 Dificuldades no Desenvolvimento da DCBD 28
2.5.1 Bases de Dados de Natureza Dinâmica 28
2.5.2 Tipos de Dados e de Bases de Dados 29
2.5.3 Incerteza 29
2.5.4 Dados Incompletos 29
2.5.5 Redundância de Dados 30
2.5.6 Dados Esparsos 30
2.5.7 Volume de Dados 30
2.5.8 Interface com o Usuário 31
3 MINERAÇÃO DE DADOS (*Data Mining*) 32
3.1 Introdução a Mineração de Dados 32
3.2 O que é Mineração de Dados? 34
3.2.1 Conceitos 34
3.2.2 Análises Orientadas pelo Usuário 36
3.2.3 O *Data Mining* 37
3.3 Fundamentos do *Data Mining* 37
3.4 Processo de Mineração de Dados 38
3.5 Tecnologia Empregada na Mineração de Dados 40
3.5.1 Retenção de Dados 41
3.5.2 Destilação de Padrões 42
3.6 Passos para a Implementação da Mineração de Dados 44
3.7 Áreas Potenciais ao uso de *Data Mining* 47
3.8 Algoritmos de Mineração de Dados 48
3.8.1 Lógica 49
3.8.2 Tabulação Cruzada 56

3.8.3 Aproximações Equacionais 59
4 Mineração de Dados na *Web* – *Web Mining* 63
4.1 Introdução à *Web Mining* 63
4.2 *Web Content Mining* (Mineração de conteúdo *Web* – páginas) 64
4.2.1 Técnicas de Recuperação de Informações 66
4.2.2 Aproximação Baseada em Agente 71
4.2.3 Aproximação de Banco de Dados 75
4.3 *Web Usage Mining* (Mineração de dados de usuários na *Web*) 76
4.3.1 Ferramentas de Descoberta de Padrões 77
4.3.2 Ferramentas de Análise de Padrão 77
4.4 Estrutura Para Mineração na *Web* 78
4.4.1 Tarefas de Pré-Processamento 78
4.4.2 Descoberta de Padrões em Transações na *Web* 81
4.5 Análise de Padrões Descobertos 85
4.5.1 Técnicas de Visualização 86
4.5.2 Técnicas de OLAP 86
4.5.3 Restrição de Dados 87
4.5.4 Análise de Usabilidade 87
4.6 Apresentação de uma Arquitetura para Mineração de Dados de Uso da *Web*. 88
4.7 Benefícios da Mineração de Dados na *Web* 91
5 TENDÊNCIAS FUTURAS EM *DATA MINING* 93
5.1. Tendências da Mineração de Dados em BD Tradicionais 93
5.1.1 Necessidades Atuais 93
5.1.2 Visão Geral 94
5.1.3 Técnicas Empregadas 96
5.1.4 Algoritmos Utilizados 96
5.1.5 Áreas de Uso 98
5.1.6 Propostas de Uso de *Data Mining* 100
5.1.7 Dificuldades para uso de DM´s 100
5.2 Tendências da Mineração de Dados na *Web* 102
5.2.1 Tendências em Web Usage Mining 102
5.2.2 Tendências em Web Content Mining 106
5.3. Fornecedores de Banco de Dados na Trilha da *Web* 108
5.4. Produtos *Data Mining* 109
6 Conclusões 114
7 Referências Bibliográficas 116

LISTA DE FIGURAS

Figura 1 – Processo Genérico de DCBD [FAY 96] 19
Figura 2 – Representação do Processo de Pré-Processamento na DCBD [FAY 96] 20
Figura 3 – Sistema Genérico de DCBD [KOR 93]..... 23
Figura 4 – Processo de Mineração de Dados [CAZ 97] 39
Figura 5 – Tecnologia Data Mining [CAZ 97]..... 40
Figura 6 – Retenção de Dados [POL 97]..... 42
Figura 7 – Processo Indutivo [POL 97]..... 52
Figura 8 – Algoritmos Genéticos (troca de padrões) [POL 97]..... 53
Figura 9 – Árvore de Decisão [POL 97]..... 55
Figura 10 – Agentes [POL 97]..... 57
Figura 11 – Rede de Confiança [POL97]..... 58
Figura 12 – Aproximações Estatísticas [POL 97] 59
Figura 13 – Modelo de Redes Neurais [POL 97]..... 60
Figura 14 – Propagação Realimentada [POL 97] 61
Figura 15 - Uma Arquitetura para Mineração de Dados de Uso de Rede [MOB 96] 89

LISTA DE TABELAS

Tabela 1 – Regras Condicionais [POL 97] 50
Tabela 2 – Tabela para Árvore de Decisão [POL 97] 54
Tabela 3 – Tabulação Cruzada [POL 97] 56
Tabela 4 – Número de hosts e domínios da Internet [NET 97] 63
Tabela 5 - Comparação entre Mecanismos de Pesquisa e Agentes de Informação Inteligentes [CAS 98] 73
Tabela 5 – Produtos Data Mining e suas Respectivas Áreas 113

LISTA DE SIGLAS

DCBD – Descoberta de Conhecimento em Bases de Dados
MIS – *Management Information Systems*
WWW – *World Wide Web*
BD – Banco de Dados
LN – Linguagem Natural
AM – Aprendizado de Máquina
SGBD – Sistemas Gerenciadores de Banco de Dados
OLAP – *On-Line Analytical Processing*
ODBC – *Open database Connectivity*
SQL – *Structured Query Language*
URL – *Uniform Resource Locator*

RESUMO

Este trabalho faz um estudo sobre a tecnologia de mineração de dados, desde sua origem dentro do processo de descoberta de conhecimento em bases de dados, até sua mais recente aplicação que é o ambiente *Web*. A tecnologia de mineração de dados, a qual vem se destacando muito atualmente, devido a grande tendência mundial pela busca de conhecimento útil escondido nas bases de dados, será abordada neste trabalho levando-se em consideração suas características, processos, algoritmos, técnicas, áreas de utilização, tendências futuras, entre outros aspectos que merecem importância no estudo desta tecnologia nova e promissora, tanto na busca de conhecimento em banco de dados tradicionais de empresas, como também na *World Wide Web*.

Palavras-chave: Banco de Dados; Descoberta de Conhecimento em Bancos (Bases) de Dados, Mineração de Dados, Mineração de Dados na *Web*, Mineração de Conteúdo na *Web*, Mineração de Dados de Usuários na *Web*, Conhecimento.

ABSTRACT

This work is a study on the technology of mining data, from its origin in the process of discovery of knowledge in data bases, to its most recent application - the Web. The technology of mining data, which is becoming famous nowadays, due to the great world tendency of searching for useful knowledge hidden in data bases, will be approached in this work, taking into account its characteristics, processes, algorithms, techniques, areas of usage, future tendencies among other aspects that deserve an important place in the study of this new and promising technology, in the search for knowledge in traditional data bases in companies as well as in the World Wide Web.

Word-key: Database; Knowledge Discovery in Databases, *Data Mining*, Web Mining, Web Content Mining, Web Usage Mining, Knowledge.

1 INTRODUÇÃO

Atualmente o ambiente de negócios cresce muito, e com ele cresce também a quantidade de informações armazenadas, tornando a tecnologia da informação o fator chave para se conseguir destaque no mercado. Ela é quem determinará a alta eficiência de uma organização e destinguirá os competidores que sobreviverão dos que cairão por terra. A melhor utilização da tecnologia da informação pelos competidores se apresenta como uma vantagem, pois esta agrega valores nas decisões a serem tomadas, sendo que estas decisões são cruciais e tem de serem mais eficientes possíveis. Uma melhor análise dos dados visa descobrir o que é realmente relevante para uma tomada de decisão.

Visando, então, esta busca desesperada pela melhor utilização desta tecnologia, organizações estão migrando cada vez mais para o processo de descoberta de conhecimento em bases de dados, cujo objetivo principal é a mineração de dados para obtenção de conhecimento útil e respectivo lucro através da aplicação desse conhecimento descoberto.

O termo descoberta de conhecimento em base de dados - DCBD (*Knowledge Discovery in Databases - KDD*) foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento em dados. O termo mineração de dados, é então a etapa de aplicação de técnicas/ferramentas para apresentar e analisar dados. A denominação "mineração de dados" tem sido utilizada por estatísticos, analistas de dados e a comunidade de sistemas de gerenciamento de informação (*management information systems - MIS*), enquanto que *KDD* tem sido utilizada por pesquisadores na área de inteligência artificial.

Além das organizações buscarem se aprimorar cada vez mais nesta área de descoberta de conhecimento em bases de dados, o crescente avanço da Internet e da multiplicação das bases de dados na *Web* fez com que estas organizações buscassem também concorrer dentro do ambiente da WWW (*World Wide Web*), dando origem ao

e-commerce (comércio eletrônico), onde grande número de empresas comerciais passaram a disponibilizar seus produtos, serviços e informações através da rede.

Devido a esta invasão da *Web* pelas organizações e ao consecutivo aumento explosivo das fontes de informação disponível neste ambiente, tornou-se necessário para as organizações utilizarem ferramentas automatizadas para achar, extrair, filtrar, e avaliar informações e recursos desejados, visando obter lucro também dentro da *Web*. Desta forma deu-se origem ao termo descoberta de conhecimento na *Web*, mais conhecido como mineração de dados na *Web* (*Web Mning*).

Um outro ponto importante na aplicação de mineração de dados na *Web*, além da procura pelo conhecimento que está camuflado nas bases de dados e, que se extraído de forma correta, pode trazer lucro às organizações, é a busca de conhecimento disponível nos bancos de dados de forma clara, ou seja, a busca por conteúdo presente na *Web*, pois o avanço da Internet trouxe a crescente insatisfação dos usuários que buscam por conhecimento na *Web*, os quais perdem muito tempo pesquisando, tendo em vista que encontram-se, via de regra, perdidos neste grande universo de informações localizadas de forma dispersa e de difícil localização. A expansão da Internet também fez com que a globalização da informação se tornasse um fato real e com isso, a luta para obte-la de forma mais rápida e eficiente cresceu muito. Tendo em vista estas necessidade, técnicas para mineração de dados na *Web* também são aplicadas neste ramo, visando uma melhor performance na busca do conhecimento desejado, pois embora navegar por meio de *hiperlinks* seja fácil e divertido, ele falha como um meio de descoberta de informações na *Web*.

O trabalho proposto tem por objetivo expor a técnica de mineração de dados, desde sua origem dentro do processo de descoberta de conhecimento em bases de dados (DCBD), passando por sua aplicação em banco de dados tradicionais até chegar a sua mais nova aplicação, ou seja, a mineração de dados no ambiente *Web*, a qual utiliza-se de adaptações das técnicas tradicionais de mineração de dados. A apresentação da Mineração de Dados na *Web* (*Web Mining*) será abordada levando-se em consideração suas duas principais formas: a mineração visando a busca de conhecimento e a mineração visando a busca de conteúdo na *Web*.

Este trabalho foi dividido em seis capítulos. O primeiro capítulo apresenta uma breve introdução sobre o conteúdo que será mostrado neste trabalho.

No segundo capítulo, é apresentado os fundamentos da área de Descoberta de Conhecimento em Bases de Dados, onde serão mostrados conceitos e definições de diversos autores nesta área. Será exposto também neste capítulo as técnicas e as fases deste processo de DCBD, com suas devidas características, dando um leve destaque ao núcleo deste processo, que é a Mineração de Dados, e por fim serão apresentadas as dificuldades que ainda são encontradas nesta área nova e promissora.

O terceiro capítulo se detêm no estudo do núcleo da mineração de dados propriamente dita. Neste capítulo serão apresentados os fundamentos da mineração de dados, primeiramente mostrando uma introdução a esta etapa tão importante do DCBD, em seguida mostrando os diferentes tipos de analises de dados, de modelos de descoberta, de tecnologias de mineração, sendo que estas tecnologias também serão mostradas através dos algoritmos de *Data Mining*, os quais variam de acordo com as condições de mineração. Aqui é apresentado o processo de mineração de dados onde esta tecnologia teve início, ou seja, aplicado em ambientes fora da Web.

O quarto capítulo apresenta o mais novo campo de aplicação desta técnica de mineração, a *Web*. Campo este que está crescendo muito, pois com o acréscimo de dados que trafegam e se acumulam na rede cresce também a necessidade de se achar um meio que consiga filtrar estes dados que estão disponibilizados de forma desorganizada na *Web*. Será mostrado neste capítulo uma pequena introdução a esta área de *Web Mining*, e também as técnicas e ferramentas que estão atuando na *Web*, tanto na mineração de conteúdo (páginas), como na mineração de informações específicas de usuários, que encontram-se camufladas nos bancos de dados.

O quinto capítulo apresenta as tendências futuras para esta tecnologia, tanto na aplicação em banco de dados tradicionais, como na Web, identificando as técnicas e ferramentas que serão utilizadas, as áreas que merecem maior estudo, e também as dificuldades que ainda serão encontradas na aplicação desta tecnologia de mineração de dados.

Por fim, serão apresentadas as conclusões e as referências bibliografias.

2 DCBD – DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

2.1 Introdução ao Processo de Descoberta em BD

Segundo Wüthrich apud Xavier, Descoberta de conhecimento é o processo capaz de gerar informações automaticamente, a partir de exemplos de uma base de dados, e representá-la de forma que os seres humanos possam entendê-la. A forma compreensível a qual foi citada na definição acima é aquela na qual existe algum formalismo simbólico, os quais poderiam ainda ser traduzidos para linguagem natural (LN) [XAV 00].

Quando se deseja que um processo de descoberta de conhecimento se realize de forma automática é necessário se ter os seguintes elementos:

- Um objeto, representado por um atributo da base de dados. Por exemplo, um atributo que indique que a empresa é bem gerenciada, possibilitando que o sistema possa descobrir regras que evidenciem uma empresa bem gerenciada.

- Um conjunto de treino que, no exemplo citado acima, conteria informações sobre diversas companhias, aquelas que são bem gerenciadas (exemplo) ou não (contra-exemplos).

Também, segundo o mesmo autor, a abordagem da DCBD se dá seguinte forma [XAV 00]:

- Simbólica, devido a utilização de regras;

- Supervisionada, porque o sistema visa encontrar um objetivo;

- Orientada a base de dados, pois como os exemplos estão representados em uma base de dados o objetivo é dado em termos de atributo ou valor de atributo sobre o qual se deseja descobrir;

- Extensível no sentido de que as regras descobertas podem ser somadas a outras regras.

O mecanismo básico para a descoberta do conhecimento é chamado aprendizado indutivo. Humanos, bem como outras criaturas inteligentes, os quais são chamados genericamente de sistemas cognitivos, procuram entender seu ambiente através de representações feitas de forma simplificada, chamadas modelos. O processo de construção destes modelos do ambiente é chamado aprendizado indutivo [HOL 94].

A DCBD busca um processo de aprendizagem de forma indutiva, ou seja, buscando representações intencionais da informação especificada em uma base de dados. Além disso, ela utiliza técnicas de aprendizado de máquina (AM), que estão ligadas diretamente a área da inteligência artificial, pois como a quantidade de hipóteses de relacionamentos, ou seja, fórmulas, regras, etc, que podem ser geradas a partir de uma base de dados é enorme, torna-se completamente inviável a busca exaustiva por este conhecimento.

2.2 A Mineração de Dados no Sistema de DCBD

De acordo com os conceitos anteriores, a Descoberta de Conhecimento em Base de Dados nada mais é do que o processo que visa extrair de uma grande base de dados aqueles conhecimentos que sejam novos, úteis e interessantes, sendo que este processo se dá de forma iterativa e interativa. O processo é dito iterativo pois é formado por uma série de etapas seqüenciais, onde se pode retornar às descobertas realizadas, ou a falta delas podem levar a novas hipóteses de descobertas. E o processo torna-se interativo, pois neste caso, o usuário pode decidir pela retomada dos processos de mineração ou a uma nova seleção de atributos, por exemplo, para validar hipóteses que surgiram durante o processo.

O processo de DCBD é composto por uma série de atividades, dentre essas atividades, a mineração de dados, por sua vez, é considerada o núcleo do processo, pois ela é responsável pela aplicação dos algoritmos de extração de padrões a partir de dados[FAY 96].

Na figura 1 abaixo, é mostrado o processo genérico de descoberta de conhecimento em bases de dados, segundo Fayyad, com suas devidas fases, ou seja, seus devidos processos aplicados as bases de dados.

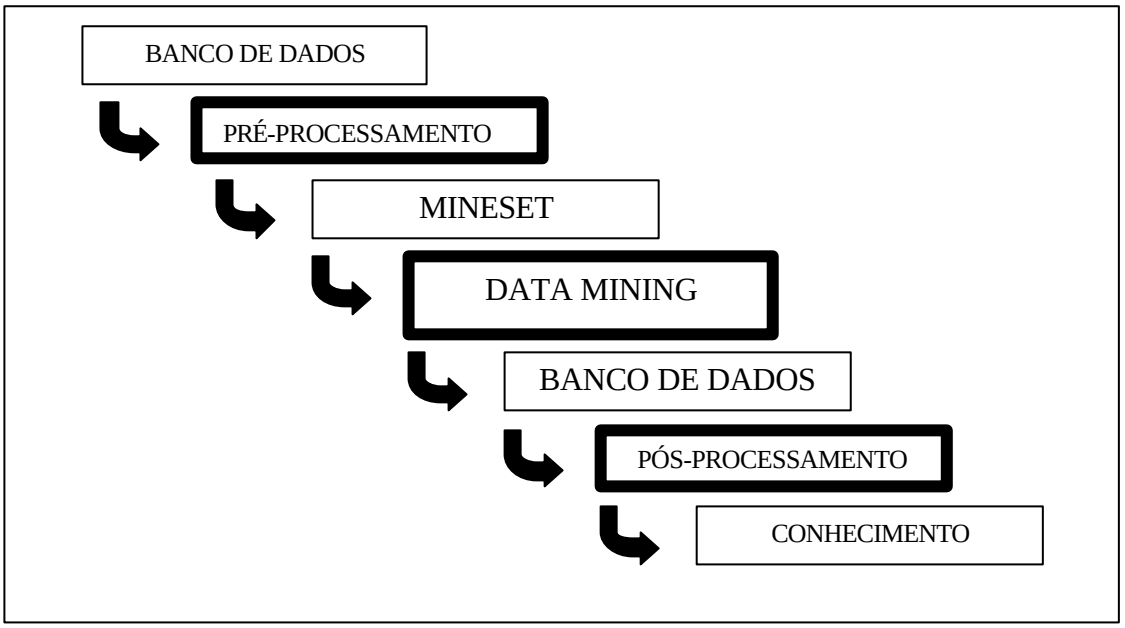

Figura 1 – Processo Genérico de DCBD [FAY 96]

a) Pré-processamento: Visa criar representações convenientes para os algoritmos de mineração, a partir da base de dados. Inclui a seleção, amostragem, transformações de representação, etc.

b) Mineração de Dados: Aplicação de Algoritmos de aprendizado aos dados pré-processados;

c) Pós-processamento: é responsável pela seleção e ordenação das descobertas interessantes que aconteceram na mineração, também cuida dos mapeamentos de representação de conhecimento e geração de relatórios.

2.2.1 Pré-Processamento

A etapa de pré-processamento envolve vários aspectos a serem considerados dentre eles podemos destacar a definição e entendimento do problema, a coleta e extração dos dados, organização dos dados e a engenharia dos dados, como mostra a seguir a figura 2[FAY 96].

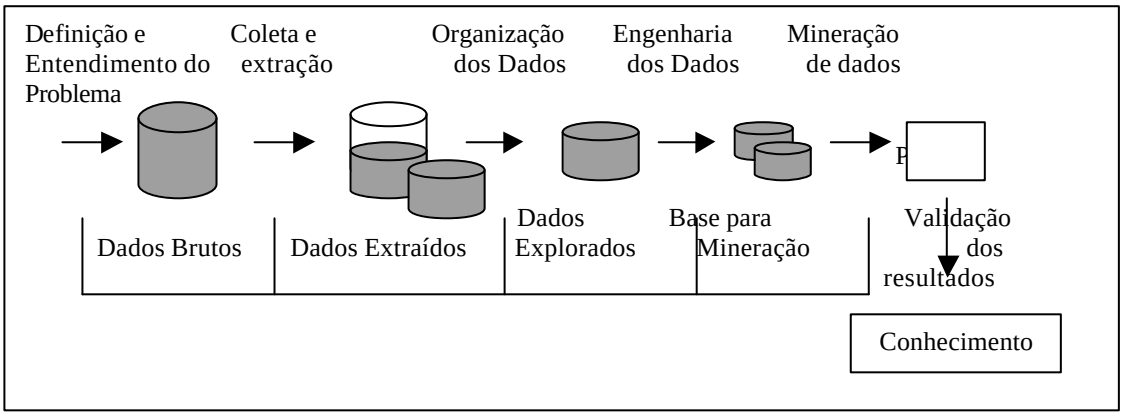

Figura 2 – Representação do Processo de Pré-Processamento na DCBD [FAY 96]

a) Definição e Entendimento do Problema: primeiramente é necessário se definir e entender bem o problema que se deseja solucionar, porém deve ser feito um exame mais detalhado desta atividade, pois se não tivermos um claro entendimento do problema e do objetivo que se quer atingir, os resultados obtidos no processo de mineração poderão ser totalmente desprezíveis. O problema então, deve ser definido de forma conjunta entre o analista de mineração e o especialista da área determinada, para que desta forma o problema seja especificado da melhor forma possível, ou seja, aquela em que os resultados obtidos possam ser medidos.

b) Coleta e Extração dos Dados: esta fase se destina a obtenção de dados relevantes para a solução do problema, sendo que estes dados podem vir de várias fontes como: banco de dados operacionais ou de um Data Warehouse, adquiridos de terceiros, estatísticas públicas e outros. Os dados também podem ser coletados através de diversas tabelas, podendo haver ainda cruzamento de dados com terceiros para posterior mineração, como por exemplo em um sistema bancário, onde o banco possui

seus dados espalhados em tabelas pertencentes a diversos sistemas como: cadastro, empréstimos, cobrança, e outros. Logo precisando ser agrupados em um único arquivo.

É preciso que duas seguintes atividades sejam desempenhadas para que se possa obter um arquivo para mineração de dados:

- Definição dos atributos: aqui se toma decisões importantes, com alto grau de subjetividade, pois pode se escolher atributos que influenciam uma determinada condição, caindo em um processo circular onde não se descobre nada de novo, não podendo também deixar todos os atributos serem tratados , pois neste caso podemos chegar a um conjunto ótimo de atributos, ficando como uma questão a ser examinada caso a caso.

- Extração física dos dados de diversas fontes: esta atividade além de ser extremamente árida e tediosa, pode trazer surpresas desagradáveis, como: ocorrência de exceções, apresentação de arquivos em formatos diferentes do que consta na documentação, falta de padrão na codificação, além do que, caso se precise incluir um arquivo mais atual pode-se ainda descobrir erros que não haviam na versão anterior, ou que passaram desapercebidos.

c) Organização dos Dados: depois de adquiridos os dados relevantes é importante que o analista se familiarize com a base de dados que será manipulada, procurando saber além do nome de cada atributo também o seu significado. Também é importante que sejam realizados alguns testes de sanidade dos dados, tendo em vista o alto risco de ocorrência de erro na fase anterior. Uma outra ação importante nesta fase é a aplicação dos algoritmos de mineração sobre estes dados ainda não totalmente corretos, onde poderemos observar dados incorretos ou obter descobertas importantes, caso haja resultados inacreditavelmente bons, ou drasticamente fora de algum limite esperado.

d) Engenharia dos Dados: depois de passar pelas fases anteriores chegamos a este ponto com toda a base de dados estática pronta para mineração, a qual

contém toda as informações necessárias no processo de mineração. Muitas bases de dados podem ser criadas no processo de engenharia de dados, em contraste com a base de mineração única criada a partir da fase de coleta e extração de dados. Isto ocorre pela necessidade de se explorar o uso de diferentes atributos, diferentes tamanhos de amostra, no caso de ser vantajoso a análise sobre um subconjunto de dados, e diferentes enfoques na definição do problema. Quando fizemos esta derivação de dados em subconjuntos estamos na verdade testando diversas hipóteses de trabalho, visando encontrar aquela que nos disponibiliza a melhor solução.

2.2.2 Mineração de Dados

A mineração de dados é a principal etapa no processo de DCBD, pois ela é o núcleo deste processo, ou seja, onde todos os algoritmos são aplicados sobre o conjunto de dados relevantes. Devido a sua importância, esta etapa do processo de DCBD será discutida detalhadamente no capítulo 3.

2.2.3 Pós-Processamento

O Pós-processamento tem a função de interpretar e validar os resultados obtidos durante a aplicação dos algoritmos na fase de mineração de dados. Esta interpretação e validação dos resultados é muito importante porque muitas vezes um padrão encontrado em uma base de dados pode não ser válido para outros casos onde a base de dados utilizadas para definir o padrão não é mesma, neste caso usamos o termo "*overfitting*", ou seja, o modelo encontrado está particularizado para a base de treinamento. Para que isto não ocorra é importante então se tomar algumas medidas. Primeiramente, é necessário se definir uma métrica para a quantidade de acertos do modelo, e isto depende do tipo de mineração de dados.

2.3. O Sistema de DCBD

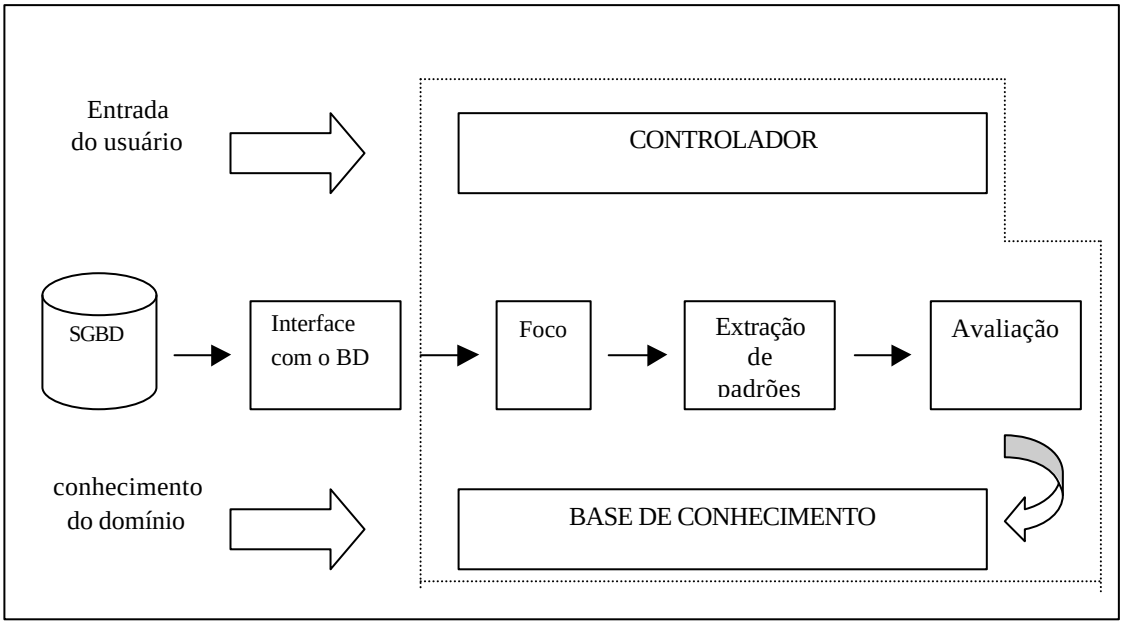

Figura 3 – Sistema Genérico de DCBD [KOR 93]

O tipo de informação fornecida pelo usuário varia entre cada sistema, onde o usuário pode estar completamente ligado ao processo de busca atuando de forma interativa, ou simplesmente fornecendo informações úteis para "focalizar" a busca. A informação pode ser fornecida pelo usuário através de uma linguagem de alto nível por menus, caixas de diálogo, etc. A fonte de dados pode ser um sistema de gerenciamento de bases de dados SGBD [KOR 93].

Na figura 3 pode-se identificar seis módulos no Sistema de DCBD [KOR 93]:

a) Controlador: responsável pelo interfaceamento e a chamada de cada um dos demais módulos, ele influencia diretamente na autonomia de um sistema de descoberta de conhecimento. O Controlador pode agir automaticamente, selecionando estratégias e orientando o foco de busca e a avaliação das descobertas ou simplesmente disparando seqüencialmente uma série de algoritmos, os quais são determinados pelo usuário. Os sistemas atuais encontram-se distantes daquilo que seria desejável em

termos de autonomia, fazendo com que o usuário atue como se fosse o próprio controlador.

b) Interface com BD: responsável pela comunicação com o banco de dados (geração de consultas). Há sistemas de DCBD que possibilitam a implementação de mecanismos de otimização de consulta mais adequados para os algoritmos de extração de padrões, pois montam sua interface com a base de dados com um formato próprio de armazenamento e gerenciamento. O ponto negativo desta interface própria é o fato da mesma gerar a necessidade de tradução dos dados a serem utilizados para o formato do sistema.

c) Base de conhecimento: contém conhecimento específico do domínio em que está sendo realizada a descoberta, tanto o conhecimento fornecido pelo usuário, quanto no conhecimento descoberto. Podem ser fornecidas ao sistema regras já conhecidas ou conhecimento que vai dar suporte à descoberta de regras. Um exemplo de conhecimento "de suporte" são as categorizações ou discretizações, por exemplo, quando uma variável contém informação quantitativa que deve ser transformada em uma informação categórica, neste caso o usuário pode dizer ao sistema o que é uma temperatura alta ou baixa, definindo limiares. Alguns algoritmos de aprendizado, podem se valer deste conhecimento para guiar o processo de busca e otimizar a sua performance.

d) Foco: indica as seções da base de dados que devem ser analisadas: O foco é responsável pela seleção dos atributos e registro acessados, baseado na estrutura da base de dados, além disso o módulo de foco identifica quais são os atributos relevantes, isso pode ser feito com base nos conhecimentos fornecidos pelo usuário ou não.

e) Extração de padrões: núcleo do sistema, o qual é composto pelos algoritmos de extração de padrões. Apesar da extração de padrões contar com o apoio de diversas áreas de pesquisa como aprendizado de máquina e inteligência artificial, as técnicas existentes ainda precisam ser aperfeiçoadas e adaptadas, pois as mesmas apresentam diversos problemas relacionados com as bases de dados. Para se amenizar

um pouco este problema estão sendo utilizadas algumas técnicas estatísticas adaptadas a algoritmos de aprendizado, além de metodologias híbridas. É importante citar também que, segundo Zytkow, mesmo de uma base de dados de tamanho "médio" pode ser extraída uma quantidade enorme de hipóteses, enquanto a maior parte destas não são interessantes para o usuário [ZYT 93];

f) Avaliação: avalia se os padrões encontrados são "fortes", "interessantes", "novos", e "úteis". Conforme esta avaliação, se organiza o conhecimento a ser apresentado ao usuário, tornando este módulo vital para que um sistema de DCBD possa ser utilizado em aplicações práticas. Dentre os critérios utilizados para a avaliação estão a precisão, cobertura, simplicidade, novidade. Ainda podem ser utilizados alguns critérios que dependam do domínio, como por exemplo: num sistema médico , priorizar as descobertas relacionadas com maior morbidez; num sistema financeiro, priorizar anomalias relacionadas com transações de maior montante.

2.4 Técnicas de Descoberta de Conhecimento

Como o exposto anteriormente a mineração de dados quando aplicada devidamente sobre a base de dados, proporciona a descoberta de padrões nesta base analisada. Estes padrões detectados são associações interessantes entre os elementos contidos na base de dados e podem ser obtidos através de diferentes técnicas, as quais serão apresentados sucintamente a seguir [XAV 00].

2.4.1 Correlação

A correlação indica que existe algum controle entre duas variáveis, ou seja, que o valor de um atributo influencia o valor de outro, podendo quantificar esta influência. A representação das correlações pode ser mostrada através de "grafos de correlação".

2.4.2 Dependência de Atributos

A dependência ocorre quando existir um valor entre dois atributos que pode vir a influenciar o outro. Podemos citar como exemplo de dependência:

a) Em um sistema de informações de uma mecânica de automóveis, poderia ser detectado que:

- determinados problemas de veículos aparecem sempre associados entre si;

- determinados procedimentos aparecem associados exclusivamente quando ocorre um problema no motor e outros quando ocorrem problema na suspensão;

b) Analisando o sistema de locações de uma vídeo locadora:

- Onde locatários de filmes de aventura acabam locando também um filme de gênero dramático .

2.4.3 Aprendizado Supervisionado

Se dá quando temos um atributo que indique a classe a qual os indivíduos na base de dados pertencem (atributo-objeto). O algoritmo de extração faz uma identificação das características comuns entre membros de uma classe e fornece uma descrição para cada classe [HOL 94].

Um tipo de descrições bastante utilizada, a qual usa uma representação a partir de regras, é a descrição característica. Como exemplo deste tipo de descrição temos:

É_animal(x), tem_asas(x), pode_voar(x) $\rightarrow$é_pássaro

Neste caso no lado esquerdo está a descrição propriamente dita, em termos de valores de atributos e no lado direito a classe a que os indivíduos com aquelas características pertencem.

Existe ainda as chamadas descrições discriminativas, onde o processo de aprendizado procura por diferenças entre as classes e apresenta os fatores que distinguem os indivíduos das diferentes classes examinadas.

2.4.4 Aprendizado não Supervisionado - Clusterização

Neste caso, não é informado a que classe pertence cada uma das entidades, cabendo ao algoritmo explorar diferentes alternativas e detectar padrões. Uma vez detectados estes padrões, são descritos os conceitos.

Neste tipo de aprendizagem, diferentes algoritmos podem optar por formas alternativas de "*clusterizar*" os dados, podendo se obter diferentes resultados a partir da mesma Base de Dados.

2.4.5 Detecção de Seqüências - Dimensão Temporal

É a detecção de regras e dependências levando-se em consideração a dimensão temporal.

Utilizando o exemplo da mecânica de automóveis, poderia se descobrir que:

- Determinado procedimento sempre procede outro;

- Determinado procedimento só pode ser repetido após um intervalo mínimo de tempo(periodicidade) – desgaste de uma peça por exemplo.

2.4.6 Detecção de Desvio

Esta técnica é utilizada para detectar anomalias em bases de dados, podendo evidenciar uma série de problemas como: quantidade de dados, fraudes ou descobrir eventos raros. Consiste na procura feita pelo algoritmo de elementos contidos no banco de dados que estão fora do conjunto de padrões determinados por dependências

seqüenciais e/ou descrições de conceitos, que podem ser obtidos automaticamente ou dados pelo usuário [FEL 97]. Analisando ainda o exemplo da mecânica de automóveis, poderiam ser encontrados registros indicando:

- Um motor trabalhando perfeitamente com um óleo “super vencido” ;

- Suspensão de um modelo de automóvel funcionando perfeitamente em outro;

2.4.7 Regressão

Os algoritmos de regressão se caracterizam por procurarem a descoberta de fórmulas, e atuam de forma semelhante aos que realizam aprendizado supervisionado. Na regressão procura-se representações que mapeiem cada entidade para um valor numérico, ao invés de se procurar regras, *frames* ou árvores capazes de mapear cada entidade para uma classe [FAY 96].

2.5 Dificuldades no Desenvolvimento da DCBD

São expostos aqui algumas metas a serem alcançadas ainda pela DCBD para a resolução de determinados problemas; metas estas que merecem uma importância considerável no caso de se desenvolver um projeto que utilize DCBD. A maioria destes problemas a serem superados estão relacionados diretamente com as características das próprias bases de dados. Existe também a dúvida ao se saber se o conhecimento que estásendo extraído é completo, pois mesmo a base de dados sendo consideravelmente grande, ela pode ser pequena no universo a qual ela se encontra e por fim não podemos deixar de levar em consideração aqueles aspectos relacionados ao próprio sistema que está sendo implementado, tais como interação com o usuário e visualização [XAV 00].

2.5.1 Bases de Dados de Natureza Dinâmica

Em muitas aplicações as bases de dados são constantemente alteradas, o que poderia levar à descobertas errôneas. A solução adotada por muitos sistemas é se fazer

uma seleção a partir da base de dados a ser analisada, e armazenando esta parte selecionada, com isso se consegue aumentar a imunidade às alterações e consequentemente melhorar o desempenho da aplicação de mineração.

2.5.2 Tipos de Dados e de Bases de Dados

Como atualmente existe muito pouca teoria relacionando tipos de bases de dados e algoritmos de aprendizado de máquina, recomenda-se que quando possível, a aplicação seja testada com vários algoritmos e, através destes testes, se encontre o mais adequado. As bases de dados podem conter informações sobre diferentes domínios, as quais podem estar armazenadas através de diferentes tipos de dados, além disso as informações podem formar escalas diferentes, tais como nominal, ordinal, intervalos ou porções, fazendo com que para cada escala seja mais eficiente a aplicação de um determinado tipo de algoritmo de extração. Diferentes tipos de dados, tais como imagens ou temperaturas também podem necessitar de tratamento diferenciado. Por esta razão, se diz que não existe um único bom "algoritmo" para aprendizado [ZYT93].

2.5.3 Incerteza

Uma vez que muitos modelos da realidade são probabilísticos, é importante que se trabalhe com formalismos de representação de conhecimento probabilístico e muitas vezes tirar proveito de técnicas estatísticas para estimar a validade do modelo.

2.5.4 Dados Incompletos

Como raramente as bases de dados existentes são projetadas tendo a descoberta de conhecimento em mente, pode ocorrer com certa freqüência em banco de dados que certas propriedades (atributos) importantes para o processo de descoberta não estejam presentes. Também se evidencia a ocorrência e dados incompletos se deve à estrutura das bases de dados relacionais. Este modelo representa dados através de tabelas, de forma que todos os registros (linhas) têm de ter um valor para todos os atributos (colunas). Quando ocorrer de um determinado atributo não se aplicar a um

determinado registro, este terá um valor nulo, que deverá ser tratado de alguma forma pelos algoritmos de extração.

2.5.5 Redundância de Dados

Muitas vezes as bases de dados contêm informações, geralmente funcionais, que são redundantes. Tais redundâncias podem ser detectadas pelo algoritmo de extração como se fossem conhecimento útil descoberto.

2.5.6 Dados Esparsos

Uma base de dados geralmente esparsa representa uma pequena amostra de todas as instâncias possíveis das classes que se deseja descrever, ou seja, as bases de dados quase sempre não conseguem abranger o número total de situações possíveis, por maior que seja esta base de dados. A conseqüência disto é que pode-se descobrir um conhecimento válido para a base de dados, mas que não seja útil no mundo real [HOL94]. Por exemplo, sobre os dados de um hospital , um sistema poderia "descobrir" que todos os indivíduos que doam sangue pesam mais que 48 kg.

2.5.7 Volume de Dados

Segundo Zytkow, a variedade de padrões que podem ser encontrados em uma base de dados, mesmo que de tamanho moderado, pode chegar a bilhões de hipóteses [ZYT93]. Além disso, as bases de dados utilizadas para descoberta de conhecimento geralmente são muito grandes [HOL94].

Em [HOL94] é citado um exemplo real utilizando uma base de dados do American Rheumatism Association Medical Information System contendo informações sobre pacientes. Como esta base de dados foi projetada visando aplicação de técnicas de aprendizado, cada visita era registrada de forma detalhada, através de 400 atributos e cada paciente tinha entre 50 e 100 visitas. Assim, cada paciente tinha um total de 20.000 a 40.000 atributos.

Conclui-se então que não há como fugir do grande volume de dados , pois este grande volume é de fundamental importância para diminuir os problemas gerados por dados esparsos e também para reduzir o efeito de eventuais inconsistências da base de dados sobre os padrões descobertos. Em [HOL94] uma base de dados típica de descoberta de conhecimento é caracterizada como sendo maior do que 100.000 registros.

2.5.8 Interface com o Usuário

Podemos destacar aqui o quanto o projeto da interface é importante, de modo que a ferramenta seja desenvolvida de acordo com o usuário, ou seja, usuários menos especializados provavelmente irão preferir o menor nível de interação possível; já engenheiros do conhecimento, por exemplo preferirão uma interface com maior interação. Esta interação existe pois, os sistemas de descoberta de conhecimento em bases de dados necessitam de configurações para a aplicação de cada usuário e cada base de dados.

3 MINERAÇÃO DE DADOS (*Data Mining*)

3.1 Introdução a Mineração de Dados

A competitividade no ambiente de negócios atual está maior do que nunca, e o fator chave que distingue os competidores que sobreviverão dos que cairão por terra é a alta eficiência. A melhor utilização da tecnologia da informação pelos competidores se apresenta como uma vantagem, pois esta agrega valores nas decisões a serem tomadas, sendo que estas decisões são cruciais e tem de serem mais eficientes possível. Uma melhor análise dos dados visa descobrir o que é realmente relevante para uma tomada de decisão.

Segundo estimativas, a quantidade de informação no mundo dobra a cada 20 anos. Computadores armazenam um volume de dados que vem crescendo a taxas elevadas. À onze anos atrás, era estimado em cinco milhões o número total de dados no mundo sendo que sua grande maioria eram pequenos arquivos em DBase. Hoje, com a automação de transações de negócios, os volumes de dados multiplicam-se. Em uma simples ligação telefônica, um teste médico, um acesso a uma página na Internet, enfim, tudo é armazenado em computadores [CAZ 97].

A NASA, por exemplo, possui mais dados do que pode analisar. Vem sendo armazenado pelo projeto genoma, milhares de bytes para cada um dos muitos bilhões de bases genéricas. Uma vasta quantidade de dados estão contidas nos censos realizados nos países, dados estes que escondem padrões de estilos de vida da população. A maioria destes dados nunca serão vistos " a olho nu " e mesmo que os fossem, não poderiam ser analisados manualmente, para isso os computadores tem a resposta [CAZ 97].

O problema não é mais o computador, mas sim qual ferramenta ele deverá utilizar. Vem sendo utilizados métodos estatísticos desenvolvidos há décadas, mas estes

não são necessariamente melhores que uma nova classe de ferramentas analíticas "inteligentes" chamadas de mineração de dados.

Mineração de dados é uma nova metodologia para melhorar a quantidade e eficiência das decisões, sejam elas científicas ou de negócios. Esta metodologia complementa e pode até substituir outras ferramentas de apoio à decisão do tipo análises estatísticas, relatórios e perguntas geradas por computador. Devido a ela ocorre o retorno dos investimentos baseando-se na análise minuciosa dos dados.

Altos retornos sobre investimentos em banco de dados e ferramentas analíticas vem sendo obtidos por muitas companhias. Projetos de Data Warehouse e de *Data Mining* tornam-se uma necessidade em qualquer negócio. Segundo *Gartner Group*, ferramentas de mineração de dados serão uma das cinco mais importantes tecnologias do próximo século [CAZ 97].

A mineração de dados é uma das etapas do processo de descoberta de conhecimento em banco de dados e têm por objetivo extrair informações implícitas e potencialmente úteis de dados [FAY 96]. Esta descoberta de conhecimento em bancos de dados, normalmente consiste em uma série de etapas, que inicia com a seleção do conjunto ou amostra dos dados que serão processados. Os dados podem ser submetidos à uma etapa de pré-processamento onde há uma transformação para um formato adequado para o algoritmo de mineração, que procura por regras ou padrões ocultos nos dados. Após esta etapa, as informações descobertas são selecionadas e formatadas para exibição, para serem interpretadas e avaliadas, de maneira que se selecione os conhecimentos úteis resultantes de todo o processo.

A mineração de dados permite aos usuários explorar e inferir informação útil partindo dos dados e descobrindo relacionamentos escondidos no banco de dados. Tendo isto em vista, fica bem claro que a mineração de dados vai muito além de uma simples consulta a um banco de dados.

Eventos passados são analisados com ferramentas típicas de regressão, como as que compõem um sistema de suporte à decisão, na mineração de dados as

análises esperadas e propostas vão muito além , suas ferramentas podem responder a questões tradicionais do mundo dos negócios, como por exemplo onde está o consumidor em potencial.

As bases de dados são exploradas em busca de regras, modelos ocultos, descobrindo informações com atributos especiais, mas falhas podem ocorrer, fazendo com que os dados venham fora das expectativas.

A mineração de dados está sendo amplamente difundida e aperfeiçoada pelo fato de ser uma ferramenta poderosa para a procura de informações em banco de dados. A seguir serão vistos os principais conceitos e técnicas utilizadas por esta tecnologia.

3.2 O que é Mineração de Dados?

Segundo Cazarine et. al., as ferramentas de mineração de dados analisam os dados automaticamente formulando hipóteses sobre estes dados, e buscando os resultados desejados, através da aplicação de algoritmos. Já as demais ferramentas analíticas que não utilizam a mineração de dados, mesmo tendo a capacidade de realizar sofisticadas perguntas pelos usuários, são dependentes destas perguntas e hipóteses, logo sendo limitadas para descobrir padrões complexos e tendências[CAZ 97].

3.2.1 Conceitos

Como um dos conceitos de mineração de dados se tem: "A mineração de dados é uma ferramenta utilizada para descobrir novas correlações, padrões e tendências entre as informações de uma empresa ou de um site por exemplo, através da análise de grandes quantidades de dados armazenados em Data Warehouse usando técnicas de reconhecimento de padrões, estatísticas e matemáticas" [CAZ 97].

A mineração de dados é uma técnica que permite buscar em uma grande base de dados, informações que, aparentemente, estão camufladas ou escondidas,

permitindo, com isso, agilidade nas tomadas de decisões. Se faz o uso de uma ferramenta de *Data Mining* não apenas pela sua agilidade e mas também pela sua própria capacidade de analisar grandes quantidades de dados proporcionando informações de maior qualidade, e embasando ainda mais as tomadas de decisão.

Segundo a IBM, *Data Mining* é um processo de extração de informações previamente desconhecidas, válidas e com capacidade de proporcionar ações, provenientes de uma grande base de dados e então usá-las na tomada de decisões de negócio cruciais.

A mineração de dados tem como objetivo extrair valiosas informações dos dados. Diferenças significativas podem aparecer nos caixas das empresas a partir de pequenas mudanças na estratégia provenientes das descobertas das ferramentas de mineração de dados. Estas ferramentas se tornam cada vez mais necessárias devido ao crescimento dos Data Warehouses (enormes depósitos de dados), vale a pena lembrar que o uso destes não se faz necessário para a utilização de uma ferramenta de mineração de dados. Os dados é que são realmente necessários.

O termo *Data Mining* é usado nos pacotes de softwares de muitas ferramentas tradicionais de análise de dados do tipo geradores de relatórios ou análises estatísticas, assim como produtos baseados em inteligência artificial também se dizem ferramentas de *Data Mining*. O importante é saber o que é o verdadeiro *Data Mining* e o que não é.

O objetivo principal do *Data Mining* é o conhecimento descoberto. Sua metodologia extrai informações preditivas das bases de dados. Com essa definição ampla, uma "*on-line analytical processing*" (OLAP), que é um pacote estatístico, pode ser considerado como uma ferramenta de *Data Mining*. Porém, estreitando-se o conceito, uma ferramenta de *Data Mining* deve descobrir as informações valiosas automaticamente. Por essa definição, a ferramenta de *Data Mining* é guiada por dados, ao contrário das outras ferramentas que são guiadas pelo usuário. A seguir será mostrado o tipo de análise orientada pelo usuário, a qual é realizadas por pseudo-ferramentas de mineração de dados [CAZ 97].

3.2.2 Análises Orientadas pelo Usuário

a) Geradores de Perguntas

Tradicionalmente, a identificação e utilização das informações e padrões escondidos é feita a partir de geradores de perguntas e sistemas de interpretação. O usuário formula uma teoria com uma hipótese e pergunta à base de dados para testar a validade da hipótese. O usuário pode gerar perguntas para a base de dados e o resultado será estabelecido em um relatório que dará uma boa visão dos resultados.

Estes geradores de perguntas são limitados, tendo em vista que são baseados em palpites dos usuários, além de que a qualidade da informação extraída depende da interpretação dos resultados, o que está sujeito a erros. Outro aspecto que dificulta este método é que a criação de hipóteses pelo usuário é limitada, em se tratando do número de variáveis que ele é capaz de manipular.

b) Ferramentas Estatísticas

Com o objetivo de extrair informações dos dados, métodos estatísticos vem sendo criados. Análises multifatoriais de variância, dentre outros métodos podem identificar as relações entre fatores que influenciam vendas ou qualquer que seja o que se está procurando.

Nesta abordagem as técnicas tendem a focar-se em tarefas onde todos os atributos tenham valores contínuos ou ordinais, alguns assumindo uma particular probabilidade de distribuição das variáveis, sendo este um dos problemas deste método. Muitos métodos assumem uma relação expressa, como combinações lineares dos atributos. Além disso, as ferramentas estatísticas freqüentemente assumem a disposição dos dados em uma curva normal, o que nem sempre pode acontecer na vida real.

3.2.3 O *Data Mining*

A geração de perguntas se dá a partir da necessidade de conhecer certos fatos. O projeto de *Data Mining* opera os dados de forma automática, ou seja, não é o usuário que gera as perguntas e sim a própria ferramenta.

Temos como exemplos de verdadeiras ferramentas de *Data Mining*: Redes neurais, árvores decisórias e programas de indução de regras, pois eles interrogam o banco de dados automaticamente em busca de padrões. Essas ferramentas de *Data Mining* oferecem um grande potencial para empresas que possuem Data Warehouse, em se tratando de que se busca descobrir padrões ao invés de simplesmente confirmá-los, isto se deve ao fato das mesmas utilizarem algoritmos estatísticos ou baseados em inteligência artificial, como ID3, C4.5, AC2, CART, CHAID, CN2, ou variações destes, os quais proporcionam um grande desempenho nestes casos[CAZ 97].

3.3 Fundamentos do *Data Mining*

Um longo processo de pesquisa e desenvolvimento de produtos deram como resultado as técnicas de mineração de dados. Esta evolução começou quando os primeiros dados de negócios começaram a ser armazenados em computadores, continuando com melhorias no acesso aos dados, e mais recente ainda, com novas tecnologias que permitem que os usuários visualizem os dados e façam alterações ao mesmo tempo.

O *Data Mining* está preparado para aplicações que envolvam basicamente três tecnologias que estão agora suficientemente desenvolvidas, quais sejam:

- **Enormes quantidades de dados:** o crescimento dos bancos de dados comerciais estão à uma taxa nunca vista

- **Computadores multiprocessados:** o desenvolvimento de computadores poderosos foi proporcional ao crescimento dos dados em bases de dados,

mas a relação custo benefício precisa ser melhorada, só assim teremos um maior consumo destes equipamentos.

- **Algoritmos de mineração de dados:** estão sendo incorporadas técnicas, que existem à pelo menos dez anos, nos algoritmos de mineração de dados, mas que somente agora tem sido implementado, aperfeiçoado, tornado-se confiável, também estão sendo constantemente testadas ferramentas mais compreensíveis e fáceis de utilizar.

Os componentes centrais da tecnologia de mineração de dados tem sido desenvolvidos por décadas, pesquisas em áreas como estatística, inteligência artificial, aprendizado de máquina e banco de dados. Hoje, o aperfeiçoamento destas técnicas, combinadas a grande performance das máquinas disponíveis para a aplicação desta tecnologia, banco de dados relacionais, e a um grande esforço para a integração dos dados, faz com que se torne possível a aplicação de técnicas de mineração de dados além de incentivar o surgimento de outras tecnologias como o Data Warehouse.

O poder computacional aumentou muito devido ao desenvolvimento de sistemas paralelos e distribuídos, tornado-se possível a análise de grandes bases de dados em um tempo bem menor do que se levava à alguns anos atrás, o que também facilitou muito o desenvolvimento de aplicações de mineração de dados.

3.4 Processo de Mineração de Dados

Partindo de uma visão de processo orientado, existem três classes de mineração de dados [CAZ 97]:

- Descobrimento
- Modelagem de prognóstico
- Análise prévia

Esta três classes estão apresentadas na figura 4.

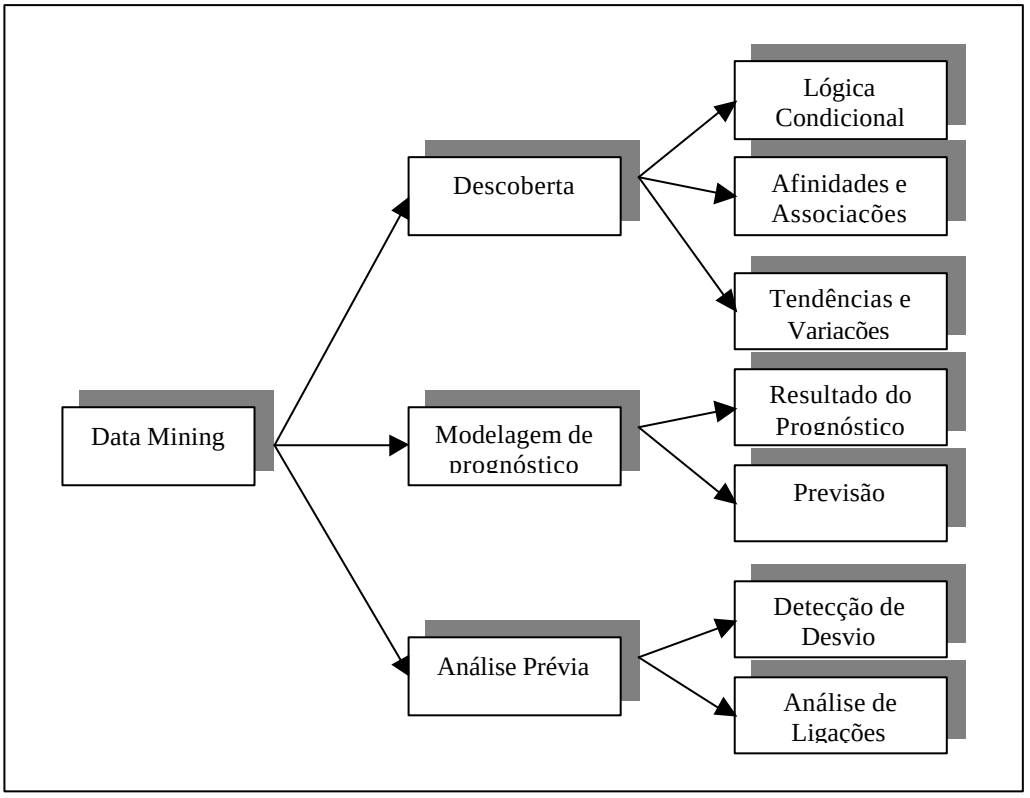

Figura 4 – Processo de Mineração de Dados [CAZ 97]

- **Descoberta:** é o processo de exame em um banco de dados para encontrar padrões escondidos sem uma idéia ou hipótese pré-determinada sobre o que são esses padrões, ou seja, o programa toma a iniciativa de encontrar aquilo que interessa aos padrões, sem que o usuário verifique se isto realmente interessa. A riqueza dos padrões que podem ser expressados e descobertos e a qualidade de informação obtida determinam a força e a utilidade da técnica de descoberta.

- **Modelagem de Prognóstico:** nesta classe os padrões descobertos no banco de dados são usados para prognosticar o futuro. Enquanto o processo de descobrimento encontra padrões em dados, o processo de modelagem de prognóstico aplica estes padrões para supor valores nos novos itens de dados.

- **Análise Prévia:** é o processo de aplicação dos padrões extraídos para encontrar anomalias ou elementos de dados raros. Para descobrir os dados raros, nós

primeiramente encontramos os dados que seguem uma norma ou os habituais, então detectamos aqueles que se desviam dos habituais dentro de um certo limiar.

Alguns desses processos serão classificados na seção 3.8. Por exemplo, há vários tipos de padrões de descobrimento tais como regras If/Then, associações, etc.

3.5 Tecnologia Empregada na Mineração de Dados

A tecnologia empregada na mineração de dados se ramifica em duas partes, a retenção de dados e a destilação de padrões, como mostrado na figura 5. Também podemos observar que as aproximações baseadas em retenção de dados podem ser implementados através de duas técnicas: a do vizinho próximo e a baseada no raciocínio, já a "destilação" de padrões se divide em três formas de abordagens: lógica, tabulação cruzada e equações, as quais possuem diferentes técnicas, como mostrado a seguir:

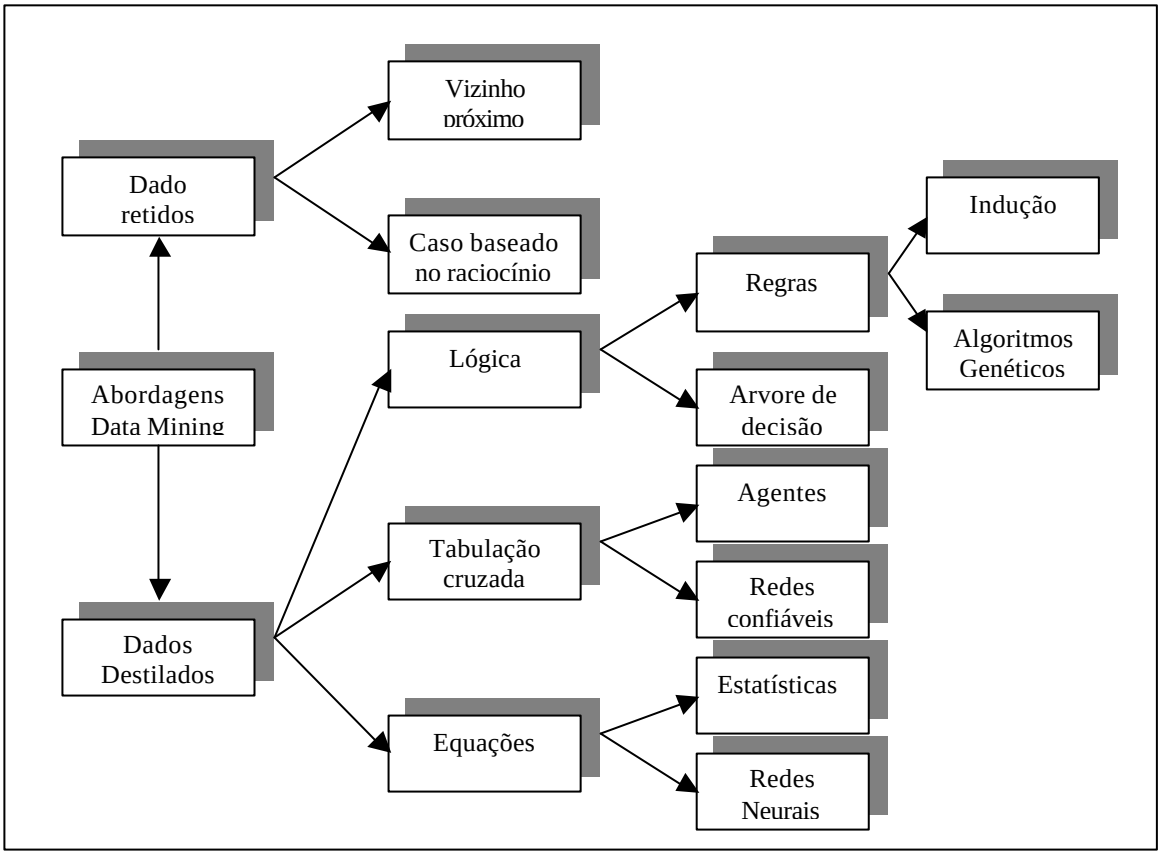

Figura 5 – Tecnologia *Data Mining* [CBV97]

3.5.1 Retenção de Dados

Ao contrário da destilação de padrões onde os dados são analisados, para se extrair padrões e logo em seguida são dispensados, na Retenção de Dados estes são mantidos para posterior combinação. Quando novos dados são apresentados, eles são combinados com o conjunto de dados anterior [POL 97].

O método mais conhecido de retenção de dados é o "método do vizinho próximo”. Neste método, o conjunto de dados é mantido, geralmente na memória, para a comparação com novos dados. Quando um novo registro está presente, por prognóstico, a "distância" entre ele e os registros semelhantes no conjunto de dados é encontrado, e os mais parecidos são identificados [POL 97]. Para exemplificar o exposto acima tomemos como exemplo a seguinte situação: queremos cadastrar um novo cliente bancário, mas queremos saber a que classe de clientes do banco ele se encontra enquadrado, então, através de suas características, ou atributos, são comparadas com todos os clientes existentes no banco. Um exemplo seria comparar a idade e o salário do futuro cliente com as idades e os salários dos clientes existentes. Neste caso, um conjunto de vizinhos são selecionados e se detecta a que grupo o novo cliente se encontrará.

Nesta técnica, o termo bastante encontrado "K-nearest neighbor" é usado para indicar que foi selecionado o topo K de vizinhos, ou seja, aquele de características mais parecidas com aquelas que estão sendo analisadas.

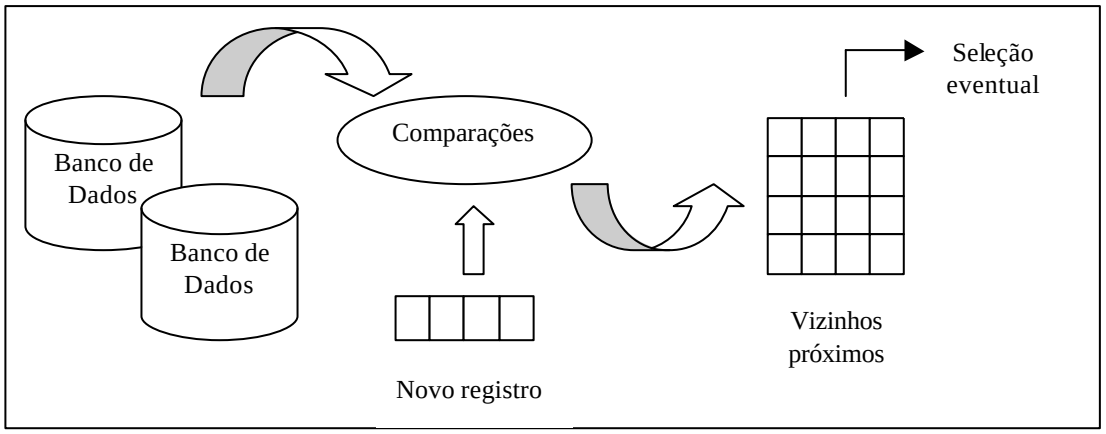

Figura 6 – Retenção de Dados [POL 97]

Um outro termo utilizado nesta técnica é o "caso baseado no raciocínio" que ocorre quando precisamos reter apenas um pequeno conjunto de "casos típicos" como base para comparações. Deve-se levar em consideração que é possível manter todos os dados em memória, porém isso torna-se complicado em casos de grande quantidade de dados.

Nesta técnica o ponto chave está na seleção dos "padrões típicos", pois se não entendermos realmente os padrões, como selecionaremos os casos típicos? e como trocaremos os padrões típicos se a base-padrão for trocada?

3.5.2 Destilação de Padrões

As tecnologias que empregam a destilação de padrões tem como objetivo principal extrair padrões a partir de uma série de dados fazendo com que estes padrões obtidos sejam utilizados para um determinado fim. Primeiramente devemos levar em consideração duas questões fundamentais para aplicação deste processo, que são: o tipo de padrão que pode ser extraído e como fazer para representar este padrão.

Os padrões necessitam ser expressos dentro de um formalismo e uma linguagem, desta forma, temos três distintas formas de abordagem [POL 97]:

- sistema lógico

- sistema equacional
- classificação cruzada

No sistema lógico podemos distribuir igualmente dados numéricos e não-numéricos, já no sistema equacional exige-se que todos os dados sejam numéricos, enquanto que o sistema de classificação cruzada, ao contrário, somente trabalha com dados não-numéricos, que é a fonte de todos os problemas. Também é importante ressaltar que o sistema equacional estima "distâncias" a partir de ocorrências, enquanto que o sistema de classificação cruzada se focaliza em co-ocorrências.

Redes Neurais são técnicas equacionais que não são visíveis totalmente ao usuário, enquanto que, internamente, trabalham com espaço numérico. Como o dado é repetidamente introduzido na estação de trabalho, os parâmetros são considerados como alterados somente após o fechamento da base.

Nem todas as aproximações baseadas na destilação de padrões fornecem informações, já que estes padrões podem ser destilados em uma linguagem "opaca" ou em um formalismo não facilmente legível por humanos - exemplo: equações complexas.

No estudo do *Data Mining* é necessário fazer distinção entre o aprendizado supervisionado e a forma livre, ou seja, percorrendo e vasculhando o banco de dados.

No aprendizado supervisionado, temos como que um "professor" que "ensina" o sistema. Por exemplo, dizendo quando uma premissa foi ou não correta. Aqui, o dado tem uma coluna específica que será usada na tentativa de descobrir algo coerente ou, ao menos, fazer predições.

No aprendizado sem supervisão, o sistema não tem um "professor"; ele simplesmente tenta encontrar os "*clusters*" (ou grupos) que lhe interessam dentro do conjunto de dados.

A maioria das aplicações comerciais baseadas em mineração de dados envolve o *Data Mining* diretamente, enquanto que a procura não supervisionada pode,

algumas vezes, ser usada na segmentação de dados e agrupamento. Por exemplo: encontrando classes de padrões em séries simultâneas.

Todos estes conceitos vistos neste tópico, serão melhor estudas no decorrer deste capítulo, dentro dos tópicos relacionados aos algoritmos de mineração de dados.

3.6 Passos para a Implementação da Mineração de Dados

Abaixo serão mostrados os passos para a extração dos conhecimentos escondidos nos bancos de dados, sejam eles um Data Warehouse, arquivos de informações de consumidores, arquivos de redes ou qualquer tipo de banco de dados [MEN 99].

1) Identificação dos objetivos

Antes de iniciar, deve-se ter em mente os objetivos organizacionais que o *Data Mining* poderá alcançar, como: encontrar relacionamento de vendas entre os produtos ou serviços, encontrar padrões específicos de compra durante o tempo, identificar potenciais tipos de consumidores e estabelecer tendência da venda de um produto.

2) Selecionar os dados

Uma vez definido os objetivos, o próximo passo é selecionar os dados que auxiliarão no alcance deste, tentando se segmentar o máximo possível o escopo dos dados a serem minerados. Para isso algumas perguntas são importantes: os dados que estão sendo obtidos são adequados para que a ferramenta de *Data Mining* consiga descrever um padrão?; é possível melhorar os dados com outras bases de dados primárias, como estilo de vida ou dados demográficos?; os dados são estáveis, ou seja, se conseguirá manter os atributos após feita a análise?; existe uma forma de relacionar esses dados, ou seja, eles têm algo em comum?; estes dados são realmente relevantes para o alcance dos objetivos?

3) Preparar os dados

Após a montagem dos dados, deve-se decidir quais atributos deverão ser convertidos em formatos utilizáveis. Nessa hora, é importante consultar os criadores e usuários dessa base de dados, estabelecer estratégias para padronizar dados, identificar e tratar variáveis redundantes, decidir o formatos dos dados, e sua transformação se necessário e inspecionar esses dados;

Essas decisões estão relacionadas com a seleção da ferramenta de *Data Mining* a ser utilizada, por exemplo caso se deseje utilizar uma rede neural ou polinomial, deve-se transformar os dados em campos.

4) Auditar os dados

Avalie a estrutura dos dados, entendendo sua natureza, para determinar a ferramenta apropriada, visando balancear o objetivo de acesso da estrutura dos dados com os objetivos e necessidades dos usuários em entender os dados.

5) Selecionar as ferramentas

Para selecionar a ferramenta adequada no processo de mineração de dados, devem ser observados os objetivos do projeto e a estrutura dos dados, e através disto procurar uma ferramenta que satisfaça essas duas condições. Para isso algumas perguntas são importantes: os dados são muito categóricos?; que ferramenta pode suportar?; a ferramenta potencial compila ODBC?; que formato de dados a ferramenta importa?

Provavelmente nenhuma ferramenta sozinha terá todas as respostas para o projeto de *Data Mining*. Algumas ferramentas integram muitas tecnologias: pacotes estatísticos, redes neurais, classificadores simbólicos.

6) Formatar a solução

Auditoria dos dados, os objetivos do projeto e a seleção da ferramenta determinarão o formato da solução. Para isso é interessante se saber qual é o formato ótimo da solução (árvores decisórios, regras, código C, SQL); quais são os formatos disponíveis; qual é o resultado da solução e o que os usuários precisam.

7) Construindo o modelo

Nesse estágio, o processo de *Data Mining* inicia. Normalmente, o primeiro passo é usar testes para construir e avaliar o modelo. Utilizar fragmentos aleatórios da base de dados para o teste, é o mais apropriado. A geração de regras de classificação, árvores decisórias, agrupamento de sub-grupos, pontuação, codificação, peso e avaliação dos dados são realizados nesse passo.

8) Validar os resultados

Compartilhar e discutir os resultados com a unidade de negócio que solicitou o projeto e o *expert* do banco de dados, discutindo-se e avaliando-se o resultado, se é necessário retornar à alguma etapa para melhorar os resultados, etc. Ter certeza se todos os resultados estão corretos e apropriados para o alcance dos objetivos da organização.

9) Entregar os resultados

Providenciar um relatório final para a unidade de negócio que solicitou o projeto. O relatório deverá documentar todo o processo do projeto de *Data Mining*, incluindo a preparação dos dados, ferramentas utilizadas, códigos fontes, regras.

10) Integrar a solução

Compartilhe os resultados com todos os usuários interessados da unidade de negócio. Deve-se incorporar os resultados nos procedimentos de negócio da companhia. As soluções de *Data Mining* devem envolver: sintaxe SQL para os usuários, código C incorporado no sistema de produção, regras integradas no sistema de apoio à decisão.

Como a ferramenta de *Data Mining* automatiza as análises, elas poderão levar a falsas conclusões se não se tomarem os devidos cuidados. Deve-se ter consciência que *Data Mining* é um processo de negócio com um objetivo específico – extrair conhecimento competitivo da base de dados histórica.

3.7 Áreas Potenciais ao uso de *Data Mining*

Aqui estão algumas áreas nas quais as ferramentas de mineração de dados vem se destacando, algumas delas já trazem a presença da mineração como fundamental no decorrer de suas atividades enquanto que outras estão recentemente migrando para esta tecnologia promissora.[VIV 98].

- Bancos de varejo: aprovação de parcelas de amortização, subscrição de empréstimos, análise e detecção de fraudes.

- Fundos de investimento: análise e previsão de performance do negócio, análise de ações

- Seguradoras: previsão de falência, análise de risco, crédito e modelos de coleta

- Internet: oportunidades de vendas cruzadas e foco na colocação de banners, construção de modelos de prospectos, análise do comportamento do consumidor, análise de tráfico, procura de conteúdo, etc. Na Web as ferramentas de *Data Mining* estão ganhando uma enorme importância, tendo em vista a tamanha quantidade de dados que trafegam pela rede e pela grande necessidade de se minerar e extrair estes dados. A aplicação de técnicas *Data Mining* na Web será discutida no próximo capítulo mais detalhadamente.

- Marketing direto: pesquisa de mercado, previsão de aceitação/desempenho do produto

- Pesquisa de mercado: seleção da mídia, análises, segmentação de produtos

- Manutenção: previsão da necessidade equipamentos de manutenção para veículos.

- Manufatura: identificação de erro na produção, controle de qualidade, planejamento de manutenção preventiva, automatização de sistemas.

- Medicina: estudos epidemiológicos, toxicologia, diagnósticos, interações de drogas, análise de fator de risco, controle de qualidade, estudos retrospectivos de medicamentos.

- Pesquisa científica: modelagem geral de todos os tipos

3.8 Algoritmos de Mineração de Dados

Existem dois objetivos que se destacam na mineração de dados que são a previsão e a descrição. A primeira delas utiliza variáveis presentes no banco de dados para prever valores desconhecidos ou futuros, já a segunda se destina a busca de padrões descrevendo os dados e apresentando-os aos usuários para que os mesmos possam fazer sua interpretação.

Existe uma gama enorme de algoritmos destinados a mineração de dados, dentre os quais se destacam: algoritmos de associação, classificação, padrões seqüenciais e agrupamento, cada um deles sendo executado de acordo com suas características. Os algoritmos de associação buscam achar todas as associações em que a presença de um conjunto de itens em uma transação implica em outros itens, já os algoritmos de classificação ou geração de perfis desenvolvem perfis de diferentes grupos, os algoritmos de padrões seqüenciais identificam tipos de padrões seqüenciais em restrições mínimas especificadas pelo usuário, e por fim, os algoritmos de agrupamento atuam segmentando o banco de dados em subconjuntos ou grupos. Os algoritmos que serão vistos a seguir foram obtidos através da fonte de pesquisa disponibilizada por Polito [POL 97].

3.8.1 Lógica

Se distingue os padrões expressos em linguagens lógicas através de duas características: Os primeiros são legíveis e compreensíveis enquanto que os outros são excelentes na representação no agrupamento de dados.

Em uma linguagem lógica, o operador central é normalmente uma variação do comando If / Then. Por exemplo: Se está chovendo, Então está nublado. Entretanto, devemos notar que enquanto a forma lógica mais comum é a lógica condicional, muitas vezes necessitamos usar outras formas lógicas como uma associação lógica com a regra When / Also. Por exemplo: Quando Windows é adquirido, o Internet Explorer também é adquirido. Ainda que as lógicas condicionais sejam conhecidas mais a fundo, outras formas de lógica também são usadas em análise dos dados, entre elas estão: lógicas de variação e de tendência, por exemplo.

Os sistemas de lógica condicional se separam em dois grupos distintos: árvores de regras ou de decisão. Regras condicionais podem ser implementadas por Indução ou Algoritmos Genéticos e várias outras formas de gerar árvores de decisão (por exemplo: CART, CHAID, C4.5, etc).

1) Regras

Neste caso os relacionamentos lógicos são representados como regras, as quais podem expressar relações condicionais ou de associação.

Como exemplo de regra condicional, temos o seguinte comando: If Condição 1Then Condição 2

Como exemplo, num banco de dados demográfico podemos ter uma regra: If Profissão = Atleta Then Idade < 30

Aqui nós comparamos os valores contidos nos campos de uma determinada tabela, isto é, temos uma representação de "atributos-valores". Profissão é o atributo e Atleta, o valor.

As regras condicionais normalmente se utilizam de tabelas com diversos atributos (isto é, campos) e valores, como mostrado na tabela 1:

{PRIVATE} **Nome**	**Profissão**	**Idade**
Pedro	Atleta	25
...	...	...

Tabela 1 – Regras Condicionais [POL 97]

As regras ainda facilitam ir além das representações de atributo-valor, pois com elas podemos ter um atributo lógico apenas comparando os valores de dois campos sem nomear explicitamente quaisquer valores. Este tipo de relação não pode ser estabelecida através de árvores de decisão ou tabulações cruzadas.

A conexão lógica é diferenciada a partir da lógica condicional, tanto em termos da linguagem da expressão como da estrutura de dados usadas. Análise por Conexão (ou Análise por Associação) é a procura por padrões e condições que descrevem vários itens, como "grupos simultâneos" ou "acontecimentos simultâneos" numa série de eventos ou transações. Uma regra de afinidade possui a seguinte forma:

When Item_1
Also Item_2.

Devemos notar que esta é uma estrutura de dados que difere da regra da lógica condicional, onde "conexões absolutas" necessitam ser estendidas para conexões dimensionais para obter melhores resultados. Uma conexão dimensional tem a forma:

IF Sistema operacional = Windows
WHEN Item = Bloco de Notas
ALSO Item = Internet Explorer

Aqui, as lógicas condicionais e as associativas estão combinadas. Esta forma de estrutura híbrida libera o real poder da lógica transparente.

As regras se destacam quando trabalham com dados numéricos e não-numéricos de uma maneira uniforme, porém ao se trabalhar com dados numéricos pode ocorrer algumas perdas de padrões, tendo em vista que algumas aproximações valem-se da quebra de campos numéricos nos códigos ou nos valores especificados, podendo, efetivamente, remover todas as considerações numéricas dos códigos.

Para exemplificar o exposto acima, consideremos um campo Idade, o qual necessita ser subdividido em 3 faixas: (1-30), (31-60), (61-100), correspondentes respectivamente à jovem, meia-idade e velho. Mesmo os dados podendo manter padrões superpostos em qualquer das faixas, pode ocorrer, por exemplo, da faixa (27-34) ser muito importante em determinada transação, logo alguns padrões e algumas aproximações baseadas em códigos de transferência podem ser perdidos.

As regras também atuam de forma satisfatória em dados multi-dimensionais, pois elas permitem que seus padrões possam ser mesclados ao longo de múltiplas dimensões e seus formatos lógicos, logo podendo se trabalhar com faixas de dados numéricos.

Regras também podem ser representadas através de árvores de decisão, mas apesar da semelhança a nível superficial, têm uma técnica diferente, pois as árvores de decisão não expressam associações ou padrões baseados em atributos, onde os valores de dois campos são comparados sem nomeação explícita de qualquer valor.

Existem duas formas de abordagem com regras que podem ser utilizadas no processo *Data Mining*: regra indutiva e algoritmos genéticos. Estas abordagens examinam, através de uma pré-computação, cada possibilidade da regra até numa série de dados que já estavam incluídas (previstas). Nesses casos, a aplicação destas abordagens não se torna prática, tendo em vista que somente uma única coluna de dados pode ser usada porque o espaço lógico é imenso.

1a) Regra Indutiva

Denomina-se regra indutiva o processo na qual se analisa uma série de dados e, a partir desta analise, se gera padrões. Pelo fato de explorar automaticamente a série de dados, como mostrado a seguir na Figura 7:

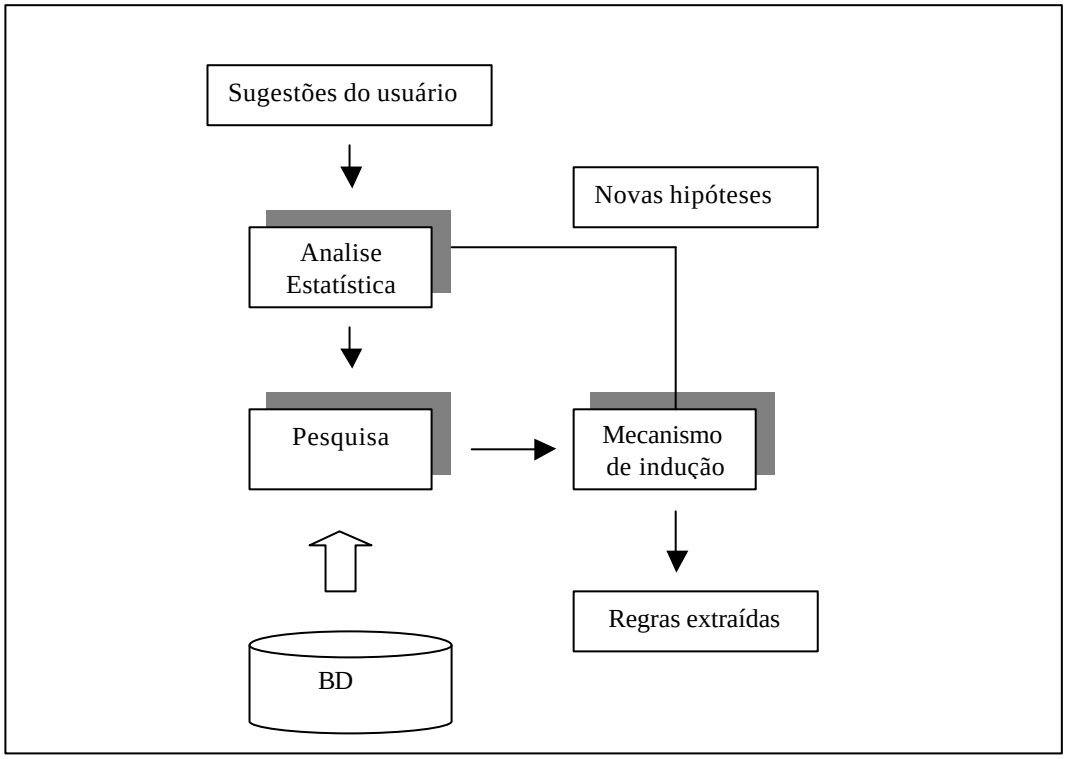

Figura 7 – Processo Indutivo [POL97]

O sistema indutivo cria hipóteses que conduzem a padrões.

O processo de aplicação de regra indutiva é semelhante àquilo que um analista faria em uma análise exploratória. Para exemplificar tomamos como exemplo um banco de dados contendo informações demográficas, primeiramente o sistema de indução pode examinar como as idades são distribuídas, a partir daí, pode-se perceber uma interessante variação para aquelas pessoas cujas profissões são listadas como atletas profissionais. Como esta observação, ou seja, esta hipótese é considerada relevante pelo sistema, o mesmo imprimirá a seguinte regra:

IF Profissão = Atleta THEN idade < 30.

Indução de regras podem descobrir regras muito gerais, as quais lidam ambas com dados numéricos e não numéricos. E as regras podem combinar declarações condicionais e de afinidade nos padrões híbridos.

É importante que se faça uma distinção entre regras cobertas e inexatas. As regras inexatas podem ter fixados a elas um padrão de margem de precisão. Entretanto, a confiança nas regras cobertas podem variar em termos de valores numéricos no corpo da regra. Deste modo regras cobertas podem produzir muito mais expressões de conhecimento compactas e conduzir à comportamentos estáveis. Uma chave de saída aqui é a habilidade de ir além dos bancos de dados formais e lidar com os padrões OLAP.

1b) Algoritmos Genéticos

Algoritmos genéticos, assim como regras indutivas, também geram regras dos conjuntos de dados, porém os algoritmos genéticos não seguem a indução de protocolo de regras de exploração orientada, ao invés disso eles contam com a técnica de "mutação" que consiste em fazer trocas nos padrões até que uma forma apropriada de modelo surja via educação seletiva, como mostrado na Figura 8.**{**PRIVATE "TYPE=PICT;ALT=Figure 6'**}**

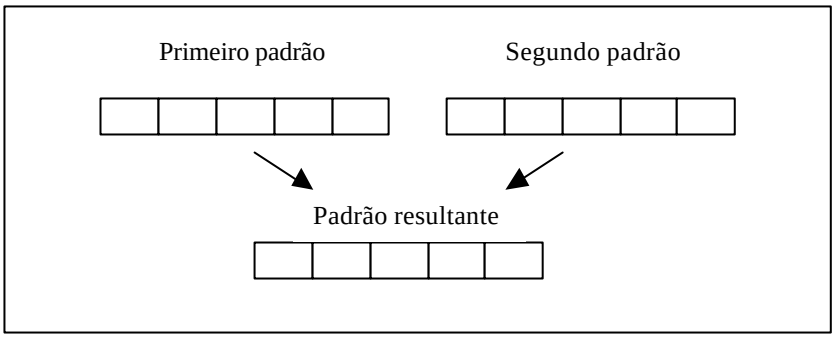

Figura 8 – Algoritmos Genéticos (troca de padrões) [POL 97]

Estes tipos de algoritmos receberam a denominação “genéticos” pois os mesmos trabalham utilizando técnicas muito semelhantes aquelas utilizadas por criadores quando cruzam material genético de plantas e/ou animais. No caso de regras, o material trocado é uma parte do modelo que a regra descreve.

A principal diferença entre os algoritmos genéticos e as regras de indução a o fato de que os algoritmos genéticos baseiam-se fundamentalmente na combinação de modelos das regras que tinham sido descobertas até então, enquanto que a indução de regra baseia-se na base de dados. Algoritmos genéticos não são apenas para geração de regra e podem ser aplicados para uma variação de outras tarefas, até mesmo processamento de textos, etc.

2) Árvores de Decisão

Árvores de decisão são utilizadas para expressar de forma simples a lógica condicional. Um sistema que utiliza este tipo de algoritmo atua, simplesmente, dividindo uma tabela em tabelas menores pela seleção de sub-conjuntos baseados em valores de um atributo dado. O algoritmo de árvore de decisão dependerá diretamente do modo como a tabela foi dividida, podendo ser: CART, CHAID, C4.5, etc.

Como exemplo, consideremos a tabela 2:

{PRIVATE}Fabricante	Estado	Cidade	Cor do Produto	Lucro
Paulo	PR	Curitiba	Azul	Alto
Paulo	BA	Salvador	Verde	Baixo
Pedro	RS	Pelotas	Azul	Alto
Pedro	BA	Salvador	Vermelho	Baixo
Lucas	RS	Pelotas	Verde	Médio
Lucas	PR	Curitiba	Vermelho	Médio

Tabela 2 – Tabela para Árvore de Decisão [POL 97]

Na figura 9 então é mostrada uma árvore de decisão originada a partir da tabela 2.

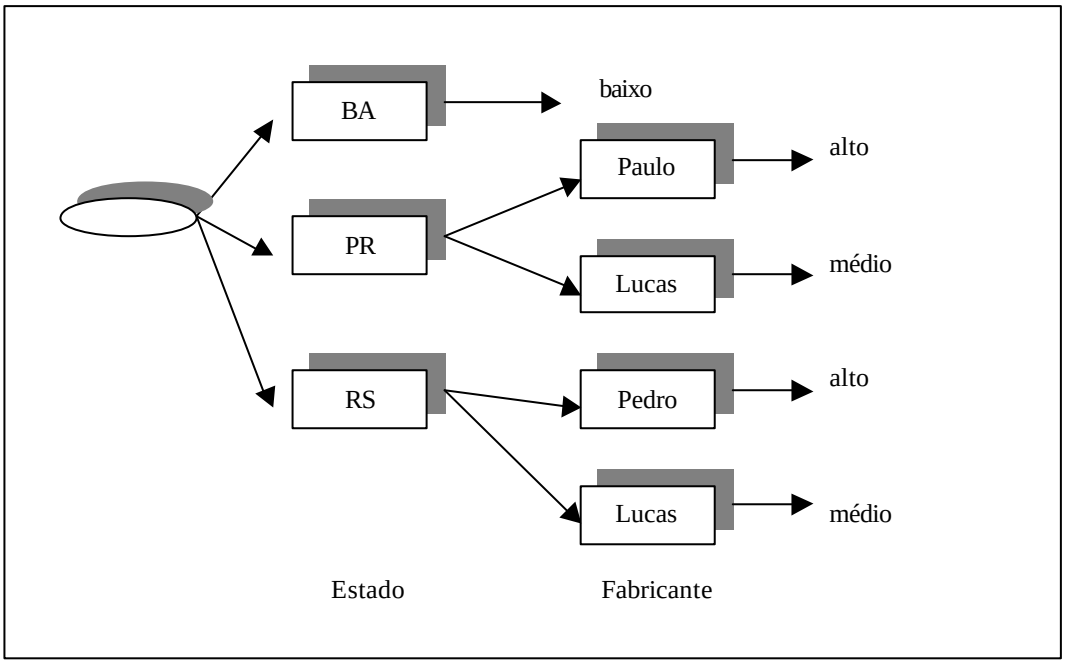

Figura 9 – Árvore de Decisão [POL97]

A árvore de decisão primeiro seleciona o atributo Estado, uma vez que em muitos casos, incluindo a tabela acima, não há um atributo melhor, e seja qual for o atributo que a árvore escolha haverá uma perda de informação, para começar a operação de divisão, e então o atributo Fabricante.

É importante observar que os seguintes fatos:

(a) “Produtos azuis são de alto lucro”.

(b) "Salvador é um lucro baixo".

nunca podem ser obtidos da tabela acima com uma árvore de decisão, uma vez que uma árvore de decisão seleciona um atributo específico por divisão a cada estágio. Neste caso então nós podemos obter ou o fato (a) ou o fato (b) da árvore, ambos não. Já se estivéssemos trabalhando com regras e tabelas cruzadas poderíamos descobrir ambos os fatos.

3.8.2 Tabulação Cruzada

Esta é uma forma de se analisar os dados muito básica e simples, a qual é muito utilizada em estatística. Para se elaborar uma tabulação cruzada de duas dimensões elabora-se uma tabela com ambos cabeçalhos de linhas e colunas como valores de atributos. As células na extensão de papel representam uma ação agregada, freqüentemente o número de co-ocorrências de atributo que avaliam-se juntos.

Considerando a tabela 2, a tabulação cruzada levando-se em consideração o nível de lucro poderia ser exposta como se segue na tabela 3:

{PRIVATE}	PR	BA	RS	Azul	Verde	Vermelho
Alto lucro	1	0	1	2	0	0
Médio lucro	1	0	1	0	1	1
Baixo lucro	0	2	0	0	1	1

Tabela 3 – Tabulação Cruzada [POL 97]

Para não deixar a tabulação cruzada muito larga, não foram incluídos na tabela acima os campos de Fabricante e Cidade. A partir desta tabela podemos verificar que a soma das co-ocorrências do Azul e de Alto lucro está acima das demais, o que nos indica um relacionamento mais forte.

Quando estamos trabalhando com um número pequeno de valores não numéricos, as tabulações cruzadas são suficientemente simples para usar e encontrar algumas relações lógicas condicionais.

Quando utilizamos as tabulações cruzadas podemos nos deparar com alguns problemas como: quando o número dos valores não numéricos sobe e quando as relações não são apenas baseadas em somas.

1) Agentes

Os agentes são variações das técnicas de tabulação cruzada. A diferença principal aqui é que os agentes executam operações verticais nas colunas da tabulação cruzada anteriormente, devolvendo afirmações de impacto.

O termo "agente" é usado algumas vezes para referir-se as tabulações cruzadas que são graficamente mostradas numa rede. Por exemplo, se considerarmos a tabulação cruzada acima, podemos definir 6 "agentes" para a meta Alto Lucro e graficamente mostrá-los como na figura 10.

{PRIVATE "TY

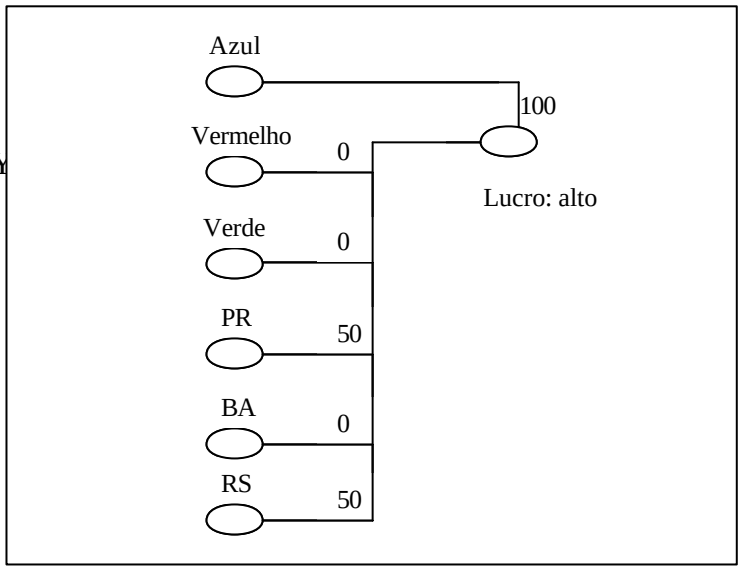

Figura 10 – Agentes [POL 97]

Aqui as estatísticas 100 e 50 são simplesmente as porcentagens com que os valores são associados com a meta, por exemplo, eles representam razões de impacto, não probabilidades.

Como outras técnicas de tabulação cruzada, quando lidando com valores numéricos, os agentes têm que separar os números em "códigos" fixos, por exemplo:

separe Idade em três classes de idade: (1 - 30), (31 - 60), (61 - 100). Naturalmente, o dado pode manter padrões que sobrepõem a qualquer uma dessas áreas, por exemplo: a área (27 - 34) a qual não será detectada pelo agente. Também temos que levar em consideração que caso as áreas selecionadas sejam muito pequenas, haverá muitas delas, fazendo com que padrões mais amplos sejam perdidos.

A tecnologia que envolve agentes, ainda será mostrada no capítulo IV, pois os mesmos são de fundamental importância na mineração de dados no ambiente Web (Web Mining).

2) Redes de Confiança

Assim como os agentes, as redes de confiança são variações das técnicas de tabulação cruzada, só que por sua vez, estas redes olham para as fileiras da tabela na forma horizontal, produzindo probabilidades.

Redes de Confiança (conhecidas como redes causais) contam também com totais de co-ocorrências, mas tanto a reprodução gráfica como a representação probabilístas são levemente diferentes dos agentes.

Para se ilustrar uma redes de confiança são utilizadas representações gráficas de distribuição probabilistas (derivadas de somas totais), os quais constam de nós (variáveis representadas) e arcos (representando dependências probabilistas) entre as variáveis de nós. A seguir, na figura 11, temos um exemplo de rede de confiança:

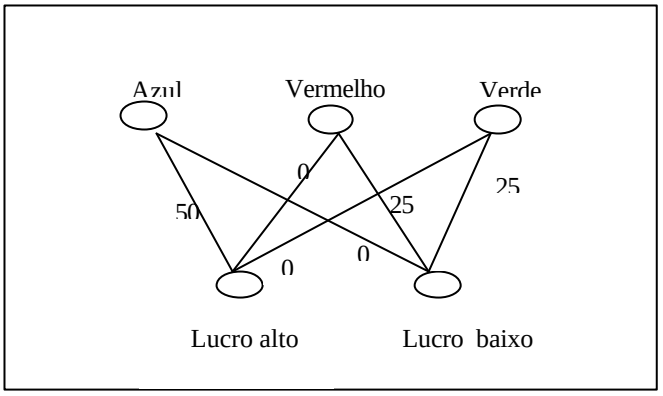

Figura 11 – Rede de Confiança [POL97]

É importante ressaltar que o gráfico da rede de confiança não é cíclico, significando que não existem ciclos.

3.8.3 Aproximações Equacionais

1) Estatísticas

Os algoritmos que se baseiam em aproximações equacionais estatísticas visam a construção de uma superfície que indique o agrupamento de determinado padrão. Tal sistema geralmente utiliza um grupo de equações para definir uma superfície dentro de um espaço numérico a partir de uma base de dados.

O melhor exemplo de conhecimento semelhante a superfície é uma linha reta dentro de duas dimensões de espaço, como na figura 11a. Esta tem a equação simples Y = (a * X) +b e conduz a aproximação linear em estatística. Como o parâmetro "a" varia nesta equação, a vertente da linha muda.

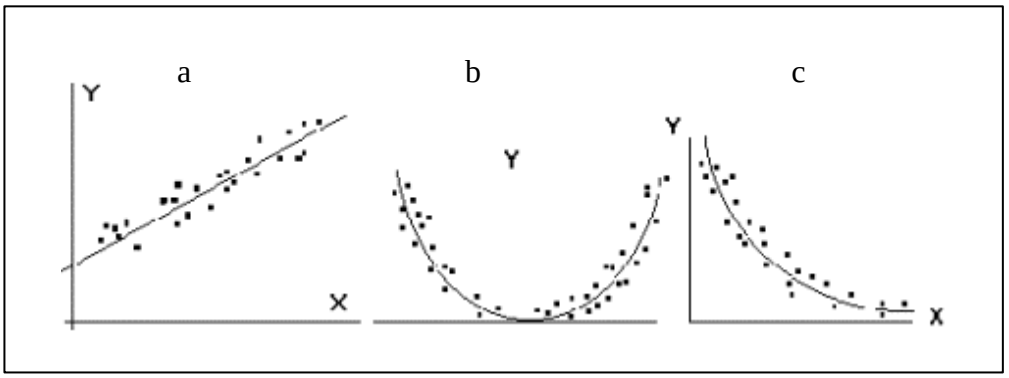

Figura 12 – Aproximações Estatísticas [POL 97]

Regressão trabalha bem quando os pontos são aproximados consistindo uma linha reta. Mas como nas figuras 11b e 11c, é também possível usar equações não-lineares para aproximar superfícies planas.

Quando as superfícies são ainda mais complexas, como exemplo Y = (X2 + X + (1/X)), ou quando há várias dimensões, a habilidade dos humanos de entender as equações e superfícies diminui muito rapidamente. Contudo, ainda é possível construir tais superfícies.

2) Redes Neurais

Redes Neurais são uma classe de modelagem de prognóstico que trabalha por ajuste repetido de parâmetro. Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios) organizados em camadas que aprendem pela modificação da conexão. Um modelo de redes neurais é mostrado na Figura 13.

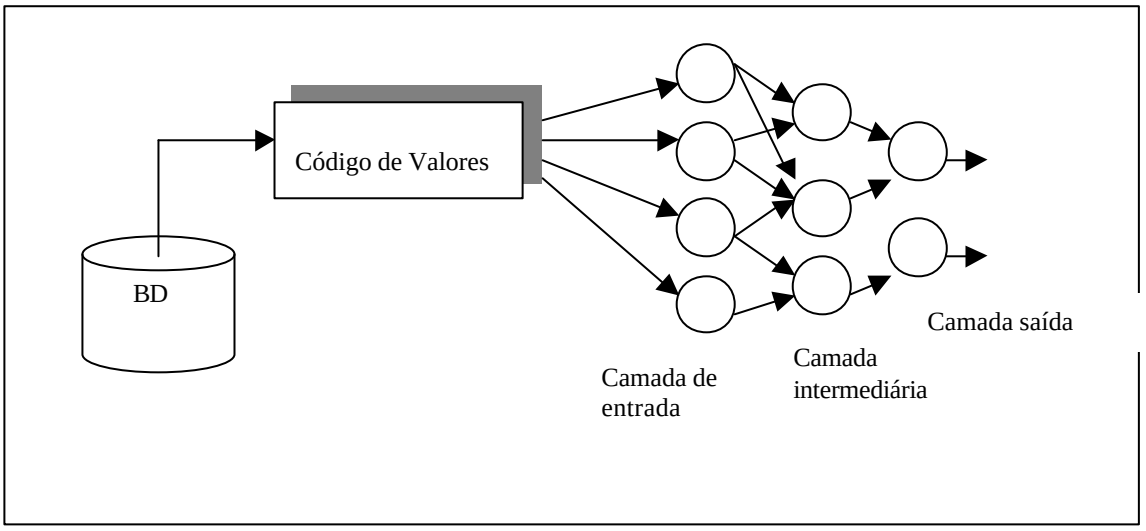

Figura 13 – Modelo de Redes Neurais [POL 97]

Redes neurais geralmente constróem superfícies equacionais complexas através de interações repetidas, cada hora ajustando os parâmetros que definem a superfície. Depois de muitas repetições, uma superfície que determina um possível padrão pode ser definida, de forma que se aproxime de muitos pontos dentro do grupo de dados.

Dentro da rede neural os neurônios tem funções específicas que são: avaliar valores de entrada, calcular o total para valores de entrada combinados, compara o total com um valor limiar, determinar o que será a saída.

Enquanto a operação de cada neurônio é razoavelmente simples, porém procedimentos complexos podem ser criados pela conexão de um conjunto de

neurônios, onde temos entradas de neurônios ligadas à uma camada intermediária (ou várias camadas intermediárias) que é então conectada com a camada de saída, como podemos observar na figura 13.

É importante fazer um "adestramento" da rede em um *dataset* de treinamento antes de se construir um modelo neural, e então depois, com a rede já treinada, partir para as predições. Também é interessante que se faça um monitoramento do *dataset* durante a fase de treinamento, para que desta forma possamos checar seu progresso.

Cada neurônio geralmente tem um conjunto de pesos que determina como o neurônio avalia a combinação dos sinais de entrada. A entrada para um neurônio pode ser positiva ou negativa. O aprendizado se faz pela modificação dos pesos usados pelo neurônio de acordo com a classificação de erros que foi feita pela rede como um todo.

Durante a fase de treinamento, a rede estabelece os pesos que determinam o comportamento da camada intermediária. Um termo popular chamado "*backpropagation*" (propagação realimentada) é usado, algumas vezes, quando os pesos são ajustados baseados nas estimativas feitas pela rede.

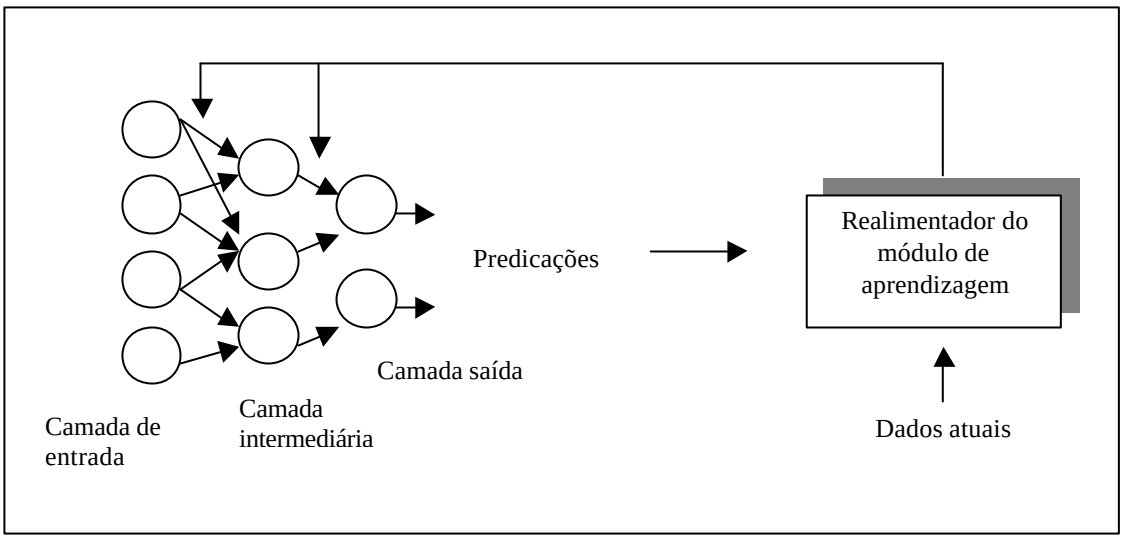

Figura 14 – Propagação Realimentada [POL 97]

Na prática, as redes neurais trabalham somente em subconjuntos e exemplos de dados e as vezes criam problemas quando rodam com grandes conjuntos de dados.

Uma solução para que redes neurais possam ser usadas em bases de dados maiores é se fazer com que os dados no depósito sejam mapeados dentro de números reais antes que a rede possa usá-los, porém isto é uma tarefa difícil para dados comerciais onde muito de seus valores não são numéricos.

Na aplicação de redes neurais o interfaceamento para um grande deposito de dados pode se tornar um problema, pois a entrada para uma rede neural deve ser através de dados numéricos, os quais ainda devem ser pesados ou medidos. Até mesmo campos numéricos (inclusive datas) são medidos ou pesados, isto significa que estes dados são mapeados dentro de valores que os tornem uniformes. Entretanto, é impossível se mapear diretamente valores que sejam numéricos, já que isto iria introduzir relacionamentos inesperados nos dados, conduzindo a erros posteriores. Para explicar o exposto acima tomemos como exemplo a seguinte situação: se nós temos 100 cidades e designamos 100 números para elas (de 1 a 100), cidades com valores 98 e 99 parecerão bem mais relacionadas do que as de números 21 e 77, ou seja, a rede irá pensar que as primeiras cidades estão de algum modo relacionadas, o que pode não ser verdade.

Portanto conclui-se que o melhor meio de usar redes neurais em grande banco de dados será pela combinação com regras, permitindo-lhes fazer prognósticos dentro de uma arquitetura híbrida.

4 Mineração de Dados na *Web* – *Web Mining*

4.1 Introdução à *Web Mining*

Nos últimos anos o número de computadores ligados a Internet tem aumentado extraordinariamente como pode ser visto pelos dados apresentados na tabela 4 sobre o crescimento de DNSs (*Domain Name Servers*) e máquinas na rede. Pode-se perceber que o número de hosts e domínios vem aumentando de maneira exponencial. Em janeiro de 1998, já eram quase 30 milhões de computadores ligados diretamente a Internet. Se considerarmos os computadores que hoje estão atrás de *firewalls* em *Intranets* ou que não ficam permanente ligados a Internet, esse número deve ser duas ou três vezes maior [NET 97].

{PRIVATE**}Data**	**Hosts**	**Domínios**
Jan 98	29.670.000	
Jul 97	19.540.000	1.301.000
Jan 97	16.146.000	828.000
Jul 96	12.881.000	488.000
Jan 96	9.472.000	240.000
Jul 95	6.642.000	120.000
Jan 95	4.852.000	71.000
Jul 94	3.212.000	46.000
Jan 94	2.217.000	30.000
Jul 93	1.776.000	26.000
Jan 93	1.313.000	21.000

Tabela 4 – Número de hosts e domínios da Internet [NET 97]

Este crescimento deve-se, principalmente, a dois fatores: primeiro, a popularização de servidores WWW (World Wide Web), ou simplesmente Web (Teia) e, segundo, o grande número de empresas comerciais que passaram a disponibilizar seus produtos, serviços e informações através da rede [NET 97]. Este é o segmento que mais tem crescido na Internet. Devido a este crescimento explosivo de fontes de informação

disponível na *Web*, tornou-se necessário para os usuários, utilizar ferramentas automatizadas para achar, extrair, filtrar, e avaliar informações e recursos desejados. Além disso, o fato da *Web* ter se tornado um importante meio de comércio eletrônico, fez com que grandes companhias investissem milhões de dólares em Internet e tecnologias de Intranet, juntamente com técnicas para descobrir padrões entre seus usuários, para que, desta forma, pudessem direcionar suas estratégias empresariais. Estes fatores fizeram com que surgisse à necessidade de se criar sistemas inteligentes que pudessem minerar estes dados na *Web*, dando origem a nova tecnologia de mineração de dados, a mineração de dados na *Web* ou *Web Mining* .

Podem se definir *Web Mining*, de forma ampla, como sendo a descoberta e análise de informação útil da *Web*. Esta definição ampla descreve duas técnicas: a primeira é a procura automática e recuperação da informação e recursos através de milhões de locais e de bancos de dados *on-lines*, à esta técnica dá-se o nome de *Web Content Mining*, a segunda visa a análise de dados de usuários objetivando encontrar padrões entre eles, e esta recebe o nome de *Web Usage Mining*.

Nesta seção serão apresentadas as técnicas de *Web Mining*, isto é, a *Web Content Mining* e a *Web Usage Mining*. Também serão descritos alguns recentes trabalhos e ferramentas relacionadas a estas técnicas.

4.2 *Web Content Mining* (Mineração de conteúdo *Web* – páginas)

Dado o cenário atual da *Web*, contendo bilhões de informações, não é de se estranhar que uma das maiores fontes de informações para pessoas e organizações seja através dela. Apesar disto, é difícil localizar informações relevantes sobre determinado assunto somente através dos *links* dos documentos hipertexto. Para permitir que a busca de informações seja mais eficiente foram criados os índices de documentos, conhecidos como *Search Engines*. Exemplos de sites que oferecem este tipo de serviço são AltaVista (www.altavista.digital.com), Excite (www.excite.com), Infoseek (www.infoseek.com.com) e Lycos (www.lycos.com). Esses tipos de *sites* pode funcionar basicamente de duas maneiras [COS 96].

- **Coleta de dados e indexação:** Agentes, conhecidos como robôs, percorrem a Web, coletam, armazenam e indexam as informações.
- **Meta-índices:** Utilizam os próprios índices de pesquisa para realizar a pesquisa.

Estes sites indexam strings em páginas HTML e, portanto, apresentam problemas relacionados com questões de relevância e abrangência.

Outros recursos que visam facilitar a pesquisa são os diretórios organizados com a intervenção humana. Apesar de eficientes, este tipo de site é difícil de se manter atualizado devido ao grande dinamismo da *Web*. Outro problema é que a visão do categorizador da informação pode ser diferente da do usuário. Exemplos de sites deste tipos são *The Mining Company* (home.miningco.com) e *Yahoo* (www.yahoo.com).

A criação desses sites de pesquisa foi um grande avanço para os usuários da *Web*, pois facilitou bastante a pesquisa na rede mas trouxe consigo o problema da relevância e abrangência. Em geral, quanto mais abrangente é uma pesquisa menos relevante são os resultados e, vice-versa, ou seja, quanto mais relevante é uma pesquisa menos abrangente são os resultados [COS 96].

O ponto central dessa questão é que a Web não foi planejada e projetada para armazenar informações de forma organizada e ordenada. Pelo contrário, sua heterogeneidade e falta de estrutura existente em suas fontes de informação, através dos documentos de hypertexto, faz com que a descoberta automatizada de conhecimento, a organização, e a administração das informação neste meio de comunicação, tornem-se atividades difíceis de se executar.

As ferramentas de procura e indexação tradicionais que atuam na *Web*, citadas acima, provêem um pouco de conforto para os usuários, porém elas geralmente não provêem informação estrutural, categorização, filtragem e interpretação de documentos.

Estão sendo realizados, atualmente, estudos e avaliações estatísticas para se comparar as ferramentas de procura mais populares. Estes estudos entusiasmaram os pesquisadores a desenvolver ferramentas mais inteligentes para recuperação de informação na *Web*, através do desenvolvimento de agentes inteligentes, como também de banco de dados e técnicas de mineração, visando, desta forma, prover um nível mais alto de organização para os dados semi-estruturados disponível na *Web*. Alguns destes esforços serão mostrados no decorrer desta seção [LEI 97].

As informações referidas como dados não-estruturados, são aquelas que encontram-se na forma de textos, imagens, gráficos, vídeos ou sons. Segundo Phillip White, chega-se calcular que 80% das informações que uma empresa utiliza não estão armazenadas em Bancos de Dados na forma de números e caracteres. Os chamados dados estruturados, aqueles convencionais, armazenados na maioria dos Sistemas de Gerência de Bancos de Dados, são mais fáceis de serem tratados por meios computacionais, porque existem linguagens formais (como SQL e QBE) que permitem sua manipulação e consulta de forma mais concisa e precisa. Já os dados não-estruturados necessitam de mecanismos computacionais diferentes dos tradicionalmente usados, para que possam ser coletados, armazenados, manipulados e consultados. Esta necessidade por ferramentas e técnicas para tratamento de dados não-estruturados se torna cada vez mais urgente na medida em que aumenta o volume de informações armazenadas, seja na Internet ou em ambientes corporativos privados [WHI 96].

Chen cita a frustração dos usuários com o problema da "sobrecarga de informações". Ela ocorre quando o usuário tem muita informação ao seu alcance, mas não tem condições de tratá-la ou de encontrar o que realmente deseja ou lhe interessa [CHE 96].

4.2.1 Técnicas de Recuperação de Informações

Segundo Jim Cowie apud Loh, a área de Recuperação de Informações procura coletar documentos e artigos com texto relevante. As ferramentas de Recuperação de Informações, geralmente, trabalham com técnicas de indexação, capazes de indicar e acessar mais rapidamente documentos de um Banco de Dados [LOH97].

Segundo Yates apud Loh, existem três tipos de indexação: indexação tradicional, indexação do texto todo, e indexação por tags [LOH97].

- **indexação tradicional:** onde uma pessoa determina os termos descritivos ou caracterizadores dos documentos, os quais farão parte do índice de busca. Este tipo pode ser encontrado na maioria das bibliotecas, sendo objeto de estudo da área de Ciência da Informação. A criação de *Thesaurus* (índice hierárquico de assuntos) é uma das preocupações desta área. Entretanto, o maior problema deste tipo de indexação é que a intervenção humana pode gerar falhas no índice, além de ser uma atividade difícil e que consome muito tempo.

Tais problemas ocorrem porque é necessário estabelecer hierarquias pré-definidas e nelas encaixar os documentos. Também será necessário analisar os documentos para determinar em que categoria se encaixam. Erros de classificação podem ocorrer quando o conteúdo dos documentos não é bem analisado (ou apenas seu título é considerado) ou quando não existem (não foram definidas) categorias intermediárias (ou seja, quando um documento não se encaixa bem em nenhuma das categorias pré-definidas, dando a impressão de que deveria haver outra categoria).

Outros problemas podem ocorrer quando o usuário do índice for recuperar documentos. Se o assunto de busca não for bem determinado ou se não existir uma categoria adequada para o que se deseja, talvez sejam recuperados documentos irrelevantes ou então nenhum documento seja recuperado.

- **indexação do texto todo:** onde todos os termos que compõem o documento fazem parte do índice. Neste segundo tipo de técnicas (indexação *full-text*) procura indexar todos os termos, sem estruturas hierárquicas entre estes. As ferramentas indexam automaticamente todas as palavras do documento, gerando índices volumosos. Para diminuir o problema do tamanho dos índices existem técnicas como árvores, listas invertidas, etc. Também para diminuir o problema do tamanho dos índices, mas principalmente para evitar a indexação de termos indesejados ou inconsistências no índice, podem ser aplicados filtros neste processo [CHE 96].

Por exemplo, as chamadas *stop-words* (palavras que não fazem diferença no conteúdo, tais como preposições e artigos) podem ser eliminadas. Datas e números podem ser normalizados (utilizando-se um padrão). Pronomes podem ser substituídos pelos nomes correspondentes. Sinônimos podem ser substituídos por um termo-padrão, e palavras compostas podem ser substituídas por um termo sinônimo simples (composto de uma palavra só).

- **indexação por tags (por partes do texto):** onde apenas algumas partes do texto são escolhidas, automaticamente, para gerar as entradas no índice (somente aquelas consideradas mais importantes ou mais caracterizadoras). Neste tipo de indexação procura-se indexar somente as partes relevantes do documento. Para tanto, as ferramentas automatizadas deverão analisar cada documento, procurando por marcas (*tags*) que identificam estas partes mais importantes.

Moulin apud Loh, esclarece como a identificação das partes significativas (aquelas a serem indexadas) pode ser feita automaticamente, declarando que os documentos possuem uma macroestrutura (composta por cabeçalhos, títulos, capítulos, etc), uma microestrutura (a qual contém o conteúdo lógico do documento e que pode ser identificada por palavras-chave que introduzem condições, exceções, referências, etc) e uma camada de domínio (contendo as demais informações) [LOH 97].

Nesta mesma linha, as ferramentas descritas por Hardy apud Loh et al. extraem palavras-chave com base em marcas (*tags*) da macroestrutura dos documentos. Há uma tabela de *tags* específica para cada tipo de arquivo (por exemplo: programas contém *procedures*, funções, comentários, etc; mensagens de correio eletrônico contém destinatário, remetente, assunto) [LOH 97].

Entretanto, estas ferramentas encontram problemas quando o documento não possui uma macroestrutura pré-concebida (tal como acontece nos textos livres armazenados em arquivos ".TXT"). Neste caso, somente as palavras das cem primeiras linhas serão indexadas, para evitar uma indexação *full-text*. A premissa utilizada por Hardy é a de que no início do documento é que se encontra sua descrição.

Para a indexação por *tags*, a maioria dos trabalhos adota o uso de gramáticas, *parsers*, expressões regulares e autômatos finitos para a definição e identificação das marcas.

O maior problema das técnicas de indexação por *tags* é que os tipos de documentos devem ter sido analisados previamente, ou seja, já se deve conhecer que tipo de conteúdo compõe os documentos e quais as marcas que identificam as partes relevantes de um documento. Então as ferramentas acabam sendo muito específicas para determinados tipos de informação e documentos.

Outro grave problema que ocorre nas técnicas tradicionais de recuperação de informações é que muitas vezes documentos importantes não são recuperados, enquanto que outros irrelevantes são apresentados ao usuário. Isto se dá porque estas técnicas estão baseadas na presença ou não de palavras nos documentos. Entretanto, pode haver documentos relevantes que não contém as palavras especificadas para a busca e pode haver documentos que contém as palavras de entrada mas que não tratam do assunto desejado.

Este problema é denominado de indexação imprecisa e ocorre porque a pessoa ou técnica que descreve e indexa os documentos pode utilizar termos diferentes de quem procura pelos documentos. Alguns trabalhos procuram solucionar tal problema com técnicas de indexação semântica ou por contexto.

Jim Cowie apud Loh, sugere que as técnicas para indexação devem gerar índices sensitivos mais intimamente relacionados ao real significado de um texto em particular e não baseados na presença de termos sem identificação do contexto [LOH 97].

Chen define contexto ou espaço conceitual como sendo um conjunto de palavras que definem um assunto ou área do conhecimento. Algumas técnicas então procuram recuperar documentos baseadas no contexto dos documentos. Além disso o mesmo autor discute técnicas baseadas na freqüência de termos em documentos. Para determinar a importância de um termo em um contexto (o quanto ele ajuda a definir um

contexto), é necessário primeiro identificar os termos que compõem os documentos, bem como quantas vezes cada termo aparece em cada documento [CHE 96].

Depois, através de fórmulas baseadas na Teoria da Informação, pode-se determinar o grau de relação entre o termo e o documento. Feito isto, o grau de relação entre os termos pode ser definido também com base em fórmulas já testadas. Ou seja, termos que aparecem nos mesmos documentos, podem ser utilizados conjuntamente para a busca de documentos deste contexto (estes termos relacionados ajudam a definir melhor o contexto).

Estas fórmulas baseiam-se na freqüência e na freqüência inversa dos termos e ajudam a definir que termos podem ser usados para recuperar determinado contexto (ou documentos deste contexto). Se um termo aparece em muitos documentos, seu grau de discriminação será baixo (pois muitos documentos serão recuperados a partir deste termo), enquanto que, se um termo aparece em poucos documentos, então diz-se que ele caracteriza bem estes documentos. Obviamente, termos que aparecem em todos os documentos (como as *stop-words*) não serão analisados.

Assim, podem ser criadas ferramentas que estabeleçam o conjunto de termos que definem um contexto, a partir de documentos textuais específicos deste contexto, previamente escolhidos (ou por trabalho humano, ou por *tags* específicas, como seções de um jornal).

Cada termo dentro do conjunto determinante pode ter um grau maior ou menor de importância, e este grau pode ser utilizado para classificar os documentos recuperados em ordem de relevância (com base nos termos presentes no documento e seus respectivos graus de importância).

Entretanto, Jim Cowie apud Loh, adverte: "modelos probabilísticos de freqüência de palavras, usados para computar a probabilidade de um texto ser relevante, dependem da disponibilidade de documentos relevantes e irrelevantes" (e este trabalho deve ser feito manualmente) [LOH 97]. Também para tratar o problema de busca contextual, há a técnica de Chakravarthy, a qual se utiliza de expansões semânticas de

palavras [CHA 95]. Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas com ela, utilizando então este conjunto para busca de documentos. O referido autor utiliza as definições de um dicionário para achar as palavras que se relacionam, eliminando stop-words e modela estas relações através de redes semânticas, criadas manualmente. Entretanto, os problemas desta técnica são saber que palavras expandir para fazer a busca e se as novas palavras acrescentadas realmente fazem parte do contexto. Segundo os experimentos de Chakravarthy, algumas das novas palavras não fazem parte do contexto, o que pode fazer com que documentos irrelevantes sejam recuperados. Quando houver mais de um contexto possível para uma dada situação, Wiess cita técnicas que utilizam modelagem de contextos alternativos, permitindo que contextos diferentes possam ser explorados em paralelo [WIE 96].

A raiz do problema de contextos diferentes está na imprecisão dos termos (termos com significados diferentes). Este problema pode ser notado tanto no momento da criação do índice, como na hora da recuperação. Isto se dá porque as pessoas utilizam vocabulários diferentes para exprimir suas intenções. Para tratar este problema, Clarisse Souza definiu técnicas para geração de linguagens (escolha de termos, expressões, representações, signos e símbolos) e para a interpretação de termos e linguagens [SOU 93].

A partir das referidas técnicas mostradas acima para recuperação da informação, podemos classificar a mineração de conteúdo na Web de duas formas: a Aproximação Baseada em Agentes e Aproximação Baseada em Banco de Dados.

4.2.2 Aproximação Baseada em Agente

Define-se agente como sendo uma peça de software que executa uma determinada tarefa empregando informação extraída de seu ambiente para agir de forma adequada no sentido de completar sua tarefa de modo bem sucedido. O software deve ser capaz de adaptar-se a eventuais modificações ocorridas em seu ambiente de modo que o resultado pretendido seja independentemente alcançado. Segundo Hermans apud Castilho, a utilização destes agentes inteligentes vem ganhando uma grande aceitação na busca de conhecimento na Web, tendo em vista que eles superam muito as técnicas utilizadas em serviços de procura atual, como Alta vista, Lycos etc, os quais são

baseados em *search enginnes* [CAS 98]. Por exemplo, os programas que coletam informações para o Lycos - um serviço de busca de informações a partir de palavras-chaves - criam resumos dos documentos que encontram constituídos pelo título, subtítulo e cabeçalhos, as cem palavras de maior *peso* (segundo um algoritmo particular de cálculo) as vinte primeiras linhas, seu tamanho em bytes e o número de palavras. A informação coletada, caracterizada por suas palavras-chaves - às vezes por alguma informação suplementar, é então inserida em uma grande base de dados. Usuários interessados em algum tipo de informação recorrem a essa base de dados por meio de consultas onde fornecem palavras-chaves relacionadas com o assunto de interesse. Este método de procura poderia se implementado com maior facilidade através da utilização de agentes.

Segundo Hermans apud Castilho, as comparações entre a aplicação de agentes inteligentes e as técnicas atuais de pesquisa na Web podem ser analisadas através da Tabela 5 a seguir [CAS 98]:

	Mecanismos de Pesquisa	**Agentes de Informação Inteligentes**
1	Uma pesquisa é realizada com base em uma ou mais palavras-chaves fornecidas pelo usuário. Isto pressupõe que o usuário é capaz de formular o conjunto mais adequado de palavras-chaves para a recuperação da informação desejada. Consultas realizadas com palavras-chaves erradas, a mais ou a menos irão produzir a recuperação de informação irrelevante (ruído) ou não irão recuperar informação relevante.	Agentes conseguem pesquisar informações de modo mais inteligente, por exemplo com o uso de ferramentas (tais como um thesaurus) que os habilitam a realizar a pesquisa baseados em termos relacionados; mesmo em conceitos. Os agentes também podem empregar tais ferramentas para refinar ou até corrigir consultas, com base em um modelo do usuário ou outras informações
2	O mapeamento da informação é realizado pela coleta de (meta-)informação sobre a informação e documentos disponíveis na Internet. Este método consome tempo demasiado, além de ocasionar o tráfego de grandes Quantidades de dados. Falta-lhe eficiência e uma melhor adequação à natureza dinâmica da Internet e à da informação que nela pode ser encontrada.	Agentes de informação individuais podem criar sua própria base de conhecimento sobre fontes de informação disponíveis na Internet, a qual é atualizada e expandida após cada pesquisa. Quando a informação (um documento) houver sido removida para um outro local o agente deve ser capaz de encontrá-la e atualizar sua base de conhecimento.
3	A pesquisa por informações é muitas vezes limitada a apenas alguns serviços da Internet, como o WWW. Pesquisar informações oferecidas por outros serviços (p. ex.: uma base de dados com acesso via telnet) é um trabalho usualmente executado de forma manual pelo usuário.	Agentes de informações podem assumir os detalhes protocolares de conexão com os diversos serviços e também a localização da informação, deixando para o usuário apenas a formulação da consulta: o *que* exatamente deve ser procurado, sem se preocupar com *onde* possa estar.

4	Os mecanismos de pesquisa nem sempre estão disponíveis. O servidor onde um serviço reside pode ter *caído* ou então estar sobrecarregado. Os usuários regulares então terão de procurar um serviço alternativo, provavelmente com operação e características diferentes.	O agente pessoal de informações está sempre à disposição de seu usuário e pode ser executado em paralelo com outras tarefas ao longo das 24 horas do dia. Além disso p ode ser programado para detectar e evitar horas de pico em diferentes fusos horários.
5	Mecanismos de pesquisa atuam de forma independente de Domínio no modo de tratar a informação coletada e no modo com que a oferece ao usuário. Os termos nos documentos coletados são extraídos de seu contexto e tratados como se fossem mera lista de palavras-chaves.	Agentes inteligentes serão capazes de pesquisar por informação com base em contextos. Irão deduzir tal contexto a partir do modelo que possuem do usuário ou por meio de outros serviços.
6	A informação na Internet comporta-se de forma extremamente dinâmica e muitas vezes os mecanismos de pesquisa referem-se a informações que foram transferidas para locais desconhecidos, ou simplesmente desapareceram. Os coletores de informações não aprendem com tais tentativas fracassadas e portanto não se auto-ajustam para evitar falhas futuras. Além disso, como realizam sua pesquisa em intervalos regulares o usuário não tem notícia de atualizações realizadas	Os agentes podem ajustar-se às preferências e desejos de seus usuários individuais. Idealmente isto poderá conduzir a agentes em permanente ajuste quanto ao que deve ser procurado com base nas reações de seus usuários ao resultado de pesquisas anteriores. Além disso os agentes são capazes de pesquisar a Internet de modo contínuo em busca de nova informação sobre tópicos em que seu usuário costuma estar interessado.

Tabela 5 - Comparação entre Mecanismos de Pesquisa e Agentes de Informação Inteligentes [CAS 98]

A aproximação baseada em agentes para mineração de dados na Web envolve o desenvolvimento de sistemas sofisticados de IA, os quais podem agir de forma autônoma ou não, em nome de um usuário particular, descobrindo e organizando informações presentes na rede.

Geralmente esta técnica se divide em três categorias de sistemas de mineração. Os exemplos de agentes, que se encontram dentro dessas três categorias, foram estabelecidos segundo Coolye et. al. [CMJ 95].:

- **Agentes Inteligentes de Procura**

Foram desenvolvidos vários agentes inteligentes para procurarem informações pertinentes, os quais usam características de um domínio particular e

possivelmente um perfil de usuário, visando organizar e interpretar a informação descoberta.

Como exemplo deste tipo de agente, temos: Harvest, FAQ-Finder e Information Manifold que trabalham ou através de pré-especificação de informações de domínio específicas sobre tipos particulares de documentos, ou em modelos codificados das fontes de informação. Outros agentes, como ShopBot e ILA (Internet Learning Agent), tentam interagir e aprender com a estrutura da fonte de informação. O ShopBot recupera informações de produtos de uma variedade de diferentes locais de venda, e através disto usa a informação recuperada sobre o domínio do produto. Por outro lado, o ILA aprende com modelos de várias fontes de informação e traduz estes em sua própria hierarquia interna de conceitos [CMJ 95].

- **Filtragem e Categorização da Informação**

Vários agentes de rede usam diversas técnicas de recuperação de informação e categorização de documentos hypertexto na *Web*, eles atuam abrindo o documento, recuperando sua informações automaticamente, filtrando-as e categorizando-as. Por exemplo, HyPursuit usa informação semântica embutida em estruturas de *links* como também conteúdo de documento para criar hierarquias de agrupamento de documentos hypertexto, estruturando um espaço de informação, já o BO (Organizador de Marcador de Páginas) combina técnicas, as quais são armazenadas obedecendo uma hierarquia, e a interação do usuário, organizando desta forma uma coleção de documentos Web baseando-se em informações conceituais.

- **Agentes de Rede Personalizados**

Outra categoria de agentes de rede inclui aqueles que obtêm ou aprendem preferências do usuário e descobrem fontes de informação de rede que correspondem a estas preferências, e possivelmente, através destes, preferências de outros indivíduos com interesses semelhantes (utilizando filtragem de informação). alguns recentes exemplos de tais agentes incluem o *WebWatcher e Paint*.

4.2.3 Aproximação de Banco de Dados

Aproximações em Banco de dados para mineração de conteúdo na *Web* enfoca técnicas que visam integrar e organizar o conteúdo heterogêneo e semi-estruturado dos dados da *Web* em coleções de dados mais estruturadas que possam disponibilizar uma maior quantidade de recursos. Esta organização pode ser feita através da formação de bancos de dados relacionais, banco de dados de padrões e de técnicas de mineração para analisar os dados presentes nestes bancos.

Abaixo serão mostrados dois tipos desta técnica [CMJ 95]:

- **Bancos de Dados Multinível**

A idéia principal na utilização desta técnica é seguinte: o mais baixo nível do banco de dados armazena informações primitivas, semi-estruturadas, presentes em vários banco de dados da Web, como documentos de hypertexto. O Nível mais alto armazena os meta dados ou generalizações obtidas dos níveis mais baixos, mantendo-as organizadas em coleções estruturadas de dados, de forma relacional ou através de bancos de dados orientados a objetos.

Por exemplo, [ZAI 95] utiliza um banco de dados multinível onde cada nível é obtido através de generalização e operações de transformação realizados nos níveis mais baixos.

- **Sistemas de Questionamento na *Web* (Linguagens)**

Existem muitos sistemas de questionamento de dados na *Web*, entre os quais se utilizam de diversos tipos de linguagens como SQL, que estrutura informações sobre documentos da rede, e linguagens naturais, usadas para se implementar os tipos de questões que são utilizadas em procura de conhecimento na *Web*. Como exemplo, podemos citar o TSIMMIS, que extrai dados heterogêneos e semi-estruturados de fontes de informação *Web* e gera uma representação de banco de dados integrada desta informação [CMJ 95].

4.3 *Web Usage Mining* (Mineração de dados de usuários na *Web*)

Web Usage Mining é a atividade de mineração de dados que visa a descoberta automática de padrões de usuários, através de seus dados de acesso a *Web*. Esta técnica vem se destacando muito atualmente, pois inúmeras organizações estão confiando cada vez mais na Internet e na *Web* para administrarem seus negócios, logo fazendo com que as estratégias tradicionais e técnicas para análise de mercado passem a ser realizadas através deste fabuloso meio de comunicação.

Organizações geram e acumulam freqüentemente volumes enormes de dados em suas operações diárias. A maioria desta informação normalmente é gerada automaticamente através de servidores de rede e é guardada em *logs* de acesso ao servidor.

A análise feita nestes arquivos de *logs* pode ajudar as organizações a detectarem uma série de padrões existentes entre os usuários, como: a estimativa de vida de clientes, estratégias de marketing direcionadas a determinados produtos, campanhas de promoção entre outras coisas.

Análise feita em dados de acesso ao servidor e dados de inscrições de usuário também podem prover valiosa informação. Por exemplo, em como melhorar a estrutura de uma página, visando um aumento no número de visitas à mesma e com isso possibilitar uma maior popularidade para a organização.

Em organizações que usam tecnologias de Intranet, estas análises podem facilitar o conhecimento entre grupos de trabalho, e com isso, fornecer a administração um maior controle da infra-estrutura organizacional.

Para aquelas organizações que utilizam a *Web* para anunciar seus produtos a análise em padrões de usuários torna-se importante pois ela ajuda a organização a direcionar seus anúncios para grupos específicos de usuários.

Em se tratando de análise de redes, a maioria das ferramentas existentes provêm de mecanismos para informar as atividades de usuários nos servidores e filtrar

várias formas de dados. Por exemplo, usando tais ferramentas é possível determinar o número de acessos ao servidor e aos arquivos individuais dentro da rede da organização, os intervalos e a duração de tempo das visitas, o domínio e as URLs visitadas no servidor da rede. Em geral, porém estas ferramentas são projetadas para melhorar o tráfego entre servidores, provendo pouca ou nenhuma análise na relação dos dados presentes nos arquivos de acesso com os diretórios do servidor de rede [CMJ 95].

Sistemas mais sofisticados e técnicas para descoberta e análise de padrões estão emergindo agora. Estas ferramentas podem ser colocadas em duas categorias principais, como mostradas abaixo [CMJ 95]:

4.3.1 Ferramentas de Descoberta de Padrões

As ferramentas de descoberta de padrões que estão surgindo atualmente, fazem enorme uso da tecnologia de IA, mineração de dados, psicologia, teoria da informação, e outras relacionadas. Por exemplo, o sistema WEBMINER proposto por Mobasher, o qual será apresentado no decorrer deste trabalho, introduz uma arquitetura geral para mineração dos dados de usuários de rede. O WEBMINER descobre automaticamente associações e padrões através dos dados presentes nos *logs* de acesso ao servidor.

O conteúdo descoberto nessas associações pode ser utilizado para se detectar diferentes caminhos seguidos pelos usuários dentro da rede, combinando padrões de caminhos tomados dentro da rede, com informações de topologia da rede, para que desta forma, possa-se categorizar páginas que tenham um acesso mais fácil por usuários [MOB 96].

Em seção subseqüente nós discutiremos algumas destas técnicas propostas em mais detalhes.

4.3.2 Ferramentas de Análise de Padrão

Uma vez que foram descobertos padrões de acesso, os analistas precisam das ferramentas e técnicas apropriadas para entender, visualizar, e interpretar estes

padrões. Como exemplo de tais ferramentas podemos citar o WebViz que visualiza padrões de caminhos na *Web* [PIT 94]. Existem também ferramentas que utilizam técnicas de OLAP com a finalidade de simplificar a análise estatística de dados presentes nos *logs* de acesso ao servidor. O sistema WEBMINER [MOB 96] propõe uma a linguagem SQL como mecanismo para examinar o conhecimento descoberto (na forma de regras de associação e padrões seqüenciais), que serão discutidas nas seções seguintes.

4.4 Estrutura Para Mineração na *Web*

Como discutido na seção acima, análise feita nos dados deixados pelo usuário ao visitarem um site é de fundamental importância, como por exemplo, para se determinar estratégias de marketing poderosas ou aperfeiçoar a estrutura lógica de um *site* na *Web*. Devido a diferentes características de clientes de servidor existentes na *Web*, de diferentes topologias, proporcionando que usuários tenham diferentes meios de acesso aos dados e a dificuldade em se identificar usuários como também sessões de usuário ou transações, é necessário se desenvolver uma nova estrutura para facilitar o processo de mineração. Deve-se levar em consideração o pré-processamento dos dados antes de se aplicar os algoritmos de mineração. Este processo inclui desenvolvimento de um modelo de dados de *logs* de acesso, técnicas para filtrar o conjunto de dados descartando aqueles que sejam irrelevantes, se agrupando acessos de página individuais em unidades semânticas, ou seja, em transações, integração de várias fontes de dados como informação de inscrição de usuários e especificando dados genéricos de algoritmos de mineração para tirar proveito da natureza específica dos arquivos de *log* [CMJ 95]. Levando-se em consideração o exposto acima, alguns aspectos devem ser observados para a estruturação da mineração, são eles:

4.4.1 Tarefas de Pré-Processamento

1) Limpeza de dados

Técnicas para limpar *logs* de servidores para eliminar os dados irrelevantes são de importância para qualquer tipo de análise de *logs* de rede, tanto quanto a

mineração de dados. As associações descobertas ou as estatísticas são úteis somente se os dados representados no *log*, deixados no servidor pelo usuário, forem precisos.

Eliminação de artigos irrelevantes pode ser realizada de forma razoável conferindo-se o sufixo do nome de arquivos em uma URL. Por exemplo, todas as entradas de *logs* com sufixos de nomes de arquivo como: gif, jpeg, jpg, etc; podem ser removidas [CMJ 95].

Um problema existente na coleta de dados de acesso é que as vezes o arquivo *log* não grava acessos importantes. Mecanismos como *caches* locais e servidores *proxy* podem modificar totalmente o deslocamento do usuário através da rede.

Uma página que só é listada uma vez em um *log* de acesso pode ter sido de fato visitada muitas vezes por usuários múltiplos. Métodos atuais para tentar superar este problema incluem o uso de *cookies*, *cache busting*, e registro explícito de usuário. Porém, estes métodos possuem desvantagens sérias. *Cookies* podem ser apagados pelo usuário, *cache busting* e inscrição de usuários são técnicas voluntárias, onde os usuários, freqüentemente, provêem falsa informação [CMJ 95].

Servidores *proxy* também possuem problemas associados com identificação de usuários. Uso de um nome de máquina para identificar os usuários exclusivamente pode resultar em vários usuários que se agrupam erroneamente como um usuário.

Um algoritmo apresentado por Pirolli mostra que cada requisição de entrada já esta incluída entre as páginas visitadas [PIR 96]. Se uma página pedida não está diretamente ligada às páginas anteriores, é considerado então, que existem usuários múltiplos para a mesma máquina. Segundo Mobasher, são usadas durações de sessão de usuário determinadas automaticamente baseado em padrões de navegação para identificar os usuários [MOB 96]. Outra maneira de se identificar o uso de um mesmo servidor de *proxy* para diferentes usuários, envolve o uso de uma combinação de endereço de IP, nome de máquina, agente de browser, e informação temporal.

2) Identificação de Transação

Após os dados do arquivo *log* serem limpos, deve-se agrupar as referências de página sucessivas em unidades lógicas que representam transações de rede ou sessões de usuário. Uma sessão de usuário é o conjunto de todas as referências de página visitadas por este usuário durante uma seção de navegação na *Web*. Ao se identificar sessões de usuários surge um problema semelhante ao problema de identificar os usuários individuais, como foi discutido anteriormente [CMJ 95].

Uma transação difere de uma sessão de usuário, pois ela pode ser uma única referência feita a uma página, enquanto que uma seção de usuário pode ter diversas referências de página.

Mobasher assume que cada referência a uma pagina ou mais páginas para a obtenção de conteúdo é definida como uma transação. Segundo ele, existem dois tipos de transações:

O primeiro tipo é aquela referente a navegação a procura de conteúdo, neste caso, a transação consiste em uma única referência ao conteúdo e todas as demais navegações serão direcionadas através de caminhos que levam ao conteúdo procurado. Estas transações podem ser usadas para mineração de padrões de caminhos na *Web*.

O segundo tipo de transação refere-se a todas referências de procura a conteúdos de uma determinada sessão de usuário. Estas transações podem ser usadas para descobrir associações entre conteúdos de páginas [MOB 96]. Uma determinada referência de página é classificada como "navegacional" , baseado no tempo gasto na página. Este tipo de referencia será mostrado mais adiante em [PIR 96], onde vários tipos de página como páginas de índice, paginas pessoais, etc. são usadas para descoberta de padrões de usuário.

Chen et al, define o conceito "máxima referência a frente" para identificar as transações. Cada transação é definida como o conjunto de páginas desde a primeira página no *log* até uma determinada página, até que uma referência para trás seja feita.

Uma transação nova é começada quando a próxima referência a frente é feita, ou seja, quando é referenciada uma pagina que não esteja incluída no conjunto de páginas da transação atual [CPY 96].

Por exemplo, uma sucessão de acesso B C D C B E F E G A H seria descrita em três transações, ou seja, B C D, depois B E F, e por fim G A H. As transações criadas com este algoritmo podem ser usadas para mineração de padrões de caminho na Web.

4.4.2 Descoberta de Padrões em Transações na *Web*

Após identificadas as transações de usuários como foi discutido na seção anterior, podem ser executados vários tipos de mineração de padrões de acesso, dependendo das necessidades do analista. Alguns destes serão discutidos abaixo:

1) Análise de caminho

Há muitos tipos diferentes de grafos que podem ser formados quando se executa uma análise de caminho, desde que o grafo utilizado represente alguma relação entre o caminho percorrido pelo usuário durante a navegação. O mais utilizado é um grafo que representa o plano físico de um *Web Site*, com páginas como nodos e *links* de hipertexto como arestas que ligam os nodos. Poderiam ser formados outros grafos baseado nos tipos de páginas com arestas que representam semelhança entre estas páginas, ou criando extremidades que dão o número de usuários que vão de uma página para outro [PIR 96]. A maioria do trabalho para traçar estes grafos se encontra na obtenção dos padrões de caminhos acessados mais freqüentemente e na obtenção das referências sucessivas. A análise de caminho é principalmente usada para se determinar a freqüência com que os caminhos são tomados em uma rede, mas além disso ela pode prover outras informações, como [CMJ 95]:

- 70% de clientes que tiveram acesso /empresa/produtos/arquivo1.html fizeram-no da seguinte maneira: começando a /empresa e procedendo por /empresa/anúcios, / empresa /produtos, e empresa/produtos/arquivo1.html;
- 80% de clientes que tiveram acesso ao local começaram de /empresa/produtos;
- 65% dos clientes acessaram o site depois de quatro ou menos referências de páginas.

A primeira regra sugere que haja informação útil em /empresa/produtos/arquivo1.html, porém esta informação não está exposta claramente, pois o usuário teve que percorrer um determinado caminho até chegar a ela.

A segunda regra simplesmente mostra que a maioria dos usuários está tendo acesso ao local por uma página diferente da página principal (assumindo-se aqui /empresa como página principal), logo poderia ser uma boa idéia, divulgar informação do diretório da página principal ou de um link direto a ela.

A última regra indica que muitos usuários não passam por mais que quatro páginas para chegar ao local desejado, aqui seria prudente guardar as informações passíveis no decorrer do caminho por estas páginas.

2) Regras de Associação

A associação de regras geralmente é utilizada em bancos de dados de transações onde cada transação trabalha com um conjunto de itens ou atributos. Aqui o problema é descobrir todas as associações e correlações entre os dados onde a presença de um item no conjunto de dados de transação implica (com um certo grau de confiança) na presença de outros itens. No contexto do Web Mining, este problema refere-se as correlações entre referências para vários arquivos disponível no servidor feitas por um cliente. Cada transação é incluída de um conjunto de URLs que o cliente acessou em uma visita ao servidor. Por exemplo, usando-se regras de associação poderíamos achar correlações como o seguinte:

- 40% de clientes que tiveram acesso a página da *Web* com URL /empresa/produtos/produto1.html, também teve acesso / empresa /produtos/produtos2.html;

- 30% de clientes que tiveram acesso /empresa/cds/lançamentos.html, fez um pedido online em /empresa/produtos/produto1.

A descoberta de conhecimento através de regras está sendo muito utilizada por organizações que baseiam seus negócios em comércio eletrônico, tendo em vista que esta técnica pode ajudar no desenvolvimento de estratégias de marketing. Além disso, regras de associação aplicadas a WWW através de arquivos de logs, podem dar uma indicação de como melhorar a organização do espaço da rede da organização [CMJ 95]. Por exemplo, se a organização descobre que 80% dos clientes que acessam /empresa/produtos e /empresa/produtos/arquivo1.html também tiveram acesso a /empresa/produtos/arquivo2.html, mas só 30% dos que acessaram somente à /empresa/produtos tiveram acesso a /empresa/produtos/arquivo2.html, então é provável que informações contidas em arquivo1.html conduza os clientes a acessarem arquivo2.html. Esta correlação poderia sugerir que as informações contidas em arquivo2.html devessem ser movidas a um nível mais alto (por exemplo, /empresa/produtos), para que desta forma, o acesso a arquivo2.html aumente.

3) Padrões Seqüenciais (padrões de tempo)

Padrões seqüências são aqueles que baseiam-se no tempo de duração das transações ocorridas em um servidor. Em transação de servidor de rede remota, uma visita feita por um cliente é registrada no arquivo de *log*, durante um certo tempo. O tempo associado com uma transação neste caso será identificado por um processo de identificação de transação executado e sua duração se dá até o momento em que os dados sejam limpos.

A descoberta de padrões seqüenciais em redes através do acesso aos arquivos de *log* permite se fazer uma organização de dados de forma que se possa

determinar padrões de usuários através de suas visitas, podendo, através da obtenção destes padrões, direcionar a atenção a um determinado grupo de clientes.

Um sistema de mineração de dados na Web, através da utilização destes padrões seqüenciais, pode fazer as seguintes observações [CMJ 95]:

- 30% de clientes que visitaram /empresa/ bicicletas/, fizeram uma procura em *Yahoo*, dentro da última semana, em acessórios para bicicletas;
- 60% de clientes que fizeram um pedido online em /empresa/produtos/produto1.html, também fizeram um pedido online em /empresa/produto/produto4 dentro de 15 dias.

Outro tipo importante de dependência de dados onde pode se usar características temporais dos dados, é as "sucessões de tempo semelhantes". Por exemplo, nós podemos estar interessados em achar características comuns de todos os clientes que visitaram um determinado arquivo dentro de um período de tempo. Ou, reciprocamente, nós podemos nos interessar por um intervalo de tempo (dentro de um dia, ou dentro de uma semana, etc.) em que um arquivo particular teve a maioria dos acessos.

4) Clusterização e Classificação

Regras de classificação na Web permitem desenvolver perfis de usuários que tenham acesso ao servidor. Isto se dá através de características de atributos em comum, como informações demográficas destes usuários ou baseado nos padrões de acesso dos mesmos, facilitando, desta forma, a designação destes usuários em grupos (*clusters*) [CMJ 95].

Por exemplo, as regras de classificação na *Web* aplicadas a arquivos de *logs* podem conduzir à descoberta de relações como as seguintes:

- funcionários do estado ou do governo que visitam o site, tendem a se interessar pela página /empresa/produtos/produto1.html;

- 50% dos usuários que fizeram um pedido online em /empresa/produtos/produto2, estavam no grupo de pessoas que possuíam idade de 20 à 25 anos e que habitavam na Costa Ocidental.

Em alguns casos, valiosas informações sobre o usuário podem ser guardadas automaticamente pelo browsers do usuário, informações estas que são armazenadas no arquivo histórico, em *cookies* etc. Outro método de obtenção de perfil de usuário, é através da inscrição *online* do mesmo.

4.5 Análise de Padrões Descobertos

Segundo Cooley et al, administradores de Web sites estão extremamente interessados em perguntas como: "Como as pessoas estão usando o site ?", "Que páginas são acessadas mais freqüentemente?", etc. para se responder a estas perguntas deve-se fazer uma análise na estrutura de hyperlinks como também nos conteúdos das páginas, levando-se em consideração: 1) a freqüência de visitas ao documento, 2) a mais recente visita ao documento, 3) quem está visitando os documentos, 4) qual a freqüência de uso de cada *hyperlink* e 5) o mais recente uso de cada *hyperlink* [CMJ95].

A descoberta de padrões de uso na *Web*, levando-se e consideração as técnicas já mostradas, não seria muito útil a menos que houvesse mecanismos e ferramentas para ajudar o analista a entender os padrões descobertos. Consequentemente, além do desenvolvimento de técnicas para mineração de informações de usuário, a fim de se obter padrões entre eles, há também a necessidade de se desenvolver técnicas e ferramentas que possibilitem uma análise dos padrões descobertos.

É esperado que estas técnicas se munam de informações de outros campos como estatística, grafos, análise de usabilidade, banco de dados e etc . Nesta seção será mostrado uma pesquisa, segundo Cooley et. al., das ferramentas e técnicas existentes. Como a análise de dados de usuários da Web é uma área muito nova, há muito pouco trabalho nisto, tornando a pesquisa não muito extensa [CMJ 95].

4.5.1. Técnicas de Visualização

Técnicas de visualização estão sendo muito usadas, pois elas ajudam as pessoas a entender vários tipos de fenômenos, tanto reais como abstratos, consequentemente sendo também aplicadas na Web, a fim de se entender melhor as informações descobertas. Pitkow desenvolveu o sistema chamado WebViz para visualizar padrões extraídos da *Web*. Este sistema utiliza o paradigma de definição de caminhos na *Web* na qual são usados conjuntos de entradas de *logs* do servidor para extrair subseqüências de padrões de caminhos na rede. WebViz permite que o analista examine a porção da rede que é de interesse filtrando as porções irrelevantes. A rede é visualizada como um grafo, onde os nodos são páginas e os *hyperlinks* as arestas.

A visualização é feita através de duas janelas. A primeira proporciona para o analista controles para ajustar as ligações, selecionar um tempo específico para realizar a análise, ajustar a animação, e reorganizar o *layout*. A segunda janela permite representar a freqüência de acesso a um documento pela largura do nodo que o representa (cor do nodo). Manipulação temporal é alcançada através de controles de *playback* [PIT 94].

4.5.2 Técnicas de OLAP

O processo analítico on-line (OLAP) é um paradigma muito poderoso para análise estratégica de bancos de dados em organizações empresariais. Algumas das características chaves deste tipo de análise estratégica incluem: 1) volume de dados muito grande, 2) apoio explícito para a dimensão temporal, 3) apoio para vários tipos de agregação de informação e 4) análise de longo alcance, onde tendências globais são mais importantes que detalhes de atributos de dados individuais [CMJ 95].

Recentes trabalhos mostram que a análise precisa de dados de usuários da Web tem muito em comum com a tecnologia do Data Warehouse, logo, técnicas de OLAP são bastante aplicáveis.

O arquivo *log*, que guarda os dados de acesso dos usuários ao servidor cresce muito com o passar do tempo, logo tornando-se impossível manter um *log* que

guarde a história de todos os acessos feitos ao servidor, especialmente porque muitos usuários usam um servidor *proxy*. Logo informações sobre os acessos serão distribuídas, e consequentemente haverá a necessidade de se integrar e manipular estes dados. Como o *log* de um servidor cresce muito rápido, pode não ser possível se fazer uma análise on-line de todo o arquivo *log*. Então, há uma necessidade de se resumir os dados de *log*, talvez de vários modos, e com isso fazer sua análise on-line. Também deve-se levar em consideração que os analistas, por razões de segurança, requeiram a seleção de alguns *logs* de forma invisível.

Estas exigências para análise de dados de usuários na *Web* se adaptam perfeitamente as técnicas de OLAP, tornando-as bastante aplicáveis na *Web*. Esta é uma tendência muito forte para o futuro desta tecnologia de *Web Mining*.

4.5.3 Restrição de Dados

Um das razões do grande sucesso da tecnologia de banco de dados relacional foi o fato do mesmo permitir que uma determinada aplicação pudesse adquirir seus dados através de condições dadas e não especificando os dados necessários de forma direta. Devido ao grande número de padrões que podem ser minerados é necessário que se indique o enfoque da análise, que pode ser dado de dois modos: primeiro, restringindo a porção do banco de dados a ser minerado e segundo, examinando e armazenando somente conhecimento valido que foi extraído pelo processo de mineração, o que pode ser feito através de uma linguagem que desempenhe tal função. A linguagem SQL foi utilizada como este mecanismo para examinar e reter os dados pertinentes no sistema WEBMINER que será mostrado mais adiante.

4.5.4 Análise de Usabilidade

Pesquisadores na área da interação do computador com o ser humano começaram a desenvolver recentemente uma ciência para estudar a usabilidade computacional. O primeiro passo dado foi desenvolver métodos de instrumentação que colecionassem dados sobre usabilidade de software a fim de medir sua performance. Então, através destes dados são construídos modelos computadorizados e simuladores

que explicam os dados que estão sendo manipulados. Finalmente, são usados vários tipos de apresentações de dados e técnicas de visualização para ajudar o analista a entender o fenômeno. Esta aproximação também pode ser usada para modelar o comportamento de usuários na rede.

Como descrito nesta seção, há uma necessidade e interesse crescente em se desenvolver técnicas e ferramentas para analisar os padrões de usuários através de suas informações deixadas na rede.

Acredita-se que as técnicas que estão tendo uma maior aceitação atualmente incluirão as seguintes características: 1) utilização de métodos empíricos, 2) grande quantidade de dados para validação, 3) usará métodos experimentais rigorosos, como análise estatística, etc, [CMJ 95].

4.6 Apresentação de uma Arquitetura para Mineração de Dados de Uso da *Web*

Será mostrada nesta seção uma arquitetura geral para mineração de dados de usuários na Web para obtenção de padrões, a qual é apresentada em [MOB 96] intitulada de WEBMINER. O WEBMINER é um sistema que implementa partes de uma arquitetura geral. A arquitetura proposta para a mineração dos dados de usuários se divide em duas partes principais: A primeira parte inclui os processos dependentes do domínio e visa disponibilizar os dados da rede de uma forma satisfatória para a transação. Isto inclui pré-processamento, identificação de transação, e componentes de integração de dados. A segunda parte inclui a aplicação da mineração de dados através das técnicas para a descoberta de padrões (regras de associação, analise de padrões, etc) A arquitetura global para o processo de mineração de dados de usuários na *Web* é mostrado na Figura 15.

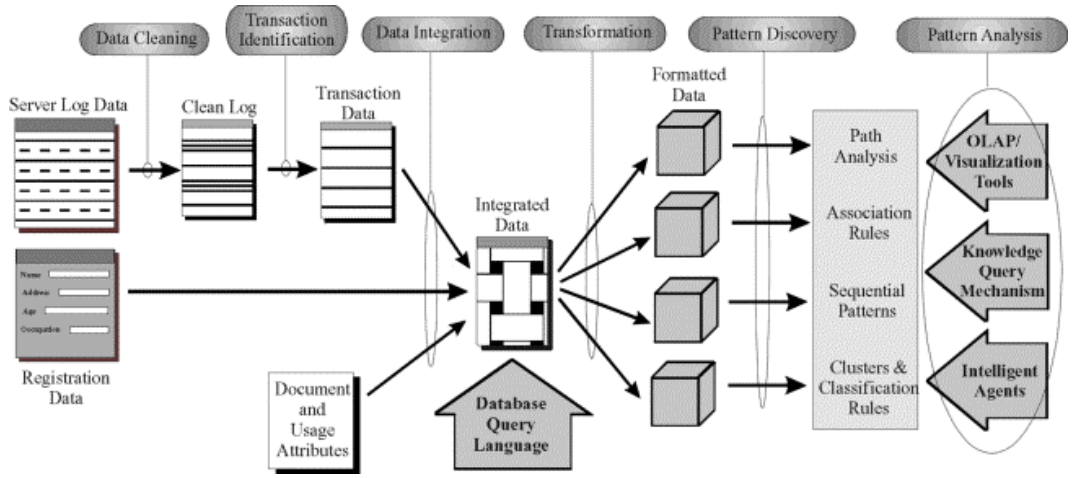

Figura 15 - Uma Arquitetura para Mineração de Dados de Uso de Rede[MOB 96]

Limpar os dados é o primeiro passo a ser executado no processo de mineração na rede. Quaisquer das técnicas de limpeza discutidas em seção anterior podem ser utilizada para isto. Atualmente, o WEBMINER usa o método simplista de verificação dos sufixos dos nomes dos arquivos. Algumas baixas tarefas de integração de dados também podem ser executadas nesta fase, como, por exemplo, combinar *logs* múltiplos.

Depois de obter os dados limpos, as entradas de *logs* devem ser divididas em agrupamentos lógicos através de uma série de módulos de identificação de transação.

O log do servidor, o qual já contém a informação limpa, ou seja, relevante, pode ser analisado de dois modos: ou como uma única transação de muitas referências de página, ou como um conjunto de muitas transações que consiste em uma única referência de página.

A meta da identificação de transação é criar agrupamentos significantes de referências para cada usuário. Estes agrupamentos podem se dar da seguinte forma: dividindo uma transação grande em transações menores ou agrupando transações pequenas originando uma maior. Estes dois processo de agrupamentos visam facilitar a tarefa de mineração de dados

Ambos os tipos de módulos de transação formados, seja pela aglutinação ou pela divisão, mantém uma lista com determinados parâmetros para controle destes dois processos, fazendo com que desta forma, o analista possa manipular os módulos de transação da forma que achar mais coerente.

Dados de *logs* de acesso podem não ser a única fonte de dados para o processo de mineração na Web. Por exemplo, a mineração de dados de inscrição de usuários em ambientes Web está sendo muito difundida atualmente, particularmente porque desta forma se obtém mais segurança e isolamento de aplicações do lado do cliente, pois consegue-se restringir o acesso do servidor a uma variedade de informações, como os IDs dos usuários, por exemplo.

Os dados obtidos na inscrição do usuário, devem ser integrados com os dados presentes no arquivo de *log* deste usuário.

Também deve-se integrar os atributos de referências de páginas em um nível mais alto, dentro de um esquema de banco de dados nivelado. Tais atributos poderiam incluir tipos de página, classificação, freqüência de acesso, informação da página, e estruturas de links.

Uma vez que a fase de transformação de dados dependentes do domínio é completada, os dados de transação resultantes devem ser formatados em um modelo de dados apropriado para a tarefa de mineração. Por exemplo, o formato dos dados para a aplicação de descoberta de regras de associação pode ser diferente do formato necessário para se aplicar a mineração através de padrões seqüenciais.

Finalmente, um mecanismo exposto permitirá ao analista prover mais controle em cima do processo de descoberta especificando várias diretrizes.

Os dados obtidos através de ferramentas e sistemas de mineração fazem com que se torne necessário a utilização de uma poderosa linguagem que se adapte as características da mineração, para que desta forma, possam ser desenvolvidas interfaces

interativas e flexíveis com o usuário. A necessidade da linguagem poder especificar um conjunto exato de dados, é uma importante diretriz a ser observada para se detectar uma boa linguagem para a mineração de dados.

O mecanismo em questão pode prover controle ao usuário em cima do processo de mineração dos dados, além de permitir ao usuário extrair apenas as regras pertinentes e úteis.

Em WEBMINER, foi implementado um mecanismo simples de manipulação para o usuário, o qual foi acrescido de algumas primitivas para uso da linguagem SQL, permitindo com isso, que o usuário especifique os padrões de interesse, ou seja, a direção que a mineração deve tomar.

Como exemplo, consideremos a situação onde o usuário está interessado nos padrões que começam a partir de uma URL A, e depois B e C nesta ordem, estes padrões podes ser expressos através da seguinte expressão regular A*B*C *. Expressão esta que será usada dentro da linguagem SQL, supõe-se também que o analista está interessado em achar todos este padrões com um percentual de confiança mínima nos dados obtidos de 90%. Além disso, assume que o analista está interessado apenas em clientes que possuam o domínio .lu, e que os dados serão analisados apenas depois de janeiro de 1999. A questão proposta acima pode ser representada da seguinte forma:

```
SELECIONE associação-rules(A*B*C *)
DE log.data
ONDE data >= 990101 E domínio = “ lu “ E confiança = 90.0
```

A informação representada da forma acima é utilizada para se reduzir o escopo, e assim o custo do processo de mineração.

4.7 Benefícios da Mineração de Dados na *Web*

Segundo Cazarini, o resultado da mudança de paradigma do comércio, dos varejos tradicionais localizados em grandes construções, para transações eletrônicas na

Internet, está influenciando dramaticamente as relações entre *e-tailers* (comerciantes eletrônicos) e seus consumidores. Não há muito contato pessoal entre varejistas e consumidores. Os consumidores, agora, são altamente móveis e demonstram lealdade somente ao valor, não importando marca ou localização geográfica. A luta principal por parte destes o *e-tailers* está em identificar e entender essa nova base de consumidores. *E-tailers* precisam aprender o mais rápido possível a reconhecer comportamentos, características individuais e preferências dos seus visitantes, em busca de tornarem-se competitivos nessa nova era de comércio, área na qual a tecnologia Web Mining vem se destacando e ganhando um espaço cada vez maior [CAZ 97].

Web Mining possibilita *e-tailers* extrairem dados online de seus clientes para entender e prever seus comportamentos. Os benefícios dos negócios suportados pelo *Web Mining* aos provedores de serviços digitais incluem personalização, filtragem colaborativa, melhora do suporte ao consumidor, definição da estratégia do produto/serviço e detecção de fraude. Em suma, a habilidade de servir às necessidades dos consumidores e apresentar o melhor e mais apropriado serviço para clientes individuais a qualquer momento.

5 TENDÊNCIAS FUTURAS EM *DATA MINING*

5.1. Tendências da Mineração de Dados em BD Tradicionais

5.1.1. Necessidades Atuais

"A informação é um dos nossos maiores tesouros...", afirma o presidente de uma das empresas mais importantes que atua no ramo de comunicação no Brasil, expondo o que não é mais novidade para ninguém, a informação é um dos bens mais preciosos de qualquer organização [INF 99]. É de fundamental importância que um administrador tenha em suas mãos o maior número possível de informações confiáveis, para tomar suas decisões da melhor forma possível.

Somente as empresas que conseguirem manipular de forma organizada, confiável e com rapidez este produto tem grandes chances de permanecerem de forma competitiva no mercado.

Acerca destas informações, os administradores desejam saber: O que querem nossos clientes? Como anda a concorrência? Que assuntos causam mais impacto na sociedade? Qual o direcionamento de nosso orçamento? - inquietações necessárias para manter sua empresa no mercado, para organizar suas informações.

Uma das soluções encontradas atualmente para colocar ordem nesta "bagunça" é a utilização do Data Warehouses, o qual permite armazenar enormes quantidades de dados, integrando-os em áreas lógicas, de assuntos de negócios, armazenando-os de forma com que fiquem acessíveis e compreensíveis as pessoas responsáveis pela tomada de decisões da empresa [FEL 99].

Porém não basta ter informação em quantidade, qualidade e organizada. O administrador necessita hoje de respostas para perguntas, tais como: Como obter o máximo de informação útil para minha empresa? Como descobrir padrões de dados e

novos conhecimentos? Como manter meu cliente? Como utilizar adequadamente e descobrir ligações entre eventos nas minhas bases de dados?

Então devido a esta necessidade, cresce a cada dia o número de empresas que aderem ao *Data Mining*, pois como já foi visto, estas ferramentas permitem prever tendências e comportamentos futuros, permitindo aos gestores tomarem decisões baseadas em fatos e não suposições. A análise automatizada e antecipada oferecida por um *Data Mining*, vai muito além da simples análise de eventos passados, que é fornecida pelas ferramentas de retrospectiva típicas de sistemas de apoio à decisão (como SQL e OLAP). [SAN 00]

5.1.2. Visão Geral

Segundo Thearling apud Edilberto Silva, , os paradigmas de negócio fizeram evoluir e mudar as questões a serem respondidas pelos Sistemas de Informações, ao longo do tempo [SIL 00]. Por exemplo:

- Data Collection (1960s) - "Qual foi o total de vendas nos últimos cinco anos?"

- Data Access (1980s) - "Quanto rendeu a nossa filial B no mês passado

- Data Warehousing & Decision Support (1990s) – “Quanto rendeu a nossa filial B no mês passado? O que isto afetou na filial C?”

- *Data Mining* (Hoje) - "O que realmente irá acontecer com as vendas na filial C no próximo mês? Por quê?"

As ferramentas de *Data Mining* podem determinar padrões de comportamento, como associação de produtos durante um processo de compras, indicando, desta forma, a presença de oportunidades e "insights" em relação àquele público consumidor. Isto se deve ao fato de que elas trabalham com descobertas

matemáticas, feitas sobre os registros corporativos já processados, e não com descobertas empíricas.

Abaixo estão alguns exemplos, de aplicações atuais de *Data Mining* que obtiveram extremo sucesso. Estes exemplos foram retirados de [INF 99].

- A rede americana Wall-Mart, pioneira no uso de *Data Mining*, descobriu ao explorar seus números que 60% das mães que compram boneca Barbie, levam também uma barra de chocolate;

- O Banco Itaú conseguiu aumentar sua taxa de retorno nas malas diretas para 30%, reduzindo a conta do correio a um quinto. Armazenando e analisando a movimentação financeira de seus 3 milhões de correntistas nos últimos 18 meses .

- A Sprint, um dos líderes no mercado de americano de telefone de longa distância, desenvolveu, com a análise de *Data Mining* em seu Data Warehouse, um método capaz de prever com 61% de segurança se um consumidor trocaria de companhia telefônica dentro de um período de dois meses. Com um marketing agressivo, conseguiu evitar a deserção de 120000 clientes e uma perda de 35 milhões de dólares em faturamento.

- O governo de Massachusetts, no Estados Unidos, compilava informações financeiras imprimindo telas e mais telas nos terminais de grande porte. Só com papel foram economizados 250000 dólares por ano. Além de permitir que o orçamento anual fosse assinado e analisado, antes do início do ano fiscal.

- Gurovitz cita também o exemplo da SERPRO no Brasil, que implantando o seu Data Warehouse e *Data Mining*, já consegue hoje cruzar e analisar informações em cinco minutos, o que antes demandavam quinze dias de trabalho [GUR 00].

Por isso e por outros inúmeros exemplos o *Data Mining* cresce em importância, principalmente nas empresas preocupadas com o futuro do negócio.

5.1.3. Técnicas Empregadas

Segundo estudo realizado por Edilberto Silva, a tecnologia atual de mineração se detêm na utilização das técnicas relacionadas abaixo, as quais tendem a serem utilizadas por muito tempo, tendo em vista os resultados obtidos através delas[SIL 00].

- **Estatística:** Será a base de todas as outras tecnologias criadas para *Data Mining*. Conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análises de discriminantes e intervalos de confiança, serão cada vez mais utilizados para realizar as pesquisas nos dados, bem como analisar e descobrir relacionamentos entre os mesmos.

- **Inteligência Artificial:** Tentar imitar a maneira de pensar do ser humano através da IA, com embasamento estatístico, tentar recuperar informações que podem auxiliar na descoberta de novas interações entre os dados será uma grande tendência nesta área. Para tal, esta técnica heurística exigirá mais processamento, resultando em um maior aproveitamento nos SGBDs, e em seus módulos de otimização de consulta.

- **Estatística & Inteligência Artificial:** Aplicação da inteligência artificial combinada a estatística, através de heurísticas e análises estatísticas. Sua função será fazer com que os dados sejam "aprendidos" pelos programas, tornando-os capaz de tomarem decisões através de "pensamentos", que se fazem mais "inteligentes" a cada vez que são acrescentados mais heurística ao seu processamento.

5.1.4. Algoritmos Utilizados

Já em relação aos algoritmos, os que tendem a serem cada vez mais utilizados nas ferramentas de mineração de dados, segundo Thearling, são [THE 00]:

- **Árvores de Decisão:** As árvores de decisão são uma evolução da técnica de *machine learning*, e continuarão sendo muito utilizadas, tendo em vista que os valores obtidos através dela são prognósticos chaves ou fatores explicativos, os quais são de grande utilidade. Além disso uma boa ferramenta baseada em Árvore de Decisão

permite que o usuário explore a árvore de acordo com a sua vontade, do mesmo modo que ele poderá encontrar grupos alvos que lhe interessem mais, e aí ampliar o dado exato associado ao seu grupo alvo.

Exemplos de atuais ferramentas: Alice d'Isoft, DataMind, Angoss Knowledge Seeker..

- **Regras de Indução:** Também bastante utilizada, as Regras de Indução constituem-se uma boa técnica pelo fato de ela poder detectar tendências dentro de grupos de dados, ou de "regras" sobre o dado, com pouca interação do usuário; O analista de negócio neste caso tem que fazer o ranking das regras, por importância, determinando aquelas que são mais importantes no modelo de Data Minig, e para os assuntos de negócios envolvidos.

Exemplos de ferramentas: IDIS, Knowledge Seeker;

- **Redes Neurais:** As redes neurais tendem a serem muito estudadas, pois oferecem o mais profundo poder de mineração de dados, porém é a técnica mais difícil de se entender. As apresentações internas, de como os padrões e os modelos foram achados nos dados, não são mostrados. Estas "caixas pretas" são inexplicáveis, o que faz desta técnica inaceitável para alguns analistas de negócios. É uma técnica excelente, muito vantajosa, mas que tem dois aspectos negativos: o primeiro é a complexidade dos modelos gerados. O segundo é o problema de exigir que o formato dos dados seja imutável; para diferentes representações dos dados surgem resultados diferentes.

Exemplos de ferramentas: SPSS Neural Connection, IBM Neural Network Utility, NeuralWare NeuralWorks Predict

- ***Market Basket Analysis:*** Devido a explosão do *e-commerce* esta técnica de localização que permite através de uma "cesta de compra", inferir regras a partir de grupos de itens ou produtos que surgem nesta venda, tem ganhado bastante importância. A análise combinatória dos produtos juntamente com a localização geográfica do cliente, permite promover investimentos em marketing específicos para determinados

produtos e regiões. Para vendas não anônimas é possível ainda promover investimentos em marketing para determinados perfis de clientes, focalizando seus traços de compra, garantindo maior retorno no investimento.

- **Algoritmos Genéticos:** Estas técnicas, por serem partes complementares das técnicas de MBR e Redes Neurais, também são bastante utilizadas.

Exemplos de ferramentas: S+, SAS, SPSS.

- ***Memory-Based Reasoning* (MBR):** Usa relações já conhecidas como "*templates*", para inferir e fazer previsões de relações a serem descobertas. A confiabilidade da opção dependerá da distância a que estiverem os "alvos vizinhos". Está ganhando grande aceitação por apresentar a vantagem de poder ser utilizada com inúmeros tipos de dados. Permite ainda "aprender" novas classificações apenas com a introdução de novas instâncias na base de dados.

- ***On-Line Analytical Processing* (OLAP):** a técnica de OLAP continua a ser utilizada também na área de mineração, mesmo não sendo uma técnica de *Data Mining*, pois ela é excelente como ferramenta para a visualização da informação das bases de dados relacionais, de um modo agradável para o utilizador e também serve para a pesquisa de padrões importantes na informação, além do que é uma técnica mais antiga e com isso mais estudada e conhecida.

5.1.5. Áreas de Uso

Segundo Edilberto Silva as três grandes áreas onde o *Data Mining* se destaca e ganham maior aceitação, são [SIL 00]:

- ***Cross-selling*:** Identificar associação entre produtos, procurando tirar proveito dessas associações.
- ***Up-selling*:** Identificar entre os clientes atuais aqueles que têm potencial para adquirir produtos que proporcionem maior lucro para a companhia. "Através da definição de um perfil de usuário, que compra um produto, podemos determinar que

outros usuários, de mesmo perfil e, que ainda não tem este produto, venham a comprá-lo";

- **Fidelização:** Combinar e descobrir fatores que causem a perda de clientes, e como forma econômica, manter estes clientes, o que é mais barato que conquistar novos.

A partir da descrição, feita por Edilberto Silva, de alguns exemplos de mineração de dados, em determinadas áreas de aplicação, podemos notar o poder de um *Data Mining*, e como elas podem determinar a melhor escolha. Nota-se que o *Data Mining*, apesar de ser uma tecnologia recente, está presente no nosso dia-a-dia através das seguintes áreas [SIL 00]:

Vendas (Marketing)

- Com análise de compra de seus clientes através dos seus cartões, a American Express oferece promoções e propagandas particulares, para seus clientes.

Saúde

- O Hospital de Problemas de Coluna Vertebral do Sul da Califórnia em Los Angeles pesquisa através de seu *Data Mining*, o IDIS da IntelligenceWare, Inc., fatores sutis que afetem o sucesso ou o insucesso das intervenções cirúrgicas à coluna vertebral.

Finanças

O foco principal de utilização de *Data Mining* na área financeira, é a prevenção e detecção de fraudes, isto pode ser verificado pelo exemplo a seguir [NOV 98].

- Security Pacific/Bank of America utiliza o *Data Mining* no suporte à decisão na área dos empréstimos bancários no intuito de prevenir a fraude.

5.1.6 Propostas de Uso de *Data Mining*

Nesta seção são propostas as seguintes aplicações para a tecnologia de *Data Mining* [SIL 00]:

- Com a devida organização dos dados o *Data Mining* pode ser capaz de analisar informações de regiões / estados, nas mais diferentes instâncias: cultura, distribuição de renda, C&T, senso demográfico, etc. Tomando-se como base que o país seria como uma grande empresa, poderíamos ter um Data Mart estadual e através da identificação de relações entre estes e o *Data Mining*, seria possível agregar e planejar estratégias de política comuns, aprender com erros ocorridos em uma área e aplicar soluções corretas de um estado para outro, além de poder auxiliar o controle e liberação de gastos públicos.

- Utilização de *Data Mining* na tomada de decisões de governantes. Em alguns países como EUA, França, Alemanha estas ferramentas já estão presentes em algumas decisões, tomadas pelos seus governantes, porém ainda não há registro do uso efetivo de *Data Mining* no governo brasileiro.

5.1.7. Dificuldades para uso de DM´s

Existem alguns empecilhos, segundo Edilberto Silva para o uso de *Data Mining*, o mais significativo hoje ainda é o alto custo, porém vários fatores podem impedir sua implantação, dentre eles destacam-se [SIL 00]:

Organização

De nada adianta investir em *Data Mining* se os dados da empresa estiverem espalhados, redundantes ou não confiáveis. Geralmente a implantação de Data Warehouse precede a implantação de ferramentas de *Data Mining*, pois possibilitam agregar informações espalhadas de diferentes estruturas, formas e plataformas. Para tanto, a definição e organização destes dados devem ser minuciosas, uma vez que tendo-se “lixo” como base de procura, certamente teremos “lixo“ como resposta.

Alto Custo

Este é sem dúvida um dos principais motivos de desistência na implantação de *Data Mining*. Na maioria das vezes, um projeto com um custo elevado por usuário, trás, a priori, resultados pouco significativos. Alguns fornecedores dessas ferramentas têm introduzido produtos com custo mais baixo, mesmo assim o preço continua limitando o uso em larga escala. A tendência é que estes preços caiam, devido o amadurecimento e divulgação da tecnologia.

Grandes Bases de Dados

Ainda é notório, segundo os fabricantes de ferramentas de *Data Mining*, ter um ambiente com terabytes de dados e poderosos servidores, como ambiente ideal de uso de um *Data Mining*, porém já têm aparecido no mercado aplicações satisfatórias de *Data Mining* em bases de dados menores.

Novidade e Complexidade

Mesmo com o surgimento de uma nova geração de ferramentas, para mineração em moderados grupos de dados, uma terceira barreira ainda permanece: A grande maioria das ferramentas ainda continua incompreensível para os usuários comuns. De fato, muitas ferramentas ainda fazem o seu trabalho em uma "caixa-preta", não permitindo que se saiba como alcançaram os seus resultados, porém estão surgindo, através de esforços de pesquisa nesta área, soluções mais compreensíveis e a tendência é que se aproximem ainda mais dos usuários.

Montar um Ambiente Ideal

Estruturar um ambiente ideal, é uma tarefa árdua, devido aos anos e anos de "desorganização" na forma de guardar as informações nas instituições. O ambiente ideal deve conter dados relevantes para as necessidades dos usuários, "limpos" (livres de erros lógicos ou de entrada de dados), consistentes, e livres de excessivas nulidades. [TEC 99]

5.2 Tendências da Mineração de Dados na *Web*

5.2.1 Tendências em *Web Usage Mining*

Segundo Cooley et. al. a tendência ao uso da mineração de dados na Web aumentar cada vez mais é enorme, tendo em vista sua grande aplicação e desempenho na área de marketing, área esta que cresce e tende a ir cada vez mais direcionada a área da Web, através do comércio eletrônico [CMJ 95].

O *web usage mining* tem um papel considerável nesta área de marketing, pois pode atuar através da personalização do conteúdo, propagandas direcionadas e construção de perfis. Além disso pode coletar informações de mercado e dados de sites que se apresentam como potenciais para marketing direto. Também através do *web usage mining*, mensagens personalizadas podem ser enviadas a pessoas individualmente, fazendo com que, desta forma, ofertas atrativas e promoções de vendas sejam exploradas com muita facilidade.

Estas facilidades fazem com que profissionais de marketing possam direcionar suas estratégias de vendas, construindo perfis de possíveis compradores e utilizá-los na busca de segmentos onde as atividades de marketing deverão ser focadas. Pela primeira vez, *e-tailers* (comerciantes eletrônicos) têm capacidade para considerar formas custo-efetivas de prover comunicação de marketing individualizadas, nas busca de alguma lealdade de seus visitantes [CAZ 97].

Atualmente o uso de *Data Mining* na Web está ganhando maior importância na área relacionada à mineração de conteúdo de uso da Web, pois esta é uma tecnologia nova que está tendo uma enorme aceitação neste ambiente, principalmente pela necessidade exposta pelas organizações de utilizarem os dados que são minerados neste processo.

Segundo Cooley et al, a tecnologia de mineração de dados na Web é uma área nova e promissora e, por esse motivo, ela apresenta uma gama enorme de assuntos em aberto para serem discutidos e pesquisados, tendo isto em vista, o autor propõe

alguns tópicos como tendências futuras a serem estudadas para esta área de mineração na Web [CMJ 95]:

1) Pré - Processamento dos Dados

Dados de usuários na Web são agrupados de várias maneiras, tendo em vista que deve-se escolher aqueles que sejam pertinentes para determinado propósito. Há então, a necessidade de se aplicar um pré-processamento nestes dados a fim de facilitar a tarefa de mineração dos mesmos.

Levando-se em consideração o exposto acima, acredita-se que os seguintes tópicos devem ser estudados:

- **Instrumentação e Conjunto de Dados:** Quando se consegue melhor a qualidade do conjunto de dados a ser analisado, se consegue também uma melhoria significativa nas fazes seguintes do processo de mineração. Entre as principais dificuldades encontradas pelos analistas no processo de mineração estão a dificuldade em se fazer uma boa análise em domínios de rede, onde os analistas necessitam de conjuntos de dados mais detalhados, e o isolamento preciso de usuários, que requer pequenos conjuntos de dados.

Este problemas tem sido influenciados pelo desenvolvimento de arquivos de *cookies* e de *caches*. Colecionar dados de perfil pode ser um alternativa para tais problemas, porém não se sabe se realmente será uma boa técnica para isto. Consequentemente, haverá uma grande necessidade de se desenvolver uma melhor instrumentação e técnicas de formação de conjuntos de dados.

- **Integração de dados:** Existem diversos tipos de dados de usuários na *Web*, *Web Server Logs*, *ReferralLog*s, *Index Server Logs*, etc. Se fazer uma integração e correlação inteligente deste tipos de dados podem revelar informações de uso da *Web* muito interessantes, as quais não se pode evidenciar através de uma análise individual sobre estes tipos de dados.

- **Identificação de transação:** Dados de usuários da *Web* estão aglomerados em vários *logs*, os quais são formados pelas mesmas características de dados. Então, existe a vantagem destes dados serem analisados de forma conjunta e de forma mais detalhada, porém há também a desvantagem de não se poder analisar diretamente estes arquivos sem antes separar estes dados em conjuntos pertinentes à análise, pois a análise pode ser feita a partir de micro tendências ou das macro tendências, logo precisando ser efetuada sobre micro ou macro conjuntos de dados. Existe grande dificuldade na escolha do conjunto de dados a serem analisados, ou seja, quando é mais adequado usar macro ou micro conjuntos de dados.

Tendo em vista o exposto acima, é considerado que haja a necessidade de se aplicar cada vez mais a Transações de Rede, as quais visam agrupar itens significativos, que geram tendências, em grupos, antes de passar os dados para o sistema de mineração.

2) O Processo de Mineração

O componente principal do *Web Mining* é a mineração propriamente dita. Como se discutiu neste trabalho, *Web Mining* adaptou técnicas utilizadas no campo da mineração de dados, bancos de dados, recuperação de informação, etc, como também desenvolvendo algumas técnicas próprias, como por exemplo análise de caminhos na *Web*, porém ainda existe muito trabalho a ser feito nesta área, se adaptando técnicas já conhecidas em processos de mineração, como também desenvolvendo novas técnicas. Cooley et. al. salienta que é muito importante que se leve em consideração os seguintes aspectos [CMJ 95]:

- **Novos Tipos de Conhecimento:** *Web Usage Mining* visa a mineração de dados para busca de informações relevantes, e faz isso através de regras de associação, sucessões temporais, agrupamentos e expressões de caminho. Devido ao fato da *Web* estar continuamente se expandido, surge a necessidade de se entender novos tipos de conhecimento sobre comportamento de usuário que precisam ser minerados.

- **Melhorias nos Algoritmos de Mineração:** a qualidade de um algoritmo de mineração pode ser medida em termos da efetividade da mineração e de sua eficiência

computacional. Porém deve-se considerar que sempre haverá a necessidade de se melhorar os algoritmos de mineração, levando-se em consideração os aspectos citados.

- **Atualização do *Web Mining*:** Naturalmente, conjuntos de dados de uso da *Web* são incrementados diariamente. Consequentemente, há uma necessidade de se desenvolver algoritmos de mineração que suportem tal crescimento dos conjuntos de dados, conseguindo manter sua eficiência.

- **Mineração de Dados Distribuída:** naturalmente, conjunto de dados de usuários na Web estão organizados de forma distribuída. É lógico que , se todos os dados presentes nos *logs* fossem integrados, se conseguiria obter ótimas informações através da mineração, mas tal integração é impraticável de se realizar, tendo em vista a grande quantidade de dados existentes na Web. Consequentemente, se faz necessário um aprimoramento dos conhecimentos de mineração, levando-se em conta esta distribuição, para que se consiga integrar uma maior quantidade de arquivos *log* e com isso se formar um modelo mais incluso.

3) Análise de Conhecimento Minerado

Os algoritmos para mineração de conhecimento atuais não são de fácil entendimento para os seres humanos e, consequentemente há uma necessidade de se desenvolver técnicas e ferramentas que ajudem a compreensão por parte do analista dos dados obtidos. A seguir serão citados alguns assuntos que devem ser considerados em relação a esta área [CMJ 95]:

- **Ferramentas de Análise de Uso da *Web*:** Há uma necessidade de se desenvolver ferramentas que incorporam métodos estatísticos, visualização, e fatores humanos, para facilitar o entendimento do conhecimento minerado. É inevitável a pesquisa por ferramentas inteligentes que possam ajudar na interpretação do conhecimento minerado.

Ferramentas estas que precisam ter conhecimento específico sobre o domínio do problema, para que então possa se fazer a filtragem baseada em atributos

estatísticos, em regras de descobertas ou padrões. O processo *Web Mining*, poderia se deter mais no uso de agentes inteligentes baseados em padrões de acesso descobertos, em topologias de redes, em heurísticas que modelem o comportamento do usuário etc,

Como um exemplo simples, consideremos um agente que (entre outras coisas) verifica a relação entre a freqüência de visitas de uma página particular e o número de caminhos de usuário que terminam naquela página. Esta relação poderia ser usada para determinar se a página visitada constitui um único ponto de entrada ou alguns poucos pontos. Isto pode sugerir que outros links para a referida página devessem ser colocados, visando um maior número de acessos.

Apesar de ser uma tecnologia nova, a qual precisa ainda ser bastante aprimorada e estudada, acredita-se muito no potencial da mineração de dados de usuários no ambiente *Web*, e grande número de organizações já migraram para esta nova etapa da mineração de dados, tendo em vista os ótimos resultados, relacionados ao lucro, que esta técnica pode trazer.

5.2.2 Tendências em *Web Content Mining*

As técnicas de Web Content Mining, com a utilização de tecnologias bem conhecidas, como banco de dados, agentes inteligentes, etc, que estão sendo aplicadas aliviam brutalmente a dificuldade em se trabalhar na recuperação da informação porém esta técnica ainda encontra grandes dificuldades de implementação. Talvez o maior problema das técnicas de extração de informação, seja a especificidade dos sistemas e ferramentas gerados. Muitos só podem ser aplicados nos casos para os quais foram criados (problemas jurídicos, textos jornalísticos, etc) ou em um domínio específico [LOH 97].

Segundo Jin Cowie apud Loh, nestes casos, há necessidade de muito trabalho e esforço manual para se conhecer conteúdos ou estruturas, pois algum ser humano deve analisar antes os padrões dos textos. Portanto, é necessário que estes textos sejam previsíveis, ou seja, possuam padrões coletáveis, e isto ainda pode trazer problemas devido às falhas possíveis numa inspeção manual [LOH 97].

Por outro lado, as técnicas que tentam ser genéricas falham por excesso. Ou seja, são vagas demais e acabam por não poderem ser aplicadas para casos específicos porque não consideram as diferenças de contexto.

Outro problema sério que também pode afetar os processos de Extração de Informações é a imprecisão dos termos utilizados nos textos, bem como nos processos de Recuperação de Informações.

Os processos de interpretação e significação na linguagem natural são sempre falhos [LOH 97]. Jin Cowie apud Loh comenta que "a escolha de palavras, frases e estruturas de sentenças afeta a precisão do significado", portanto dificultando a extração de informações [LOH 97].

Segundo Simon Parsons apud Loh, além disto tudo, há os problemas com o tratamento de informações imperfeitas, ele afirma que, para decidir que informações recuperar e extrair, é preciso modelar as informações como elas aparecem em seu estado natural, no mundo real [LOH97].

Segundo o mesmo autor, há 5 tipos de informações imperfeitas:

- **informações incompletas:** quando faltam valores para determinados atributos;
- **informações imprecisas:** quando há variações na granularidade dos valores (algumas vezes, os valores não podem ser medidos com precisão); por exemplo, o momento no tempo em que ocorreu um evento pode ser medido em anos, meses, dias, turnos, horas, minutos, etc;

- **informações incertas:** ocorre porque as decisões são tomadas com base em fatos que não se sabe se são verdade ou não, devido às contradições nas fontes da informação ou às representações falhas;

- **informações vagas:** decorrentes da imprecisão de termos;

- **informações inconsistentes:** quando há valores conflitantes.

O autor ainda sugere algumas técnicas para tratamento das informações imperfeitas, como mostrado a seguir:

O uso de lógica difusa (*fuzzy logic*) ou de conjuntos difusos (*fuzzy sets*) permitem trabalhar com variações de critérios de pertinência. Assim, pode-se classificar um elemento em conjuntos distintos, mas com graus de pertinência diferentes. Por exemplo, um elemento pode ser "bom" com grau X e ser "ruim" com grau Y. Noutro exemplo, pode-se interpretar que "18 horas" é um horário da noite e da tarde ao mesmo tempo.

Outra técnica sugerida é a lógica probabilística associada com sistemas de crença, para que os fatos possam ser avaliados quanto à sua certeza ou não [LOH 97].

Além destes problemas todos, outro que recebe menos atenção (mas não é menos importante) é a falha nos processos de digitação dos textos. Muitas palavras são incluídas em textos com a grafia errada. Antes que se possa utilizar técnicas de recuperação e extração de informações, será necessário corrigir tais palavras.

5.3. Fornecedores de Banco de Dados na Trilha da *Web*

Nesta seção, são citados alguns exemplos, baseados em Manzonini, de empresas fornecedoras de Banco de Dados que mudaram seus rumos em direção a *Web*, pelo fato da Internet estar indo cada vez mais para o pelotão de frente, a fim de atacar o mercado corporativo.

São citados aqui alguns exemplos: em julho, em seu tradicional evento de usuários nos Estados Unidos, a Informix lançou o Informix Internet Foundation.2000, seu banco de dados pronto para a Web. No final de agosto, a Sybase apresentará o EIP (Enterprise Information Portal). A IBM acaba de lançar a versão 6.1 do DB2 com todos os pré-requisitos para atuar no universo Web. A Progress aposta numa série de

ferramentas, integradas ao BD, para atacar esse novo mundo. E a Oracle, líder do mercado, joga suas fichas no Oracle 8i. [MAN 99]. Abaixo estão algumas características dos sistemas de banco de dados de cada um dos fabricantes mencionados:

- **IBM:** o banco de dados vem com uma série de funcionalidades embutidas, como suporte a Java, JDBC, linguagem Perl e integração com os principais servidores Internet do mercado.

- **Informix:** acaba de lançar o Informix Internet Foundation.2000, BD voltado para a Web. Ele será o alicerce para rodar as aplicações I-reach e I-sell. Foco em soluções.

- **Oracle**: o Oracle 8i é a estratégia da empresa para o universo Web. Faz parte de um plano de transformar a companhia numa provedora de soluções para o segmento de e-business

- **Progress:** com suas ferramentas, como o Webspeed e o Apptivity, quer ajudar as software houses a portar suas soluções para a Web

- **Sybase:** a estratégia Web está sendo delineada. Em agosto, apresenta o EIP (Enterprise Information Portal). Em outubro, chega seu BD pronto para a Web.

5.4. Produtos *Data Mining*

Existem inúmeras ferramentas para *Data Mining*, bem como diversos fornecedores e fabricantes. As principais ferramentas e respectivos fabricantes que estão atuando no mercado são listados abaixo na tabela 5, segundo o catálogo atualizado em 27/03/2000 pela *Systens Incorporated* (http://www.lgisystems.com), separados por categoria.

{PRIVATE}Case-Based Reasoning	
CBR-Works 4	TECINNO - http://www.tecinno.de
Emerald Empower	Emerald Intelligence - http://www.emeraldi.com
Induce-It	Inductive Solutions, Inc. - http://www.inductive.com
KATE-CBR	AcknoSoft - http://www.acknosoft.fr
ReCall	Alice – http://www.alice-soft.com
The Easy Reasoner	The Haley Enterprise, Inc. - http://www.haley.com
Data Visualization	
Alterian Nucleus	Alterian Limited - http://www.alterian.com
Axum	Mathsoft - http://www.mathsoft.com
ChartWorks Server	Visual Mining - http://www.visualmining.com
Cognos Visualizer	Cognos Corporation - http://www.cognos.com
Coronado	Portola Dimensional Systems - http://www.portolasystems.com
CrossGraphs	Belmont Research Incorporated - http://www.belmont.com
D.B. Express	Computer Concepts Corp. - http://www.computerconcepts.com
Data Desk	Data Description - http://www.datadesk.com
Data Visualization Workstation	Computer Science Innovations, Inc. - http://picard.csihq.com/CSI
DataMiner 3D	Dimension 5 - http://www.dimension5.sk
DataScope	Mindmaker - http://www.mindmaker.com
DataVista	VI/Visualize, Inc. - http://www.visualizetech.com
Drill Down	Vasanth Desai - http://members.tripod.com/~vpdesai
Facet	Facet Decision Systems, Inc. – http://www.facetps.com
Gific	LMI – http://www.gific.com
Heatmaps	NeoVision Hypersystems - http://www.neovision.com
IDL	Research Systems, Inc. - http://www.rsinc.com
Iglass	i2 Inc – http://www.i2group.com
Illuminator	Visualize, Inc. - http://www.visualizetech.com
In3D	Visible Decisions, Inc. - http://www.vdi.com
Jwave	Visual Numerics, Inc. - http://www.vni.com
MAX	Maximal Innovative Intelligence - http://www.maxsw.com
Metaphor Mixer	Maxus Systems - http://www.maxussystems.com
ModelSheet	iWorker - http://www.iworker.com
Nested Vision3D	NVision Software Systems - http://www.nvss.nb.ca
NetMap	Alta Analytics - http://www.altaanalytics.com
OpenViz	Advanced Visual Systems - http://www.avs.com
Parallel Visual Explorer	IBM – http://www.ibm.com
Portal	Electrohouse - http://em7.com
SAS System	SAS Institute - http://www.sas.com
SOMine	Eudaptics Software GmbH - http://www.eudaptics.co.at
Spotfire	Spotfire AB - http://www.spotfire.com
Table Lens	Inxight Software - http://www.inxight.com
TempleMVV	Mihalisin Associates, Inc. - http://www.corptech.com/

TowerView	High Tower Software - http://www.high-tower.com
Visual Insights	Lucent Technologies - http://www.visualinsights.com
VisuaLinks	Visual Analytics - http://www.visualanalytics.com
Visualization Data Explorer	IBM – http://eagle.almaden.ibm.com/dx
VisualMine	Artificial Intelligence Software SpA - http://www.visualmine.com
VRCharts	AlterVue System http://www.vrcharts.com
VRDev	EM7 – http://em7.com
Fuzzy Query and Analysis	
CubiCalc	HyperLogic Corporation - http://www.electriciti.com/hl/cbc.html
FuziCalc	FuziWare, Inc. - http://www.corptech.com
Fuzzy Data Explorer	Metus Systems - http://www.metus.com
FuzzyDecisionDesk	Fuzzy Logik Systeme GmbH - http://www.fuzzy.de
FuzzyQuery	Sonalysts - http://www.sonalysts.com/fuzzy.htm
FuzzyTECH for Business	Inform Software - http://www.fuzzytech.com
Z Search	Zaptron - http://www.zaptron.com
Knowlege Discovery	
Aira	Hycones IT - http://www.hycones.com.br
ALICE d'Isoft	Alice – http://www.alice-soft.com
AnswerTree	SPSS - http://www.spss.com
AT Sigma Data Chopper	Advanced Technologies - http://www.atsigma.com
Athena Knowledge Server	Triada – http://www.triada.com
Atlas	Sofresud - http://www.sofresud.com
BusinessMiner	Business Objects, Inc. - http://www.businessobjects.com
Capri	MineIT Software Ltd - http://www.mineit.com
CCM Data Correlation Model	Applied Technical Systems - http://www.apptechsys.com
C5.0	RuleQuest Research - http://www.rulequest.com
Clementine	SPSS – http://www.spss.com
Clustan	Clustan Ltd. - http://www.clustan.com
Data Surveyor	Data Distilleries - http://www.ddi.nl
DataBase Mining Marksman	HNC Software Inc. - http://www.hncs.com
DataDetective	Sentient Machine Research - http://www.smr.nl
DataEngine	MIT GmbH - http://www.mitgmbh.de
DataLogic/R	REDUCT & Lobbe Technolgies Inc. - http://ourworld.compuserve.com/homepages/reduct
DataMite	Logic Programming Associates – http://www.lpa.co.uk
DataSMARTS	Cognition Technologies Corporation
DataX	Zaptron - http://www.zaptron.com
DBMiner	DBMiner Technology - http://www.dbminer.com
Decision Force	Pro-Action - http://www.proaction.gr
Decision List Learner	Schenley Park Research - http://www.schenley.com
Decision Series	Accrue Software - http://www.accrue.com
Discoverer	TriVida - http://www.trivida.com

Discovery Server	Pilot Software, Inc. - http://www.pilotsw.com
Dowser	Redshed Software - http://www.redshed.tsx.org
Enterprise Miner	SAS – http://www.sas.com
ESX	Information Acumen - http://www.infoacumen.com
Galvano	PMSI - http://ctw.citeweb.net/desactive/index.htm?pmsi
Heatseeker	WhiteCross Systems - http://www.whitecross.com
Intelligent Miner	IBM – http://www1.ibmlink.ibm.com
K.wiz	K.wiz Solutions - http://www.kwiz-solutions.com
KDD Explorer	SRA – http://www.knowledgediscovery.com
Kepler	Dialogis - http://www.dialogis.de
Knowledge Access Suite	Information Discovery - http://www.datamining.com
KnowledgeMiner	Script Software - http://www.scriptsoftware.com/km
KnowledgeSTUDIO	ANGOSS Software International - http://www.angoss.com
Mineset	Silicon Graphics - http://www.sgi.com/Products/software/MineSet
ModelQuest Enterprise	AbTech Corporation - http://www.abtech.com
Ngram Transform-DB	Triada - http://www.triada.com
Nuggets	*Data Mining* Technologies, Inc. – http://www.data-mine.com
ODBCMINE	Intelligent Systems Research - http://www.xmlmethods.com
Omega	KiQ – http://www.kiq.com
Oracle Darwin	Oracle - http://www.oracle.com
Orchestrate	Torrent Systems Inc. - http://www.torrent.com
Pattern	Magnify, Inc. - http://www.magnify.com
PV/FutureView	Continuum Software, Inc. - http://www.continuumsi.com
S-Plus	MathSoft - http://www.mathsoft.com
Scenario	Cognos - http://www.cognos.com
SmartMiner	GRIMMER Logiciels - http://www.grimmersoft.com
SphinxVision	ASOC AG - http://www.asoc.de
SuperQuery	Azmy Thinkware Inc. - http://www.azmy.com
Syllogic *Data Mining* Tool	Syllogic - http://www.syllogic.nl
Synera	Synera - http://www.syneracorp.com
TeraMiner Stats	NCR – http://www.ncr.com
ThinkBase	Science in Finance - http://www.thinkbase.com
To-The-Point	Bennely Software - http://www.topsoft.com/main
VisiRex	CorMac Technologies - http://www.cormactech.com
WizWhy	WizSoft, Inc. - http://www.wizsoft.com
Xpert Rule Analyser, Profiler	Attar Software - http://www.attar.com
Zoom 'n View	SkyGate Development ApS - http://www.skygate.dk
Neural Networks	
BackPack	Z-Solutions - http://www.zsolutions.com
Braincel	Promised Land Technologies – http://promland.com
BrainMaker	California Scientific Software – http://www.calsci.com

BrainSheet	Bitstar International - http://www.bitstar.com
Easynet for Windows	Renascence Partners Ltd. - http://www.rpl.com
4Thought	Cognos - http://www.cognos.com
NeuFrame/NeuroFuzzy	NCS – http://www.ncs.co.uk
Neural Connection	SPSS Inc. - http://www.spss.com
Neural Network Browser	Apreal, Inc. - http://members.aol.com/apreal
Neural Technologies	Neural Technologies - http://www.neuralt.com
NeuralNet Tutor	Advanced Technology Transfer Group - http://www.attg.com
Neuralyst For Excel	Cheshire Engineering Corporation - http://www.cheshireeng.com
NeuroGenetic Optimizer	BioComp Systems - http://www.biocompsystems.com
NeuroShell	Ward Systems Group, Inc - http://www.wardsystems.com
NeuroSolutions	NeuroDimension, Inc. - http://www.nd.com
NNB	Apreal - http://members.aol.com/apreal/NNB
Partek	Partek Incorporated - http://www.partek.com
Pattern Recognition Workbench	Unica - http://www.unica-usa.com
SENN	Siemens Nixdorf - http://www.sni.de
Simulator	Matrica - http://www.matrica.org
Synes	Synes - http://www.synes.com [DWINFOCENTER, 2000]

Tabela 5 – Produtos Data Mining e suas Respectivas Áreas

6 Conclusões

É indiscutível o crescente avanço das ferramentas de Data Mining. Elas estão se tornando cada vez mais importantes no mundo dos negócios, tendo em vista que são "inteligentes" o bastante para adquirirem, com enorme precisão e eficiência, relações entre informações, que para nós, seriam muito mais complexas e onerosas.

Sua solidificação é certa, pois o número de informações que são manipuladas pelas organizações cresce muito a cada dia. Com isto ocorre um grande acúmulo de dados nos bancos de dados da empresas, os quais necessitam de mineração. O Data Mining é uma nova forma de explorar estes dados, ajudando a própria empresa a manter-se no mercado.

Não devemos, porém, tratar o Data Mining como a solução para uma empresa evoluir, ou como a solução para "acabar com as concorrentes". Antes de sair correndo atrás desta nova tecnologia para implantá-la, é necessário organizar bem os dados da empresa, planejar bem o escopo do negócio a ser estudado, além de dados confiáveis. Em muitos casos, para solucionar este problema, deve-se primeiro aplicar o Data Warehouse.

Deve-se levar em conta também a relação custo/benefício da implantação, em função de ser uma tecnologia de efeitos progressivos, isto é, os resultados são vistos pouco a pouco.

Também devemos levar em consideração o fator humano para o sucesso de um solução de Data Mining, apesar de ser uma tecnologia que atua praticamente de forma automatizada, a decisão final sempre caberá a pessoa envolvida no negócio, pois de nada adianta ter um ambiente com computadores de última geração espalhados pelas mesas dos gerentes da empresa, interligados a um Data Warehouse, sem que os mesmos consigam extrair todo o potencial e riqueza de informações.

Dentro do ambiente Web podemos notar também, o grande crescimento da aplicação de técnicas Data Mining, pois a popularidade da *Web* continua aumentando, e

com isso cresce a necessidade em se desenvolver ferramentas e técnicas que ajudarão a melhorar sua utilidade global. Desde que uma das metas principais da *Web* é agir como um recurso de informação distribuída pelo mundo inteiro, vários esforços são somados para o desenvolvimento de técnicas que facilitem esta distribuição de informação.

O termo *Web* Mining foi usado para se referir a tipos diferentes de técnicas que abrangem uma variedade de assuntos, porém a tecnologia denominada *Web* Mining, por ser extremamente recente, ainda possui um vocabulário pouco específico, o que gera confusão por parte das pessoas. Há, então, uma necessidade de se desenvolver um vocabulário comum para todos estes esforços.

O objetivo deste trabalho foi apresentar a tecnologia de Data Mining na busca de conhecimento na *Web*. O processo de mineração de dados vem se destacando muito nas áreas em que vem sendo empregada, devido a grande necessidade existente de se analisar dados que estão dispostos, algumas vezes, de forma desorganizada ou de forma implícita dentro de um Banco de Dados, e que, se extraídos de forma correta, serão de grande utilidade. Através deste trabalho foi estudado a tecnologia de mineração de dados, visando sua exposição desde sua área inicial de atuação, ou seja, aplicada a banco de dados tradicionais de organizações até sua atual utilização na *Web*, que originou-se a partir das técnicas que se destacaram na primeira aplicação citada. Além de proporcionar a consolidação de conhecimentos de banco de dados, algoritmos, estruturas de dados, etc., este trabalho permitiu a elaboração de material de pesquisa desta tão atual e promissora tecnologia, a qual merece muita atenção tanto por parte das empresas fornecedoras deste tipo de serviço, como por parte de pesquisadores, tendo em vista que a consolidação desta tecnologia já é dada como certa por grande número de analistas da área.

7 Referências Bibliográficas

[CAS 98] CASTILHO, J. M. V. – *Agentes de Informação Inteligentes* – Trabalho do Curso de Informática da Universidade Católica de Pelotas – Pelotas/ RS 1998.

[CAZ 97] CAZARINE, A; BARBOSA, C; VANCIN, F et. al. – *Data Mining* – Trabalho de Mestrado em Engenharia de Produção USP/São Paulo, 1997.

[CHA 95] CHAKRAVARTHY, ANIL S.; HAASE, KENNETH B. *NetSerf: using semantic knowledge to find Internet information archives*. SIGIR, 1995.

[CHE 96] CHEN, HSINCHUN et al. *A concept space approach to addressing the vocabulary problem in scientific information retrieval: na experiment on the worm community system*. MIS Department, University of Arizona, 2 de Julho de 1996. Disponível em http://ai.bpa.arizona.edu/papers/

[COS 96] COSTA, PATRÍCIA; FREITAS, JULIANA. *Search Engines*. Trabalho da disciplina Internet, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 1996.

[CPY 96] CHEN M.S.; PARK J.S.; YU P.S. *Data mining for path traversal patterns in a web environment*. In Proceedings of the 16th International Conference on Distributed Computing Systems, 1996.

[FAY 96] FAYYAD, USAMA M. et al. *Advances in Knoledge Discovery and Data Mining*. Menlo Park, Califórnia EUA: AAAI Press, 1996.

[FEL 97] FELDENS, M. A; CASTILHO, JOSÈ M. V. Engenharia da Descoberta do Conhecimento em Bases de Dados: Estudo e aplicação na Área da Saúde. Porto Alegre, CPGCC/UFRGS, Março de 1997.

[FEL 99] FELDENS, M. A. *Metodologias para Implementação da Inteligência do Negócio: desenvolvimento de sistema de informação para Database Marketing*, Caxias do Sul: Revista do CCET, 1999. Disponível em http://i.am/mfeldens

[GUR 00] GUROVITZ, HELIO - *DataWarehouse – Data Mining*. Disponível http://www.datawarehouse.inf.br/, artigo: "O que cerveja tem a ver com fralda",

[HOL 94] HOLSMEIMER M; SIEBES, A. *Data Mining: The Search for Knowlodge in Databases*. Disponível por ftp em ftp.cwi.nl no arquivo /pub/CWIreports/AA/CS-R9406.os.Z. 1999.

[INF 99] INFO EXAME, Revista de Informática, *O abc da mineração de dados*, São Paulo, nº.154, Janeiro/1999, p. 117., disponível em http://www.infoexame.com.br

[KOR 93] KORTH, H. F; SILBERSCHATS, A. *Sistemas de Banco de Dados*. São Paulo: Makron Books, 1993.

[LEI 97] LEIGTON H. VEMON. *Precision among WWW search services: Alta Vista, Excite, Hotbot, Infoseek, Lycos*. Disponível em http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm, 1997.

[LOH 97] LOH, S. – *Descoberta de Conhecimento em Bases de Dados Textuais* - 1997

[MAN 99] Manzonini, R. *Internet Transforma Operações dos Fornecedores* – Computerworld – Edição 303, 1999

[MEN 99] MENA, JESUS. *Data Mining Your Web Site*, Digital Prees, July 1999.

[MOB 96] MOBASHER B.; JAIN N.; HAN E.; SRIVASTAVA J.. *Web mining: Pattern discovery from world wide web transactions*. University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.

[NET 97] NETWORK WIZARDS. *Internet domain survey*. Network Wizard Homepage, http://www.nw.com/zone/WWW/top.html, 1997.

[NOV 98] NOVABASE, BPP - *Data Mining* – BPP. Disponível em http://www.novabase.pt/nbsd/nbsd_proj.html

[PIT 94] PITKOW J. *Webviz: A tool for world-wide web access log analysis*. In First International WWW Conference, 1994.

[POL 97] POLITO, M – *Data Mining* – Trabalho apresentado na disciplina de Banco de Dados do Curso de Análise de Sistemas da UERJ – Rio de Janeiro, 1997.

[PIR 96] PIROLLI P.; PITKOW J.; RAO R.: *Extracting usable structures from the web*. In Proc. of 1996 Conference on Human Factors in Computing Systems, Vancouver, British Columbia, Canada, 1996.

[SAN 00] SANTOS , JOSÉ, HENRIQUES, NUNO A.C, REIS, VANDA. *Data Mining e Data Warehousing* - FCT, Janeiro de 1999. Disponível em http://students.fct.unl.pt/users/nach/dmdw/prologo

[SIL 00] SILVA, E. M. *Avaliação do Estado da Arte e Produtos Data Mining* – Pesquisa enviada para o Bulletin Board da disciplina de TSI – Universidade Católica de Brasília / UCB 2000.

[SOU 93] SOUZA, CLARISSE S. *The semiotic engineering of user interface languages*. International Journal of Man-Machine Studies. Academic Press Ltd. 1993.

[TEC 99] TECNIQUES, COURSE. *Data Mining Tecniques*. Disponível em http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes.htm

[THE 00] THEARLING, Kurt - *"Data Mining, CRM, Decision Support, and Database Marketing* (Kurt Thearling)", Disponível em http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm

[VIV 98] VIVAS, BRUNO A. *Técnicas Data Mining* – UFMG/ BH –1998. Disponível em http://www.lch.dcc.ufmg.br/node8.html#

[XAV 00] XAVIER, P. M. *Mineração de Dados: Tecnologias e Ferra*mentas – Trabalho de Conclusão do Curso de Informática da UCPel – 2000

[WHI 96] WHITE, P. *Uma empresa que sabe aonde quer chegar*. Entrevista na Revista Informática Exame Especial, ano 11, n.6, Set 1996.

[WIE 96] WIEBE, J; HIRST, G; HORTON, D. *Language use in context*. Communications of the ACM, v.39, n.1, Jan 96.

[ZAI 94] ZAIANE O. R.. *Resource and knowledge discovery in global information systems: A preliminary design and experiment*. In Proc. of the First Int'l Conference on Knowledge Discovery and Data Mining, pages 331--336, Montreal, Quebec, 1995.

[ZYT 93] ZYTKOW, J. M; ZEMBOWOCZ, R *Databases Exploration in Search of Regularities*. Journal of Intelligent Information Systems, Dordrecht, Holland, v.2,n.1, 1993.