UNICAMP

Técnicas e ferramentas para a extração inteligente e automática de conhecimento em banco de dados

Newton Roy Pampa Quispe

Dissertação apresentada na Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas para a obtenção do Título de Mestre em Engenharia Elétrica

Orientador: Prof. Dr. Vitor Baranauskas

Banca Examinadora:

Prof. Dr. Vitor Baranauskas .....
Prof. Dr. Mauricio Ribeiro Baldan .....
Prof. Dr. Helder José Ceragioli .....
Prof. Dr. Ioshiaki Doi .....

Departamento de Semicondutores, Instrumentos e Fotônica
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas
2003

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

P191t

Pampa Quispe, Newton Roy
Técnicas e ferramentas para a extração inteligente e automática de conhecimento em banco de dados / Newton Roy Pampa Quispe.--Campinas, SP: [s.n.], 2003.

Orientador: Vitor Baranauskas
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Garimpagem de dados. 2. Banco de dados. I. Baranauskas, Vitor. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Dedico este trabalho
a meus sacrificados e amados pais,
Florentino e Nieves

Agradecimentos

A Deus pela vida e amor.

Ao meu orientador Prof. Dr. Vítor Baranauskas, pelo exemplo de dedicação a ciência e pelas diretrizes. O meu obrigado pela liberdade concedida, pela confiança depositada e por ter possibilitado a minha iniciação em um novo domínio.

Ao Laboratório de Avaliação de Qualidade de Software (LAQS) do Centro Nacional de Pesquisa Renato Archer (CenPRA), ao Sr. Alberto Passos, a Sra. Regina Thienne e a Sra. María Villalobos, pela oportunidade de estagiar e aplicar conceitos teóricos na Área de Tecnologia de Qualidade de Software (ATQS).

Aos meus amados pais, Sr. Florentino e Sra. Nieves pelo sacrifício ao longo da minha vida e a valiosa educação cristã recebida. Aos meus esforçados, otimistas e valiosos irmãos: Néstor, Walter, María, Noé e Marta pelos incentivos recebidos, conforto, apoio, otimismo e confiança em Deus.

Aos meus irmãos e irmãs na fé da Igreja Adventista do Sétimo Dia de "Vila São Pedro", pelas orações.

Aos meus parentes e amigos da minha terra peruana, motivo inspirador de minhas realizações.

Aos meus caros amigos e colegas do Laboratório do Departamento de Semicondutores, Instrumentos e Fotônica (DSIF) pela amizade.

Ao Departamento de Semicondutores, Instrumentos e Fotônica da FEEC da UNICAMP pela recepção no programa de pós-graduação

Meus agradecimentos especiais ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) do Brasil, pelo financiamento dos estudos que envolveram o projeto de mestrado.

Muito Obrigado

Resumo

Esta dissertação apresenta uma visão geral e exploratória de técnicas e ferramentas para a extração inteligente e automática de conhecimento em grandes bancos de dados. Esta área é uma ciência emergente que aplica modernas tecnologias computacionais e estatísticas, para fornecer elementos para descobrir relações significativas, mas desconhecidas, em banco de dados. Esta pesquisa envolve técnicas e ferramentas combinadas de diferentes disciplinas como: sistemas de base de dados, almoxarifado de dados, estatística, aprendizagem de maquina, visualização de dados, recuperação de informação, entre outras, além do processo interativo e iterativo com o usuário. Apresentamos um exemplo simples de aplicação, relacionado com a avaliação da qualidade do produto de software.

Palavras chaves: Descoberta de conhecimento em banco de dados, garimpagem de dados, sistemas de bancos de dados.

Abstract

This thesis presents an exploratory overview of techniques and tools for the intelligent and automatic extraction of hidden knowledge from database systems. This is an emerging science that applies modern computational and statistics technologies to supply the needs of discovery significant hidden relationships in large database systems. This research involves tools and combined techniques of many different areas such as database systems, data warehouse, statistics, machines learning, data visualization, information recovery among others, and the interactive process with the user. We present an application example related to the quality of software products.

Keywords: Knowledge discovery in database (KDD), data mining, database systems.

Conteúdo

Capítulo I

Introdução

1 Generalidades

Modernas tecnologias computacionais e estatísticas têm sido desenvolvidas para suprir as necessidades de descobrir relações significativas, mas desconhecidas, em Bancos de Dados (BD). Esta área de pesquisa é denominada de Extração Inteligente e Automática de Conhecimento em Bancos de Dados (EIACBD). A EIACBD focaliza o desenvolvimento de métodos de extração eficientes e que possam ser utilizados em escala de bancos de dados de diversas dimensões (escalonamento). A EIACBD é uma forma de expressar o processo de descoberta de conhecimento em BD (*Knowledge Discovery in Database* - KDD) que também é conhecido como "garimpagem de dados" (*Data Mining*). O KDD foi proposto em 1989 para referir-se às etapas que produzem conhecimentos a partir de dados relacionados, sendo a garimpagem de dados a etapa que transforma dados em informações. O KDD refere-se ao processo de extração da informação relevante ou de padrões nos dados contidos em grandes BD e que sejam: não-triviais, implícitos, previamente desconhecidos e potencialmente úteis [Fayyad et al., 1996].

O desafio principal da extração de conhecimento é processar automaticamente e inteligentemente grandes quantidades de dados crus, identificar os padrões mais significantes e representativos, e apresentar estes modelos ou padrões como conhecimento apropriado para alcançar os objetivos do usuário.

Este trabalho tem como objetivo apresentar uma breve visão geral das técnicas e ferramentas de extração de conhecimento desenvolvidas por vários grupos, com ênfase nas técnicas implementadas em sistemas de extração de conhecimento em BD. As técnicas e ferramentas que são apresentadas foram selecionadas por sua capacidade de automatização, inteligência, eficiência e escalonamento.

A utilização de ferramentas inteligentes pode resolver problemas complexos como exemplo o incremento do espaço de soluções. A automatização dos processos pode reduzir o trabalho operativo do usuário, mas a interação humana é requerida com algum grau. Esta é a razão para considerar o usuário como parte do sistema, porem, o alvo do sistema automático e inteligente é o de reduzir ao mínimo a ação humana no sistema.

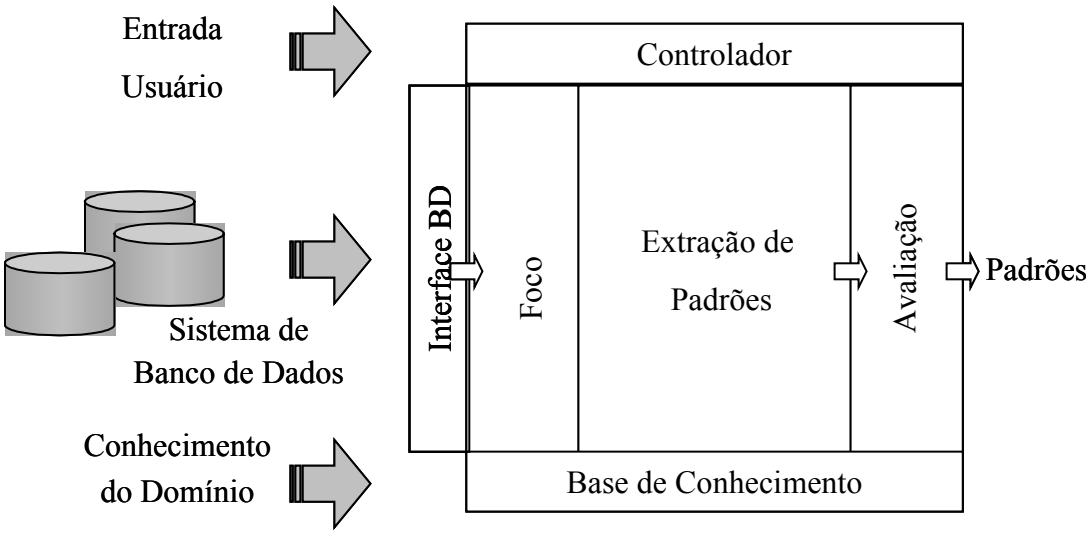

Figura I.1: *Modelo de um sistema de extração de conhecimento segundo [Matheus et al., 1993]*

A Figura I.1 apresenta o esquema básico das partes componentes de um sistema de extração de conhecimento em BD [Matheus et al., 1993]:

- *Controlador:* controla a solicitação e parametriza os componentes do sistema,
- *Interface de BD*: acessa, gera e processa consultas a BD,
- *Base de conhecimento*: armazenamento de informação especifica de domínio,
- *Foco*: determina a porção de dados a ser analisados,
- *Extração de padrões*: coleção de algoritmos de extração de conhecimento,
- *Avaliação*: avalia a qualidade de padrão extraído (interesse, utilidade, etc).

Os dados são a matéria prima bruta do conhecimento. No momento que o usuário atribui algum significado especial aos dados, estes passam a ser uma "*informação*", e quando os especialistas geram regras, surge o "*conhecimento*". A Figura I.2 apresenta um

esquema típico de hierarquia entre dados, informação e conhecimento, existentes em um BD, observando seu volume e valor hierárquico

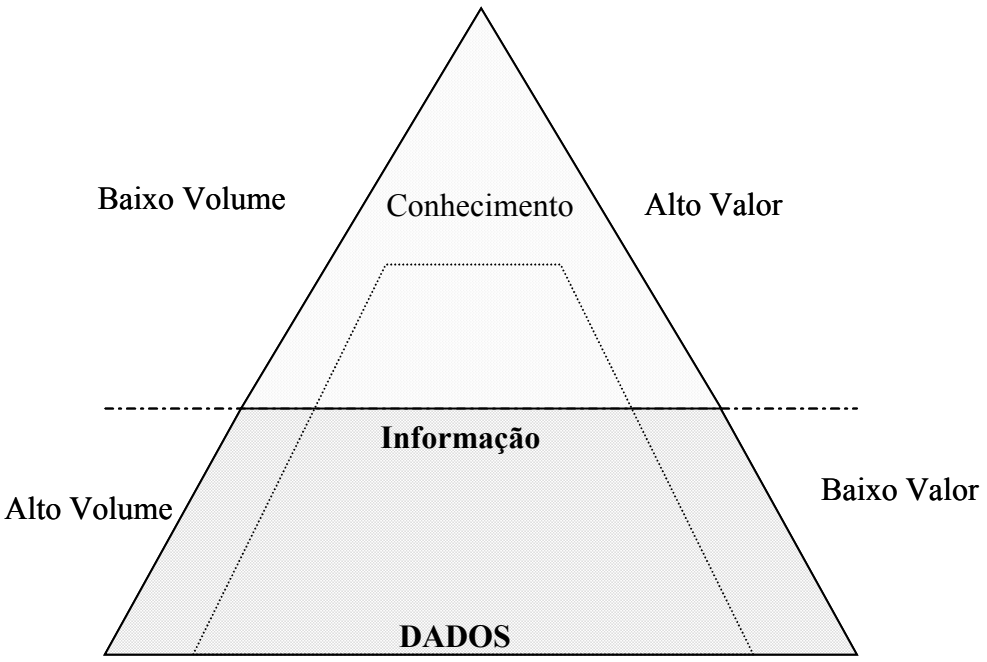

Figura I.2: *Pirâmide do processo de conhecimento*

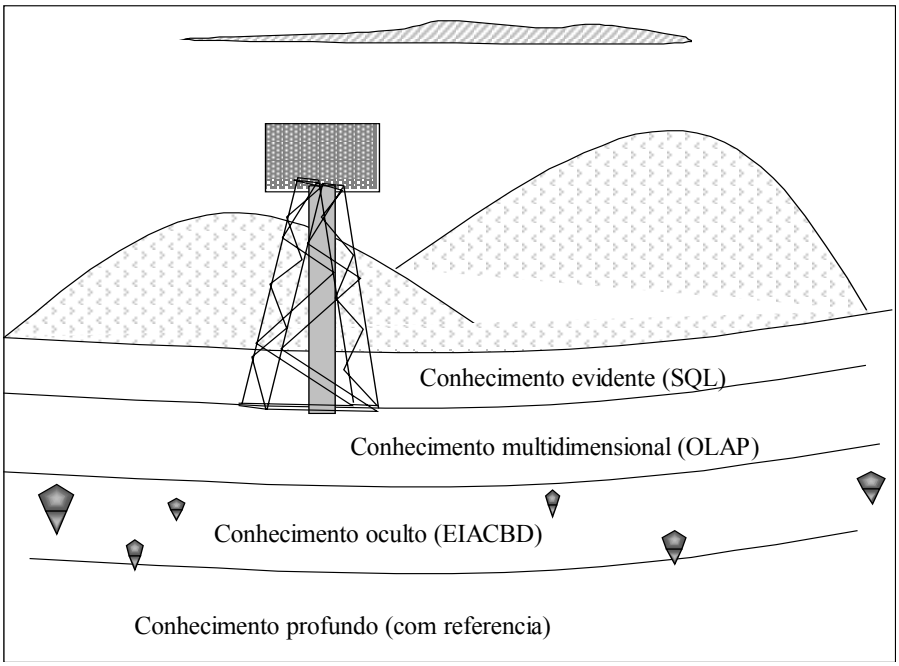

Figura I.3*: Níveis de conhecimento*

Podemos também esquematizar os níveis de profundidade do conhecimento, conforme ilustrado na Figura I.3. O *conhecimento evidente* pode ser obtido através da linguagem de consultas estruturadas (*Structure Query Language* - SQL), o *conhecimento multidimensional* pode ser obtido com o processamento analítico em tempo real (*On-Line Analysis Processing* - OLAP), o *conhecimento oculto* pode ser extraído com a EIACBD e o *conhecimento profundo* só pode ser extraído com referências ou pistas adicionais.

1.1 Motivação

Vários fatores contribuíram para o desenvolvimento de técnicas e ferramentas inteligentes e automáticas, tais como:

- A emergência de uma quantidade muito grande de dados, devido à geração automática, coleção, gravação digital automatizada por computador, arquivos centralizados e simulações.
- O notável decréscimo do custo dos dispositivos de armazenamento massivo de dados.
- O desenvolvimento rápido de técnicas e tecnologias de sistemas de administração de informação.
- Os avanços na tecnologia de computação, como computadores rápidos e arquiteturas de processamento paralelo.
- O desenvolvimento constante de técnicas de aprendizagem automática, e a inteligência artificial.
- A presença possível de ruídos, dados fora do padrão, informações perdidas.
- O descobrimento de conhecimento.

1.2 Visões de classificação de extração de conhecimento

Com o propósito da aplicação

- Verificação: serve para confirmar uma determinada suspeita ou hipótese, como por exemplo: “a possibilidade de que nos finais de semana, ocorra maior número de acidentes de transito”.

- Descobrimento: examinar os dados tentando descobrir relacionamentos que não foram previstos pelos usuários e apresenta-los. O descobrimento pode ser feito para dois propósitos.
    - Descrição: sistemas que procuram padrões para apresentar descrições ao usuário de forma compreensível.
    - Predição: sistemas que procuram padrões para prever o comportamento futuro de algumas entidades.

Com o propósito da diferenciação

- Tipos de dados a serem garimpados:
  Banco de dados: relacional, transacional, objeto-orientado, objeto-relacional, ativo, espacial, séries-temporais, textual, multimídia, heterogêneo, *web*, etc.
- Tipos de conhecimentos a serem descobertos
  Caracterização, discriminação, associação, classificação, segmentação, tendência, analise do desvio, etc.
- Tipos de técnicas a serem utilizadas
  Orientado a: banco de dados, armazenagem de dados, estatística, aprendizagem de maquina, visualização de dados, etc.
- Tipos das aplicações
  Varejo, telecomunicações, operação bancaria, analise de fraude, garimpagem de DNA, analise de estoque de mercado, mineração da *web*, etc.

1.3 Aplicações

Aplicações comerciais

- Análise de mercado e gerenciamento: Marketing alvo, gerenciamento de relação com o cliente, análise de cesta de mercado, venda transversal, segmentação de mercado. Por exemplo, a *British Broadcasting Corporation* (BBC) do Reino Unido emprega um sistema para prever a audiência televisiva para um determinado programa, assim como o tempo ótimo de exibição [Brachman et al., 1996]. O sistema utiliza redes neurais e

arvores de decisão, aplicadas a dados históricos da empresa, para determinar os fatores que influenciam o sucesso do programa escolhido.

- Análise de risco e gerenciamento: Prognostico de vendas, retenção da clientela, aumento da clientela, controle de qualidade, análise do competidor. Por exemplo, uma companhia espanhola descobriu o perfil que apresentava os clientes que deixavam a companhia e com isto pode desenvolver um tratamento personalizado para seus clientes atuais que possuíam características similares aos clientes que deixaram a companhia.
- Detecção e prevenção de fraudes: Por exemplo, em 2001 muitas instituições financeiras perderam mais de dois bilhões de dólares americanos em fraudes com cartões de crédito e débito. Foram então desenvolvidos sistemas inteligentes que testam as transações dos proprietários de cartões e dados financeiros para detectar e minimizar as fraudes. O sistema *Falcon Fraud Manager*, por exemplo, permitiu poupar mais de 600 milhões de dólares por ano e atualmente protege aproximadamente 65% de todas as transações com cartões de credito no mundo.

Aplicações científicas

- Astronomia: Por exemplo, o projeto *Second Palomar Observatory Sky Survey (POSS-II)* armazenou em seis anos, três *terabytes* de imagens que continham aproximadamente dois milhões de objetos no céu, com o objetivo de formar um catalogo dos objetos. Foi então desenvolvido o sistema *Sky Image Cataloging and Analisys Tool* (SKYCAT) baseado na técnica de agrupamento (*clustering*) e árvores de decisão, que pode classificar os objetos em estrelas, planetas, sistemas, galáxias etc., além de ajudarem os astrônomos na descoberta de 16 novos *quasares* [Fayyad et al., 1996].
- Biologia: Desenvolvimentos significantes da biologia computacional e da bio-informática permitiram a descoberta de padrões interessantes de bio-seqüências, incluindo DNA, RNA, e de proteínas. A meta é achar sucessões ou padrões repetidos, escondidos no DNA, ou em outros bio-dados de grandes BD. O sistema DNA-*Miner* é um protótipo que processa dados do DNA e executa o garimpo de padrões seqüenciais para análise. [Han et al., 2001].
- Medicina: Companhias farmacêuticas armazenaram a composição e propriedades de combinações de substâncias químicas em BDs muito grandes e estão usando o KDD

para procurar substâncias promissoras que possam formar a base de novas drogas. No âmbito hospitalar existem os sistemas de apoio à atividade medica, indo deste o sistema de gestão dos dados dos pacientes até sistemas de apoio ao diagnostico. Porem, a integração destes sistemas não é ainda eficiente.

- Geologia: Existem modelos de KDD para análise das mudanças do clima global, e para a busca de possíveis padrões de espaço-temporal, que são importantes indicativos de possíveis ciclones [Fayyad et al., 1996].

Aplicações acadêmicas

- Acompanhamento de formandos: Aplicando a técnica KDD descobriu-se que quatro parâmetros principais determinavam a inserção no mercado do trabalho dos estudantes recém titulados de certa universidade, sendo que três dos quatro parâmetros não dependiam da universidade [Rodas, 2001].

1.4 A confluência com outras disciplinas

A EIACBD fundamenta-se na geração de hardware e software, e inclui analises estatísticas, tecnologias de BD, aprendizagem de maquina, ciência da informação, visualização de dados, e outras disciplinas, conforme esquematizado na Figura I.4.

Sem a estatística não seria possível ter-se a EIACBD. A análise estatística clássica desempenha um papel fundamental, pois introduz os conceitos de formas de distribuição, variância, análise de regressão, desvio simples, análise de conjuntos, discriminantes, intervalos de confiança, etc., que são usados para medir e estudar os dados e seus relacionamentos.

A inteligência artificial é construída a partir de fundamentos de heurística. A heurística é uma ciência que tenta imitar o comportamento humano durante a resolução dos problemas, ou seja, ela trabalha em contraponto à estatística [Kremer, 1999].

O aprendizado de máquina é a união entre a estatística e a inteligência artificial, procurando fazer com que os computadores tenham a capacidade de aprender por meio dos dados que manipulam e tomar decisões baseadas nas características destes dados [Kremer, 1999].

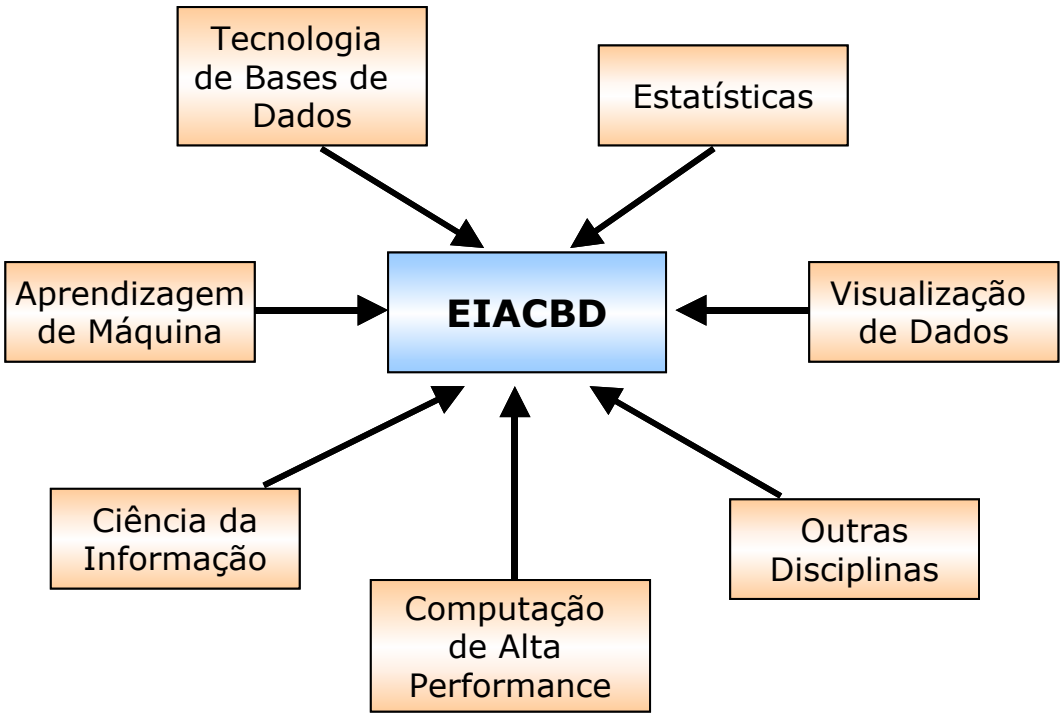

Figura I.4: *A confluência de EIACBD com outras disciplinas*

[Santos et al., 1999] indicam que as técnicas de KDD resultaram de longa evolução no processo de pesquisa e desenvolvimento e que, neste momento, tem o suporte de três tecnologias principais: armazenamento de dados, computadores de alta performance e algoritmos de garimpagem de dados. O principal componente desta tecnologia é a inteligência artificial que está em pleno desenvolvimento há décadas.

1.5 As extensões de extração de conhecimento

- Extração de conhecimento de banco de dados espacial *(Spatial Data Mining*): É uma extensão de KDD voltada para domínios de aplicação onde a consideração da dimensão espacial é essencial na extração de conhecimentos.
- Extração de conhecimento de banco de dados multimídia *(Multimedia Data Mining*): É uma extensão de KDD voltada para domínios de dados de multimídia, como: textos, imagens, gráficos, animação, vídeo, áudio e composição de multimídia [Adjeroh and Nwosou, 1997].

- Extração de conhecimento de banco de dados seqüenciais e series-temporais *(Time-series and sequence Data Mining*): Voltado para domínios de dados seqüências ou series-temporais.
- Extração de conhecimento de *web* (*Web Mining):* Aplica as técnicas KDD para documentos e serviços de *web* [Kosala and Blockeel, 2000]. As técnicas de *web minig* analisam e processam os dados gerados pelos serviços da internet (endereços IP, navegador). Os servidores automaticamente armazenam em uma memória de acessos (*log*), e produzem informações significativas. Os conteúdos da internet consistem de vários tipos de dados, como texto, imagem, vídeo, meta-dados e hiper-enlaces. Pesquisas recentes usaram o termo *multimedia data mining* como uma instância de *web mining* [Zaine et al., 1998]. A w*eb mining* pode ser classificada em três domínios de extração de conhecimento, de acordo com a natureza dos dados: garimpo do conteúdo da *web* (*web content mining*), garimpo da estrutura da w*eb (web struture mining*) e garimpo do uso da *web (web usage mining*).
- Extração de conhecimento de texto (*Text Mining*): Refere-se ao exame de uma coleção de documentos para descobrir informações sem ter referência inicial [Nasukawa and Nagano, 2001]. O campo de estudo é amplo, e as técnicas de categorização de texto, processamento de linguagem natural, extração e recuperação da informação ou aprendizagem automática e outras, facilitam o trabalho do *text mining*.

1.6 Os desafios na área de extração de conhecimento

- Bancos de dados volumosos: Os BDs com centenas de tabelas, campos, milhões de registros estão tornando-se cada vez mais comuns e para isto é necessária a aplicação de algoritmos mais eficientes, com processamento massivo ou paralelo [Fayyad et al., 1996].
- Alta dimensionalidade: Além do grande número de registros os BDs também podem conter grande número de campos (atributos, variáveis). As soluções para estes sistemas incluem métodos para reduzir a dimensionalidade efetiva e o uso de conhecimento prévio para identificar variáveis irrelevantes [Fayyad et al., 1996].
- Sobre-ajuste (*Overfitting*): Quando o algoritmo busca os melhores parâmetros para um modelo particular que usa um conjunto limitado de dados, não só pode modelar os

padrões gerais nos dados, mas também modelar qualquer ruído específico ao conjunto de dados, resultando em baixo desempenho do modelo em dados de teste. As possíveis soluções incluem a validação cruzada, regularização e outras estratégias estatísticas sofisticadas [Fayyad et al., 1996].

- Avaliação de significação estatística: O problema do sobre-ajuste acontece quando o sistema estiver procurando com base em muitos modelos possíveis. Por exemplo, se um sistema testa modelos em nível de significância a 0,001, então em média, com dados puramente aleatórios, N/1000 destes modelos serão aceitos como significantes. Um modo para lidar com este problema é usar métodos que ajustam a estatística de teste como uma função de busca, por exemplo, o ajuste de Bonferroni para testes independentes ou testes aleatórios [Fayyad et al., 1996].
- Dados e conhecimentos dinâmicos: Os dados podem alterar-se rapidamente e com isto tornar inválidos os padrões previamente descobertos. Além disso, as variáveis medidas podem ser modificadas ou apagadas com o passar do tempo. As possíveis soluções para esta questão incluem métodos com incremento para atualização temporal dos padrões [Fayyad et al., 1996].
- Dados ruidosos e perdidos: Este problema é especialmente crucial nos BDs empresariais. Dados do censo norte-americano têm taxas de erros tão grandes quanto 20% em alguns campos. Atributos importantes podem ser perdidos se o BD não for projetado para descoberta de conhecimento. Estratégias estatísticas mais sofisticadas para identificar variáveis escondidas e dependentes podem ser aplicadas nestes casos [Fayyad et al., 1996].
- Relações complexas entre campos: Os valores ou atributos estruturados hierarquicamente, as relações entre atributos, e meios mais sofisticados para representar conhecimento sobre os conteúdos de BDs podem requerer algoritmos que possam usar efetivamente tal informação. Historicamente, os algoritmos garimpagem de dados foram desenvolvidos para registros simples, mas novas técnicas para derivar relação entre variáveis podem ser desenvolvidas [Fayyad et al., 1996].
- Compreensão de padrões: Em muitas aplicações é importante tornar as descobertas mais compreensíveis para humanos. A possível solução inclui representação de gráficos, regras estruturadas, geração de linguagem natural, e técnicas para visualização

de dados e conhecimento. As estratégias de refinamento de regras podem ser usadas para focalizar um problema relacionado e o conhecimento descoberto pode ser implicitamente ou explicitamente redundante.

- Interação com usuário e conhecimento prévio: Muitos métodos de KDD não são verdadeiramente interativos e não podem incorporar facilmente conhecimentos prévios. O uso de conhecimento de domínio é importante em todos os passos do processo de KDD. Por exemplo, a solução bayesiana usa probabilidades prévias em cima de dados e distribuições como uma forma de codificar o conhecimento prévio.
- Integração com outros sistemas: Um sistema de descoberta isolado pode não ser muito útil. Sistemas de integração típicos incluem a integração com sistemas de administração de BD (por exemplo, por uma interface de consulta), integração com; planilhas eletrônicas, ferramentas de visualização e leitura por sensores em tempo real.
- Suporte a novas tecnologias de dados: Com a evolução de tecnologias de armazenamento, os dados armazenados passaram a conter, alem de textos e números; objetos gráficos, multimídia, dados dinâmicos, dados temporais, entre outros. Gerenciadores de BD orientados a objetos podem tratar deste tipo de problema de armazenamento, facilitando a geração de meta-dados.
- Ambiente de rede e sistemas distribuídos: O rápido crescimento de recursos disponíveis na internet demanda uma grande necessidade por pesquisas para o desenvolvimento de ferramentas, técnicas e sistemas que possam permitir a realização do processo de KDD no ambiente conectado e distribuído. Ainda, a tendência da área de KDD é guiar-se para um descobrimento de conhecimento "colaborativo" envolvendo um grande time de analistas e especialistas do domínio espalhados que possam utilizar os BDs disponíveis na rede. As pesquisas atuais de agentes inteligentes são um começo para se conseguir atingir os desafios impostos à área de KDD pelas novas tecnologias de *web* e BD Multimídia.

2 Objetivo

O objetivo da presente dissertação é a definição dos processos de EIACBD com o estudo das ferramentas envolvidas, incluindo o estudo das técnicas de garimpagem de dados, concluindo com a aplicação destas técnicas e ferramentas para um caso real.

3 Estrutura da tese

Esta dissertação esta organizada em sete capítulos. No Capítulo I apresenta-se o contexto no qual esta envolvida a EIACBD, com suas aplicações, desafios e extensões.

No Capitulo II descrevem-se as ferramentas ou elementos envolvidos no processo de EIACBD, suas funções, e atuações nos diferentes processos.

No Capitulo III define-se as tarefas que se realizam nas diferentes etapas e processos da EIACBD; inicialmente a compreensão de domínio seguido pelo pré-processamento, processamento e pós-processamento.

No Capitulo IV aborda-se a garimpagem de dados de forma ampla, onde os dados são transformados em informações através de técnicas (algoritmos) de analise de dados.

No Capitulo V descreve-se a experiência da aplicação de EIACBD na avaliação da qualidade de produtos de software. Os dados foram armazenados em um BD sob o modelo relacional de dados, seguindo todos os passos do processo, até a apresentação dos resultados.

No Capitulo VI apresentam-se as conclusões e propostas de futuros trabalhos.

Capítulo II

Ferramentas para extração de conhecimento

1 Introdução

A EIACBD depende de uma nova geração de técnicas e ferramentas de análises de dados para cumprir determinadas funções no processo. As ferramentas ou elementos fornecem as condições de interação do usuário com os dados através de diferentes *softwares* (armazenagem de dados, recuperação de dados, preparação de dados, análise de dados e apresentação de resultados), e também na construção da estrutura de armazenamento e recuperação de dados para simplificação das tarefas.

2 Sistema de Banco de Dados (SBD)

Um SBD é usualmente uma aplicação que serve de suporte a outras aplicações, tais como folha de pagamento, controle de pessoal, informações bancárias, etc. O SBD é constituído por um Banco de Dados (BD) e pelo Sistema Gestor de Banco de Dados (SGBD) [Ricarte, 1996]. A Figura II.1 apresenta um esquema de SBD com ambiente simplificado.

Tradicionalmente os SBDs estavam dirigidos para aplicações na área comercial, entretanto, nos últimos anos os SBD tem sido utilizados para organizar a armazenagem de informação de diferentes aplicações científicas.

A independência da aplicação com respeito à armazenagem de dados, abre a possibilidade de compartilhar dados entre aplicações, possibilitando sua utilização por exemplo, em projetos de engenharia (CAD/CAE – *Computer Aided Design, Computer Aided Engineering*), automação industrial e de escritórios (CAM –*Computer Aided*

Manufacturing, AO – *Office Automation*), sistemas especialistas de auxílio ao trabalho cooperativo (CSCW – *Computer Supported Cooperative Systems*), sistemas especialistas de suporte a tomada de decisões (DSS – *Decision Support Systems*). Muitos destes domínios "não-convencionais" exigem SBDs com maior volume de dados, manipulação de dados "não alfanuméricos" (informação multimídia), etc.

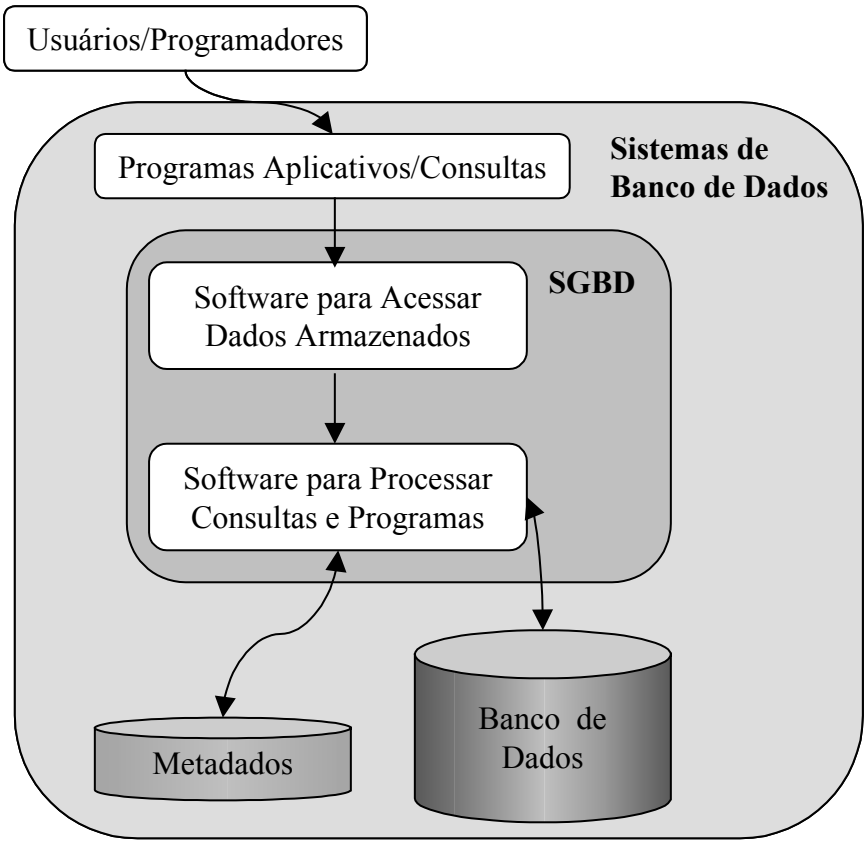

Figura II.1: *Ambiente simplificado de um sistema de banco de dados*

2.1 Banco de Dados (BD)

Um banco de dados é por exemplo, uma coleção de informações operacionais armazenadas e utilizadas pelo sistema de aplicações de uma empresa específica [Batini and Lenzerini, 1986].

Segundo [Ricarte, 1996] o banco de dados também pode ser uma coleção de dados relacionados que pode ser armazenada sob alguma forma física. Os dados armazenados

representam algum aspecto especifico do mundo real, e apresentam algum grau de coerência lógico entre seus componentes, sendo modelado, construído e povoado com dados para uma proposta específica.

A possibilidade de aplicações em domínios "não convencionais" promove o desenvolvimento de bancos de dados para outras formas de coleção de dados , por exemplo, do tipo, relacional, transacional, objeto-orientado, objeto-relacional, ativo, espacial, series-temporais, textual, multimídia, heterogêneo, etc.

2.2 Sistema Gerenciador de Banco de Dados (SGBD)

Um SGBD consiste no software que permite criar, manter e manipular BDs para diversas aplicações. A criação e manutenção de BD são tarefas de uma pessoa ou grupo de pessoas, normalmente referenciadas por administrador de BD. A manipulação de BD (atualizações e consultas) é realizada direta ou indiretamente (programas aplicativos) pelo usuário de BD. O SGBD pode ser de propósito geral ou especifico para alguma aplicação [Ricarte, 1996].

Para que um SGBD possa desempenhar satisfatoriamente sua função ele deve oferecer um grupo de interfaces que garantam uma perfeita comunicação entre o núcleo e seus usuários, de maneira direta ou através de aplicações. Para atingir esse objetivo de modo que seus usuários não tenham contato com as complexas estruturas necessárias para armazenamento e recuperação de dados, o ambiente em que um SGBD deve operar em três níveis de abstração, conforme apresentado na Figura II.2 [CODASYL – DBTG 1971, ANSI/SPARC 1975]. O BD pode ser visto através três níveis de abstratos, ocultando desta forma suas partes estruturais para facilitar seu uso por profissionais sem muito conhecimento de computação.

- Nível Físico: Nível de abstração de base, no qual as estruturas complexas são descritas em detalhe, apresentando como os dados estão armazenados.
- Nível Conceitual: Nível intermediário, em que se definem quais dados farão parte do BD. Esta definição se dá através de estruturas menos complexas que ficam sob a responsabilidade de profissionais da informática.

- Nível de Visão: Nível superior de abstração, que apresenta apenas parte do BD. Neste nível poderão ser criadas "mascaras" para adequar aos dados conceitualmente armazenados em formatos de visualização esperados pelos usuários.

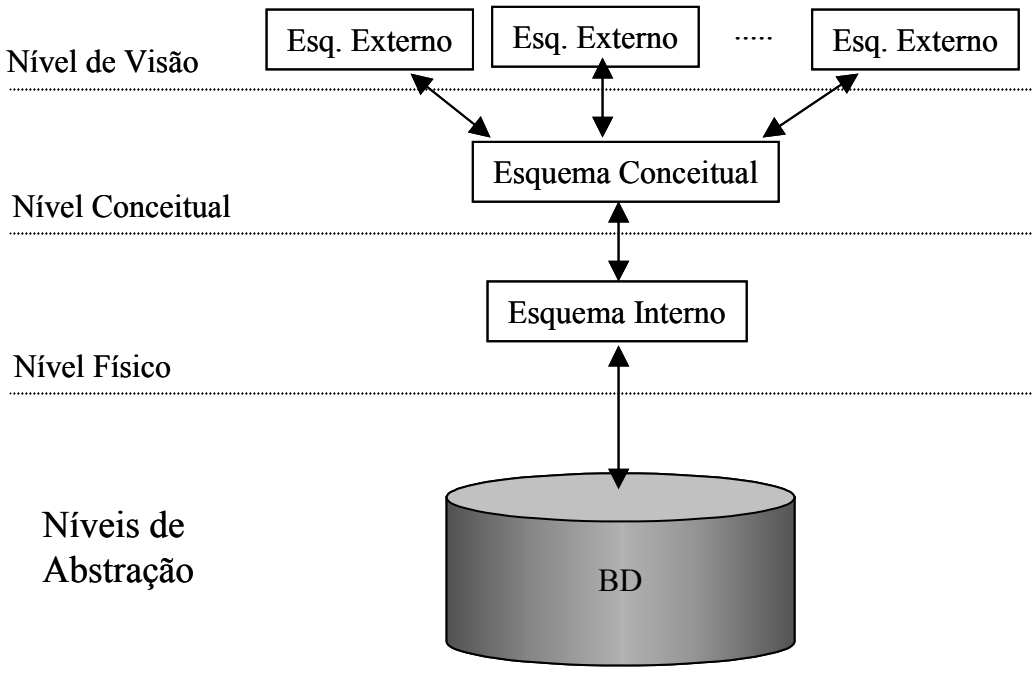

Figura II.2: *Níveis de abstração de dados*

Para garantir aos usuários a interação com o SGBD aos níveis em questão, a estrutura do SGBD deve garantir a existência de interfaces nos níveis esperados. A estrutura básica esperada de um SGBD e seus componentes básicos deve conter os seguintes ítens [Korth and Silberschatz, 1995]:

- Gerenciador de arquivos: Responsável pela administração do meio físico encarregado da armazenagem dos dados e pelas estruturas de dados utilizados para representação dos dados nesse meio físico;
- Gerenciador do banco de dados: Responsável de prover aos programas aplicativos de consultas submetidos ao sistema, interface para recuperação e armazenamento dos dados mantidos pelo gerenciador de arquivos;
- Processador de consultas: Traduz as necessidades dos usuários em linguagem de consulta própria, para instruções que possam ser entendidas e interpretadas pelo

gerenciador de BD. Em alguns sistemas é dotado de mecanismo que ao traduzir, racionaliza o formato da consulta garantindo melhor tempo de resposta;

- Pré-compilador da linguagem de manipulação de dados: Elemento complementar e altamente sincronizado com o processador de consultas, o pré-compilador da linguagem de manipulação de dados é responsável por garantir e fazer com que as solicitações de armazenamento e recuperação de dados feitas ao gerenciador do BD a partir de programas aplicativos escritos em linguagem própria possam usufruir os recursos do SGBD;
- Compilador da linguagem de definição dos dados: Responsável de converter os comandos para definição dos dados que se pretende armazenar e recuperar através do SGBD;
- Outros elementos auxiliares presentes em um SGBD serão responsáveis da manutenção das estruturas físicas, diretamente relacionadas ao serviço de gerenciamento de arquivos tais como o dicionário de dados e índices.

Modelagem de dados

Os SGBDs são construídos para implementar arranjos e distribuições específicas sobre os dados de tal modo que o mundo real possa ser retratado em suas estruturas mais internas. O processo de modelagem dos dados se da desde os níveis mais reais (formas como são vistos pelo mundo) até os níveis mais estruturados possíveis (ponto em que o dado é efetivamente armazenado pelo sistema).

Varias propostas de modelagem foram apresentados para os diversos níveis de abstração de dados para implementar SGBDs. É importante ressaltar que propostas de modelagem consistentes foram concebidas para o projeto lógico, tais como o Modelo Entidade-Relacionamento, apresentado por [Chen, 1990]. Porém, esses modelos para projetos lógicos não são suficientes para uma implementação de um SGBD real.

Dentre os modelos que se propõem a sustentar a implementação de um SGBD, quatro se destacam em função de sus consistência e formalismo. Esses modelos são descritos a seguir em ordem cronológica de apresentação à comunidade de usuários.

- Modelo Hierárquico: São os que requerem que os tipos de registros de dados sejam organizados em uma forma hierárquica. Estes modelos impõem algumas restrições, pois

nem sempre existe uma relação de hierarquia natural entre os registros que se pretende fazer. Atualmente, com utilização bastante restrita, sua grande ferramenta de implementação foi certamente o Sistema de Gerencia da Informação (IMS – *Information Management System*) da IBM.

- Modelo de Rede: Proporcionam capacidades mais complexas de estruturas de dados do que os sistemas hierárquicos. Existe uma abertura para que um determinado tipo de registro tenha diversos tipos de registros. Poucos são os SGBD utilizados na atualidade que o usam como modelo de implementação, mas importantes ferramentas foram construídas sobre ele, tais como: o IDS da Honeywell, o DMS-1100 da UNIVAC e o IDMS da Cullinane.
- Modelo Relacional: É amplamente empregado na atualidade. Sua estrutura básica e sua flexibilidade garantem grande aceitação, Foi implementado através do "System R" desenvolvido pela IBM [Codd, 1970]. Para o modelo relacional, uma linha da tabela representa um relacionamento entre valores. Desta forma, uma tabela é um conjunto de relacionamentos do mesmo tipo. Cada coluna da tabela é uma característica do relacionamento e recebe o nome de atributo. O conjunto de valores possíveis para o atributo é denominado domínio. Este modelo apresenta forte coesão, pois tem fundamentos matemáticos.
- Modelo Orientado a Objetos: São modelos que conceitualmente se utilizavam na modelagem lógica dos BD. São implementáveis em alguns sistemas gerenciadores experimentais em sua forma pura e, em algumas ferramentas comerciais, em formato híbrido com modelos relacionais. Apresenta facilidades para manipulação de dados não convencionais como em: voz, imagem, imagens dinâmicas, etc.

Armazenamento e recuperação de dados

Os SGBDs, quando implementam determinados modelos, embora guardem suas características próprias e articuladas, por conta da adesão ao modelo, tornam-se padronizados. Esta padronização permite a construção de projetos comuns para a distribuição dos dados que se deseja manter e acabam por viabilizar a construção de linguagens de consultas especificas, que também seguem o mesmo padrão.

O modelo relacional de dados não se preocupa apenas com padrões para o armazenamento dos dados, mas também com modelos para recupera-los (única razão de armazenagem) e a parte do modelo relacional que se encarrega disto é chamada de álgebra relacional. Da álgebra relacional deriva a conhecida a Linguagem Estruturada de Consultas (SQL - *Structure Query Language*). Nem sempre a proposta de modelagem relacional, que certamente é a mais difundida atualmente, fornece facilidades para aplicação de EIACBD. No processo de fragmentação imposto pelo modelo, para várias situações, poderá ser necessário o pré-processamento dos dados à um estagio em que os dados possam ser melhor interpretados, mesmo com algum grau de redundância,.

3 Armazém de Dados (*Data Warehouse* - DW)

O Armazém de dados (*Data Warehouse* DW) é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para auxiliar o processo gerencial de tomada de decisão [Inomn, 1997]. Para entender melhor o que é um DW é importante fazer uma comparação com o conceito tradicional de BD. Conforme [Batini and Lenzerini, 1986], um BD é uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de uma empresa específica. Os dados mantidos por um domínio são chamados de "operacionais" ou "primitivos". Os dados no BD são "dados operacionais", distinguindo-se de dados de entrada, dados de saída e outros tipos de dados. A Figura II.3 apresenta o esquema de interação dos componentes de um DW.

Levando em consideração esta definição sobre dados operacionais, pode-se dizer que um DW é uma coleção de dados derivados dos dados operacionais para sistemas de auxílio à decisão. Estes dados derivados são, muitas vezes, denominados de dados "gerenciais", "informacionais" ou "analíticos" [Inmon, 1996].

Os BD operacionais armazenam as informações necessárias para as operações momentâneas do domínio. São utilizados por todos os usuários para registrar e executar operações pré-definidas, por isso seus dados podem sofrer constantes mudanças conforme as necessidades atuais do domínio. Para não ocorrer redundância nos dados as informações

históricas não ficam armazenadas por muito tempo, este portanto este tipo de BD não exige grande capacidade de armazenamento.

Na Tabela II.1 estão relacionadas algumas diferenças entre BD operacionais e DW bem como as diferenças dos dados que eles manipulam segundo os seguintes autores: [Inmon, 1996] [Barquini, 1996] [Kimball, 1996] [Oneil, 1997].

Tabela II.1: *Comparação entre Banco de Dados Operacionais e Data Warehouse*

Características	**Bancos de dados**	*Data Warehouse*
Objetivo	Operações eventuais do domínio	Analisar o domínio
Uso	Operativo	Informativo
Tipo de processamento	OLTP	OLAP
Unidade de trabalho	Inclusão, alteração, exclusão.	Consulta
Número de usuários	Milhares	Centenas
Tipo de usuário	Operadores	Gerenciadores
Interação do usuário	Somente pré-definida	Pré-definida e *ad-hoc*
Condições dos dados	Dados operacionais	Dados Analíticos
Volume	Megabytes – gigabytes	Gigabytes – terabytes
Histórico	Mensal, Bimestral.	Annual
Granularidade	Detalhados	Detalhados e resumidos
Redundância	Não ocorre	Ocorre
Estrutura	Estática	Variável
Manutenção desejada	Às vezes	Freqüentemente
Acesso a registros	Dezenas	Milhares
Atualização	Contínua (tempo real)	Periódica (em *batch*)
Integridade	Eventualmente	A cada atualização
Número de índices	Poucos/simples	Muitos/complexos
Intenção dos índices	Localizar um registro	Aperfeiçoar consultas

O DW armazena dados analíticos, destinados às necessidades da gerência no processo de tomada de decisões. Isto pode envolver consultas complexas que necessitem

acessar um grande número de registros, por isso é importante a existência de muitos índices criados para acessar as informações da maneira mais rápida possível. O DW armazena informações históricas de muitos anos e por isso deve ter uma grande capacidade de processamento e armazenamento dos dados que se encontram de duas maneiras, detalhados e resumidos.

Com base nestes conceitos podemos concluir que o DW não é um fim, mas sim um meio que os domínios dispõem para analisar informações históricas podendo utilizá-las para a melhoria dos processos de EIACBD. Os DWs são resumos de dados retirados de múltiplos sistemas de computação normalmente utilizados para vários anos e que continuam em operação. O DW é construído para que tais dados possam ser armazenados e acessados de forma que não sejam limitados por tabelas e linhas estritamente relacionais.

Os dados sob um DW podem ser compostos por um ou mais sistemas distintos e sempre estarão separados de qualquer outro sistema transacional, ou seja, deve existir um local físico onde os dados desses sistemas serão armazenados.

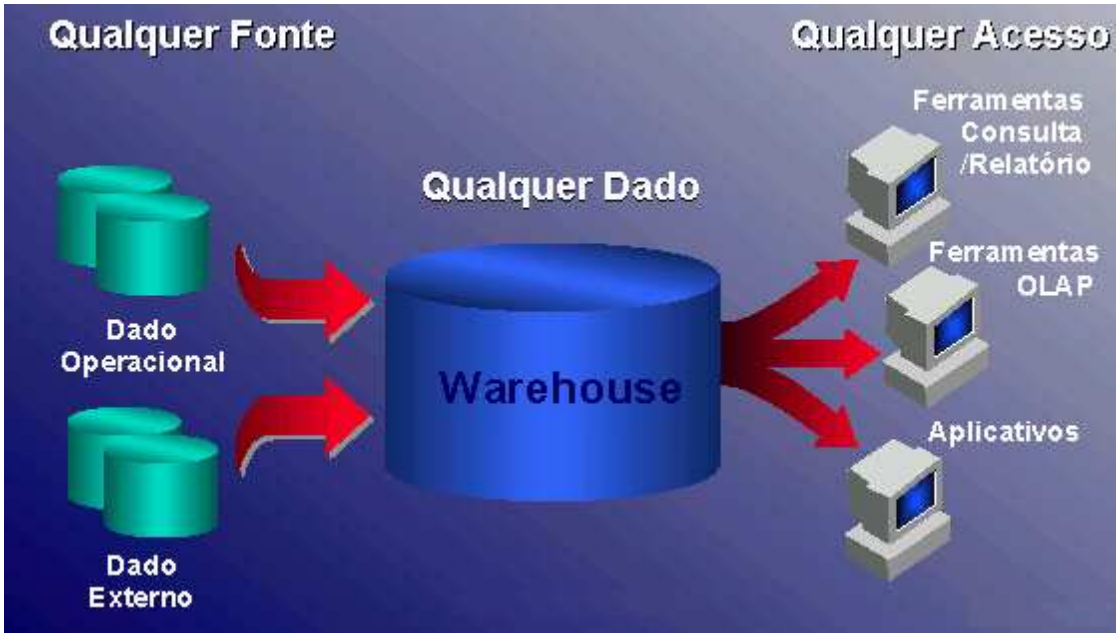

Figura II.3: *Componentes de um Data Warehouse* [Cazarini, 2002].

Sub-armazém de dados (*Data Marts*)

Algumas aplicações utilizam-se armazém de dados pequenos, denominados de *Data Marts*.. As principais vantagens dos *Data Marts* é a redução do tempo de implementação e o fator preço. Conforme [Inmon 1997], os *Data Marts* são subconjuntos de dados do domínio armazenados fisicamente em mais de um local,, geralmente divididos entre os departamentos da empresa. Existem diferentes alternativas de se implementar os *Data Marts*, porém a proposta original é aquela onde os *Data Marts* são desenvolvidos a partir de um DW central.

Nesta arquitetura, grupos de usuários acessam diretamente os *Data Marts* de seus respectivos departamentos. Somente aquelas análises que necessitam de uma visão global dos domínios são realizadas com o DW. Os *Data Marts* se diferenciam do DW pelos seguintes fatores [Inmon, 1997]:

- São personalizados: Atendem às necessidades de um departamento específico ou grupos de usuários;
- Menor volume de dados: Por atenderem a um único departamento, armazenam menor volume de dados;
- Histórico limitado: Os *Data Marts* raramente mantém o mesmo período histórico que o DW. Geralmente os DW mantém históricos que abrange de 5 a 10 anos, enquanto que os *Data Marts* devem optar em manter o mesmo período, porém com os dados em um nível maior de granularidade, ou um menor período, com os dados armazenados no mesmo nível de granularidade do DW;
- Dados resumidos: Os *Data Marts* geralmente não mantém os dados no mesmo nível de granularidade do DW, ou seja, os dados são, quase sempre, resumidos quando passam do DW para os *Data Marts*.

Um dos problemas dos *Data Marts* é o grande risco de desvio do modelo original, pois pode acontecer um crescimento desestruturado. Por ser muito utilizado e estar em constante aperfeiçoamento pode ocorrer a replicação das mesmas informações em vários locais o que dificultará uma futura integração dos *Data Marts* num único DW.

4 Processamento Analítico *On-Line* (*On-Line Analytical Processing* - OLAP)

Uma ferramenta de Processamento Analítico em Tempo Real (OLAP) é constituída por um conjunto de tecnologias especialmente projetadas para auxiliar o processo de consultas, análises e cálculos mais sofisticados nos dados corporativos, estejam estes armazenados em um DW ou não.

O OLAP permite aos seus usuários ganharem perspicácia nas consultas e análises dos dados, através de um acesso consistente, interativo e rápido em uma grande variedade de possíveis visões dos dados [Forsman, 1996]. Esta ferramenta transforma dados crús em informações que são facilmente compreendidas pelos usuários e refletem a real dimensionalidade dos negócios da empresa. Segundo este autor, a ferramenta aumenta a produtividade dos usuários e sua flexibilidade permitem maior auto-suficiência. No OLAP as respostas não são automáticas. Trata-se de um processo interativo, onde o usuário formula hipóteses, faz consultas, recebe informações, verifica um dado específico em profundidade e faz comparações [Carvalho, 1997]. Ajudam os usuários a sintetizarem as informações sobre o domínio, através de comparações, visões personalizadas, análises estatísticas, previsões e simulações. A maioria das ferramentas OLAP é implementada para ambientes multi-usuário e arquitetura cliente/servidor, o que proporciona respostas rápidas e consistentes às consultas iterativas executadas pelos usuários, independentemente da complexidade da consulta [Figueiredo, 1998].

A principal característica das ferramentas OLAP é permitir uma visão conceitual e multidimensional dos dados de um domínio [Figueiredo, 1998; Pendse, 1998a e Tyo, 1996]. Esta visão é muito mais útil para os usuários que a visão tradicional baseada em tabelas (modelo entidade-relacionamento), utilizada nos sistemas de processamento de transação (*On Line Transaction Processing* - OLTP). A visão multidimensional dos dados é um conceito que pode parecer algo completamente abstrato e irreal, porém é mais natural, fácil e intuitiva, permitindo a visão dos eventos do domínio em diferentes perspectivas, níveis de abstração ou visão lógica dos dados, assim, transformando os usuários em exploradores de informações.

Características das ferramentas OLAP

As doze regras para o processamento analítico on-line podem ser expressas como [Codd, 1995]:

1. Visão conceitual multidimensional.
2. Transparência.
3. Acessibilidade.
4. Informações de desempenho consistente.
5. Arquitetura Cliente-Servidor.
6. Dimensionalidade genérica.
7. Manipulação dinâmica de matrizes.
8. Suporte para multi-usuários.
9. Operações ilimitadas em referências cruzadas.
10. Manipulação intuitiva de dados
11. Consultas flexíveis.
12. Níveis de dimensões e agregações ilimitados.

OLAP Relacional (ROLAP)

O ROLAP é uma simulação da tecnologia OLAP, feita em BD relacionais. Possui a grande vantagem de não ter restrições no volume de armazenamento de dados pois utiliza a estrutura relacional, [Carvalho, 1997].

A principal vantagem de se adotar uma ferramenta ROLAP é a utilização de uma tecnologia estabelecida, de arquitetura aberta e padronizada, como é a relacional, beneficiando-se da diversidade de plataformas, escalonamento e paralelismo de *hardware* [Figueiredo, 1998]. Segundo esta autora, quanto às limitações, cita-se o pobre conjunto de funções para análises dimensionais e a inadequação ao esquema estrela (*star scheme*) para se realizar a manipulação dos dados.

As ferramentas ROLAP podem realizar o processamento dos dados para efetuar consultas, análises ou cálculos no modelo dimensional de duas formas, e também, dependendo da ferramenta e do suporte de *hardware*, gerar passos múltiplos e complexos em linguagem SQL [Gentia Software, 1998]:

- No próprio servidor OLAP ou carregando os dados necessários em outro equipamento, que pode ser outro servidor ou os equipamentos dos clientes no ambiente cliente/servidor.
- Isto ocorre porque a linguagem SQL não pode executar as atividades do modelo dimensional diretamente [Gentia Software, 1998]. Diferentes fornecedores de ferramentas ROLAP usam técnicas diferentes para superar esta dificuldade e alguns conseguiram alcançar um nível surpreendente de funções de atividades do modelo dimensional com a linguagem SQL, mas às custas de ter que realizar o processamento com consideráveis múltiplos passos e utilizando diversas tabelas temporárias.

OLAP Multidimensional (MOLAP)

A modelagem multidimensional é a técnica utilizada para se ter uma visão multidimensional dos dados, com que os dados são modelados em uma estrutura multidimensional conhecida por cubo. As dimensões do cubo representam os componentes dos negócios da empresa. A célula resultante da interseção das dimensões é chamada de medida e geralmente representa dados numéricos.

O MOLAP é uma classe de sistema que permite a execução de análises bastante sofisticadas, usando Bancos de Dados Multidimensionais (*Multidimensional Database* - MDB ou MDDB) [Figueiredo, 1998 e Pendse 1998b]. Na ferramenta MOLAP, os dados são mantidos em estruturas de dados do tipo "array" de maneira a prover um ótimo desempenho no acesso a qualquer dado. O tipo de acesso e agregação dos dados faz que esta ferramenta tenha um excelente desempenho. Além de ser rápida, outra grande vantagem é o conjunto de funções de análises que oferece [Figueiredo, 1998 e Pendse 1998b].

OLAP Híbrido (HOLAP)

Os fabricantes dos produtos OLAP também passaram a utilizar um sistema híbrido que usa o ROLAP e o MOLAP. Este novo tipo de OLAP é chamado de HOLAP. Desta forma, os produtos ROLAP estão incorporando MBD, para poder oferecer aos seus clientes as vantagens das duas tecnologias.

As ferramentas HOLAP são inteligentes e selecionam automaticamente a tecnologia mais adequada, de acordo com a atividade que será executada, proporcionando-lhe o máximo desempenho.

OLAP para *Web* (WOLAP)

Já existe uma migração da tecnologia OLAP para o ambiente da Internet, a nova versão da ferramenta está sendo chamada de "WebOLAP" ou "WOLAP". As facilidades são: a possibilidade de plataformas independentes para dar suporte a usuários distantes, aplicações de recursos de grupo (*groupware*), facilidade de aprendizado e de manutenção [Carickhoff, 1997]. As dificuldades são: as limitações dos recursos da Internet, as interfaces e as funcionalidades, quando comparados com o ambiente cliente/servidor.

5 Visualização de dados

As técnicas e ferramentas para visualização de dados são "instrumentos" necessários ao processo de extração de conhecimento. Elas podem ser utilizadas durante a execução das etapas do processo de extração de conhecimento melhorando a compreensão dos resultados obtidos ate a comunicação entre os usuários [Rezende and Pugliesi, 1998].

As técnicas de visualização de dados estimulam o sentido da percepção e inteligência humana, incrementando a capacidade de compreensão de novos padrões ou modelos nos dados. Poderosas ferramentas de visualização que consigam gerar visualizações em diversas metáforas (arvores, regras, gráficos 3D/2D, espectro) combinadas com técnicas de EIACBD podem melhorar em muito o processo KDD [Oliveira and Rezende, 1998].

A visualização de dados é a simples apresentação dos dados de forma gráfica, ela torna possível ao analista se aprofundar o conhecimento intuitivo do dado e assim poder trabalhar melhor com a garimpagem de dados. A garimpagem de dados permite ao analista enfocar certos padrões e tendências e explorar os dados com profundidade usando visualização. A própria visualização de dados pode ser esmagada pelo volume de dados em

uma base de dados, mas em conjunto com garimpagem de dados pode ajudar a exploração dos dados [Dilly, 1995].

Esta técnica permite ao analista aprofundar a compreensão intuitiva dos dados pela apresentação de um gráfico ao usuário que percebe mais rapidamente o processo. Por exemplo, uma imagem gráfica mostrando quatro variáveis apresentando a quantidade de informação de cada um através de grupos de dados dispostos como picos ou vales.

A visualização dos dados fornece ao usuário algumas habilidades [Berson, 1997]:

a) Habilidade de comparar os dados:
b) Habilidade de controlar escalas (olhar a partir de um certo nível detalhe ou detalhá-lo ainda mais);
c) Habilidade de mapear a visualização inversa para detalhar o dado que criou essa visualização;
d) Habilidade de filtrar dados para olhar somente seus subconjuntos ou sub-regiões em um determinado tempo.

6 Estatística

A Estatística é a área da matemática que: estuda, coleta, organiza e interpreta dados numéricos, especialmente na análise de características da população de dados, por inferências a partir de amostras. As técnicas de estatística possuem muita importância dentro do processo de EICBD e vários métodos utilizados em EIACBD tiveram suas origens da estatística [Elder and Pregibon, 1996 e Glymour et al., 1998].

As áreas de extração de conhecimento e de estatística estão fortemente relacionadas. As duas disciplinas têm como objetivo encontrar padrões e regularidades nos dados.

As técnicas estatísticas auxiliam diversas etapas do processo de EIACBD, além de grande parte dos algoritmos de aprendizado de maquina fazerem uso de mecanismos disponíveis na estatística para realizar a descoberta de padrões, calcular aproximações, medias, taxas de erros e desvios [Elder and Pregibon, 1996 e Rocha, 1999].

7 Agentes de *Software* (AS)

A natureza dinâmica e distribuída de algumas aplicações exige que os aplicativos de *software* não apenas respondam as requisições de informação, mas também que estes consigam antecipar, adaptar e buscar novas formas para auxiliar o usuário. Os projetos de sistemas computacionais tendem a tornar-se cada vez mais complexos uma vez que estes precisam, por exemplo, especificar interfaces entre diferentes aplicativos, possuir características distribuídas e de persistência. Estes sistemas também devem auxiliar a execução e gerenciamento de programas distribuídos [Bradshaw 1997].

Uma ferramenta típica de AS é integrada por diferentes áreas, como Inteligência artificial (IA), Sistemas Distribuídos (SD), Interface Humano Computador (IHC), etc. Têm como principais objetivos: 1) atender os novos requisitos exigidos por determinadas aplicações, 2) facilitar a interação usuário/maquina e 3) a construção de Sistemas Inteligentes Distribuídos (SID).

Os AS são considerados como sendo resultante das pesquisas desenvolvidas pelas comunidades de agentes inteligentes e de agentes de *software*. A primeira está envolvida com a construção de sistemas inteligentes capazes de reagir a eventos tanto em ambientes do mundo real (robôs) quanto do mundo virtual (programas). A segunda comunidade está preocupada com a propriedade de se mover dentro de uma rede de computadores distribuindo informações, para outros agentes que possam ser, tanto humanos quanto aplicações de *software* [Hendler, 1996].

Os agentes inteligentes podem ser programas de IA cuja finalidade é agir em diversos ambientes de importância para os humanos, sendo divididos em duas categorias: agentes físicos e agentes de informação [Mães, 1994 e Hendler, 1996]. Os agentes físicos trabalham em um ambiente onde seja difícil colocar um ser humano (espaço) ou que seja perigoso (núcleo de um reator nuclear). Os agentes de informação (*softbots* ou *software robots*), atuam em um mundo virtual onde exista uma grande quantidade de informações espalhadas por diversos computadores (internet).

Esta ferramenta é pouco utilizada, têm uma perspectiva promissora na resolução de problemas complexos na *web mining*.

8 Sistemas Paralelos (SP)

O desenvolvimento de sistemas paralelos de BD tem apresentado tendências de sucesso [DeWitt and Gray, 1992]. Os SP dependem muito pouco de componentes específicos e podem ser construídos a partir de processadores, memórias e dispositivos de armazenagem convencionais. O armazém de dados de tamanho grande, precisa de hardware de BD com processamento paralelo [Fayyad et al., 1996].

O processamento paralelo segura a chave para extrair o potencial máximo de extração de conhecimento, e é um dos meios principais para efetivar eficazmente determinados processos [Stolfo and Chan, 1995]. A união das tecnologias de sistemas paralelos e extração de conhecimento é muito desejável. Podem produzir resultados inovadores e ter profundo impacto no futuro da computação. A crescente disponibilidade de *hardware* e *software* paralelo aumenta a velocidade do processo de extração de conhecimento.

Os conjuntos de dados gigantescos disponíveis e sua alta dimensionalidade demandam aplicações de extração de conhecimento de grande escala e portanto mais recursos computacionais. A computação paralela com seu alto-desempenho está se tornando um componente essencial de solução nestes casos. Além disso, a qualidade dos resultados de extração de conhecimento freqüentemente depende diretamente do volume de recursos computacionais disponível. As aplicações de extração de conhecimento serão certamente serão as principais usuárias dos sistemas futuros de supercomputação [Mahesh et al., 1998].

Capítulo III

O processo de extração de conhecimento

1 Introdução

O processo de EIACBD é interativo e iterativo. A Figura III.1 apresenta um esquema da visão geral da seqüência dos processos envolvidos.

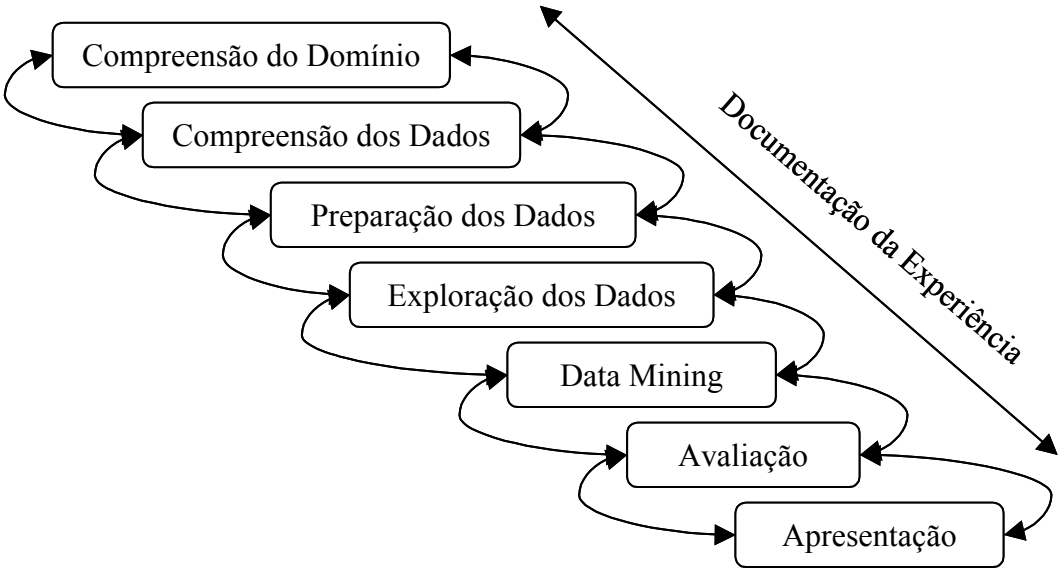

Figura III.1: *Esquema do processo de extração de conhecimento segundo* [Reinartz, 1999]

A Figura III.2 apresenta o esquema seqüencial das principais tarefas nas três etapas essenciais: pré-processamento, processamento e pós–processamento.. Na etapa de pré-processamento realizam-se as tarefas de: preparação de dados, integração de dados, seleção de dados, limpeza de dados, e transformação de dados. A ordem e seqüência das tarefas nesta etapa são modeladas com o objetivo de formar um conjunto de dados preparados para aplicar o algoritmo de garimpagem de dados (*data mining*). Na etapa de processamento aplica-se o algoritmo de garimpagem aos dados preparados extraindo-se os padrões. Na

etapa de pós-processamento realizam-se a: obtenção, interpretação, e avaliação de padrões e apresentação de conhecimento. Na seqüência do processo a ordem das tarefas difere geralmente para os objetivos propostos, podendo ocorrer seqüências recorrentes (*loops*) entre tarefas ou exoneração de algumas tarefas.

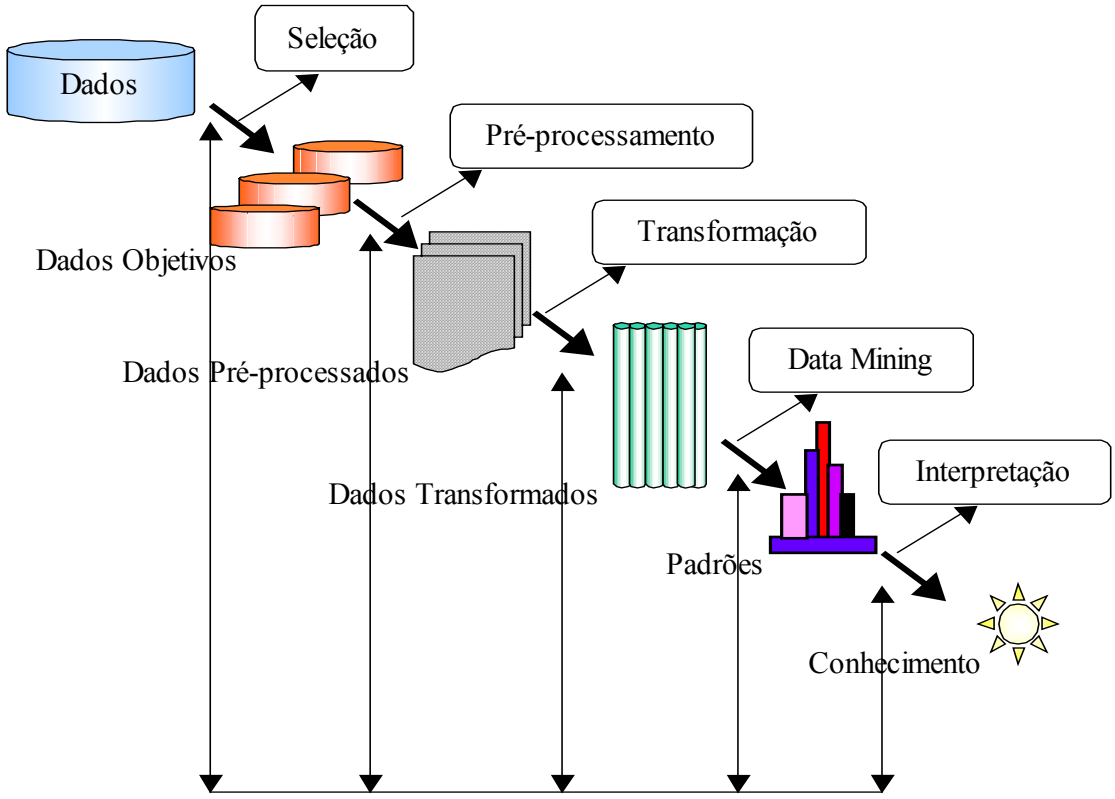

Figura III.2: *Principais tarefas no processo de EIACBD*

2 Definição e compreensão do domínio de aplicação

Nesta etapa preliminar destaca-se o conhecimento prévio dos objetivos de um usuário final e o conhecimento do domínio. Fazendo uma estratégia de relação entre objetivos do usuário e o conjunto de dados do domínio se obtém a descrição do modelo do processo. A Figura III.3 apresenta o conjunto de dados selecionados, relacionados à solução do problema, que será utilizado como matéria prima a ser manipulada, garimpada por um algoritmo, que pode gerar padrões para a solução do problema.

Compreender o domínio do problema é importante acompanhar todos os processos de extração de conhecimento e avaliar os sub-resultados obtidos em cada tarefa, e assim prosseguir ou corrigir o processo.

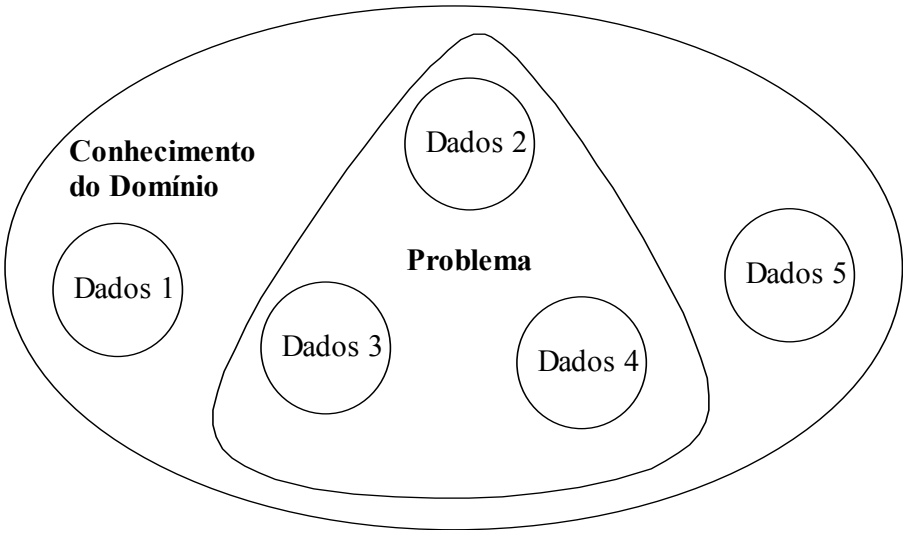

Figura III.3: *Relação de conhecimento de domínio, dados e o problema* [George, 1997]

Nesta etapa pré-liminar são realizadas:

- A compreensão do domínio (conhecimento da área)
- O conhecimento dos dados relevantes
- A definição do problema em termos de domínio (tipo de conhecimento a ser descoberto)
- A definição de objetivos da aplicação e metas especificas (propósito da aplicação)
- A definição do modelo da solução desejada (técnica a serem utilizadas, aplicação a serem adaptadas)

3 Pré-Processamento

Correspondem as atividades que visam gerar um grupo de dados representativos convenientemente organizados e estruturados para ser garimpados pelo algoritmo selecionado (*software data mining*).

3.1 Integração de dados

Múltiplas fontes de dados heterogêneos podem ser integradas em um único BD. Diferentes sistemas computacionais provavelmente encontrarão diferentes formas de armazenar seus dados, neste caso será necessário homogeneizar e integrar as BD. O processo de homogeneização dos dados é complexo na medida em que depende da adoção de padrões de tratamento para dados de diferentes fontes (sistemas de *software* ou de *hardware*).

3.2 Seleção de dados

Identifica um subconjunto de atributos onde será garimpado, facilitando o trabalho dos algoritmos de garimpagem. Nem todos os podem interessar ao processo de garimpo, de acordo com as necessidades dos sistemas de informação, conforme esquematizado na Figura III.3. As técnicas de garimpagem de dados dependem muito da seleção adequada dos atributos relevantes. Para ter um melhor subconjunto de atributos utilizam-se técnicas de seleção de atributos como técnicas baseadas na teoria da informação [*Relieff*, GD], técnicas de determinação de provas (*Rough Sets*), ou outras.

3.3 Limpeza de dados

Deve se projetar uma estratégia adequada de manipulação de dados ruidosos, errôneos, perdidos ou irrelevantes.

As organizações que implementam seus BD através de ferramentas específicas do tipo SGBD, e que seguem todas as regras e padrões estabelecidos por elas, certamente terão pouco trabalho de limpeza dos dados, restando apenas a eliminação de redundâncias e ajustes em atributos com valores fora do padrão (*outliers*)..

3.4 Transformação de dados

Os dados são transformados ou consolidados em formatos apropriados para garimpar. A transformação ou conversão de variáveis para casos ou de casos para variáveis podem ser necessários para uma melhor manipulação pela técnica de garimpagem de dados.

3.5 Discretização de dados

Dividem-se os valores contínuos dos atributos (inteiros ou reais) numa lista de intervalos representados por um código, convertendo-se valores contínuos em discretos

Cada intervalo resulta num valor discreto do atributo. Por exemplo, mostra-se uma possível discretização para o atributo IDADE: {0...18}→ faixa 1; {19...25} →faixa 2; {26...40}→ faixa 3 e assim por diante. Nesse exemplo, os valores contínuos das idades foram discretizadas em três faixas.

As vantagens da discretização dos atributos é a melhora da compreensão do conhecimento descoberto; redução do tempo de processamento pelo algoritmo garimpador, diminuição do espaço de busca; facilitação do algoritmo de tomada de decisões globais já que os valores dos atributos foram englobados em faixas. Porém a discretização reduz a medida de qualidade de um conhecimento descoberto, podendo assim, perder detalhes relevantes sobre as informações extraídas.

3.6 Agregação de dados

Consiste em agregar aos dados existentes mais informações de modo que essas agregações contribuam no processo de descoberta de conhecimento. Essas informações geralmente podem ser conhecidas.

3.7 Derivação de dados

Adicionam-se novos dados derivados por uma operação ou de séries de operações de dados existentes na tabela de dados. As operações de derivação também podem ser baseadas em conhecimentos passados ou referências.

3.8 Redução de dados

Reduz-se o número de variáveis a considerar. Normalmente é aplicado a domínios onde a meta é se agregar ou amalgamar a informação contida em um conjunto de dados de grande tamanho para um conjunto de dados controlável. Os métodos de redução de dados podem incluir tabulação simples, agregação (com estatística descritiva) ou técnicas mais sofisticadas como agrupação.

4 Processamento

Os dados são analisados por um algoritmo e transformados em informações (resultados, padrões) úteis que serão avaliados no processo seguinte.

4.1 Escolha da função de garimpagem de dados

Decide-se a função que o algoritmo de garimpagem realizará na tabela pré-processada, conforme o objetivo proposto. As funções estão descritas no Capitulo IV, parágrafo 2.

4.2 Seleção de algoritmo de garimpagem de dados

Seleciona-se o método apropriado a ser usado para encontrar padrões nos dados. Isto inclui decidir quais modelos e parâmetros podem ser apropriados para emparelhar a um método de garimpagem de dados particular com os critérios globais do processo de EIACBD. As técnicas garimpagem de dados estão descritas no Capitulo IV, parágrafo 3.

4.3 Garimpagem de Dados (*Data Mining*)

Executa-se a análise dos dados armazenados através de um programa computacional (software *Data Mining*) baseado em algoritmos de inteligência artificial e estatística para analisar os dados e encontrar padrões de interesse, na forma de representação particular ou de um conjunto de representações.

5 Pós – Processamento

São feitas a avaliação e interpretação dos padrões, para serem representados em forma de conhecimento compreensível e confiável ao usuário, e para serem incorporadas ao conhecimento anterior.

5.1 Interpretação e avaliação dos resultados

São feitas a avaliação e interpretação dos padrões, para identificar os padrões verdadeiramente interessantes que representam o conhecimento baseado em algumas medidas de interesse.

Depois desta tarefa é possível regressar aos passos anteriores que envolvem a repetição de processos, talvez com outros dados, outros algoritmos ou outras estratégias.

5.2 Apresentação e incorporação de conhecimento

São usadas técnicas de representação e visualização de conhecimento para apresentar o conhecimento de forma compreensível ao usuário.

Finalmente pode ser incorporado o conhecimento ao sistema para melhorar o conhecimento anterior, o que pode incluir a resolução de conflitos potenciais com o conhecimento existente.

Capítulo IV

Garimpagem de dados

1 Definição

A garimpagem de dados é a analise inteligente e automática de dados para descobrir padrões ou regularidades em grandes conjuntos de dados, através de técnicas que envolvem métodos matemáticos, algoritmos baseados em conceitos biológicos, processo lingüístico, e heurísticas

Os pesquisadores da área de inteligência artificial e estatística, e os físicos que trabalham no domínio de dinâmicas não lineares contribuíram para o desenvolvimento de novos conjuntos de métodos lógicos. Entretanto, estes métodos exigem máquinas de alto desempenho e por isto somente recentemente puderam ser implementados, despertando interesse também de pesquisadores de outras áreas. A garimpagem de dados utiliza esses métodos, para que, a partir de um conjunto de dados, seja possível descobrir uma representação otimizada da sua estrutura. Por exemplo, temos abaixo um exemplo de padrão e modelo.

Padrão $f(x) = 3x^2 + x$

Modelo $f(x) = ax^2 + bx$

2 Funções

Dependendo dos objetivos da aplicação, a garimpagem de dados pode realizar múltiplas funções. Cada função tem como base um conjunto de algoritmos que são

utilizados na extração de relações relevantes em um, ou vários conjuntos de dados. Descreveremos estas funções nos próximos parágrafos.

2.1 Associação

As associações visam determinar relacionamentos entre conjuntos de itens ou regras de associação, tais como: "90% das pessoas que compram pão também compram leite". Exemplos de uso de regras de associação são: projetos de catálogos, segmentação de clientela baseada em padrões de compra, etc. Neste caso os itens podem ser agrupados, por exemplo, em catálogos ou espaços físicos, de modo a induzir a venda dos artigos relacionados.

A associação identifica afinidades entre itens de um subconjunto de dados. Essas afinidades são expressas na forma de regras: “85% de todos os registros que contém os itens A, B, e C também contém D e E”. A porcentagem de ocorrência (85 no caso) representa o fator de confiança da regra, e costuma ser usado para eliminar tendências fracas, mantendo apenas as regras mais fortes. Dependências funcionais podem ser vistas como regras de associação [Aldana, 2000].

A associação trata-se de uma função tipicamente endereçada à análise de mercado, onde o objetivo é encontrar tendências de associação dentro de um grande número de registros de compras, por exemplo, expressas como transações. Essas tendências de associação podem ajudar a entender e explorar padrões de compra naturais, e podem ser usadas para ajustar mostruários, modificar prateleiras ou propagandas, e introduzir atividades promocionais específicas.

2.2 Classificação

Classificação é uma função que consiste na aplicação de um conjunto de exemplos pré-classificados para desenvolver um modelo capaz de classificar uma população maior de registros. A classificação é uma função de previsão que pode ser usada para encontrar um modelo que classifique um item de dado entre várias classes previamente definidas. Os algoritmos de classificação incluem: árvores de decisão, estatística, redes neurais, algoritmos genéticos, etc., e começam com um treinamento a partir de “transações-exemplo”. O algoritmo classificador usa estes exemplos para determinar um conjunto de

parâmetros, codificados em um modelo, que será mais tarde utilizado para a discriminação dos dados restantes [Aurélio et al., 1999].

Uma vez que o algoritmo classificador foi desenvolvido de forma eficiente, ele será usado de forma preditiva para classificar novos registros naquelas mesmas classes pré-definidas. Por exemplo, um classificador pode ser treinado para identificar empréstimos arriscados, a partir das informações cadastrais de milhares de interessados, e usado como suporte a decisão no momento de conceder um empréstimo a alguém.

Mais informações sobre a função de classificação podem ser encontrados em [Metha et al., 1993], [Shafer et al., 1996]

2.3 Regressão

Esta função que mapeia um item de dado numérico, possibilitando a previsão das variáveis dos valores reais. As aplicações de regressão são muitas, por exemplo: Prever a quantidade de biomassa presente em uma floresta através de sensores, estimar a probabilidade de sobrevivência de um paciente baseando-se em um conjunto de testes de diagnóstico, prever a demanda do consumidor para um novo produto, e prever as séries de tempo onde as variáveis de ingresso podem ser intervalos de irregulares.

2.4 Sumarização e caracterização

A sumarização visa obter uma descrição compacta de um conjunto de dados. As técnicas de resumo são freqüentemente aplicadas em analise de exploração de dados e geração automática de relatórios. A sumarização geralmente não é uma função usada para resolver problemas, mas tende achar a característica do conjunto de dados e identificar qualquer anomalia. Porém, pode ser usada para comparar as freqüências de atributos categóricos entre dois conjuntos de dados.

A caracterização descreve as qualidades relevantes a partir de analises quantitativas, propiciando uma descrição compacta do conjunto, podendo generalizar, sumarizar e possivelmente contrastar características de dados. As funções de sumarização e caracterização tendem a ser complementares.

2.5 Analises de desvio e tendência

São funções focalizadas em descobrir mudanças mais significativas nos dados através de medidas anteriores ou de valores normativos. O objetivo é modelar o processo, gerando uma seqüência ou relatando tendências do processo ao longo do tempo, buscando e caracterizando a tendência de evolução, padrões seqüenciais, e dados desviados.

Os desvios cobrem uma ampla variedade de padrões potencialmente interessantes como: dados anômalos que não se ajustam a uma classe de padrão, dados fora do grupo que aparecem na franja de outros padrões, classes que exibem valores comuns, mas diferentes na sua derivação (pai e filho), valor ou conjunto de valores variáveis em um período de tempo com relação a outro período passado, discrepâncias entre valores observados e valores esperados previstos por um modelo [Matheus, 1993].

2.6 Agregação e segmentação (*Clustering*)

É a função descritiva cujo objetivo é identificar um conjunto de categorias finitas ou agrupamentos naturais. Estas categorias podem ser mutuamente exclusivas ou podem consistir em uma representação hierárquica.

A proposta de segmentação é basicamente endereçada a problemas de agrupação, na qual se faz um "corte" de um grande número de atributos em um pequeno conjunto de grupos ou de segmentos relativos. A segmentação é realizada automaticamente por algoritmos que identificam características em comum e particionam o espaço *n*-dimensional definido pelos atributos.

Muitas vezes a segmentação é uma das primeiras tarefas dentro de um processo de garimpagem de dados para identificar grupos de registros correlacionados, que serão usados como ponto de partida para seguintes explorações. O exemplo clássico é o de segmentação demográfica, que serve de início para uma determinação características de um grupo social, visando desde hábitos de compras, utilização de meios de transporte, etc.

2.7 Modelagem de dependência

É a função que visa encontrar um modelo para descrever dependências significantes entre variáveis ou atributos. Os modelos de dependências existem em dois níveis: o nível

estrutural do modelo, onde são especificadas (em forma gráfica) quais variáveis são localmente dependentes uma das outras, e o nível quantitativo do modelo, onde são especificadas as forças das dependências através de algumas escalas numéricas. Por exemplo, a rede de dependência probabilística usa independência condicional para especificar o aspecto estrutural do modelo e probabilidades ou correlações para especificar as forças da dependência. Redes de dependência probabilística estão sendo aplicadas quanto o desenvolvimento de sistemas diagnóstico-médico probabilístico em BD, recuperação de informação, e modelagem do genoma humano.

3 Técnicas

As técnicas de garimpagem de dados podem ser divididas em três componentes principais:

a) Representação do modelo: A linguagem de representação do modelo é usada para descrever os padrões.
b) Avaliação do modelo: Os critérios da avaliação do modelo são declarações quantitativas ou funções de ajuste, de como um padrão particular (um modelo e seus parâmetros) conhecendo as metas do processo de extração de conhecimento.
c) Método de busca: O método de busca consiste de dois processos; busca de parâmetro e busca de modelo. Na busca de parâmetro, o algoritmo procura os parâmetros que otimizam os critérios de avaliação de modelo desenhado. A busca do modelo acontece de forma seqüencial em anel (*loops*) acima dos métodos de busca de parâmetro.

A Tabela IV.1 apresenta-se uma classificação das técnicas de garimpagem segundo as funções que realizam.

Tabela IV.1: *Funções das técnicas*

Funções	Técnicas
Função de Segmentação	• Shepherd • Fuzzy C-Mean • Redes de Agrupamento (*Kohonen)*
Função de Classificação	• Arvore de Decisão (ID3, C4.5, SLIQ) • Redes Neurais Artificiais • Algoritmos Genéticos • Conjunto de Regras • *K-Nearest Neighbor*
Função de Regressão	• Regressão Linear • Regressão Quadrática • Regressão de Vetor de Suporte • Regressão Logística • Analises de Tendência com Regressão Não-Linear • Redes Neurais Artificiais • *K-Nearest Neighbor*

3.1 Redes Neurais Artificiais (RNA)

Uma RNA é uma técnica computacional que constrói modelos matemáticos, emula sistemas neurais ou biológicos, e tem capacidades de: aprendizagem, generalização, associação e abstração. As RNAs tentam aprender padrões diretamente dos dados através de um processo iterativo de apresentações dos dados (experiências), dessa forma, uma RNA procura por relacionamentos construir automaticamente modelos e corrigir seus próprios erros. Uma RNA simples possui nodos (sistemas de neurônios) e conexões ponderadas (sinapses, pesos). Numa RNA os nodos são arrumados em camadas, com conexões entre elas [Holland, 1992 e Dhar and Stein, 1997].

Para entender como uma RNA aprende é necessário saber como os pesos da rede afetam sua saída. O aprendizado de uma RNA envolve os ajustes dos pesos. A Figura IV.1 mostra o esquema de um neurônio artificial criado a partir do modelo simplificado do neurônio biológico [Von Zuben e De Castro, 2003]. O neurônio artificial possui várias entradas, que podem ser estímulos do sistema ou saídas de outros neurônios [Perelmuter, 1996].

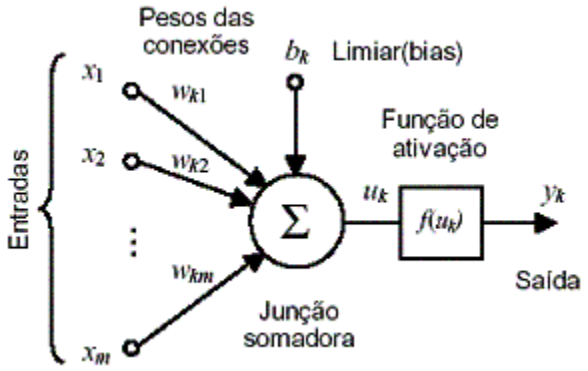

Figura IV.1: *Neurônio artificial segundo* [Von Zuben e De Castro, 2003]

O neurônio artificial é dividido em duas seções funcionais. A primeira seção combina todas as entradas que alimenta o neurônio. Essa etapa indica como as entradas serão computadas (regra de propagação). A segunda seção recebe esse valor e faz um cálculo determinando o grau de importância da soma ponderada utilizando uma função de transferência, ou função de ativação (sigmóide e tangente hiperbólica). Essa função determina que grau uma soma causará uma excitação ou inibição do neurônio, pois fornecem as características de não linearidade para uma RNA [Aurélio et al., 1999].

Uma RNA ajusta seus pesos na fase de treinamento. Sendo fornecido um dado de observação, este é processado, e será produzida uma resposta. O resultado fornecido é comparado com uma saída desejada (saída correta). Se a rede acerta essa saída, então ela não faz nada, entretanto se o resultado não está correto, ocorrem ajuste dos pesos de modo que o erro seja minimizado.

A Figura IV.2 apresenta uma representação conceitual da arquitetura de uma RNA simples. Os círculos representam os nodos e as linhas representam os pesos das conexões.

Por convenção, a camada que recebe os dados é chamada de camada de entrada e a camada que mostra o resultado é chamada de camada de saída, e as camadas intermédias estão escondidas, como também se realiza o processamento interno.

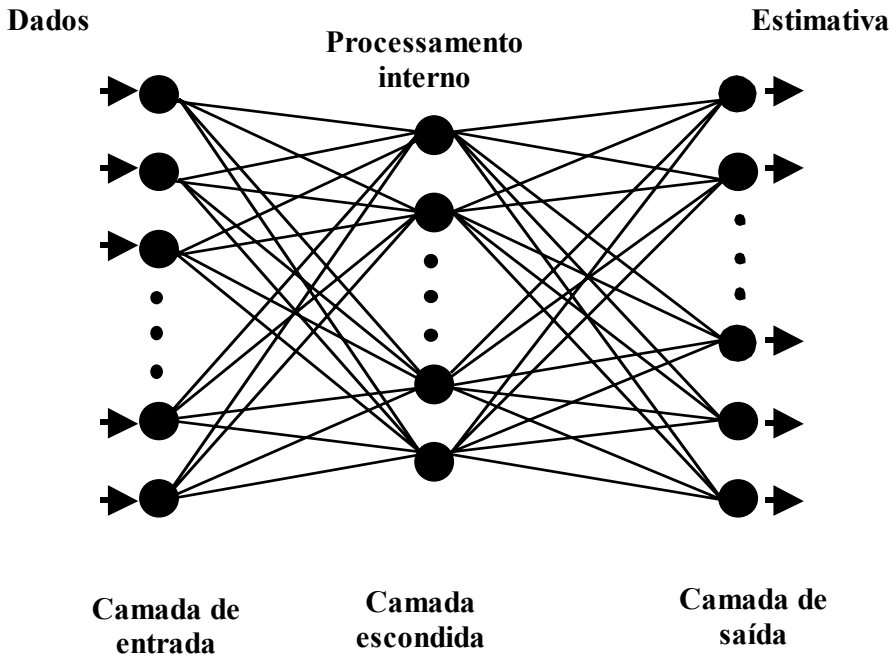

Figura IV.2: *Arquitetura de uma rede neural artificial*

As topologias mais comuns das RNAs são as de múltiplas camadas realimentadas (*feed-forward*) e as redes recorrentes. O aprendizado de uma RNA pode ser dividido em 3 grupos:

a) Sem treinamento: Os valores dos pesos sinápticos são estabelecidos a priori (ajustados em um único passo), por exemplo, Redes de Hopfield [Dayhoff, 1990];
b) Treinamento supervisionado: A rede é treinada através do fornecimento dos valores de entrada e dos seus respectivos valores de saída desejados (procura minimizar o erro médio quadrado);
c) Treinamento não-supervisionado: O sistema extrai as características dos dados fornecidos, agrupando-os em classes (*clusters*).

A Figura IV.3 apresenta o modelo de aplicação de uma RNA na extração de conhecimento, onde a arquitetura de uma RNA recebe uma tupla (atributos) pela primeira camada (camada de entrada)

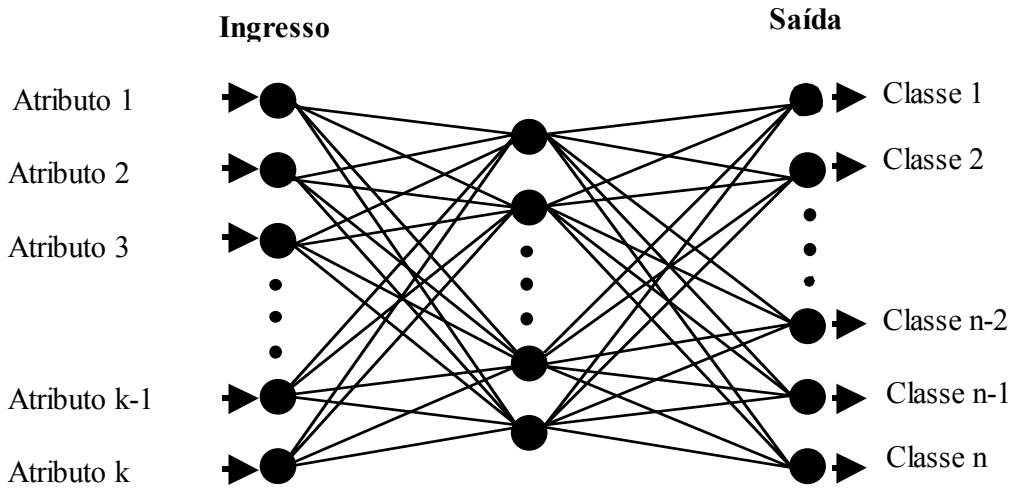

Figura IV.3: *Modelo de uma rede neural artificial para garimpagem de dados*

Fatores positivos da RNA:

- Possuem alta capacidade de adaptação,
- Tem tolerância à falhas,
- Possuem capacidade de resolver problemas práticos sem a necessidade da definição de listas, de regras ou de modelos precisos.

Fatores negativos da RNA:

- Precisam de computadores de alta performance,
- A interpretação dos resultados é complexa.

3.2 Analise de Cesto de Compras (ACC) - (*Market Basket Analysis*)

Pode ser entendida como a análise das transações produzidas por um conjunto de indivíduos, sobre o qual se possa extrair algum conhecimento (como se comportam) [Rodrigues, 2001].

Para formar uma imagem do que é ACC, imaginemos, um carrinho de compras com produtos adquiridos, por alguém. Um cesto informa-nos sobre um cliente, os clientes não são todos iguais. Cada cliente compra um conjunto diferente durante uma semana. A ACC utiliza a informação sobre o que os clientes adquirem, e fornece critérios que indiquem quem são eles, e porque fazem determinadas aquisições. A ACC fornece critérios sobre a mercadoria, dizendo quais produtos tendem a ser comprados, ela pode sugerir novas disposições na loja e determinar quais produtos devem ser colocados mais apropriadamente.

Fatores positivos do ACC:

- Os resultados da análise de compras estão associados a regras expressados em linguagem.
- A análise de cesto de compras pode avaliar quantidades variáveis de dados.
- Os cálculos necessários para aplicar a análise do cesto de compras são relativamente mais simples.

Fatores negativos do ACC:

- Os cálculos necessários para gerar regras de associação aumentam exponencialmente com o numero de itens e com a complexidade das regras em consideração.
- É uma técnica especializada para itens em transação, nem todos os problemas ajustam-se a essa descrição.
- Dificuldade na determinação do conjunto certo de itens a ser usados na análise.
- A ACC funciona melhor quando todos os itens têm aproximadamente a mesma freqüência nos dados.

3.3 Algoritmo Genético (AG)

O Algoritmo Genético (AG) é inspirado na teoria biológica, no principio darwiniano da evolução das espécies e na genética aplicada a problemas complexos de otimização [Goldberg, 1989]. São algoritmos probabilísticos que fornecem um mecanismo de busca paralela e adaptativa baseados no principio de sobrevivência dos mais aptos e na

reprodução. Constituem uma técnica estocástica e probabilística de busca e otimização, altamente paralela.

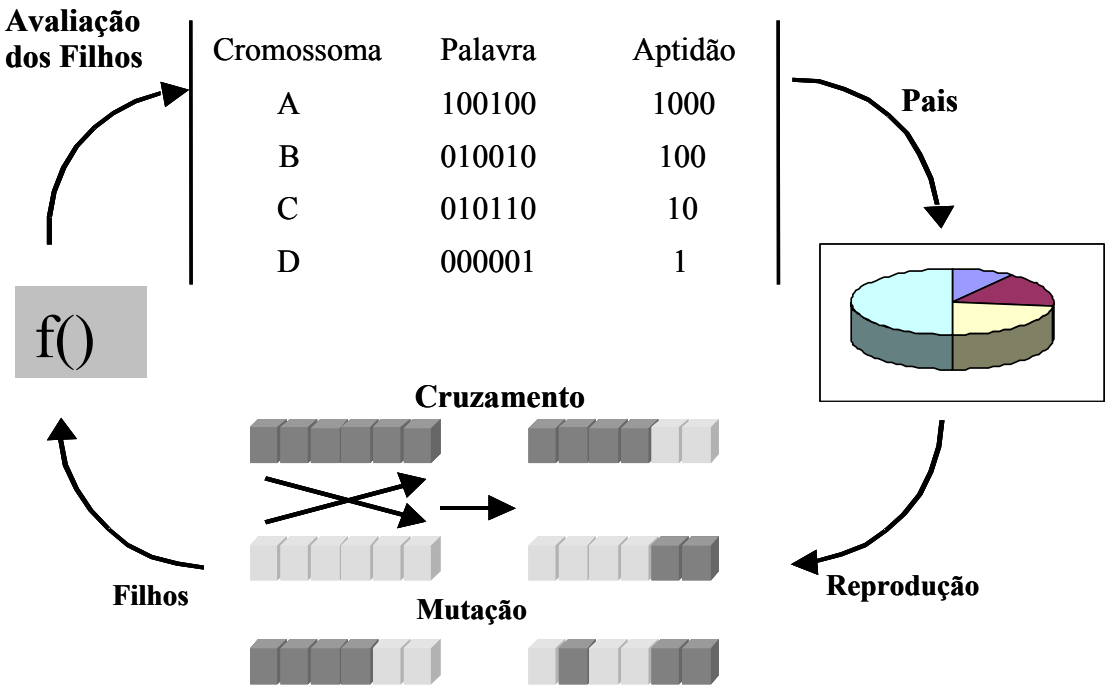

Figura IV.4: *Ciclo básico do algoritmo genético*

Os AGs tornaram-se particularmente interessantes devido ao fato de não ser necessário descrever como encontrar uma boa solução, tais soluções podem ser encontradas de forma paralela, avaliando-se e percebendo-se em que direções devem estar localizadas as melhores soluções, ou soluções potenciais.

Os problemas de otimização tipicamente envolvem três componentes: variáveis, restrições e objetivos. As variáveis descrevem os vários aspectos do problema. As restrições monitoram os valores que as variáveis podem ter. As funções objetivas são utilizadas para avaliar a solução, geralmente envolvem a maximização ou a minimização de algum recurso, e medem a qualidade de uma regra gerada.

As variáveis, as restrições e as funções objetivas, descritas em um problema de otimização definem a geografia básica do espaço de busca, e determinam que técnicas podem ser usadas. Técnicas baseadas em heurísticas como AG não podem garantir a

solução ótima, porem conseguem soluções aproximadas (aceitáveis ou sub-ótimas). Os AG são aplicados em problemas complexos com muitas variáveis e restrições ou com grandes espaços de busca.

A menor unidade de um AG é chamada de gene. Um gene representa uma unidade de informação do domínio do problema, ou no âmbito de garimpagem de dados, um valor de um atributo. Uma serie de genes, ou um cromossoma representa uma possível solução para o problema (uma regra candidata).

As soluções ou cromossomas são avaliadas por uma função capaz de medir através de um valor as qualidades dessas soluções.

Para que um cromossoma seja avaliado é necessário converter o cromossoma numa solução para o problema, isto é, decodificar os cromossomas.

Uma vez que o cromossoma foi decodificado, o modulo de avaliação determina a quantidade de soluções boas ou ruins.

O AG emprega uma população de cromossomas, executando assim, uma busca de forma paralela. Após a criação da população inicial (aleatória), tem inicio num processo iterativo de refinamento ou evolução de soluções inicias. O AG cria novas soluções através da combinação e refinamento das informações dos cromossomos usando três operações: seleção, cruzamento e mutação, essas operações produzem novas soluções que formam uma nova população. Cada nova população e chamada de geração. Os três operadores agem no refinamento das soluções a quando combinados com os módulos de decodificação e avaliação, podem resolver uma vasta variedade de problemas

Durante a seleção o AG escolhe os cromossomas, privilegiando aqueles com maiores aptidões para permanecer e se multiplicar na população. Durante o cruzamento o AG utiliza as informações dos cromossomas selecionados, formando outros cromossomas, enquanto que durante a mutação o AG busca eventualmente melhora-las.

Uma forma comum de seleção é uma onde cada cromossomas tem a probabilidade de permanecer na próxima geração proporcional a sua aptidão (método de roleta). Cromossomas com maiores aptidões possuem maior espaço na roleta e conseqüentemente possuem maiores chances de serem escolhidos para o cruzamento e a mutação.

Durante o cruzamento, dois cromossomas basicamente trocam algumas de suas informações gene a gene. Ou seja, o cruzamento permite a combinação dos elementos de

uma solução com os de outra. O operador de cruzamento intercambia as informações, porem não é capaz de gerar diversidade numa população, essa diversidade é obtida através do operador de mutação. Sua função é alterar o valor de um gene por um outro valor qualquer no domínio da aplicação.

A mutação permite explorar novas áreas a procura de melhores soluções.

Na aplicação de garimpagem de dados pode ser caracterizado por numa regra de associação [Agrawal, 1993]. Esta regra representada num cromossomo, é da forma: "*Se (A1 e A2 e A3 e ...e An) então P"*, onde os itens ou atributos do BD {*A1,A2,A3,...,An*} estão representados simbolicamente como condições para que a conclusão da regra caracterizada por (*P*), seja verdadeira. Representando dessa forma regras pode-se então procurar obter padrões que caracterizem um BD através da evolução genética. A busca dessas regras é o ponto crucial desta técnica, visto que o espaço de busca é em geral muito grande e não conhecido, a principio [Rodrigues, 2001].

Diante deste cenário, onde tecnologias de BD, garimpagem de dados e visualização de dados estão intrinsecamente ligadas, a descoberta de padrões em forma de regras de associação; utilizando AG, mostra-se um método mais promissor de extrair conhecimento.

O AGs tem sido empregados em garimpagem de dados para as tarefas de classificação e descrição de registros de um BD, alem da seleção de atributos do BD que melhor caracterizem o objetivo da tarefa de extração de conhecimento proposta [Kira and Rendell, 1992 e Koller and Sahami, 1996].

Fatores positivos do AG:

- Os AGs não são inerentemente limitados aos tipos de dados de que eles necessitam.
- Em muitos casos, a informação necessária dos dados é o valor ótimo de um ou mais de seus parâmetros.
- A aplicação mais comum dos AGs esta na combinação com as redes neurais. Eles são aplicáveis em varias áreas. Eles também podem ser usados para determinar a melhor topologia e reduzir o numero de entradas mais importantes. Quando usados com as redes neurais, os AG estão sempre incorporados diretamente em pacotes comerciais e os detalhes dos algoritmos estão escondidos dos usuários.

Fatores negativos do AG:

- A função nem mesmo precisa ser conhecida pelos algoritmos genéticos para produzir bons resultados. Tem aparência de ser uma caixa preta, cujos detalhes estão escondidos da técnica, tais como o treinamento de uma rede neural.

3.4 Raciocínio Baseado em Memória (RBM) - (*Memory-Based Reasoning*)

Nesta técnica as experiências passadas são base para tomar decisões. Quando alguém conhece uma pessoa nova faz comparação com um rosto parecido. Os médicos diagnosticam doenças baseando-se em sintomas característicos em casos similares, especialistas encontram fraudes de seguro, confiando nas similaridades com casos anteriores de fraude e nas diferenças em relação aos casos sem fraude. O primeiro passo é identificar casos similares de experiência, então se aplica a informação desses casos ao problema atual. Esta é a essência do RBM. O RBM é uma técnica direcionada na coleta de dados que igualmente explora a experiência através da manutenção de uma BD de registros conhecidos. O RBM encontra registros similares a um novo registro e os utiliza para classificar e fazer previsões [Rodrigues, 2001]

Fatores positivos do RBM:

- A lista de registros similares fornece uma explicação de como o RBM chega a um resultado especifico.
- É uma técnica muito generalizada que não depende da representação subjacente de dados, RBM depende somente da existência de duas funções, a função distancia e a função de combinação, e não da representação dos dados.
- O desempenho do RBM depende mais do tamanho de conjunto de instrução do que numero de campos nos registros. Isso o torna prático para utilização quando outras técnicas, como redes neurais.
- O conjunto de instruções define quão bem o RBM funciona. Assim que novas categorias são introduzidas, novos registros para aquelas categorias podem ser diretamente adicionados ao conjunto de treinamento para que o RBM retire informações. Esse mecanismo contrasta com as redes neurais ou as arvores de decisão, que requerem um extenso período de readaptação para ingerir novas informações.

Fatores negativos do RBM:

- A desvantagem no desempenho do RBM geralmente ocorre durante a fase de previsão, ao invés da fase de instrução. Embora seja possível agilizar a memória do conjunto de instrução para aperfeiçoar o desempenho da previsão, essa fase é sempre mais cara porque, encontrar os vizinhos mais próximos envolve a aplicação da função distancia para todos os campos no registro e para todos os registros no conjunto de instrução. Em contraste, as arvores de decisão e redes neurais incorporam o conjunto de instrução em seus modelos, e depois descartam o conjunto de instruções.
- O conjunto de instruções do RBM utilizado pelo RBM é o modelo e quanto maior o conjunto de instruções melhor o resultado. Embora haja algumas técnicas para reduzir o numero de registros no conjunto de instruções os registros restantes devem ainda ser representados. Em contraste, o tamanho de um modelo de rede neural depende apenas da topologia da rede e não tem qualquer dependência em relação ao tamanho do conjunto de instruções.
- Os resultados do RBM dependem da escolha especifica da função distancia, da função combinação e do número de vizinhos escolhidos (k). Escolhas diferentes podem afetar os resultados.

3.5 Rede Bayesiana (RB)

[Hruschka e Silva, 1996] definem a probabilidade bayesiana como uma teoria consistente e que permite a representação de conhecimento certos e incertos via distribuição de probabilidade conjunta. Tal distribuição conjunta pode ser representada pelo produto de distribuições condicionadas.

Uma variável é condicionada a uma ou mais variáveis numa relação casual. Uma distribuição pode ser representada por um grafo orientado. No grafo, cada nodo representa uma variável do modelo e os arcos ligam as variáveis que estão em relação direta causa/efeito. Estas estruturas gráficas, com a quantificação de crença nas variáveis e seus relacionamentos, denominam-se Redes Bayesianas ou Redes Causais.

Uma representação de conhecimento utilizando numa arquitetura baseada em RB combina o melhor de duas áreas: o conhecimento do domínio do especialista e a estatística dos dados [Heckerman, 1996] e [Aliferes and Cooper, 1994].

As RBs são utilizadas no processo de extração de conhecimento que segue os seguintes passos: Inicia codificando-se o conhecimento existente de um especialista ou um conjunto de especialista numa RB. Logo se utiliza um BD para atualizar esse conhecimento, criando uma ou mais novas RBs, finalmente inclui um refinamento do conhecimento original do especialista e algumas vezes da identificação de novos relacionamentos.

A descoberta de conhecimento realizada por RB é similar a descoberta por redes neurais artificiais.

Conseqüentemente pode-se interpretar e entender o conhecimento codificado na representação mais facilmente. A interpretação de uma probabilidade como uma freqüência numa serie de repetições de experiências é tradicionalmente relacionada como sendo uma interpretação objetiva. Em contraste, a interpretação de uma probabilidade como um grau de certeza é chamado de subjetiva, ou interpretação bayesiana. Nesta última interpretação, a probabilidade ou certeza geralmente dependerá do estado de conhecimento da pessoa que fornece aquela probabilidade. Essa probabilidade pode ser escrita como P(e/E), significando "probabilidade de e dado E". o símbolo "E" representa o estado de conhecimento da pessoa que prove a probabilidade "e" um evento. Uma variável representa uma distinção sobre o mundo. Ela toma valores de uma coleção de estados mutuamente exclusivos ou coletivamente exaustivos, onde cada estado corresponde a algum evento. Uma RB é um modelo para algum problema de domínio ou universo, o qual consiste de um conjunto de variáveis. Uma RB num determinado domínio "U" representa uma função distribuição de probabilidade P(U/E).

Fatores positivos:

- Habilidade de reduzir o cálculo, usando somente variáveis obtidas de um objeto e seus vizinhos em uma estrutura de grafo para representar problemas do mundo real, nos quais existam relações de causa e conseqüência entre as variáveis.
- Pode-se facilmente codificar conhecimento de um especialista em RBs e usar esse conhecimento para aumentar a eficiência e a qualidade do conhecimento descoberto,
- Os nodos dos arcos em uma RB treinada, geralmente correspondem a distinções de variáveis e relacionamentos causais.

Fatores negativos:

- Exige grande esforço computacional para o cálculo das distribuições de probabilidades geradas pela explosão combinatória.

3.6 Arvore de Decisão (AD)

Os arvores de decisão são ferramentas poderosas e populares para classificação e diagnóstico. O atrativo dos métodos baseados em arvores está no fato de que, em contraste com as redes neurais, as arvores de decisão representam regras. Estas regras podem prontamente ser expressas em linguagem coloquial, de modo que qualquer pessoa possa compreendê-las.

A arvore de decisão é uma estrutura em que cada nodo interno que sai deste nodo identifica um dos atributos de previsão, cada linha que sai deste nodo identifica um valor que poderá ser assumido por tal nodo; cada folha identifica o resultado da previsão ou objetivo.

ID3 [Quinlan, 1986] , C4.5 [Quinlan, 1993], SLIQ [Metha et al., 1993] SPRINT [Shafer et al., 1996], são técnicas de construção da arvore de decisão.

Fatores positivos da AD:

- Habilidade das ADs para gerar regras que podem ser traduzidas para uma linguagem compreensível ou para SQL. É o ponto mais forte dessa técnica.
- Pode parecer óbvia a indução de regra e a árvore de decisão em particular, é uma escolha excelente em domínios onde realmente existam regras a serem encontradas. [Berry and Linoff, 1997] apresentam aplicações dessa técnica em trabalho desenvolvido para a empresa *Caterpilhar INC.*
- A árvore de decisão pode tomar varias formas. Na prática os algoritmos usados para produzir árvores de decisão geralmente produzem árvores com um fator de ramificação baixo e de teste simplificado.
- Os métodos da AD são igualmente aptos para lidar com variáveis discretas e contínuas.
- Os algoritmos das ADs colocam o campo que realiza o melhor trabalho de divisão dos registros de treinamento no nodo principal da árvore.
- Eficiência computacional e simplicidade de construção.

Fatores negativos da AD:

- Os algoritmos das árvores de decisão são menos apropriados para tarefas de estimativas nas quais o objetivo seja prever o valor de uma variável continua, tais como imposto, pressão ou taxa de interesse. As árvores de decisão também são problemáticas para dados em séries temporais.
- Alguns algoritmos de árvore de decisão podem somente lidar com classe de valor binário (sim/não). Outros são capazes de determinar registros em um número de classes arbitrárias, mas estão propensos a erro quando o numero de exemplos de treinamento por classes se torna pequeno o que pode acontecer muito rápido em uma árvore com muitos níveis e/ou muitos ramos por nodos.\

3.7 Lógica Indutiva (LI)

Segundo [Raedt, 1992 e Lavrac et al., 1991] a lógica indutiva (LI) pode ser vista como uma máquina de aprendizado em lógica de primeira ordem, onde as relações são apresentadas no contexto de BD dedutivos. [Ullman, 1988] indica que a LI é relevante para a descoberta de conhecimento em BD relacionais e dedutivos, pois pode descrever padrões envolvendo mais de uma relação.

Quando se procura aprendizagem em grandes BDs, a redução da complexidade é extremamente importante. Dois extremos na construção do processo de extração de conhecimento são identificados; o primeiro está na escolha de uma linguagem de hipóteses muito simples, e o outro extremo oposto esta na seleção de um pequeno conjunto de dados. Algoritmos simples de aprendizado detectam hierarquias que são utilizadas para estruturar o espaço de hipóteses para um algoritmo de aprendizado mais complexo. Essas características podem ser obtidas combinando a LI diretamente com o SGBD relacional [Morik and Brockhausen, 1997].

Desenvolvimentos em aprendizado indutivo focalizam o problema da construção de uma definição lógica para uma relação [Quinlan, 1990] através do conhecimento de duplas que pertencem ou não a essa relação. Desse modo novas relações podem ser especificadas por um menor numero de duplas, as quais são então generalizadas para induzir uma definição lógica [Quinlan, 1990] e [Lavrac et al., 1991].

Capítulo V

Estudo de aplicação

1 Introdução

Desenvolvemos uma aplicação das técnicas e ferramentas de EIACBD na avaliação de qualidade de produtos de *software*, através da participação em um projeto desenvolvido na Divisão de Qualificação de Software DQS do CenPRA.

1.1 Objetivos da aplicação

1. Aplicar técnicas e ferramentas de extração automática e inteligente de conhecimento de BD relacional para avaliação da qualidade de produtos de *software*.
2. Diagnosticar o estado dos produtos de *software* pelas suas características e componentes nas diferentes categorias e períodos.

2 Qualidade de Produto de *Software* (QPS)

2.1 Aspectos gerais do *software*

O *software* é um produto diferenciado de outros tipos de produtos manufaturados. Ele é caracterizado como um produto complexo, intangível, não manufaturado em série, sem desgaste, etc. O custo final é basicamente o custo do projeto e desenvolvimento [Villalobos, 2000].

2.2 Definições de qualidade de *software*

Segundo a ISO8402, a qualidade é a totalidade das características de uma entidade que lhe confere a capacidade de satisfazer necessidades explícitas e implícitas.

A totalidade das características de um produto de *software* que tem a capacidade de satisfazer as necessidades; por exemplo, deve estar em conformidade com suas especificações. As características compostas do *software* determinam o grau com que o *software* em uso irá satisfazer as expectativas do cliente.

2.3 Normas de qualidade do produto de *software*

- ISO/IEC 9126, Inf. Tech - Sw Prod Eval - *Quality Characteristics and Guidelines for their Use, 1991.*
- NBR 13596, Tecnologia de Informação - Avaliação de Produto de *Software* - Características de Qualidade e Diretrizes para o seu Uso, ABNT - Abril 1996 (versão brasileira da Norma ISO/IEC 9126, 1991. Esta última é a versão brasileira da primeira).
- NBR ISO/IEC 12119 - Tecnologia da Informação - Pacotes de *Software* - Teste e Requisitos de Qualidade.
- ISO/DIS 9241: *Ergonomic requirements for office work with visual display terminals (VDTs); Part 10: Dialogue principles; Part 11: Guidance on usability; Part 12: Presentation of information; Part 14: Menu dialogues; Part 16: Direct manipulation dialogues.*
- ANSI IEEE 1063 - *Standard for software user documentation*

2.4 Componentes do produto de *software*

O Método de Avaliação de Qualidade de *Software* - MEDE-PROS ®, desenvolvido na DQS do CenPRA considera que um produto de *software* é composto: pelo *Software* e Interface (que conformam a aplicação), pela Documentação do Usuário, Documento de Descrição do Produto e pela Embalagem. Ou seja, a lista de verificação do MEDE-PROS ® avalia estes cinco componentes.

O *software* e a Interface podem ser avaliados principalmente com base na norma ISO/IEC 9126, sendo que na Interface também é considerada a norma de ergonomia ISO/IEC 9241. A documentação do usuário pode ser avaliada principalmente com base nas normas NBR ISO/IEC 12119, ANSI/IEEE 1063 e ISO/IEC 9126. A descrição do produto e a embalagem podem ser avaliadas com base na norma NBR ISO/IEC 12119.

2.5 Modelo de qualidade do produto de *software*

Consideramos no projeto a proposta da norma NBR 13596 ou ISO/IEC 9126 que é contemplar todos os aspectos da qualidade de produtos de *software* com a definição das seis características de qualidade de produtos de *software*: Confiabilidade, Funcionalidade, Usabilidade, Eficiência, Portabilidade, Manutenibilidade, as quais são por sua vez subdivididas em sub-características. Especifica-se assim um modelo de qualidade de *software* que categoriza atributos de qualidade para este produto.As características são manifestadas externamente quando o *software* é executado, e é o resultado de atributos internos de qualidade.

- Funcionalidade: conjunto de atributos que evidenciam a existência de um conjunto de funções e suas propriedades específicas. As funções são as que satisfazem as necessidades explícitas e implícitas.

Sub-características: Adequação, Precisão, Interoperabilidade, Conformidade, Segurança de acesso.

- Confiabilidade: conjunto de atributos que evidenciam a capacidade do *software* de manter seu nível de desempenho sob condições estabelecidas durante um período de tempo estabelecido.

Sub-características: Maturidade, Tolerância a falhas, Recuperabilidade.

- Usabilidade: conjunto de atributos que evidenciam o esforço necessário para se poder utilizar o *software*, bem como o julgamento individual desse uso, por um conjunto explícito ou implícito de usuários.

Sub-características: Inteligibilidade, Aprendizagem, Operacionalidade.

- Eficiência: conjunto de atributos que evidenciam o relacionamento entre o nível de desempenho do *software* e a quantidade de recursos usados, sob condições estabelecidas.

Sub-características: Comportamento em relação ao tempo, Comportamento em relação aos recursos.

- Portabilidade: conjunto de atributos que evidenciam a capacidade do *software* de ser transferido de um ambiente para outro.

  Sub-características: Adaptabilidade, Capacidade para ser instalado, Conformidade, Capacidade para substituir.

- Manutenibilidade: conjunto de atributos que evidenciam o esforço necessário para fazer modificações especificadas no *software*.

  Sub-caracteristicas: Analisabilidade, Modificabilidade, Estabilidade, Testabilidade.

3 Aplicação das técnicas e ferramentas de EIACBD na avaliação de qualidade de produtos de *software*

3.1 Definição e compreensão do domínio de aplicação

O domínio da aplicação é a área de avaliação de qualidade de *software*. Durante três períodos realizaram-se avaliações de produtos de *software* no Brasil (ASSESPRO 1995, ASSESPRO 1996 e ASSESPRO 1998) Organizados pela Associação de Empresas Produtoras de Software (ASSESPRO), em cada evento se classificaram os produtos de *software* em categorias como apresentado na Tabela V.1.

Os dados armazenados em um BD relacional foram obtidos através de um Sistema de Administração de Avaliações (SISAVAL 3.0). O SISAVAL é um sistema computacional que automatiza a administração das avaliações de produto de software realizadas pela DQS/CenPRA.

Os dados pertinentes às avaliações (método utilizado, avaliador que executou a avaliação, resultados das avaliações, etc.) são armazenados no BD e manipulados por programas do computador. Estes dados serão processados e transformados em informações úteis.

Tabela V.1: *Descrição das categorias*

CATEGORIAS	DESCRIÇÃO
C1: Suporte à Documentação e . Planejamento	Sistemas destinados a composição de documentos, organização e manipulação de dados (valores, palavras, textos, imagens, tabelas, etc.). Ex.: Editores de texto, Planilhas, Dicionários, Gerenciadores de Projeto, Programas de Editoração, Formatadores de Relatórios, Etc.
C2: Software Básico e de Apoio ao Desenvolvimento	Sistemas que apresentam funções de controle básico do computador e de seus periféricos; sistemas destinados ao suporte e desenvolvimento de programas aplicativos. Ex.: Sistemas Operacionais, Compiladores, Redes, Servidores, Geradores de Aplicativos, Ferramentas CASE, Gerenciadores de Atividades Computacionais e Controle de Acessos, etc.
C3: Sistemas de Engenharia e Ferramentas Gráficas	Sistemas que utilizam cálculos de engenharia ou cálculos complexos; e sistemas que reúnem ferramentas gráficas na execução de suas funções; alem de sistemas de automação industrial, monitoração e controle de processos. Ex.: CAD`S, Geradores de Desenhos e Gráficos, Processadores de Imagens, Bibliotecas de Apoio a Programação de CLP ou CN, Sistemas de Controle de Trafego, Sistemas de Controle de Processo Industrial, Sistemas Especialistas, etc.
C4: Sistemas de Informação Específicos,	Sistemas convencionais destinados à realização das atividades fim do usuário nas áreas administrativa, comercial e financeiros em organizações e diversos ramos das atividades. Ex.: Folha de pagamento, Controle de Estoque, Vendas, Contabilidade, Faturamento, Controle Fiscal, etc.
C5: Sistemas de Informação Integrados e	Sistemas destinados à realização de atividades fim do usuário nas áreas de atividades não convencionais ou especificas, e que permitam ou não a troca automatizada de dados com o ambiente. Ex.: Controle de ponto e/ou acesso, Controle de Produção, Sistemas com Funções Integradas (faturamento, controle financeiro, estoque, contas apagar, etc.), sistemas especifico apara escolas, bancos saúde, jurídico, comercio, etc.
C6: Educação e Entretenimento	Sistemas que visam disseminar conhecimento como cultura geral ou especifica e programas de entretenimento, que fornece informações de forma organizada e que utilizam ou não recursos de multimídia. Ex.: Cursos, Treinamentos, Jogos, Enciclopédias, Programas de Alfabetização, etc.

Características do sistema

A primeira versão do Sistema de Administração de Avaliações (SISAVAL 3.0) foi especificada e desenvolvida em 1995 pela DQS/CenPRA. Atualmente o sistema esta desenvolvida em ORACLE. A Figura V.1 mostra a Base de Dados relacional do conjunto de dados a serem processados.

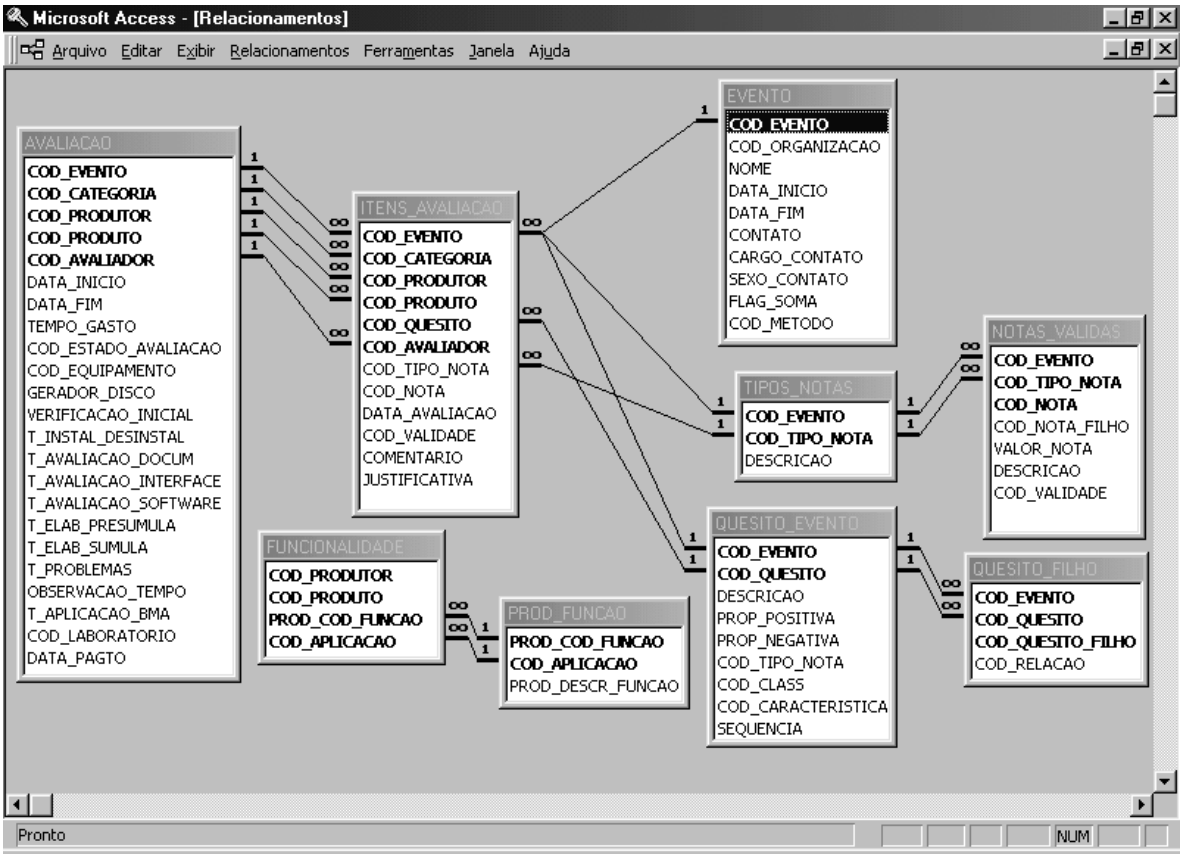

Figura V.1: *Banco de dados relacional*

3.1.1 Criação de objetivos para os conjuntos de dados

- OBJETIVO 1: Diagnosticar o estado dos produtos de *software* pela suas características definidas pela [ISO/IEC 9126 ou NBR 13596], nas diferentes categorias e eventos conforme descrevemos a seguir:

  Características: Confiabilidade (CN), Completitude (CP), Funcionalidade (FU), Usabilidade (US), Eficiência (EF) e Portabilidade (PO).

Categorias: Suporte à documentação e ao planejamento (C1), Software básico e de apoio ao desenvolvimento (C2), Sistemas de engenharia e ferramentas gráficas (C3), Sistemas de informação específicos (C4), Sistemas de informação integrados (C5) e Educação e entretenimento (C6).

Eventos: Assespro 1995 (Ass95), Assespro 1996 (Ass96) e Assespro 1998 (Ass98).

- OBJETIVO 2: Descrever os componentes do produto de *software*.

  Componentes: Embalagem (EMB), Descrição de produto (DEP), Documentação (DOC), Interface (INT) e Software (SOF).

Para cada um dos objetivos propostos pode ser necessário desenhar diferentes modelos de processamento e diferentes algoritmos de garimpagem de dados. Em seguida descreveremos as tarefas necessárias para implementar os objetivos propostos.

4 Objetivo 1

4.1 Pré-Processamento

a) Criação de conjunto de dados significantes, integração, e seleção

Todas as tabelas foram integradas em um BD, menos a tabela componente, integrada posteriormente. Selecionamos as tabelas mais significativas para o objetivo proposto tais como: Itens_Avaliação, Avaliação, Eventos, Tipos_Notas, Notas_Validas, Questao_Evento, Quesito_Filho, Funcionalidade, Prod_Funcao.

Selecionaram-se os atributos relevantes de diferentes tabelas para ter finalmente o grupo de dados em uma tabela para o primeiro objetivo proposto. Não foi necessário utilizar uma técnica de seleção de atributos pelo atributo único que continha os valores da avaliação. Utilizou-se o programa "Microsoft Access 97" e o programa estatístico "SPSS 11.5 for Windows" para integrar os atributos selecionados. Os resultados forma compostos na tabela "assxx" como os seguintes atributos: cod_even, cod_cate, cod_prtr, cod_prod, cod_ques, cod_aval, cod_tnot, val_nota, cod_clas, cod_cara, des_clas, des_cara.

b) Padronização de dados

Devido que em cada evento os intervalos das notas eram diferentes, padronizamos os intervalos dos diferentes eventos, de modo que as notas padronizadas estejam entre 0 e 1. Assim, as notas do atributo val_nota da tabela "assxx" é armazenado como tabela "assx1" com o atributo padronizado nota_pd3, dos diferentes períodos de avaliação utilizando a transformação no programa estatístico SPSS 11.5. Utilizou-se a seguinte equação de padronização:

$$nota\_pd3 = \frac{val\_nota - \min\_ev}{\max\_ev - \min\_ev},$$

onde:

val_nota: nota da questão

min_ev: nota mínima do evento

max_ev: nota máxima do evento

c) Redução e transformação de dados

Inicialmente a tabela conformada anteriormente teve 45.732 registros com 12 atributos. Existiram dois avaliadores para uma cada questão no evento Ass98, foi calculada pela média das notas dadas pelos avaliadores nas questões, excluindo-se o atributo "Avaliador", ficando então 28782 registros, com 11 atributos.

Em diferentes eventos teve-se diverso critério de avaliação, notas por questão, questões por características, e características em componentes os quais foram avaliadas. Inicialmente, teve-se que identificar as características que estão representadas nos componentes e agrupar questões por característica para cada produto.

Cada questão avalia uma característica em um componente mas cada componente e característica, podem ser avaliadas por mais de uma questão. Desta forma foi calculada a média dos valores das notas das questões com o mesmo componente e característica, excluindo o atributo questão, ficando 1858 registros com 10 atributos.

Observação: Este seria o melhor conjunto de dados para trabalhar, pois as características são aplicáveis na dependência dos componentes e são melhor entendidas quando se tem a visão do componente em que se aplicam. Mas, infelizmente, devido ao fato que as

ferramentas utilizadas eram apenas demonstrativas ("*DEMOS*"), a capacidade de processamento dos dados estava limitada. Por isto teve-se que trabalhar com dados mais reduzidos, eliminando separadamente o componente ou a característica.

d) Redução de dados

Teve-se que fazer a média das notas em função das características, excluindo assim o atributo componente, ficando então 833 registros.

e) Reestruturação de dados

Reestruturou-se a tabela de casos para variáveis, os casos da "característica" passaram a ser variáveis. Este processo teve que re-arranjar 7 variáveis representando: as 6 características definidas (US - Usabilidade, CN - Confiabilidade, CP - Completitude, FU - Funcionalidade, PO - Portabilidade, EF - Efetividade) e uma não definida (XX). Ficaram então 119 registros com os seguintes atributos: cod_even, cod_cate, cod_prtr, cod_prod, CN, CP, EF, FU, PO, US, XX. A variável XX representa um quesito de avaliação do mercado, o qual só foi aplicado no período (Ass95) ficando finalmente um registro por produto, evento e categoria em que participou, e as notas médias que obteve em cada característica avaliada.

f) Limpeza de dados

Nesta tarefa se identificou um registro sem nenhum valor nos atributos, que foi apagado, ficando 118 registros.

4.2 Processamento

a) Escolha da função de garimpagem de dados

Para o primeiro objetivo que é diagnosticar, resumiu-se as características nas diferentes categorias e períodos.

b) Seleção de técnica de garimpagem de dados

A função de sumarização como associação foram processados por técnicas estatísticas e redes neurais.

c) Garimpagem de dados

Os *softwares* de garimpagem de dados utilizados foram o "DataScope" versão 5.1 *trial* e o Polyanalyst 4.5. Estes *softwares* são descritos a seguir.

O "DataScope versão 5.1.21.10 (*versão de teste da Cygron*). é um instrumento de garimpagem de dados que permite analisar visualmente os índices de um BD arbitrário, e extrair o conhecimento escondido em dados. Os dados podem ser importados através do *Microsoft ODBC* de uma única tabela do BD relacional. O DataScope usa técnicas poderosas de visualização, e intuição na análise de dados. Permite o reconhecimento de padrão, avaliação simultânea dos dados de muitos pontos de vista, tradução dos números para raciocínio subjetivo e pergunta dos dados sem comandos ou fórmulas."

O "PolyAnalyst 4.5 é um *software* mais compreensivo, apropriado e mais versátil para ferramentas de garimpagem de dados avançados. O PolyAnalyst incorpora as últimas tecnologias na descoberta automatizada do conhecimento e pode analisar dados estruturados e não estruturados (textos)".

Apresentamos os resultados obtidos pelos *software*s DataScope e Polyanalyst nas tabelas e gráficos a seguir.

4.3 Pós-processamento

a) Avaliação de padrões

No espectrograma a freqüência dos valores das notas dos produtos está representada pelo comprimento das linhas horizontais dispostas sobre as verticais. Cada linha vertical representa as características específicas do produto.

Figura V.2: *Distribuição dos valores das notas por característica*

A partir da Figura V.2 e Tabela V.2 pode-se observar que a "Eficência" tem melhor avaliação devido ao bom nível de desempenho dos computadores. Apenas dois produtos que se mostraram fora do grupo. A "Funcionalidade" é a segunda característica melhor avaliada, pois é a característica mais básica que um produto de *software* deve atender, isto é se ele tivesse uma funcionalidade ruim, o produto não sobreveviria no mercado. Em terceiro e quarto lugar a "Usabilidade" e a "Portabilidade" tem-se observado que só em algumas categoria é imprescidivel como em outras necessaria. A "Confiabilidade" e a "Completitude" mostram a vulnerabilidade dos dados e falta de completitude da documentação, que quase sempre é deixada para o final do desenvolvimento, quando os prazos para o lançamento do produto no mercado estão esgotando.

Por outro lado, temos fortes índices de que o método (MEDE-PROS) consegue distinguir um bom produto de um produto regular, pois existe uma gama muito ampla dos resultados, podendo-se distinguir os produtos ruins.

Tabela V.2: *Média das características*

Características	CN	EF	PO	US	CP	FU	XX
Média	0.60	0.89	0.69	0.68	0.57	0.75	0.74

No espectrograma abaixo a freqüência dos valores das notas dos produtos está representada pelas linhas horizontais, as linhas verticais representam as categorias dos produtos.

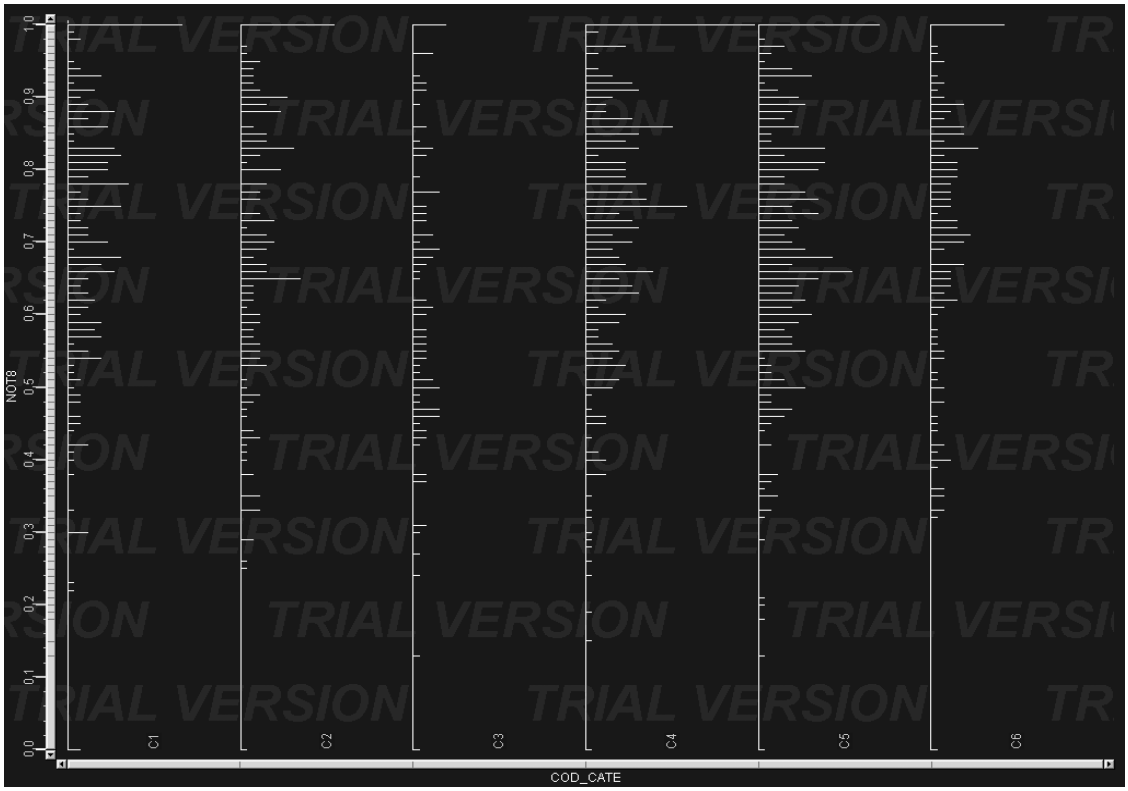

Figura V.3: *Distribuição dos valores das notas por categorias*

Embora as diferenças são pouco significativas, observa-se na Figura V.3. que: a categoria C4 (Sistemas de Informação Específicos) foi melhor avaliada possivelmente devido que foi uma categoria de produto de *software* que foram lançados no mercado antes que outros produtos de outras categorias. Os produtos dessa categoria tiveram grande demanda ate hoje em diferentes setores, predominantemente no setor comercial. As categorias C5 (Sistemas de Informação Integrados) e C1 (Suporte à Documentação e ao Planejamento) também tem influência predominante do setor comercial, seguidas pelas categorias C6 (Educação e Entretenimento) e C2 (Software Básico e de Apoio ao

Desenvolvimento). Estas categorias ainda estão em amadurecimento. Na Tabela V.3 pode-se descrever as categorias e suas médias.

Tabela V.3: *Média das categorias*

Categoria	C1	C2	C3	C4	C5	C6
Media	0.72	0.71	0.63	0.73	0.70	0.72

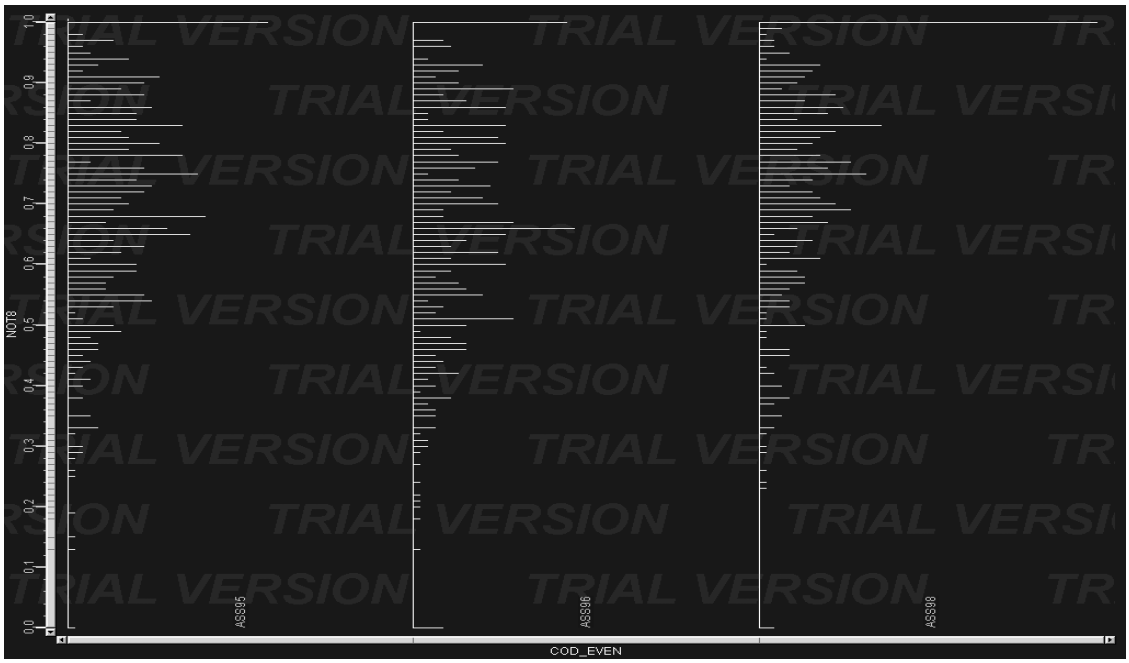

Figura V.4: *Distribuição dos valores das notas por evento*

Na Figura V.4. pode-se observar que no evento o Assespro98 obteve a melhor avaliação, seguida pelo Assespro95 e Assespro96. A diferença, entretanto, não é considerável e a melhoria dos *softwares* é pouco significativa.

Tabela V.4: *Média dos eventos*

Eventos	Ass95	Ass96	Ass98
Media	0.71	0.67	0.75

Tabela V.5: *Resultados estatísticos das características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	114	0.5997	0.2058	0.13	1	0.87	0.62
CP	118	0.569	0.1377	0.2	0.9	0.7	0.55
EF	113	0.8864	0.1317	0.15	1	0.85	0.92
FU	115	0.7511	0.1157	0.45	0.96	0.51	0.77
PO	117	0.6884	0.1329	0.33	0.93	0.6	0.7
US	115	0.6814	0.1326	0.31	0.94	0.63	0.69
XX	118	0.7383	0.1744	0	0.99	0.99	0.78

Na Tabela V.5. observa-se que a média da eficiência teve melhor avaliação, mas sua mediana foi mas alta, confirmando que vários valores das notas estiveram acima da média. A “Completitude“ está entre a média e mediana.

Na Figura V.5. O gráfico em barras bidimensional (2D Bar); o tamanho dos quadros representa a freqüência das questões e o brilho representa o valor médio das notas, este tipo de gráfico pode ser comparado a um mapa de nível.

x = Eventos, y = Media das notas obtidas.

Figura V.5: *Representação das características por eventos*

Tabela V.6: *Média dos eventos / características*

	Confiabilidade	Completitude	Eficiência	Funcionalidade	Portabilidade	Usabilidade
Ass95:	0.68	0.60	0.90	0.75	0.72	0.65
Ass96:	0.46	0.52	0.89	0.72	0.63	0.68
Ass98:	0.67	0.59	0.86	0.79	0.71	0.72

Na Figura V.5 e Tabela V.6 pode se obsevar que :

A confiabilidade nos eventos 95 e 98 foi boa, sendo abaixo da média no evento 96.

A Completitude nos tres eventos foi boa tendo uma breve descida no último evento

A Eficiência nos três eventos foi excelente.

A Funcionalidade nos três eventos foram próximas de excelente

A Probabilidade nos três eventos foi boa.

A Usabilidade nos três eventos foi boa, observando-se um aumento nos últimos eventos em relação aos anteriores.

Tabela V.7 *Média das características / categorias*

	Confiabilidade	Completitude	Eficiência	Funcionalidade	Portabilidade	Usabilidade
C1:	0.60	0.55	0.91	0.76	0.72	0.71
C2:	0.61	0.58	0.91	0.74	0.68	0.65
C3:	0.45	0.46	0.80	0.66	0.62	0.59
C4:	0.65	0.61	0.87	0.77	0.72	0.71
C5:	0.59	0.57	0.91	0.75	0.66	0.67
C6:	0.61	0.56	0.87	0.78	0.67	0.70

Na Figura V.6, o gráfico de barras bidimensional (2D Bar) sendo o tamanho dos quadros a representação da freqüência das questões e o brilho a representação do valor médio das notas. Este tipo de gráfico também pode ser comparado a um mapa de nível onde x = Características e y = Categorias

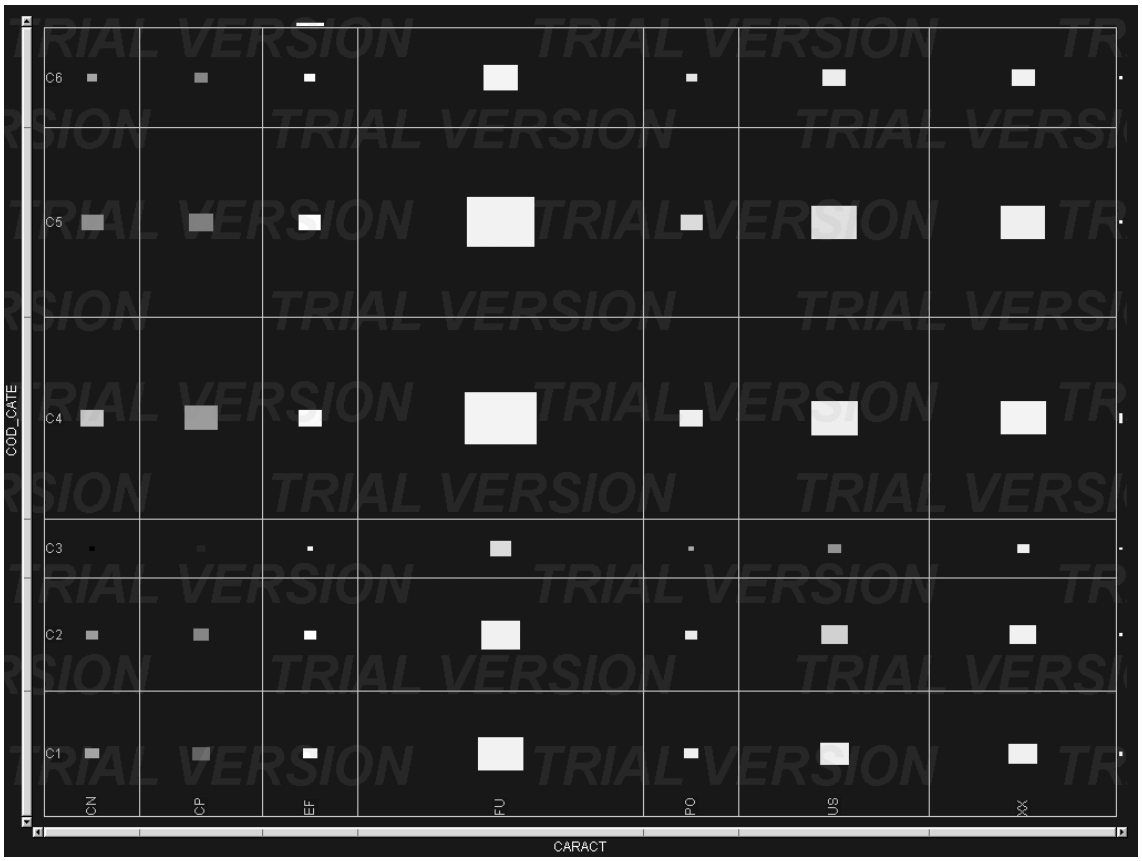

Figura V.6: *Representação das características por categorias*

A partir da Figura V.6 e Tabelas: V.7, V.8, V.9, V.10, V.11, V.12, V.13 pode-se observar que

A Confiabilidade é boa em algumas categorias e regular em outras, não é homogênea em todas as categorias.

A Completitude é boa em quase todas as categorias, exceto na categoria C3.

A Eficiência é excelente em todas as categorias.

A Funcionalidade é boa em todas as categorias.

A Portabilidade é boa em todas as categorias.

A Usabilidade é boa em todas as categorias.

Tabela V.8: *Média da Categoria C1: Suporte a Documentação e Planejamento / Características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	19	0.6105	0.1739	0.22	0.84	0.62	0.65
CP	19	0.5547	0.1221	0.38	0.85	0.47	0.54
EF	19	0.9116	0.09227	0.75	1	0.25	0.91
FU	19	0.7574	0.122	0.45	0.96	0.51	0.78
PO	19	0.7216	0.1113	0.55	0.93	0.38	0.73
US	19	0.7147	0.1134	0.51	0.9	0.39	0.71
XX	19	0.7363	0.2251	0	0.99	0.99	0.78

Tabela V.9: *Média da Categoria C2: Software Básico e de Apoio ao Desenvolvimento / Características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	16	0.6037	0.2449	0.25	1	0.75	0.67
CP	17	0.5794	0.1486	0.29	0.86	0.57	0.55
EF	16	0.9144	0.09986	0.66	1	0.34	0.94
FU	17	0.7394	0.1375	0.47	0.93	0.46	0.78
PO	17	0.6818	0.1538	0.33	0.9	0.57	0.71
US	17	0.6471	0.1758	0.31	0.87	0.56	0.72
XX	17	0.7529	0.1168	0.52	0.91	0.39	0.77

Tabela V.10: *Média da Categoria C3: Sistemas de Engenharia e Ferramentas Gráficas / Características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	9	0.4467	0.1753	0.13	0.71	0.58	0.5
CP	9	0.4644	0.1239	0.27	0.63	0.36	0.49
EF	9	0.7989	0.1882	0.45	1	0.55	0.86
FU	9	0.6633	0.1089	0.47	0.79	0.32	0.72
PO	9	0.6156	0.158	0.37	0.82	0.45	0.68
US	9	0.5944	0.1024	0.43	0.73	0.3	0.62
XX	9	0.7122	0.1683	0.42	0.96	0.54	0.68

Tabela V.11: *Média da Categoria C4: Sistemas de Informação Específicos / Características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	28	0.6454	0.1722	0.32	1	0.68	0.63
CP	29	0.6066	0.1185	0.32	0.79	0.47	0.61
EF	27	0.8678	0.172	0.15	1	0.85	0.92
FU	28	0.7679	0.1238	0.48	0.96	0.48	0.78
PO	29	0.7241	0.113	0.45	0.91	0.46	0.74
US	28	0.7111	0.1452	0.34	0.94	0.6	0.75
XX	29	0.7541	0.2034	0	0.96	0.96	0.8

Tabela V.12: *Média da Categoria C5: Sistemas de Informação Integrados / Características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	29	0.5879	0.2209	0.18	1	0.82	0.6
CP	29	0.5714	0.1567	0.2	0.9	0.7	0.56
EF	29	0.9083	0.1033	0.62	1	0.38	0.94
FU	29	0.751	0.08674	0.52	0.9	0.38	0.77
PO	29	0.6638	0.1373	0.38	0.9	0.52	0.67
US	29	0.6686	0.1051	0.49	0.89	0.4	0.67
XX	29	0.7183	0.1519	0.27	0.93	0.66	0.76

Tabela V.13: *Média da Categoria C6: Educação e Entretenimento / Características*

Numerical and integer attributes:	Values	Mean	Std.Dev	Min	Max	Range	Median
CN	13	0.6131	0.2326	0.35	1	0.65	0.62
CP	15	0.5607	0.1332	0.32	0.79	0.47	0.55
EF	13	0.8654	0.1178	0.67	1	0.33	0.88
FU	13	0.7823	0.1097	0.56	0.9	0.34	0.78
PO	14	0.675	0.135	0.33	0.91	0.58	0.69
US	13	0.7023	0.1193	0.44	0.87	0.43	0.7
XX	15	0.748	0.1622	0.39	0.93	0.54	0.8

Intensidade de "relacionamentos pares" das características

Tabela V.14: *Intensidade de relacionamentos pares*

Características	Intensidade
US - FU	0.6683
PO – FU	0.5145

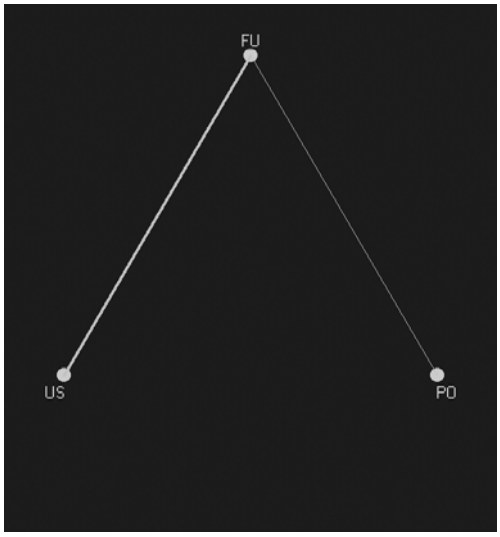

Figura V.7: *Intensidade de relacionamentos pares: US FU, PO FU*

Na Figura V.7 a intensidade de "relacionamento par" se observa entre a US FU (Usabilidade, Funcionalidade) e a PO FU (Portabilidade, Funcionalidade). A linha com mais brilho representa maior intensidade. A relação US FU tem maior índice de dependência. A dispersão dos pontos é apresentada na Figura V.8, onde os pontos aproximam-se de uma faixa linear, seguido da relação PO FU, representado pela linha cinza, com distribuição de pontos (ilustrado na Figura V.9). A faixa de maior dispersão pode ser interpretada como menor intensidade de relacionamento, isto é, quanto mais estreita a faixa maior intensidade de dependência.

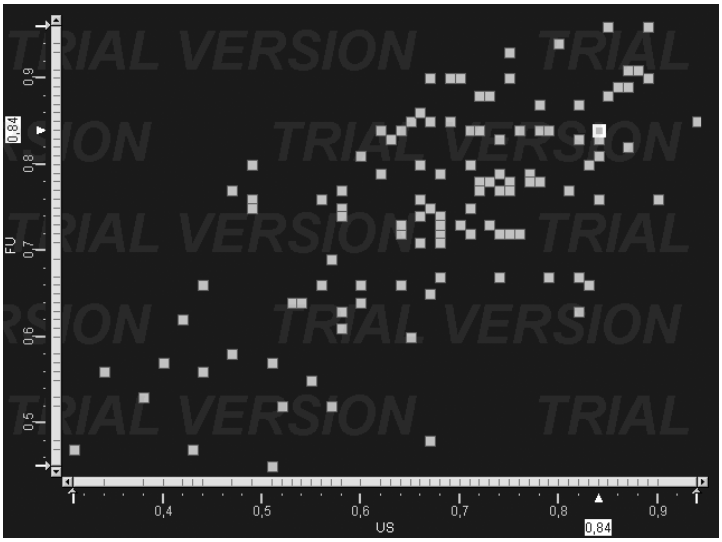

Figura V.8: *Distribução* US FU

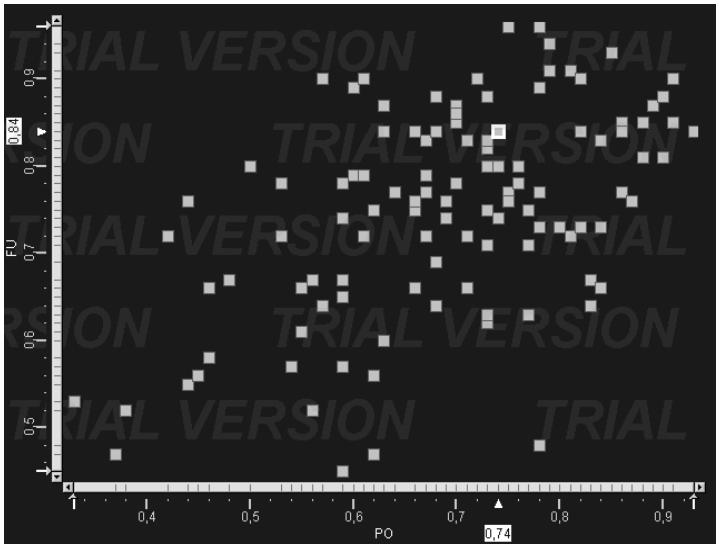

Figura V.9: *Distribução* PO FU

Intensidade de "relacionamentos triplos" das características

Tabela V.15: *Intensidade de relacionamentos triplos*

Características	Intensidade
US - FU - PO	0.5850
US - CN - FU	0.5442

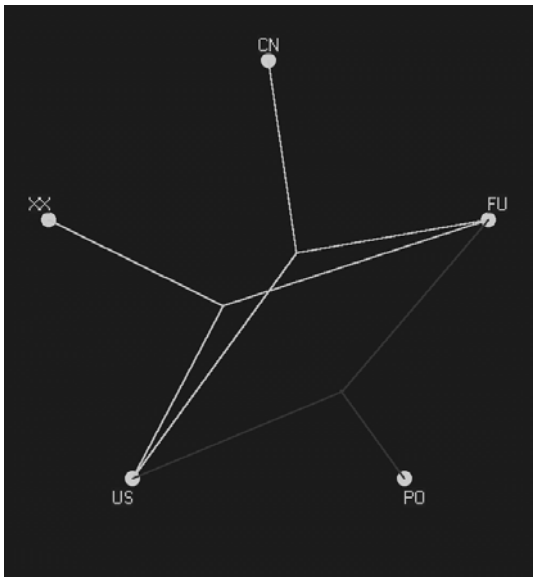

Figura V.10: *Intensidade de relacionamento triplo US FU PO*

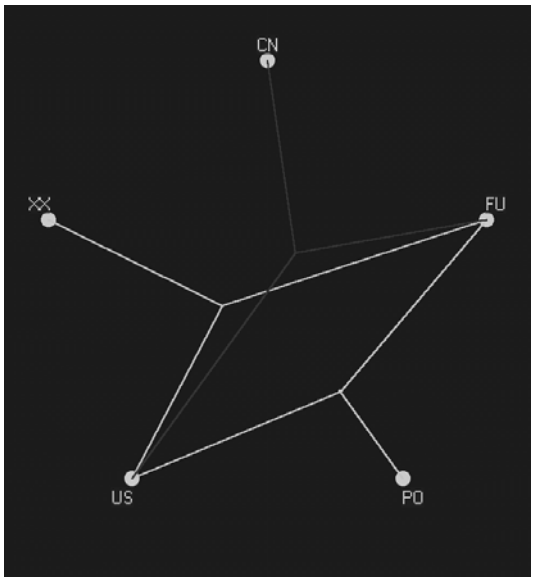

Figura V.11: *Intensidade de relacionamento triplo US CN FU*

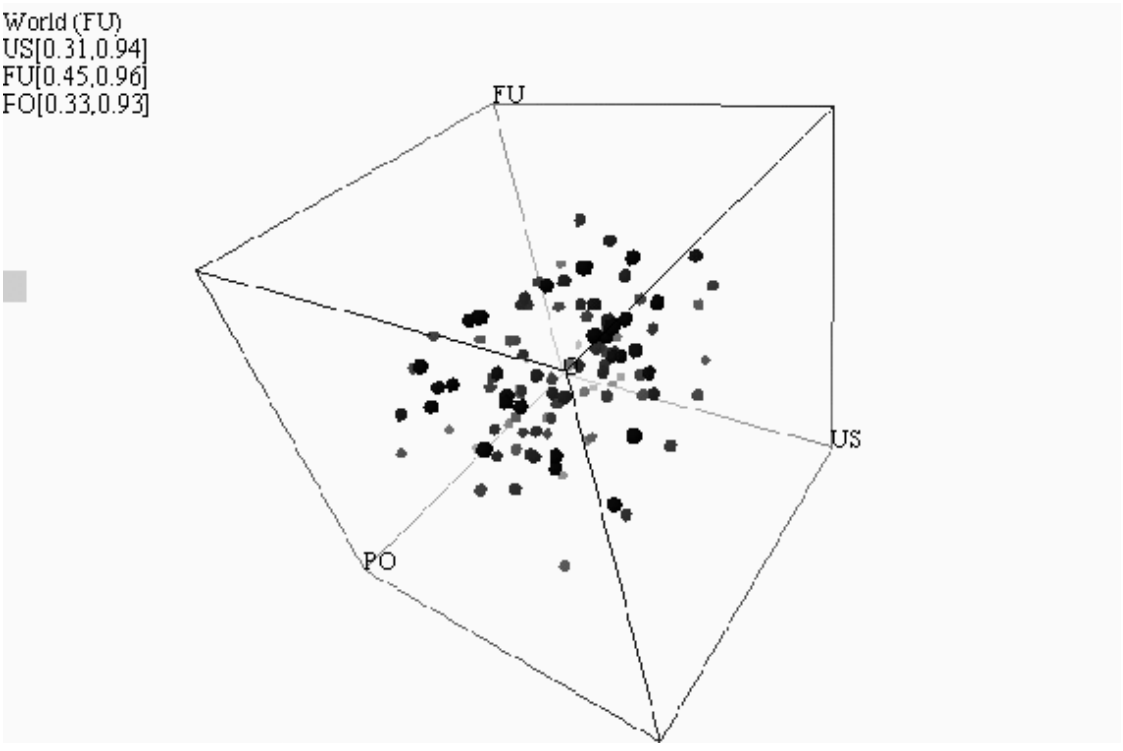

Figura V.12: *Gráfico 3D, distribuição US FU PO*

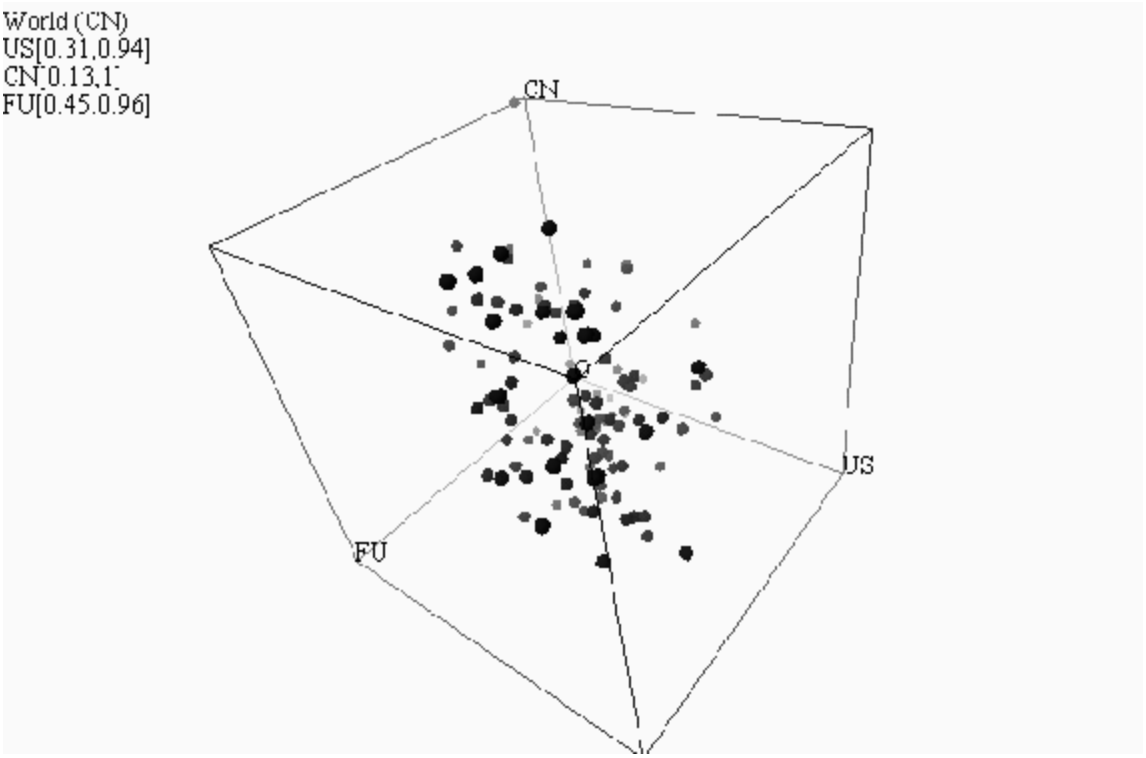

Figura V.13: *Gráfico 3D, distribução US CN FU*

Na Figura V.10 as linhas cinza representan os relacionamentos das caracteristicas com intensidade de "relacionamento triplo" entre as caracteristicas US FU PO (Usabilidade, Funcionalidade e Portabilidade) . Os pontos distribuidos desta relação são apresentados na Figura V.12. Quando os pontos estão mais próximos da linha central diagonal (forma de cilindro) estes tem mais intensidade de relacionamento. A Funcionalidade é predominante no dois relacionamentos, confirmando-se que esta característica é esencial no produto, para que ele se mantenha no mercado.

b) Apresentação de conhecimento e uso de conhecimento descoberto

- Segundo a Figura V.2 a característica melhor avaliada foi a Eficiência, comprovada nas Figuras V.5 e V.6 em eventos e categorias.
- Segundo a Figura V.3, os produtos da categoria C4 (Sistemas de Informação específicos) tiveram maior desenvolvimento.
- Segundo a Figura V.4 as avaliações por eventos tiveram diferenças pouco significativas.

- Segundo as Figuras V.7, V.8 e V.9 percebe-se que a intensidade de "relacionamento par" é maior na relação Usabilidade/Funcionalidade, seguido da relação Portabilidade/Funcionalidade.
- Segundo as Figuras V.10, V.11, V.12 e V.13 observa-se que a intensidade de "relacionamento triplo" é forte entre Usabilidade/Portabilidade/Funcionalidade seguido do relacionamento Usabilidade/Confiabilidade/Funcionalidade.

5 Objetivo 2

As etapas iniciais do processo, como a compreensão de domínio e a definição de objetivos foram realizadas. O Objetivo 2 é descrever os produtos de *softwar*e pelos seus componentes.

5.1 Pré-Processamento

Para o Objetivo 2 utilizou-se as mesmas tabelas do Objetivo 1 na etapa de pré-processamento, com a diferença de não se reestruturar a tabela de casos para variáveis, reduzidos em 1338 tuplas.

5.2 Processamento

a) Escolha da função de garimpagem de dados

Para o segundo objetivo será utilizada a função de sumarização.

b) Seleção de técnica de garimpagem de dados

A função de sumarização pode utilizar diversas técnicas, como métodos de sumarização estatística. Utilizamos o *software* de garimpagem de dados "DataScope" versão teste.

5.3 Pós-Processamento

a) Avaliação de Padrões

Os valores das notas dos produtos estão representados pelas barras horizontais, e cada linha vertical sobre a horizontal representa os componentes do produto.

Figura V.14: *Distribuição dos valores das notas por componentes*

Na Figura V.14 pode-se observar que o componente Software (SOF) teve melhor avaliação pelo fato de ser a essência do produto com as características de Funcionalidade, Confiabilidade, Efetividade, e Portabilidade. A Funcionalidade é fundamental para cumprir os objetivos propostos do produto e usuário. Seguido pelo componente Documentação (DOC) o qual tem que prover recursos de aprendizagem. Em seguida temos a componente Interface (INT) que prove a interface do produto com o usuário., seguido por outros componentes como Descrição do Produto (DEP) e Embalagem (EMB). Os produtores também parecem colocar maior esforço nessa ordem, pelas suas percepções.

Tabela V.16 *Média dos componentes*

Componentes	SOF	DOC	BEM	PEG	DEP	INT	COM
Médias	0.74	0.73	0.71	0.71	0.55	0.72	0.66

No gráfico em barras bidimensional (2D Bar) apresentado na Figura V.15, o tamanho dos quadros representa a freqüência das questões e o brilho representa o valor médio das notas.

x = Características, y = Componentes.

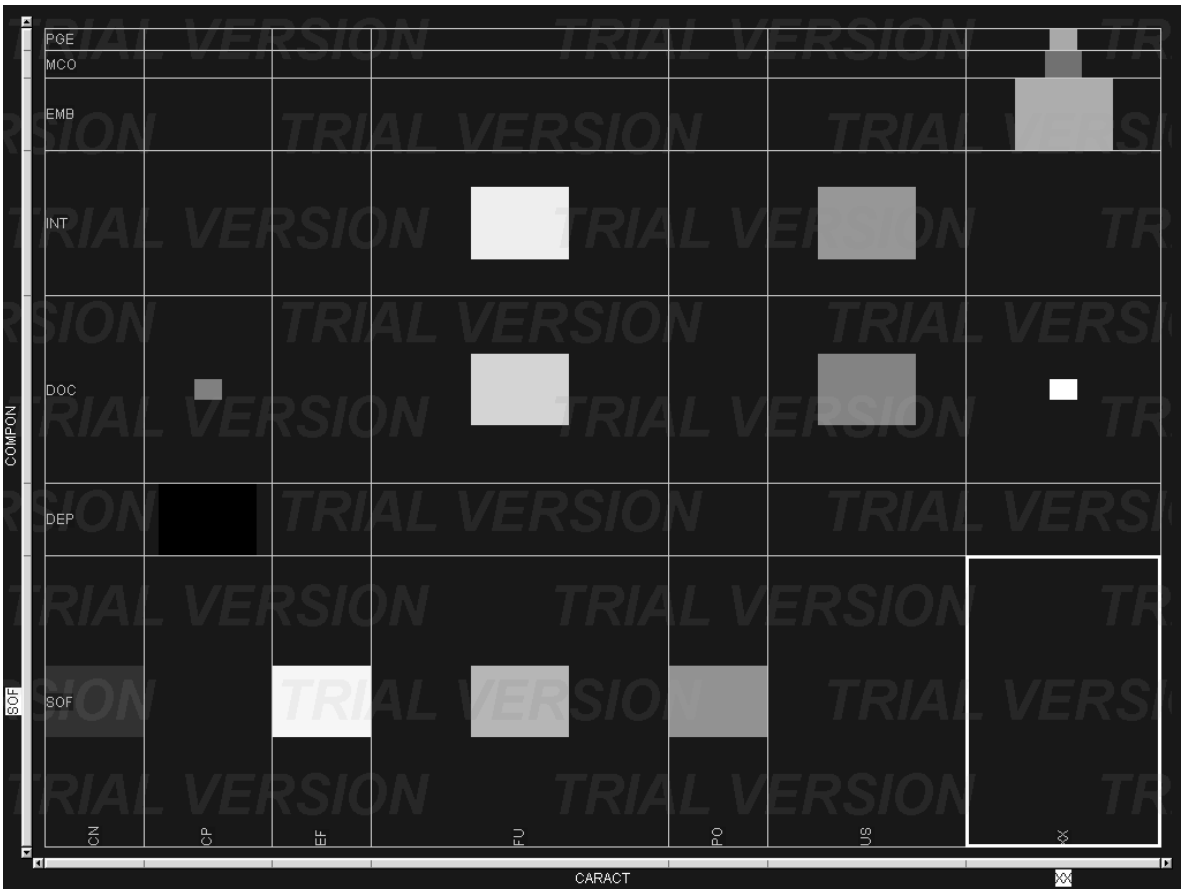

Figura V.15: *Representação dos componentes / caracteristicas*

Na Figura V.15. Pode-se observar na Figura V.15 que as caracteristicas estão representadas em componentes, constituíndo a Tabela V.17.

Tabela V.17: *Representação das caracteristicas em componentes*

Caracteristica	Componentes
CN	SOF.
CP	DEP e DOC.
EF	SOF.
FU	SOF, DOC e INT.
PO	SOF.
US	DOC e INT

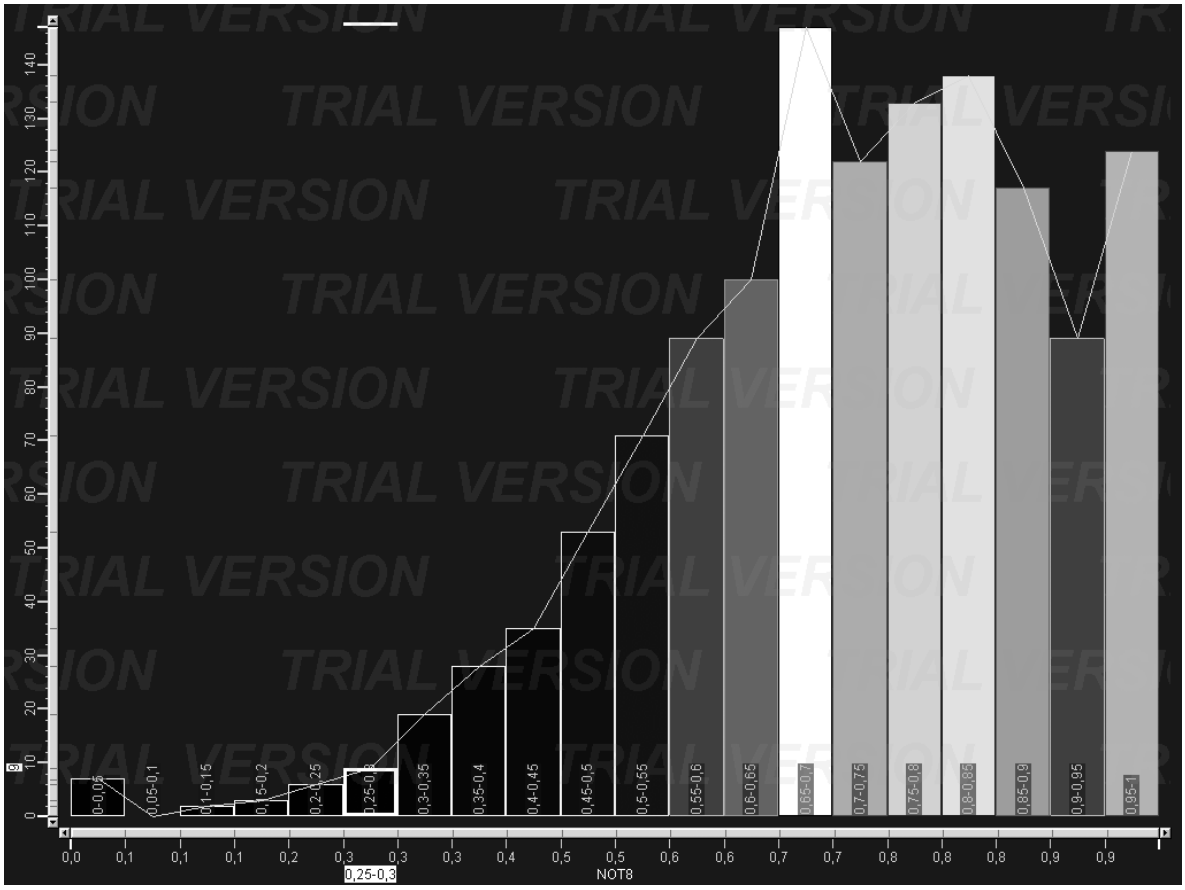

Figura V.16 *Distribuição dos valores discretizados*

No histograma apresentado na Figura V.16 se pode observar a tendência dos valores das notas por produto para o lado direito da média, pelo brilho e altura da barra.

b) Apresentação de conhecimento e uso de conhecimento descoberto

- Os componentes melhor desenvolvidos ou melhor apresentados foram o Software (SOF), Documentação (DOC) e seguido pela Interface (INT), Embalagem (BEM), e Descrição do Produto (DEP).
- As caracteristicas do produto de *software* estão baseados principalmente em dois componentes; Software (SOF) e Documentação (DOC) seguido da Inteface (INT) e Descrição do produto (DEP).
- O estado dos produtos de *software* estão em um nível aceitavel de uso.

6 Conclusões

- Com relação ao primeiro objetivo do projeto, que foi de aplicar técnicas e ferramentas de extração inteligente e automática de conhecimento a BD relacional na avaliação da qualidade de produtos de *software*.
    - Usaram-se diferentes técnicas e ferramentas de: preparação de dados (estatística, recuperação da informação), análise inteligente de dados (estatística, visualização de dados, inteligência artificial): sistemas de BD (BD relacional, SGBD).
    - Para sumarizar e associar, as técnicas empregadas foram a estatística, redes neurais que são técnicas incluídas no *software* de garimpagem de dados "*DataScope 5.1 Trial*" e "*Polyanalyst 4.5*"
- Com relação ao segundo objetivo do projeto, de diagnosticar o estado dos produtos de *software*.
    - O estado do produto de *software* se encontra em um nível aceitável de uso.
    - A Funcionalidade tem-se notado decisiva no componente do produto de *software*, porém a Efetividade teve melhor avaliação, percebendo-se que os produtores colocaram maior ênfase no componente que representa esta característica.

Capítulo VI

Conclusões e futuros trabalhos

Os processos de EIACBD podem ser descritos por uma etapa preliminar e três etapas de processamento. Na etapa preliminar devem ser conhecidos o âmbito do domínio de aplicação e a definição dos objetivos da aplicação, de modo a permitir a modelagem do processo como também a escolha prévia das técnicas e ferramentas para manipulação dos dados. Nas etapas de processamento os dados são preparados através das diferentes tarefas necessárias, e analisados exaustivamente, transformando-se em informações. Finalmente na etapa de pós-processamento os dados são interpretados e apresentados como "conhecimentos". Podem também ocorrer processos iterativos para verificar ou descobrir outros "conhecimentos".

Observa-se também que integrando o SBD relacional e o *Data Warehouse*, os primeiros quatro processos de: limpeza de dados, integração de dados, seleção de dados, e transformação de dados, podem ser realizados por esta integração, ou pela execução de algumas operações de OLAP no *Data Warehouse*. Observa-se também que os processos de garimpagem de dados, avaliação de padrão, e apresentação de conhecimento às vezes podem ser integrados em um processo, de modo iterativo.

As técnicas e ferramentas estudadas e aplicadas na EIACBD de dados estruturados tiveram excelente desempenho, e possivelmente também possam ser aplicadas eficientemente em grandes BD distribuídos.

As técnicas e ferramentas estudadas não podem ser aplicáveis da mesma forma ou com o mesmo desempenho em dados semi-estruturados e não-estruturados. Esta é uma área de pesquisa que requer maior investimento interdisciplinar.

Com a construção do *Data Warehouse* para a EIACBD o processamento analítico *on-line* (OLAP) é aprimorado. A integração de OLAP e tecnologias de cubos de dados, e os mecanismos de garimpagem analíticos *on-line* contribuem significativamente para a análise de BDs multidimensionais.

Na aplicação que apresentamos, como "caso de estudo" foram descritas as técnicas e ferramentas aplicadas nas diferentes tarefas. A utilização apropriada das técnicas e ferramentas favorece consideravelmente a extração de conhecimento, porém sua utilização inapropriada pode transfigurar a modelagem do processo, traduzindo-se inclusive na obtenção de conhecimentos não-confiáveis ou não-compreensíveis. A qualidade do conhecimento que está sendo extraído depende então ainda da análise do especialista do domínio.

Durante o transcorrer do presente trabalho aprendemos também que existem diversas outras técnicas e ferramentas complementares que podem aprimorar os procedimentos de análise dos dados. A utilização destas técnicas pode ser interessante para a realização de trabalhos futuros nesta área. Uma das técnicas é a extensão do KDD serial para o KDD paralelo e distribuído, principalmente para problemas mais complexos, em tamanho, dimensionalidade e complexidade. Outra técnica seria o KDD cooperativo, principalmente para fontes distribuídas de dados, como por exemplo, no ensino a distância auxiliado por computador. Outra área crescente é a integração de técnicas e ferramentas aplicadas para BDs de multimídia. Enfim, a garimpagem eficiente de dados para gerar informações e conhecimentos é uma área de pesquisa interdisciplinar e multidisciplinar, em pleno desenvolvimento.

Referencias bibliográficas

[Adjeroh and Nwosou, 1997] Adjeroh, D.A.and Nwosou, K.C. Multimedia Database Management-Requirements and Issues. *Multimedia Database System, IEEE,* 1997.

[Adriaans and Zantige, 1996] Adriaans, P. and Zantige, D. Data Mining. Addison-Wesley, Harlow, UK, 1996.

[Agrawal et al., 1993] Agrawal, R.; Imielinski. T. and Swami, A. Mining Association Rules Between Sets of Items in Large Databases. *Proc. Int. Conf. Management of Data* (SIGMOD-93), 1993.

[Aldana, 2000] Aldana, W.A. Data Mining Industry: Emerging Trends and New Opportunities. Master thesis, Department of electrical Engineering and Computer Science at the Massachusetts Institute of Technology, 2000.

[Aliferes and Cooper, 1994] Aliferes, C. and Cooper, G. An Evaluation of an Algorithm for Inductive Learning of Bayesian Belief Networks Using Simulated Data Set. *In Proceeding of Tenth Conference on Uncertainty in Artificial Intelligence*, 8-14, San Francisco: Morgan Kaufman, 1994.

[Aurélio et al., 1999] Aurélio, M.; Vellasco, M. e Henrique L.C. Descoberta de Conhecimento e Mineração de dados. Notas de aula, Engenharia Elétrica, PUC-RIO, 1999.

[Barquini, 1996] Barquini, R. Planning and designing the Warehouse, New Jersey, Prentice-Hall, 311 pg, 1996.

[Batini and Lenzerini, 1986] Batini, C. and Lenzerini, M. (). Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*. New York, v.18, nº 4, pg.323-364, December, 1986.

[Berry and Linoff, 1997] Berry, M.J.A. and Linoff, G. Data Mining Techniques for Marketing, Sales, and Customer Support, New York, NY: John Wiley and Sons, 1997.

[Berson, 1997] Berson, A. Data Warehousing, Data Mining and OLAP. New York: Mc Graw-Hill, 1997.

[Bigus, 1996] Bigus, J.P. Data Mining With Neural Networks: Solving Business Problems-Form Application Development to Decision Support, ISBN 00070057796, New York; Mc Graw-Hill, 1996.

[Brachman et al., 1996] Brachman, R. J.; Khabaza, T.; Kloesgen, W.; Piatetsky-Shapiro, G. and Simoudis, E. Mining Business Databases. *Communications of the ACM*, Vol. 39 pág. 42-48, 1996

[Carickhoff, 1997] Carickhoff, R. A New Face for OLAP, DBMS Internet Systems, January 1997.

[Carvalho, 1997] Carvalho, J. OLAP sem Segredos. *Computer World*, Rio de Janeiro, v.6, n. 236, p. 28-31, 24 nov, 1997.

[Cazarini, 2002] Cazarini E.W. Projeto de um Data Warehouse. USP, 2002.

[Chandauri and Dayal, 1996] Chandauri, S. and Dayal, U. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD, 1996.

[Chen, 1990] Chen, P. Gerenciando Banco de Dados. São Paulo, McGraw-Hill, 1990.

[Chen et al., 1996] Chen, M.S.; Han, J. and Yu, P.S. Data Mining: An Overview From a Database Perspective. *IEEE Transaction on Knowledge and Data Engineering*, 8(6): 866-883, 1996.

[Codd, 1970] Codd, E.F. A Relational Model of Data for Large Shared Data Banks. CACM 13(6): 377-387, 1970.

[Codd, 1995] CODD, E. F. Twelve Rules for On Line Analytical Processing, *Computer World*, Abril, 1995.

[Dayhoff, 1990] Dayhoff, J. Neural Network Architectures: An Introduction, Van Nostrand Reinhold, New York, NY, 1990.

[Dhar and Stein, 1997] Dhar, V. and Stein, R. Seven Methods for Transforming Corporate data into Business Intelligence, Prentice-Hall, 1997.

[DeWitt and Gray, 1992] DeWitt, D. J.; Gray J. Parallel Databases System: The Future of High Performance Database Systems. *Communications of the ACM,* 1992.

[Dilly, 1995] Dilly, R. *Data Mining* an Introduction. (1995), Student Notes, access 23/09/2002. Available URL:www.qub.ac.uk/courses/datamining/stu-notes/dm-book_1.html

[Diniz, 2000] Diniz, C.A.R. Data Mining: Uma introdução. São Paulo: ABE, 2000.

[Elder and Pregibon, 1996] Elder J. and Pregibon, D. A Statistical Perspective on Knowledge Discovery in Databases. Pages 83-116. In (Fayyad et al., 1996), 1996.

[Fayyad et al., 1996] Fayyad, U. M.; Piatestky-Shapiro, G.; Smith P. and Uthurusamy, R. Advances in Knowledge Discovery and Data Mining. AAIPress, The Mit Press, 1996.

[Fayyad, 1998] Fayyad, U. M. Mining Databases: Towards Algorithms for Knowledge Discovery. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 21, no. 1, 1998.

[Félix, 1998] Félix, L. C. M. Data Mining no Processo de Extração de Conhecimento de Base de Dados. Instituto de Ciências Matemáticas e de Computação, São Carlos, Universidade de São Paulo, 1998.

[Ferreira, 1998] Ferreira B.C.A. Uma Analise da Nova Geração de Sistemas de Apoio à Decisão. Escola de São Carlos da Universidade de São Paulo/ Engenharia de Produção. 1998.

[Ferrer et al., 2000] Ferrer, J.; Aguilar, J. y Peña, J. Data Mining. Departamento de Lenguajes y Sistemas Informáticos, facultad de Informática, Universidad de Sevilla, 2000.

[Figueiredo, 1998] Figueiredo, A. MOLAP x ROLAP: Embate de Tecnologias para Data Warehouse, *Developer's Magazine*, Ano 2 - Numero 18, Fev/1998.

[Forsman, 1996] Forsman, S. OLAP Council White Paper, OLAP Council, San Rafael CA, 1996.

[Gentia Software, 1998] Gentia Software. (1998), access (23/04/2003). URL:http://www.prnewswire.com/gh/cnoc/comp/126953.html

[George, 1997] George H.J. Enhancements to the Data Mining Process. Department of Computer Science, Stanford University, 1997.

[Glymour, et al., 1997] Glymour, C.; Mandingan, D. and Smyth, P. Statistical Themes and Lesson for Data Mining. Data Mining and Knowledge Discovery, 1997.

[Goldberg,1989] Goldberg, D.E. Genetic Algorithms in Search, Optimization and Machine Learning. MA: Addison-Wesley, NY, 1994.

[Han and Kambler, 2000] Han, J. and Kambler M. Data Mining: Concepts and Techniques. Simon Fraser University, 2000.

[Han et al., 2001] Han, J.; Jamil, H.; Lu, Y.; Chen, L.; Liao, Y. and Pei, J. DNA-Miner: A System Prototype for Mining DNA Sequences. Simon Fraser University and Mississippi State University, 2001.

[Hand et al., 2000] Hand, D.J.; Mannila, H. and Smyth, P. Principles of Data Mining. MIT Press, 2000.

[Heckerman, 1996] Heckerman, D. Bayesian Networks for Knowledge Discovery. (Advances Knowledge and data mining, Fayyad U.M. 1996), 1996.

[Hendler, 1996] Hendler, J.A. Intelligent Agents: Where AI Meets Information Technology. IEEE Expert pages 20-23, 1996.

[Holland, 1992] Holland, J. H. Adaptations in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.

[Holsheimer and Siebes, 1994] Holsheimer, M. and Siebes, A.P.J.M. Data Miming: The Search for Knowledge in Databases. Computer Science/Department of Algorithmic and Arquitecture, 1994.

[Hruschka e Silva, 1996] Hruschka, J. E. R. e Silva, W. propagação de Crença e Aprendizado em Redes Bayesianas. Relatório de pesquisa CIC/UnB – 03/96, Departamento de Ciências da Computação, Universidade de Brasília, 1996.

[Inmon, 1996] Inmon, W. H. Building the Operations Data Store, Ed. Jonh Wiley & Sons, 1996.

[Inmon, 1997] Inmon, W. H. Como Construir uma Data Warehouse. Campus, Rio de Janeiro, 1997.

[Kimball, 1996] Kimball, R. The Data Warehouse Toolkit, John Wiley & Sons Inc., New York, 1996.

[Kira and Rendell, 1992] Kira, K. and Rendell, L. The Features Selection Problem: Traditional Methods and a New Algorithm. Proc 10th Int. Conf. Artificial Intelligence (AAAI-92), 1992.

[Koller and Sahami, 1996] Koller, D. and Sahami, M. Toward optimal features selection. Proc 13 th Int. Conf. Machine Learning, 1996.

[Korth and Silberschatz, 1995] Korth H.F. and Silberschatz A. Sistemas de Banco de Dados. 2.ed. São Paulo, Makron Books. 1995.

[Kosala and Blockeel, 2000] Kosala, R. and Blockeel, B. Web Mining Research: A Survey. SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining. ACM Press, 2000.

[Kremer, 1999] Kremer, R. Sistema de apoio à decisão para previsões genéricas utilizando técnicas de data mining. Blumenau, Trabalho de conclusão de curso (Ciência da Computação) - Centro de Ciências Exatas, Universidade Regional de Blumenau, 1999.

[Lavrac et al., 1991] Lavrac, N.; Dzeroski, S. and Grovelnik, M. Learning Nonrecursive Definitions of Relation with LINUS. *Proc. of the Fifth European Working Session on Learning,* 1991.

[Maes, 1994] Maes, P. Agent Adaptive Autonomous Agent. Artificial Life Journal, 1(1):135-162, 1994.

[Mahesh et al., 1998] Mahesh, V.; Eui-Hong H.; Karypis, G.; Kumar, V. Parallel Algorithm in Data Mining. University of Minnesota, Minneapolis, MN 55455, USA, 1998.

[Matheus et al., 1993] Matheus, C. J.; Chan P. K.; Piatesky-Shapiro, G. System for Knowledge Discovery in Databases. GTE Laboratories Incorporated. IEEE TKDE special issue on Learning and Discovery in Knowledge-Based Databases. 1993.

[Metha et al., 1996] Metha, M.; Agraval R. and Rissanen, J. SLIQ: A Fast Scalable Classifier for Data Mining, IBM Almaden Research Center, 1996.

[Michalski and Kaufman, 1997] Michalski, R. and Kaufman, K. Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach, 1997.

[Morik and Brockhausen, 1997] Morik, K. and Brockhausen, P. A Multistrategy Approach to Relational Knowledge Discovery in Databases. Machine Learning, 27, page 287. Kluber Academic Publisher, Boston. 1997.

[Muggleton, 1992] Mugleton, S. Inductive Logic programming: Techniques and Applications. Chi Chester, UK: Ellis Harwood, 1992.

[Nasukawa and Nagano, 2001] Nasukawa, T. and Nagano, T. Text Analysis and Knowledge Mining System. IBM System Journal, knowledge management, 2001.

[Oliveira e Rezende, 1998] Oliveira, R. B. T. e Rezende, S. O. Ferramentas de Visualização de Dados do Mineset. Technical Report 71, ICMC-USP, 1998.

[Oneil, 1997] Oneil, B. Oracle Data Warehousing. Indianapolis, Sams Publishing, 1997.

[Pazzani and Kibler 1992] Pazzani, M. and Kibler, D. The Utility of Knowledge in Inductive Learning. Machine Learning, page 57-94, 1992.

[Pendse, 1998a] Pendse, N. What is OLAP?, The OLAP Report, January 11, 1998.

[Pendse , 1998b] Pendse, N. As Inovações do OLAP. Byte Brasil, São Paulo, v. 7, n. 3, p. 94-98, 1998.

[Perelmuter, 1996] Perelmuter, G. Redes Neurais Aplicadas ao Reconhecimento de Imagens Bi-dimensionais. Dissertação de Mestrado, DEE, PUC – Rio, 1996.

[Piateski and Frawley, 1991] Piateski-Shapiro, G. and Frawley, W. J. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

[Quinlan, 1986] Quinlan, J. R. Induction of Decision Trees. Machine Learning, 1:81-106, 1986.

[Quinlan, 1990] Quinlan, J. Learning Logical Definitions from Relations. Machine Learning, 5(3): 239-266, 1990.

[Quinlan, 1993] Quinlan, J. R. C4.5: programs for machine Learning. Morgan Kaufman, San mateo, CA, 1993.

[Raedt, 1992] Raedt, De L. Interactive Theory Revision: An Inductive Logic Programming Approach. London Academic Press, 1992.

[Reinartz, 1999] Reinartz, T. Focusing Solution for Data mining LNAI – 1623 Springer Verlag, 1999.

[Rezende e Pugliesi, 1998] Rezende, S.O. e Pugliesi, J.B. Aquisição de Conhecimento Explicito ou Manual. Reporte técnico 37, ICMC-USP, 1998.

[Ricarte, 1996] Ricarte I. L. M. Sistemas Paralelos de Banco de dados: Arquiteturas e Algoritmos. Notas de aula, FEEC, Universidade Estadual de Campinas, 1996.

[Rocha, 1999] Rocha, C. A. J. Redes Bayesianas Para Extração de Conhecimento de Bases de Dados, Considerando a Incorporação de Conhecimento de Fundo e o Tratamento de Dados Incompletos. Dissertação de Mestrado, ICMC - Universidade de São Paulo, 1999.

[Rodas, 2001] Rodas, J. Un Ejercicio de Análisis Utilizando *Rougth Sets* en un Dominio de Educación Superior Mediante el Proceso KDD. Barcelona: Departamento de Lenguaje y Sistemas Informáticos, Universidad Politécnica de Catalunia, España, 2001.

[Rodrigues, 2001] Rodrigues F. Data Mining: Conceitos, Técnicas e Aplicação. Dissertação de Mestrado, Escola Politécnica, Universidade de São Paulo, 2001.

[Santos et al., 1999] Santos, J.; Henriques, N. e Reis, V. Data mining e Data Warehousing. Dissertação (Licenciatura em Engenharia Informática) - Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, 1999.

[Sahami et al., 1998] Sahami, M., Dumais, S. Heckeman, D. And Hortvist, E. Bayesian Approach to Filtering Junk: Papers from the 1998 Workshop. AAAI Technical Report, 1998.

[Shafer et al., 1996] Shafer, J.; Agraval, R.; Mehta, M. SPRINT: A scalable parallel classifier for data mining. *In Proc. Of the 22nd VLDB Conference*, 1996.

[Stolfo and Chan, 1995] Stolfo, S.J. and Chan, P.K. An Illustrative Scenario of

Metalearning Agents for Scalable Data mining on the Internet. Dpto. Of Computer Science, Columbia University, 1995.

[Swanson and Smalhairser, 1994] Swanson, D.R. and Smalhairser, N.R. Assessing a Gap in the Biomedical Literature: Magnesium Deficiency and Neurologic Disease. Neuroscience Research Communications, 1994.

[Thuraisinham, 1999] Thuraisinham, B. M. Data Mining: Technologies, Techniques, Tools, and Trends, ISBN 0849318157, CRC Press, 1999.

[Tyo, 1998] Tyo, J. Desktop OLAP tools: if the tool fits, Use It – On line analytical processing tools offer ease of use for data retrieval and analysis with minimal user training techweb news, 1996.

[Tukey, 1977] Tukey, J. Exploratory Data Analysis, Reading, MA, Addison-Wesley, 1977.

[Ullman, 1988] Ullman, J. Principles of Databases and Knowledge Base System. Volume I. Rockville, Mass.: Computer Science Press, 1988.

[Vesato, 2000] Vesato, J. Using SOM in Data Mining. Department of Computer Science and Engineering, Helsinky University of Technology, Finland, 2000.

[Villalobos, 2000] Villalobos A., M. T. Avaliação da Qualidade de Produtos de Software. Divisão de Qualificação de Software, CenPRA, 2000.

[Von Zuben e De Castro, 2003] Von Zuben, F.J. e De Castro, L. Redes Neurais. Notas de aula, FEEC, Universidade Estadual de Campinas, 2003.

[Weiss and Indurkhya, 1998] Weiss, S.M. and Indurkhya, N. Predictive Data Mining: A Practical Guide, San Francisco, CA: Morgan Kaufmann Publishers, 1998.

[Witten and Frank, 1999] Witten, I.H. and Frank, E. Data Mining Practical Machine Learning Tools and Techniques with JAVA Implementations. San Francisco, CA: Morgan Kaufmann, 1999.

[Zaine et al., 1998] Zaine, O.R.; Han, J.; Li, Z.N.; Chee, S.H. and Chiang, J.Y. Multimedia Miner: A System Prototype for Multimedia Data Mining. *Proceeding of International Conference on Management of Data*. ACM SIGMOD, 1998.

Apêndice A

Lista de figuras

Apêndice B

Lista de tabelas

Apêndice C

Lista de abreviaturas e siglas

ACC: Analises de Cesto de Compras

AD: Árvore de Decisão

AG: Algoritmo Genético

BD: Base de Dados

DM: Data Mining

DW: Data Warehouse

EAICBD: Extração Inteligente e Automática de Conhecimento em Bases de Dados.

GD: Garimpagem de Dados

KDD: Knowledge Discovery in Databases

MDB ou MDDB: Multidimensional Database

MOLAP: Multidimensional On-Line Analytical Processing

OLAP: On Line Analyze Processing

RBM: Raciocínio Baseado em Memória

RNA: Redes Neurais Artificiais.

ROLAP: Relational On-Line Analytical Processing

SBD: Sistemas de Bases de Dados

SGBD: Sistema Gerenciador de Banco de Dados

SQL: Structure Query Language