Universidade Federal de Pelotas
Instituto de Física e Matemática
Curso de Bacharelado em Ciência da Computação

O PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS: UM ESTUDO DE CASO SOBRE OS DADOS DA UFPEL

Daniela Francisco Brauner

Pelotas - RS
2003

Daniela Francisco Brauner

O PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS: UM ESTUDO DE CASO SOBRE OS DADOS DA UFPEL

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação do Instituto de Física e Matemática da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Profª.Drª.Gertrudes A. Dandolini
Universidade Federal de Pelotas
Co-Orientador: Prof. Dr. João Artur de Souza
Universidade Federal de Pelotas

Pelotas - RS
2003

O PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS: UM ESTUDO DE CASO SOBRE OS DADOS DA UFPEL

por

DANIELA FRANCISCO BRAUNER

Monografia defendida e aprovada em 21 de fevereiro de 2003 pela banca examinadora constituída pelos seguintes integrantes:

_____

Prof[a] Dr[a]. Gertrudes Aparecida Dandolini/Orientadora (UFPel)

_____

Prof Dr. João Artur de Souza/Co-Orientador (UFPel)

_____

Prof. Dr. Ricardo Azambuja da Silveira (UFPel)

_____

Prof. Msc. Paulo Ricardo Prestes Porto (UFPel)

Dedico este trabalho aos meus pais, pela dedicação, doação, carinho, segurança e amor que sempre dedicaram a mim.

AGRADECIMENTOS

À Deus, pela vida.

Aos orientadores, pela amizade, força e confiança.

Aos professores, pela sabedoria.

Aos colegas, pelo companheirismo.

Aos colegas de trabalho, pela colaboração.

Aos amigos, pelas visitas, telefonemas e compreensão.

Ao namorado, pelos carinhos...

À minha irmã, pela compreensão da minha ausência.

E, principalmente, aos meus pais, por TUDO!

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANSI	*American National Standards Institute*
BD	Banco de Dados
DBA	*Database Administrator*
DCBD	Descoberta de Conhecimento em Bancos de Dados
DDL	*Data Definition Language*
DER	Diagrama Entidade-Relacionamento
DML	*Data Manipulation Language*
DW	*Data Warehouse*
ISO	*International Standards Organization*
ODBC	*Open database Connectivity*
OLAP	*Online Analytical Processing*
PHP	*Personal Home Page*
RBF	*Radial Base Function*
SGBD	Sistemas Gerenciadores de Banco de Dados
SQL	*Structured Query Language*

RESUMO

Este trabalho apresenta um estudo inicial da aplicação do processo de Descoberta de Conhecimento em Banco de Dados (DCBD) nos dados da Universidade Federal de Pelotas (UFPel). Baseado num estudo comparativo de diferentes abordagens do processo de DCBD foi proposto um modelo para este estudo de caso. Para implementação do modelo proposto, foi desenvolvida uma ferramenta que realiza as tarefas de migração e análise dos dados a serem utilizados. Esta ferramenta foi implementada em PHP, acessando os bancos de dados via ODBC e o banco de dados MySQL. Ela é composta de três módulos: módulo de migração, que realiza a tarefa de migração dos dados de um banco de dados de origem para o MySQL; o módulo de análise, que faz a análise da estrutura das tabelas e da ocorrência dos dados; e o módulo de informações, que disponibiliza as informações relevantes descobertas durante o processo. O processo de DCBD proposto mostrou-se eficiente para este estudo de caso, disponibilizando informações relevantes, despertando e motivando a continuidade da pesquisa neste estudo de caso.

Palavras-chave: Mineração de dados, banco de dados, descoberta de conhecimento, sistema acadêmico, redes neurais.

ABSTRACT

This work presents an initial study of the application of the process of Knowledge Discovery in Database (KDD) in the data of the Universidade Federal de Pelotas (UFPel). Based on a comparative study of different views of the KDD process, a model was proposed for this case study. For implementation of the proposed model, a tool was developed in order to accomplish the migration tasks and analyse the data that will be used. This tool was implemented in PHP, accessing the databases through ODBC and the MySQL database. It is composed of three modules: migration module, that accomplishes the task of data migrating from an origin database to MySQL; the analysis module, that makes the analysis of the structure of the tables and of the occurrence of the data; and the information module, that shows the important information discovered during the process. The proposed KDD process turned to be efficient for this case study, providing important information, rousing curiosity and motivating the continuity of the research in this case study.

Key-word: Data mining, database, knowledge discovery, academic system, neural nets.

1 INTRODUÇÃO

Atualmente, os sistemas são implementados com a finalidade de auxiliar as tarefas humanas em qualquer área de atuação. Estes atuam gerando e coletando dados operacionais, ou seja, do dia-a-dia transacional. Grandes empresas, organizações e instituições adotam sistemas computacionais com armazenamento de informações em bancos de dados para os mais diversos fins: sistemas financeiros, controle de estoque, sistemas cadastrais, entre outros.

Os avanços no armazenamento de informações, disponibilizados pelos Sistemas Gerenciadores de Bancos de Dados (SGBDs), tais como: velocidade, facilidade de acesso e baixo custo, impulsionaram a geração e aumento no volume de dados armazenados. Nestas grandes bases, com o passar dos anos, torna-se impossível a análise dos dados de forma manual, tornando-se uma tarefa difícil de ser realizada utilizando-se métodos tradicionais. Estas informações, de extrema relevância para a organização, devem ser utilizadas e manipuladas da melhor forma possível a fim de prover informações importantes para auxiliar na tomada de decisões.

Para que esse grande volume de dados não seja analisado manualmente, existem ferramentas e técnicas de Mineração de Dados que são fundamentadas na idéia de adquirir conhecimento e descobrir tendências, baseadas na análise em busca de padrões nos dados armazenados em bancos de dados.

Com a proliferação da utilização da tecnologia de armazenamento de dados, a utilização de ferramentas baseadas em técnicas de mineração de dados vem crescendo na mesma proporção. O objetivo destas ferramentas é a extração de informações ocultas de grandes quantidades de dados armazenados para a descoberta de conhecimento.

A área de descoberta de conhecimento tem despertado interesse tanto no meio comercial quanto científico e está sendo amplamente difundida e aperfeiçoada pelo fato de fornecer resultados promissores para qualquer domínio de aplicação.

1.1 Motivação

Durante os últimos anos tem sido verificado um crescimento substancial da quantidade de dados armazenados pelas organizações. A análise destes dados,

produzidos e armazenados em larga escala, realizada por especialistas através de métodos manuais tradicionais, é praticamente impossível.

Por outro lado, a informação vem desempenhando um papel fundamental em todos os setores da sociedade e a grande quantidade de dados equivale a um maior potencial de informação. Entretanto, algumas informações não estão caracterizadas explicitamente em um banco de dados, uma vez que, sendo dados operacionais, não interessam quando estudados individualmente. Diante deste cenário, surge a necessidade de explorar estes dados para extração de conhecimento implícito para utilizá-lo no âmbito do problema [GRO 97] e [THE 97].

A Universidade Federal de Pelotas, como entidade irradiadora de informação não só no meio acadêmico, mas também para a comunidade a qual está inserida, deve estar munida de recursos que possibilitem a análise adequada dos seus dados operacionais para extração e disseminação de informação.

Conforme levantamento realizado junto ao Centro de Informática e demais unidades da UFPel, a Instituição não dispõe de nenhum sistema que se utilize de técnicas de descoberta de conhecimento sobre as bases de dados operacionais. Constatou-se que tal estudo seria de extrema importância institucional oferecendo informações anteriormente desconhecidas que se encontram em dados históricos da Instituição, bem como informações relevantes na tomada de decisão provendo agilidade à diversos procedimentos e ações da Administração da Universidade.

1.2 Objetivos

O objetivo principal deste trabalho é o estudo das técnicas de descoberta de conhecimento através de um estudo de caso sobre os dados da Universidade Federal de Pelotas em busca de conhecimento implícito nesses dados.

Os objetivos específicos são:

- Avaliar as bases de dados da UFPel em busca dos dados relevantes para descoberta de conhecimento.
- Estudar algumas abordagens do processo de descoberta de conhecimento em banco de dados com a finalidade de escolher a que melhor se adapte ao estudo de caso.
- Aplicar uma ou mais técnicas de Mineração de Dados nos dados.

- Avaliar as técnicas utilizadas.
- Analisar os possíveis resultados.

1.3 Contribuição esperada

Como conseqüência do estudo do processo de descoberta de conhecimento aplicado aos dados da Instituição, espera-se contribuir significativamente para a comunidade da Universidade Federal de Pelotas. Com os estudos realizados, espera-se fornecer a oportunidade de se obter informações relevantes que auxiliem no processo de tomada de decisão. Com isso, o planejamento estratégico e operacional da Instituição poderá encontrar alternativas para solucionar problemas de seu planejamento e organização. Assim, tanto a Instituição como a comunidade científica, serão beneficiadas com o estudo das técnicas de mineração de dados incorporadas no auxílio à tomada de decisões.

Este trabalho propõe o estudo inicial da aplicação de técnicas de descoberta de conhecimento em banco de dados através de um estudo de caso realizado sobre os dados da Universidade Federal de Pelotas (UFPel). Este estudo contribuirá significativamente com o meio acadêmico através de um exemplo de aplicação real do processo de DCBD, possibilitando a futura ampliação do estudo dando continuidade a esta pesquisa.

O objetivo principal da aplicação do processo de DCBD nestas bases é investigar padrões de comportamento da comunidade da UFPel de acordo com algum padrão até então desconhecido.

1.4 Organização do trabalho

Este trabalho está dividido em cinco capítulos.

No segundo capítulo, são apresentadas diferentes técnicas de armazenamento de dados, dando destaque aos Sistemas Gerenciadores de Bancos de Dados, utilizados neste estudo de caso.

O terceito capítulo aborda as diferentes ferramentas de manipulação de dados. Neste capítulo, são definidas as técnicas mais utilizadas, como as ferramentas de

consulta e os métodos não tão tradicionais, como mineração de dados e o Processo Analítico On-line (*Online Analytical Processing – OLAP*).

O quarto capítulo apresenta as diferentes abordagens para o Processo de Descoberta de Conhecimento em Bancos de Dados (DCBD) fazendo uma comparação entre as visões dos autores dos processos, para a criação de um modelo de processo para ser utilizado no decorrer deste estudo. Neste capítulo, também serão brevemente mostradas as técnicas mais utilizadas para mineração de dados.

O quinto capítulo se detém ao estudo de caso propriamente dito. Com a definição do modelo do processo de DCBD utilizado para padronizar as etapas utilizadas na prática no estudo de caso proposto neste trabalho. Ainda neste capítulo é descrita a aplicação das fases do processo ao estudo de caso, são explícitas as técnicas e ferramentas utilizadas durante as fases do processo e os resultados obtidos. Além disso, apresenta a ferramenta desenvolvida para auxiliar no processo de migração e análise dos dados, bem como sua aplicação nas etapas deste estudo de caso.

Por fim, são apresentadas as conlusões no capítulo sete, juntamente com as propostas para trabalhos futuros.

2 ARMAZENAMENTO DE DADOS

A informação é a matéria prima de maior importância de qualquer organização, seja esta uma empresa, instituição ou orgão governamental. Atualmente, o sucesso e prestígio são medidos através da detenção de conhecimento, aquelas organizações que possuem a habilidade na manipulação de suas informações, são os detentores da matéria prima essencial para o seu desenvolvimento.

Atualmente, milhares de organizações contam com a tecnologia de armazenamento de dados, que contribui para que as informações sejam mais facilmente localizadas, compreendidas e melhor utilizadas. Um Sistema Gerenciador de Bancos de Dados (SGBD) proporciona à organização este controle das informações, gerenciando o armazenamento os dados operacionais.

2.1 Sistema de Arquivos X SGBD

Uma tecnologia bastante utilizada nos primórdios do armazenamento de dados era o sistema de arquivos. Vários programas e aplicações manipulavam arquivos permanentes e à medida que novos dados eram requisitados, novos arquivos e aplicativos tinham que ser criados para atender tais necessidades. O armazenamento era uma tarefa simples, porém a obtenção dos dados organizacionais apresentava inúmeras desvantagens [DAT 91].

No sistema de arquivos, cada aplicativo manipulava arquivos permanentes, gerando redundância nos dados, ou seja, a existência de uma mesma entrada de dados em diversos arquivos. Com o passar dos anos e as atualizações destas aplicações, novos arquivos eram criados com dados atualizados, provocando a inconsistência de dados. Além disso, os aplicativos, muitas vezes por serem programados por pessoas diferentes, armazenavam os dados de forma diferente, não permitindo a integração dos dados entre as aplicações. Como conseqüência, esta forma de armazenamento gerava sérios problemas para a organização, tais como: redundância, inconsistência e isolamento dos dados, além de problemas de segurança e integridade. Para solucionar estes e outros problemas, surgiram os Sistemas Gerenciadores de Bancos de Dados (SGBDs).

2.2 Sistemas Gerenciadores de Bancos de Dados – SGBDs

Um SGBD consiste em uma coleção de dados inter-relacionados e um conjunto de programas para acessá-los. [KOR 99]. O objetivo principal de um SGBD é prover aos usuários uma visão abstrata dos dados, omitindo detalhes de armazenamento físico destes dados, proporcionando um ambiente conveniente e eficiente para definição, armazenamento, recuperação e alteração de dados.

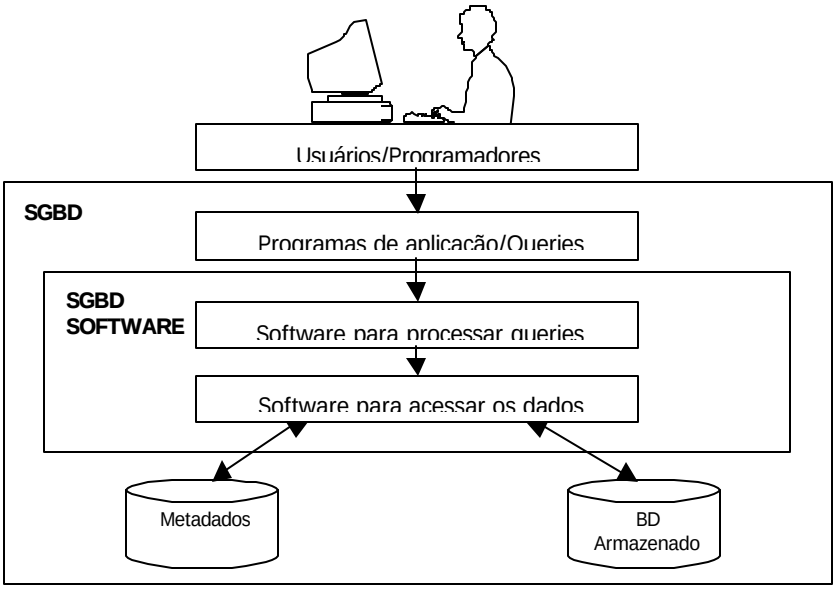

Figura 2.1: Um ambiente simplificado de SGBD

Além do próprio banco de dados, um ambiente de SGBD é composto do software para processar e acessar estes dados, juntamente com as consultas e manipulações dos dados realizadas pelos programadores. A Fig.2.1 ilustra um ambiente simplificado de banco de dados.

A facilidade e rapidez de manipulação e gerenciamento dos relacionamentos dos dados em um banco de dados são os principais motivos da difusão dos SGBDs. Além disso, um sistema de banco de dados proporciona o controle centralizado dos dados operacionais da organização, ou seja, existe um administrador de banco de dados (*Database Administrator - DBA*). O DBA possui responsabilidade central sobre os dados operacionais da organização. Além de possuir alto grau técnico deve ter a capacidade de entender e interpretar as necessidades da organização à nível de gerência executiva e operacional.

2.2.1 Tipos de SGBDs

- **SGBD Relacional:** representa os dados e seus relacionamentos por um conjunto de tabelas e suas respectivas colunas (campos). Os relacionamentos são feitos de acordo com os valores dos registros contidos nos campos das tabelas.

- **SGBD de Rede:** os dados são representados por coleções de registros e os relacionamentos por ligações entre os registros. Estes registros no banco de dados são organizados como coleções de grafos arbitrários.

- **SGBD Hierárquico:** os dados e relacionamentos são representados por registros e ligações. Os registros são organizados como coleções de árvores.

A Tab.2.1 mostra alguns exemplos de cada tipo de bancos de dados e respectivos fabricantes.

TABELA 2.1: Alguns exemplos de produtos de banco de dados

Modelo	Sistema	Fabricante
Relacional	DB2	IBM
	ORACLE	ORACLE Corp.
	SQL Server	Microsoft
	Access	Microsoft
	MySQL	MySQL
	Interbase	Borland
Hierárquico	IMS	IBM
	System 2000	Intel
Rede	IDMS	Cullinet
	DMS 1100	Sperry

Fonte: DATE, C. J.,p.23 [DAT 91]

2.2.2 Vantagens dos SGBDs

- **Compartilhamento de dados:** Diversos aplicativos podem utilizar um mesmo banco de dados para manipular informações em comum, ocorrendo a redução da redundância de dados. Se o DBA estiver ciente das necessidades de dados das aplicações, este pode prover acesso aos mesmos dados à duas aplicações distintas e à novos aplicativos.

- **Evitar a inconsistência de dados:** Um banco de dados inconsistente é aquele que possui duas entradas para o mesmo dado. Quando ocorre o compartilhamento de um mesmo banco de dados, aplicativos distintos correm o risco de duplicar as informações, gerando entradas repetidas para o mesmo dado. Para evitar a inconsistência deve-se eliminar a redundância, ou apenas controlá-la, se esta for realmente necessária, assegurando que quando um dado for atualizado, o outro automaticamente também será.

- **Restrições de Segurança:** O DBA pode restringir o acesso ao banco de dados, definindo meios únicos de acesso e/ou estabelecendo diferentes controles para cada tipo de usuário, definindo acessos somente de atualização ou por exemplo somente para consulta. O controle centralizado dos SGBDs requer também um bom sistema de segurança, pois os usuários terão acesso aos dados operacionais da organização como um todo, podendo modificar estes dados gerando a redundância ou inconsistência.

- **Integridade dos dados:** A inconsistência entre duas entradas de um mesmo dado é um exemplo de falta de integridade. A integridade de dados assegura que os dados contidos no banco de dados estejam corretos, ou seja, o DBA pode definir regras de integridade a serem realizadas a cada atualização dos dados. Um exemplo é um SGBD Relacional, onde a integridade referencial é verificada a cada entrada de dados, ou seja, é verificado se o novo dado que está sendo inserido existe na tabela de referência.

- **Independência dos dados:** Uma aplicação está dependente dos dados quando não pode-se modificar a forma e o local de armazenamento. Utilizando SGBDs as aplicações ficam independentes da estrutura de armazenamento e da estratégia de acesso à estes dados. O SGBD gerencia a forma de acesso dependendo de sua arquitetura interna.

2.2.3 Abstração e modelo de dados

Um SGBD é composto de uma coleção de arquivos inter-relacionados e um conjunto de programas para acessá-los e modificá-los. Para o usuário, o SGBD provê uma visão abstrata dos dados, isto é, o sistema omite detalhes de como os dados são armazenados e mantidos. Para permitir que os aplicativos possam acessar o SGBD para manipulação dos dados, estes devem ser buscados eficientemente. Portanto, os SGBDs simplificam a interação com o usuário através de três níveis de abstração; os quais escondem a complexidade do banco de dados aos usuários [KOR 99].

- **Nível físico:** nível mais baixo de abstração. Descreve como os dados estão armazenados fisicamente, englobando estruturas complexas de baixo nível tais como: dados armazenados em um bloco de posições de memória consecutiva (palavras ou bytes).

- **Nível conceitual:** descreve quais os dados que estão armazenados e seus relacionamentos. Neste nível o banco de dados é descrito através de estruturas simples, tais como os modelos de dados.

- **Nível de visões:** descreve partes do banco de dados, de acordo com as necessidades de cada usuário, individualmente, ou seja, cada usuário tem acesso a um conjunto de tabelas que pode ou não ser visualizado e/ou alterado por outro usuário.

Para descrever as informações armazenadas em uma base de dados, utiliza-se um modelo, onde é feita uma descrição formal das estruturas. Esta modelagem é feita de acordo com os níveis de abstração [HEU 2000].

- **Modelo físico:** informa detalhes de armazenamento físico. É de extrema importância para o DBA, pois serve para fazer a sintonia do banco de dados, procurando otimização e performance. A notação para a modelagem física depende do SGBD utilizado. Alguns realizam grande parte da tarefa de otimização automaticamente.

- **Modelo Conceitual:** descreve o banco de dados de forma independente do SGBD utilizado. A técnica mais utilizada de modelagem conceitual é a abordagem Entidade-Relacionamento (E-R), onde o banco de dados é representado através do Diagrama Entidade-Relacionamento (DER).

- **Modelo lógico:** descreve o banco de dados no nível de abstração visto pelo usuário do SGBD. A modelagem é feita de acordo com o SGBD utilizado. Em um SGBD relacional, onde os dados são organizados em forma de tabelas, o seu modelo lógico define quais são as tabelas contidas no banco e suas respectivas colunas.

A figura Fig.2.2, mostra a relação existente entre os niveis de abstração e a modelagem em banco de dados.

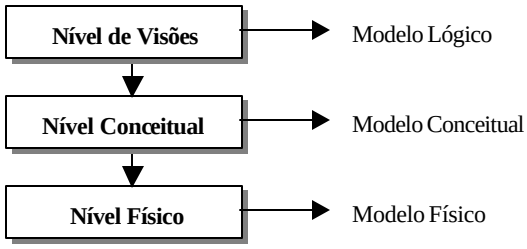

Figura 2.2: Abstração e modelagem de dados

2.2.4 SGBD Relacional

Atualmente, os SGBDs relacionais são predominantes até mesmo em plataformas de grande porte. Nestes ambientes, os SGBDs relacionais estão gradativamente substituindo os bancos de dados de outras tecnologias (hierárquica, redes) [HEU 2000].

Na abordagem relacional, os dados em um banco de dados são armazenados em tabelas (relações) que são conjuntos não ordenados de registros (linhas) compostas por uma série de colunas (campos). Para estabelecer as restrições de integridade e estabelecer os relacionamentos entre as tabelas, utiliza-se um campo identificado como chave. Uma chave primária é uma ou várias colunas que não possuem valores duplicados dentro de uma tabela. Uma chave estrangeira é uma ou várias colunas em que os valores estejam identificados necessariamente como chave primária de outra

tabela. A chave estrangeira é o mecanismo que define os relacionamentos em um banco de dados relacional.

Um banco de dados relacional é composto por diversas tabelas que armazenam os dados operacionais dos sistemas em funcionamento no dia-a-dia da empresa ou organização. Dependendo do tipo e da quantidade de dados que se pretende armazenar, são criados diversos bancos de dados, um para cada domínio de aplicação.

2.2.5 SGBD otimizado para *data warehousing*

Utilizando bancos de dados em SGBDs relacionais, os sistemas em funcionamento nas organizações, armazenam os dados operacionais do dia-a-dia transacional de acordo com a modelagem adotada no banco de dados. Muitas vezes, um banco de dados possui tabelas que sofrem constantes atualizações. Sendo assim, seus dados são sobrepostos com dados atualizados, perdendo as informações históricas.

Para o armazenamento de dados históricos, é utilizado um SGBD com tecnologia apropriada para armazenamento de um armazén de dados (*Data Warehouse - DW*).

Um DW é uma base de dados histórica da organização, contendo um conjunto de dados baseados em assuntos, integrado, variável em relação ao tempo e não volátil [INM 97].

- **Baseado em assuntos:** os dados são organizados de acordo com o assunto relacionado à aplicação. Por exemplo: uma instituição de ensino utilizando um *data warehouse* pode armazenar seus dados por assuntos: dados acadêmicos, patrimoniais e recursos humanos.

- **Integrado:** quando os dados estão em aplicativos separados no ambiente operacional, a codificação é frequentemente inconsistente. Por exemplo: dados de sexo podem ser identificados em um aplicativo como “m” ou “f” e em outro por “0” ou “1”. Quando são transformados para o ambiente de *data warehouse*, eles assumem uma padronização.

- **Variável em relação ao tempo:** Um *data warehouse* armazena dados históricos. Estes dados não são atualizados, ou seja, não são sobrescritos.

Por exemplo, o endereço de um aluno pode ter mudado e ter sofrido atualização no ambiente operacional, mas no *data warehouse* possuímos esta informação histórica.

- **Não volátil:** os dados não são atualizados nem modificados uma vez armazenados no *data warehouse*, são apenas carregados e acessados.

Data warehousing refere-se ao processo de criação de um *data warehouse.* Este processo faz a coleta, limpeza e armazenamento dos dados operacionais por assunto. Um *data warehouse* permanece em funcionamento paralelo aos bancos de dados operacionais da organização. O diferencial, é que em um *data warehouse* não acontecem atualizações nos dados já armazenados, estes são apenas carregados e acessados por ferramentas específicas.

Os principais processos envolvidos na construção e manutenção de um *data warehouse* são [INM 97]:

- **Extração dos dados:** significa entendimento das bases de dados operacionais disponíveis, bem como compreensão dos seus significados e extração de um conjunto de dados relevantes para o domínio da aplicação;

- **Transformação dos dados:** depois de extraídos, os dados devem sofrer transformações para tornar viável sua leitura, devem ser limpos para garantir a integridade da informação e deve ser feita a verificação da qualidade para assegurar a relevância e veracidade dos dados;

- **Publicação dos Dados:** significa a carga de novos dados no *data warehouse* oriundos das bases de dados operacionais. Após a carga dos novos dados, os usuários do *data warehouse* devem ser notificados que novos dados estão disponíveis para acesso;

- **Acesso:** é o processo de recuperação de informação, seja por meio de ferramentas de relatório ou sistemas de suporte à decisão. Ambas tecnologias acessam os dados através do processamento analítico on-line (*Online Analytical Processing - OLAP*);

- **Backup e Recuperação:** deve-se planejar a estratégia de backup e recuperação, analisando as necessidades da aplicação. Entre estas: custo/benefício deste processo, tempo que a base pode permanecer off-line enquanto acontecem estes procedimentos, espaço de armazenamento ocupado pelas cópias de segurança e hardware específico necessário a execução do backup e recuperação.

Um SGBD otimizado para *data warehousing* armazena uma base de dados com dados tratados adequadamente, garantindo consistência, confiabilidade e alta qualidade nos resultados das consultas.

3 MANIPULAÇÃO DOS DADOS

Com a tecnologia de armazenamento de dados em Sistemas Gerenciadores de Bancos de Dados e todas as suas vantagens descritas anteriormente, a maioria das organizações, por menor que sejam, está optando por esta tecnologia como uma alternativa mais ágil e segura de armazenarem informações. Porém, não adianta ter os dados simplesmente armazenados, torna-se necessária uma adequada manipulação dessas informações de acordo com as necessidades da organização.

3.1 Ferramentas de consultas

Em um banco de dados relacional utiliza-se a linguagem estruturada de consulta (*Structured Query Language - SQL*) para definição e manipulação do banco de dados. A versão em uso é ANSI/ISO SQL com seus padrões criados pelo Instituto de Padrões Nacionais Americano (*American National Standards Institute – ANSI*) e pela Organização Internacional de Padrões (*International Standards Organization – ISO*) [KOR 99]. Este é o padrão utilizado em todos os SGBDs. Mesmo aqueles que definem seus próprios padrões SQL, aceitam o padrão ANSI/ISO SQL para acesso aos seus bancos de dados.

A linguagem SQL oferece várias funcionalidades: manipulação, definição e controle de dados. Para consultas, inserções, atualizações e remoções de linhas em tabelas do banco de dados, são utilizados os comandos da Linguagem de Manipulação de Dados (*Data Manipulation Language – DML*). Para criar, alterar ou remover tabelas no banco de dados são utilizados os comandos da Linguagem de Definição de Dados (*Data Definition Language – DDL*).

A maioria dos SGBDs incorporam a linguagem SQL para consultar, acessar e manipular os dados, levando o usuário direto onde ele quer ir para encontrar as informações nos dados de acordo com os padrões pré-definidos pelo usuário. Vários programas são desenvolvidos utilizando ferramentas de consultas SQL para devolver informações aos usuários.

Com estas ferramentas de consultas à bancos de dados, consegue-se somente responder perguntas com hipóteses previamente formuladas - consultas

fechadas. Mais precisamente, o usuário constrói hipóteses sobre associações entre itens de um banco de dados, e o sistema então verifica sua veracidade. A Fig.3.1, mostra um exemplo de uma consulta SQL onde são devolvidos todos os nomes de alunos do curso de Ciência da Computação (identificados na tabela de alunos pelo código do curso 3910), com idade superior à 23 anos.

```
SELECT Nome
FROM Alunos
WHERE CodCurso = 3910
AND Idade > 23
```

Figura 3.1: Exemplo de consulta SQL

O problema com consultas fechadas é que, em muitas situações, os usuários não conseguem formular hipóteses, ou apenas conseguem formular meias hipóteses. Os usuários precisam descobrir padrões nos dados para então gerar hipóteses possivelmente corretas para então comprovar o comportamento do restante dos dados em relação ao padrão descoberto.

3.2 Mineração de Dados (*Data Mining*)

Com o crescente volume de dados armazenados em bancos de dados, torna-se ainda mais difícil a formulação de hipóteses para posterior verificação através de consultas SQL. Por isso, atualmente, técnicas são aplicadas para automatizarem o processo de descoberta de padrões e tendências sobre os dados. As técnicas de mineração de dados, são aplicadas com essa finalidade aos dados de um banco de dados em uma etapa do processo de Descoberta de Conhecimento em Banco de Dados (DCBD), descrito detalhadamente no Capítulo 4.

O usuário final sempre confunde a diferença entre ferramentas de consulta, que permitem o usuário formular perguntas sobre os dados e técnicas de mineração de dados. As técnicas de mineração de dados permitem ao usuário encontrar padrões novos e interessantes sobre os dados armazenados no banco de dados. Estes padrões são comprovados através de consultas SQL que respondem estas perguntas específicas verificando a existência dos padrões [GRO 97].

É importante salientar que as ferramentas de consulta e as técnicas de mineração de dados são complementares, pois a mineração dos dados não substitui uma ferramenta de consulta [ADR 97], mas oferece condições para que as informações sejam descobertas em grandes volumes de dados. Os padrões encontrados são utilizados para prever futuras tendências e comportamentos, permitindo novos processos de tomada de decisão, baseado principalmente no conhecimento desconhecido, frequentemente desprezado, contido nos bancos de dados.

3.3 OLAP: Processamento Analítico On-line

O Processamento Analítico On-Line (*Online Analytical Processing - OLAP*) faz análises multidimensionais dos dados armazenados em um armazém de dados. Estas análises buscam padrões em diferentes níveis de abstração, ou seja, uma visão lógica dos dados.

OLAP é uma análise interativa, permitindo ilimitadas visões através de agregações em todas as dimensões possíveis. Permite obter informações e mostrá-las em tabelas de 2D e 3D, mapas e gráficos. Além disso, derivam-se análises estatísticas (razões, médias, variâncias) envolvendo quaisquer medidas ou dados numéricos entre muitas dimensões. Uma consulta OLAP é executada com um tempo de resposta pequeno, pois é manipulada em um SGBD com otimização para *data warehousing*.

DWs e OLAP são tecnologias complementares. Um DW armazena e gerencia os dados. OLAP transforma os dados de um DW em informação estratégica. OLAP abrange desde a navegação básica e cálculos até análises de séries de tempo e modelagem complexa [FOR 97].

4 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS

A grande quantidade de dados armazenados em um banco de dados equivale a um volume pontencial de informação sobre uma Organização. Porém, esta informação contida nos dados não é explícita, até mesmo porque sendo dados operacionais, do dia a dia da Organização, não interessam quando estudados separadamente.

A simples consulta SQL, abordada no capítulo anterior, extrai dados operacionais para o usuário final. Para obter-se efetivamente o conhecimento sobre a Organização, tornam-se necessárias consultas avançadas e a identificação e reconhecimento de padrões para aplicação direta de consultas em um banco de dados para que retornem resultados interessantes. Estes resultados disponibilizarão informações potencialmente úteis. Utilizando-se técnicas de mineração de dados torna-se possível a exploração desses dados para descoberta de padrões nos dados. Estas técnicas compõem uma etapa do processo de Descoberta de Conhecimento em Bancos de Dados – DCBD (*Knowledge Discovery Database – KDD*).

Este processo tem por objetivo a descoberta de conhecimento em bases de dados, sendo um processo não trivial de identificação de padrões dos dados a fim de extrair informações implícitas e potencialmente úteis dos bancos de dados [FAY 96a]. A finalidade do processo de DCBD é fazer com que os padrões sejam facilmente entendidos pelos usuários, facilitando o entendimento dos dados armazenados a fim de gerar conhecimento.

4.1 DCBD x Mineração de Dados

A literatura sobre 'Mineração de Dados' e 'Descoberta de Conhecimento em Bancos de Dados', propõe diferentes abordagens na definição desses termos. A procura por padrões nos dados é definida de diferentes formas de acordo com o autor. Estas diferentes abordagens são apresentadas a seguir na seção 4.2.

As abordagens mais voltadas à estatísticos, profissionais de tecnologia da informação, analistas de dados e de negócios, chamam o processo de DCBD de

'Mineração de Dados'. Na visão dos profissionais da inteligência artificial o processo é denominado de Descoberta de Conhecimento em Bancos de Dados e Mineração de Dados é apenas a etapa onde são utilizados os algoritmos para extração de conhecimento.

4.2 Etapas do processo de DCBD

Independente da abordagem, o processo é definido por um conjunto de etapas, envolvendo desde o entendimento do domínio da aplicação até a interpretação e consolidação dos resultados. Entre as etapas deste processo exite uma etapa cuja finalidade é a aplicação de técnicas que utilizam algoritmos para extração de conhecimento de um conjunto de dados já preparado para tal.

4.2.1 Etapas segundo FAYYAD

Esta abordagem define a iteratividade das etapas e a interatividade do usuário ao processo. A cada etapa o usuário analisa as informações geradas e procura incorporar sua experiência e tomar decisões para obter resultados cada vez melhores. O processo é composto de cinco etapas (conforme a Fig.4.1) [FAY 96a].

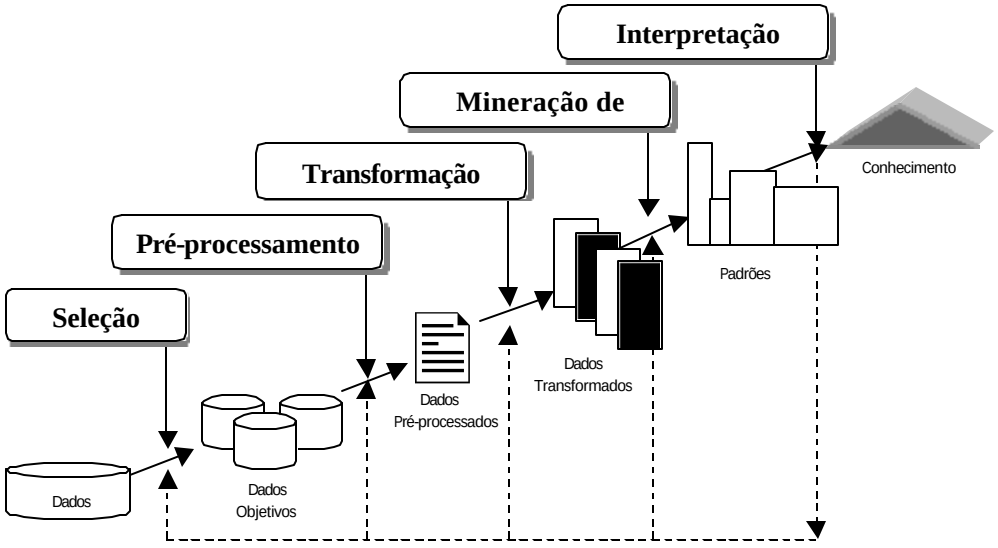

Figura 4.1: Etapas do processo de DCBD definidas por FAYYAD

Fonte: FAYYAD, Usama et al., p.10 [FAY 96a]

Antes do inicio do processo de DCBD, faz-se uma análise do problema a ser resolvido pelo processo de descoberta de conhecimento. O perfeito entendimento do problema é importante para definir corretamente os objetivos do processo de DCBD. A partir daí é feita uma análise dos dados disponíveis para solução do problema.

- **Seleção dos dados:** Após definido o objetivo, parte-se para a etapa de seleção dos dados, onde é feito um subconjunto de dados selecionados a partir da(s) base(s) de dados disponíveis. Este subconjunto conterá apenas aqueles dados relevantes para a solução do problema. O sucesso do processo depende da escolha correta dos dados que formam o conjunto de dados alvo, pois é neste subconjunto que, mais adiante no processo, serão aplicados os algoritmos para descoberta de conhecimento.

- **Pré-processamento:** Após a etapa de seleção, inicia a limpeza e pré-processamento dos dados. Os dados quando armazenados num banco de dados, muitas vezes aparecem com alguns problemas aparentes, tais como: informações incompletas, dados redundantes, ruído e incerteza. Nesta etapa, devem ser estudadas e aplicadas as estratégias para tratamento desses dados.

- **Transformação:** Dependendo do objetivo da tarefa, os dados armazenados no banco de dados podem não ser suficientes. Podem ser necessárias outras informações que poderão ser geradas a partir dos dados armazenados no banco de dados. Para isso, utilizam-se métodos de transformação para gerar outros dados relevantes. Além disso, os dados devem estar no formato exigido pelos algoritmos escolhidos na etapa de mineração. Portanto, o ideal seria, primeiramente, definir a técnica e o algoritmo minerador que serão utilizados para então transformar os dados para o formato adequado ao algoritmo escolhido.

- **Mineração de dados (*Data Mining*):** Etapa caracterizada pela busca de padrões nos dados. Nesta etapa, é escolhido o método e são definidos os algoritmos que realizarão a busca pelo conhecimento implícito e útil do banco de dados. É a fase mais importante do processo de DCBD onde dados

são transformados em informação. Por isso, é importante que seja realizada quando os dados estiverem corretos e a tarefa seja adequada para alcançar o objetivo.

- **Interpretação dos resultados:** Esta é a última etapa da DCBD, onde é realizada a interpretação dos resultados obtidos após a aplicação do algoritmo minerador. A principal meta dessa fase é melhorar a compreensão do conhecimento obtido, em forma de relatórios demonstrativos, com a documentação e explicação das informações relevantes descobertas no processo de DCBD. Os resultados do processo de descoberta do conhecimento podem ser mostrados de forma que possibilite uma análise criteriosa para identificar a necessidade de retornar a qualquer uma das etapas anteriores do processo de DCBD, caso os resultados não sejam satisfatórios.

4.2.2 Etapas segundo ADRIAANS e ZANTINGE

Esta abordagen do processo de DCBD baseia-se na necessidade das organizações em obterem continuamente novas informações sobre seus dados, por isso não deve ser executado apenas uma vez, mas repetido sempre que novas necessidades de informações aparecerem.

Portanto, nesta abordagem não existe uma etapa específica para entendimento dos dados. É pressuposto que já exista um conhecimento prévio do domínio do banco de dados e, consequentemente, do objetivo do processo. O processo é composto por seis etapas (conforme a Fig.4.2) [ADR 97].

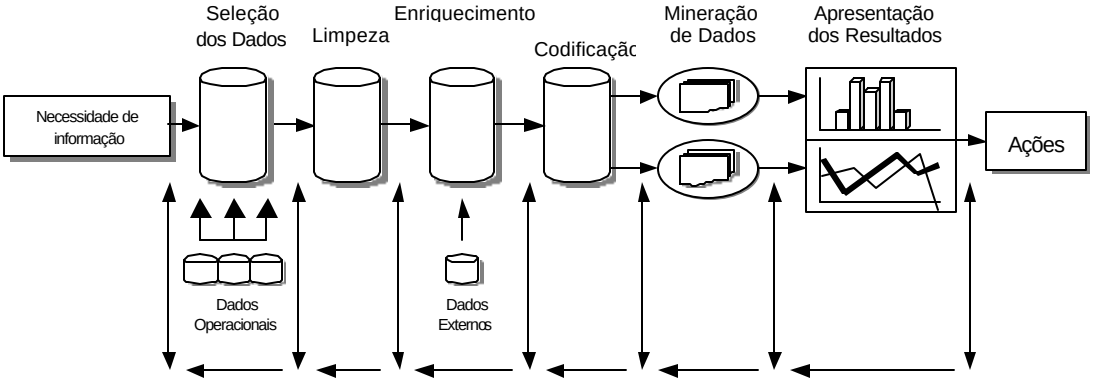

Figura 4.2: Etapas do processo de DCBD definidas por ADRIAANS e ZANTINGE

Fonte: ADRIAANS, Pieter e ZANTINGE, Dolf , p.38 [ADR97]

- **Seleção dos dados:** Nesta etapa ocorre uma análise de todos os dados operacionais do banco de dados e são selecionados apenas aqueles que são necessários para alcançar os objetivos do processo. Podem ser feitas novas seleções quando houver outra iteração, ou seja, pode-se incluir dados anteriormente descartados, pois o processo é iterativo, permitindo a retomada de qualquer etapa, independente da etapa em que se encontra.

- **Limpeza:** Nesta etapa são utilizadas diversas operações de limpeza nos dados, como por exemplo: a remoção de dados duplicados e a correção dos dados. A etapa de limpeza pode ser executada inúmeras vezes, já que é impossível prever com antecedência todos os problemas de qualidade existente na base.

- **Enriquecimento:** Algumas informações podem ser incluídas ao banco de dados para que seja possível atingir os objetivos do processo. Estes dados podem estar disponíveis em outros locais, ou até mesmo podem ser gerados a partir de dados existentes no banco de dados e transformados para obtermos a informação.

- **Codificação:** A forma que os dados estão armazenados nos bancos de dados pode não ser a representação mais apropriada para a utilização no processo de DCBD. Geralmente, os dados têm sua representação apropriada ao

contexto da aplicação. Por exemplo, um atributo com valores literais pode não ser adequado a determinados algoritmos mineradores utilizados na etapa de mineração de dados. Para adequá-lo pode ser necessário normalizar estes valores dentro de um determinado intervalo. A codificação é um procedimento criativo, existem diversas maneiras de codificação, assim é difícil descrevê-las, pois cada caso deve ser analisado individualmente e sua codificação pode variar de acordo com a escolha do algoritmo minerador da próxima etapa.

- **Mineração de dados:** Esta é a etapa onde os dados são manipulados para que seja extraído o conhecimento. É a etapa que mais exige dos recursos computacionais. O autor afirma que, utilizando inicialmente uma ferramenta de consulta SQL, pode ser possível ter uma visão geral dos dados para então partir para uma análise menos trivial. Nesta primeira tarefa, 80% do conhecimento é extraído e já pode revelar alguma informação interessante. Entretanto, as informações extraídas por estas consultas podem não ser suficientes, surgindo a necessidade de utilizarmos técnicas avançadas.

- **Apresentação dos resultados:** Finalizada a etapa de mineração de dados, resultam informações num formato específico de acordo com a técnica utilizada. Deve-se levar em conta que os dados podem estar codificados ou mesmo que o método utilizado na etapa de mineração gere, como saída, informações em algum formalismo ou representação muito específicos. Estes resultados devem ser exibidos de forma clara para que sejam de fácil entendimento para quem irá utilizá-los, geralmente pessoas que necessariamente não interpretarão os resultados tão facilmente quanto aquela que conduziu o processo de DCBD.

4.2.3 Etapas do modelo CRISP-DM 1.0

Nesta abordagem, o termo 'Mineração de Dados' (*Data Mininig - DM)* descreve o processo completo de descoberta de conhecimento em bancos de dados. O objetivo é criar um modelo que auxilie na tomada de decisão, prevendo o

comportamento futuro baseado na análise de atividades passadas. Esta abordagem é mais voltada aos profissionais de negócios.

A modelagem CRISP-DM (*CRoss-Industry Standard Process for Data Mining*), foi desenvolvida para servir como uma metodologia padrão, que visa identificar as diferentes fases na implantação de um projeto para mineração de dados em bancos de dados empresariais. Esta metodologia foge aos conceitos anteriores, pois descreve o processo realizado no dia-a-dia de profissionais de negócios.

A implementação de um projeto de mineração de dados consiste no desenvolvimento de etapas, cuja sequência não é rígida, sempre que necessário podem acontecer retornos e avanços entre as diferentes etapas. Esta iteratividade depende dos resultados disponibilizados pela fase anterior.

O processo é composto por seis etapas (conforme a Fig.4.3) [CHA 2000].

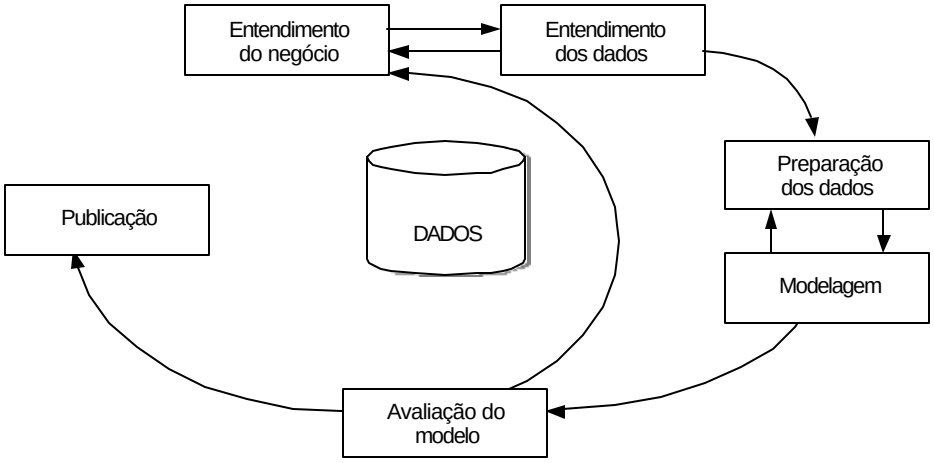

Figura 4.3: Etapas segundo o modelo CRISP-DM 1.0

Fonte: CHAPMAN, Pete et al., p.13 [CHA 2000]

- **Entendimento do negócio:** Nesta fase identifica-se as metas que deseja-se alcançar com a aplicação desta deste processo. Assim, este objetivo é convertido para uma aplicação de mineração de dados e é definido um plano de ataque ao problema.

- **Entendimento dos dados:** Esta fase tem como atividade principal extrair uma amostra dos dados disponíveis para avaliar os possíveis problemas de qualidade e para detectar subconjuntos relevantes para formular hipóteses de descoberta de informações ocultas.

- **Preparação dos dados:** A fase de preparação engloba as atividades de construção de um conjunto de dados para posterior aplicação do modelo. Geralmente, esta etapa, é refeita múltiplas vezes e em nenhuma ordem pré-definida. São realizadas tarefas de seleção, transformação e limpeza no conjunto de dados para uso pelos algorítmos de mineração de dados.

- **Modelagem:** Seleção de um ou mais algorítmos a serem utililizados no efetivo processamento do modelo. Alguns algorítmos necessitam dos dados em formatos específicos, o que acaba causando vários retornos à fase de preparação dos dados.

- **Avaliação do modelo:** Ao final da fase de modelagem, vários modelos devem ter sido avaliados sob a perspectiva do analista responsável. Agora, o objetivo é avaliar os modelos com a visão do negócio, se certificando que não existem falhas ou contradições com relação às regras reais do negócio.

- **Publicação:** Após a validação do modelo é preciso tornar a informação gerada acessível. Isto pode ser feito de várias maneiras, desde a criação de um software específico, até a publicação de um relatório para uso interno da empresa.

4.3 Comparação entre as diferentes abordagens do processo de DCBD

Em virtude das diversas abordagens presentes na bibliografia, optou-se por realizar um estudo comparativo entre as três propostas estudadas. Através da análise comparativa, mostrada na Tab.4.1, foram detectadas algumas similaridades. Porém nenhuma das abordagens adequou-se ao estudo de caso proposto. Consequentemente, propõe-se um modelo do processo integrando estas diferentes abordagens. No Capitulo 5, serão descritas as etapas do modelo proposto para este estudo de caso.

As diferentes visões do processo de DCBD (ou 'Mineração de Dados') em etapas, sugerem existir uma trajetória linear das etapas. No entanto, isso não se verifica, uma vez que em cada etapa pode ser identificada a necessidade de retorno para uma das

etapas anteriores. Por exemplo, se já numa etapa avançada, onde técnicas já estão sendo aplicadas para descoberta de conhecimento, for verificada a necessidade de um dado que não havia sido previsto anteriormente, isso pode levar à retomada à fase de seleção.

Assim, o processo ocorre de forma iterativa, pois cada etapa pode possuir interseções com as demais. Da mesma forma, este processo é interativo, onde a cada etapa as informações geradas são analisadas e enriquecidas com outras informações para obter-se resultados melhores.

TABELA 4.1 : Comparação entre os processos

Etapas	Fayyad et al.	Adriaans et al.	CRISP-DM 1.0
Definição do objetivo	O objetivo é definido para solucionar o problema que gerou a necessidade de implementação de um processo de descoberta de conhecimento.	O objetivo surge da constante necessidade das organizações em obterem novas informações sobre seus dados.	É definido na primeira fase, onde o responsável pelo processo realiza o entendimento do domínio do negócio para então definir o objetivo do processo e critérios de sucesso.
Entendimento dos dados	Ainda durante a fase de definição do objetivo é realizado o entendimento do domínio e avaliação das possibilidades de sucesso de acordo com os dados disponíveis.	Não existe a etapa de compreensão dos dados. É pressuposto que o objetivo proposto pelo usuário já está vinculado aos dados a serem selecionados.	Descreve os dados disponíveis, avaliando quantidade e problemas de qualidade dos dados. Detecta dados relevantes para formular hipóteses.
Seleção dos dados	É feito um subconjunto de dados selecionados a partir da(s) base(s) de dados disponíveis. Este subconjunto conterá apenas aqueles dados relevantes para a solução do problema.	Ocorre uma análise dos dados e são selecionados apenas os dados relevantes com a finalidade de reduzir o volume do conjunto de dados.	Tarefa realizada na etapa de preparação dos dados, onde é selecionado um conjunto que pode ser modificado a cada nova iteração.
Limpeza	Realizada na etapa de pré-processamento, porém o autor não define como deve ser conduzida a tarefa de limpeza.	Utiliza algumas operações de limpeza nos dados: deduplicação de dados, correção,etc. Executada inúmeras vezes, já que é impossível prever todos os problemas de qualidade.	Também realizada na etapa de preparação dos dados, todas as tarefas de limpeza realizadas são documentadas.
Transformação	Utiliza algumas técnicas para formatação dos dados, para adequa-los ao algoritmo de mineração. O autor indica que seja primeiramente definida a técnica de mineração e escolhido o algoritmo que será utilizado.	Trata de como incorporar dados externos à base (enriquecimento) e como transformar os dados (codificação) para que possam ser utilizados na etapa de mineração. O autor apresenta métodos de codificação dos dados.	Outra tarefa realizada durante a etapa de preparação dos dados. Onde os dados são formatados para entrada no modelo gerado pelo algoritmo minerador, porém sem alteração do seu significado.
Mineração dos dados	Considerada a fase mais importante do processo, onde é escolhido o método, definido(s) o(s) algoritmo(s) e realizada a busca pelo conhecimento no banco de dados.	Definição da técnica e dos algoritmos, porém, é garantido que 80% do conhecimento pode ser extraído por consultas SQL e complementado com a utilização de técnicas e algoritmos avançados.	Tarefa de modelagem, onde é definida a técnica, construído o modelo, testado e aplicado no banco de dados.
Resultados	Ocorre a interpretação dos resultados obtidos com o objetivo de melhorar a compreensão do conhecimento. Caso não tenha alcançado os objetivos o processo é retomado.	Os resultados são interpretados com o objetivo de melhorar a compreensão, pois os dados podem estar codificados dependendo do método utilizado na etapa de mineração.	Na fase de avaliação é feita uma validação dos resultados de acordo com as regras reais do domínio da aplicação.

4.4 Mineração de Dados

Embora alguns autores adotem essa nomenclatura para o processo como um todo, considera-se neste trabalho, Mineração de Dados como uma etapa do processo de Descoberta de Conhecimento em Bancos de Dados. Esta é a etapa onde ocorre efetivamente a descoberta de conhecimento.

Indiferente da abordagem adotada para o processo de DCBD, a etapa de Mineração de Dados possui alguns objetivos básicos que devem ser pré-determinados. Nesta fase é escolhida a técnica que se deseja utilizar para construir o modelo para a descoberta de padrões. Esta escolha depende fundamentalmente do objetivo da aplicação. O modelo gerado será baseado em técnicas de Mineração de Dados que utilizam algoritmos específicos para a descoberta de conhecimento.

4.4.1 Objetivos básicos da Mineração de Dados

Os objetivos básicos da Mineração de Dados são a **predição** e a **descrição**, e são diretamente relacionados com o tipo de conhecimento que deseja descobrir e o objetivo da aplicação [AVI 98].

A predição utiliza dados existentes na base de dados para previsão de valores desconhecidos ou futuros de outros dados de interesse. A descrição descobre padrões que descrevem os dados e são facilmente interpretáveis pelo usuário [FAY 96a]. Para alcançar estes objetivos são utilizadas técnicas de modelagem. Existem técnicas capazes de modelar, ao mesmo tempo os dois objetivos.

4.4.2 Construção do Modelo

Um conceito importante, é o de **padrão**. Encontrar um padrão nos dados, significa selecionar conjuntos de dados que apareçam freqüentemente no banco de dados. Consequentemente, é importante provar a necessidade de usar ou não todos os dados para tirar conclusões sobre o que poderia estar acontecendo com o restante dos dados.

Outro conceito associado é a definição de **modelos**. Os modelos são gerados a partir dos padrões. Ou seja, se foi encontrado um padrão em alguns dados, deve-se ter

a confirmação de que o modelo construído para a previsão ou descrição validará esse padrão para qualquer amostra de dados [BER 99].

- **Modelo:** uma descrição construída a partir do banco de dados original que pode ser aplicado aos dados para fazer previsões.

- **Padrão:** um evento, ou combinação de eventos que ocorre freqüentemente em um banco de dados.

A construção de um modelo a partir dos padrões encontrados nos dados é chamada **modelagem**. Este modelo será aplicado ao restante dos dados comprovando a existência dos padrões em toda a base. As técnicas de modelagem surgiram de diversas áreas de aprendizado de máquina, processamento de sinais, computação evolutiva e estatística. Esta diversidade é a principal responsável da Mineração de Dados ser específica para cada objetivo de aplicação.

Para cada tipo de aplicação do processo de DCBD (marketing, medicina, análises corporativas ou institucionais), é feita a seleção da técnica de modelagem a ser realizada. Pois é necessária a criação de um modelo específico para cada tipo de informação que se deseja extrair. Por isso, é necessário o conhecimento da relação existente entre o objetivo da aplicação com os objetivos básicos da Mineração de Dados.

4.5 Técnicas de modelagem

As técnicas de modelagem para Mineração de Dados utilizam algoritmos para gerar um modelo. Nenhuma técnica resolve todos os problemas, a utilização de diversos tipos de técnicas e algoritmos torna-se necessária para proporcionar melhores resultados [BER 97].

Entre os algoritmos utilizados podem ser citados: árvores de decisão, regras de associação, redes neurais, algoritmos genéticos, OLAP, ferramentas de consultas e visualização, e técnicas estatísticas. As técnicas são aplicadas à objetivos distintos, que se baseiam na utilização de um ou mais algoritmos específicos. Algumas técnicas de modelagem são mostradas a seguir:

- **Classificação:** É uma técnica de modelagem com o objetivo de gerar um modelo que classifique os dados. Em geral, os algoritmos de classificação começam com um conjunto de treinamento. Este conjunto é composto de dados classificados para determinar o conjunto de parâmetros que constitui o modelo, que será mais tarde utilizado para a classificação do restante dos dados.

  Uma algoritmo classificador eficiente, será usado de forma preditiva para classificar novos registros nas classes pré-definidas. Árvores de decisão e redes neurais artificiais são exemplos de algoritmos de classificação.

- **Agrupamento ou Clusterização:** É uma técnica de modelagem que gera um modelo descritivo, que divide a base da dados em subconjuntos de dados com características em comum. Essa segmentação é realizada automaticamente por algoritmos e muitas vezes nas primeiras etapas dentro da fase de Mineração de Dados. Esta tarefa identifica grupos de registros correlatos, que serão usados como ponto de partida para futuras explorações.

- **Regras de Associação:** A premissa básica para associação é encontrar associações relevantes entre atributos de uma linha da tabela do banco de dados, isto é, determinam relacionamentos entre conjuntos de itens. Esta tarefa é geralmente aplicada em supermercados, no planejamento de promoções de vendas, segmentação de clientela baseada em padrões de compra e marketing direto.

- **Padrões Seqüenciais:** Seqüências são um tipo especial de associação, onde os itens associados são resultantes de transações diferentes, ou seja, os dados de entrada são tipicamente uma lista de transações seqüenciais.

  As seqüências visam determinar padrões de ordenação entre dados, tais como: ordenação temporais ou ordenação por classificação. A ordenação por séries de tempo visa definir grupos com séries de tempo similares, ou seja, mesmo padrão de comportamento num determinado intervalo de

tempo. A análise de seqüências pode identificar padrões temporais utilizados para predizer acontecimentos futuros.

Algoritmos de padrões seqüenciais são especialmente úteis para companhias de catálogos. É aplicável também em empresas de investimento financeiro, que serão capazes de analisar seqüências de eventos que afetam os preços dos instrumentos financeiros [BER 99].

4.6 Aplicação e avaliação do Modelo

Após a construção do modelo, é necessário gerar um procedimento para testar a qualidade e validade do modelo. Por exemplo, se a técnica escolhida utiliza um algoritmo para classificação, é comum utilizar taxas de erros como medida de qualidade do modelo. Portanto, separa-se os dados em conjunto de treinamento e conjunto de teste, construímos o modelo a partir do conjunto de treinamento e estima-se a qualidade no conjunto de testes.

As diferentes técnicas são comparadas de acordo com algumas características [KOK 2000]:

- **Eficiência:** O esforço computacional necessário para obter uma boa generalização dos dados e a qualidade da regra geral gerada, medida de acordo com a performance da técnica em dados desconhecidos.

- **Interpretação dos resultados:** As técnicas diferem em suas linguagens. A saída de cada uma tem uma forma específica, que pode ou não ser entendida pelo usuário. Tecnicamente isto faz pouca diferença, mas na prática, o entendimento humano de uma regra é desejável para qualquer aplicação.

5 O ESTUDO DE CASO

A partir da análise das diferentes visões do processo de DCBD [FAY 96a], [ADR 97], [CHA 2000], constatou-se que nenhuma delas se adequou para este estudo de caso. Em virtude disso, houve a necessidade da modelagem de um processo para aplicação neste estudo de caso, baseado nas três propostas, conforme mostrado na Fig.5.1.

O modelo proposto inicia com a etapa de entendimento, onde é analisado o domínio da aplicação, bem como os dados disponíveis. Na seqüência, parte-se para a definição do objetivo do processo. A seguir vem a etapa de preparação dos dados, onde são realizadas as etapas de seleção, limpeza e transformação dos dados. A etapa seguinte é a mineração dos dados propriamente dita, onde são definidas as técnicas e algoritmos que serão utilizados no processo. A última etapa é onde acontece a interpretação dos resultados da mineração dos dados, transformando os resultados obtidos em conhecimento para a Organização.

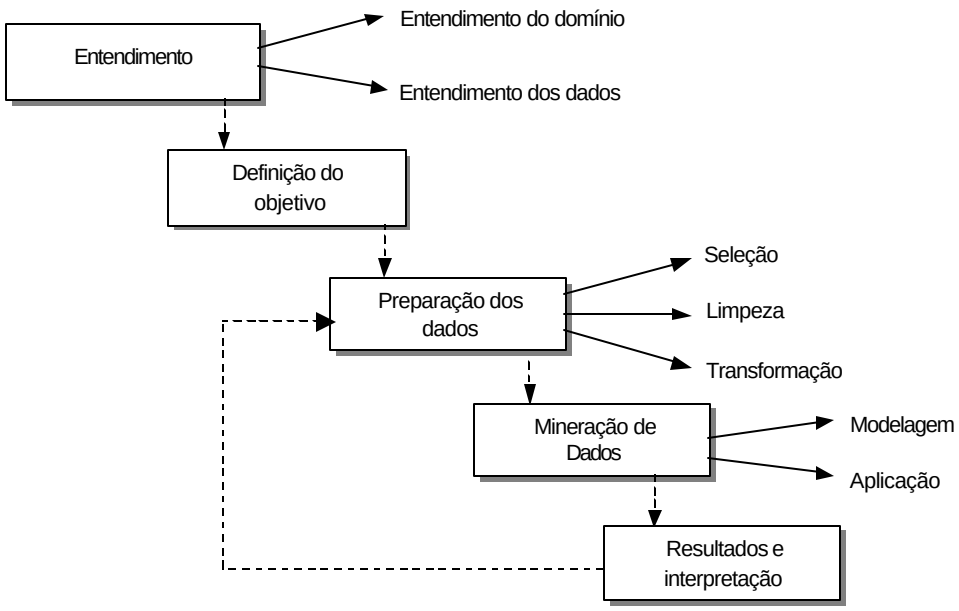

Figura 5.1: Processo de DCBD proposto

5.1 Entendimento

Para definir o objetivo do processo de DCBD é necessário o estudo do domínio da aplicação, para detectar as informações relevantes disponíveis e identificar as necessidades do usuário final. No entanto, raramente o usuário consegue expressar sua necessidade de uma forma diretamente implementável, isto é, para a execução do processo torna-se necessário traduzir esta necessidade em termos do processo. É preciso verificar quais informações estão ligadas a esta necessidade, ou seja, quais são os dados relevantes, se eles existem e como estão formatados.

Assim, torna-se necessário um conhecimento aprofundado do domínio e dos dados em questão. Assim, será possível definir o objetivo da aplicação do processo de DCBD e então selecionar os dados relevantes para prepará-los para posterior mineração.

5.1.1 Análise e entendimento do domínio da aplicação

A Universidade Federal de Pelotas (UFPel) foi criada em 1969, em resposta a demandas crescentes por uma instituição de ensino superior na região. A Tab.5.1 mostra alguns dados institucionais atuais da UFPel. A Instituição oferece, aproximadamente 43 cursos de graduação e 5970 alunos matriculados em 2002.

TABELA 5.1: Informações institucionais da UFPel em 2002

Dados do ano de 2002	Qtde.
Unidades Administrativas	21
Departamentos	73
Cursos de Graduação	43
Alunos Matriculados	5970
Quadro docente permanente	701
Quadro docente substituto	179

O quadro docente da UFPel, conta com aproximadamente, 880 professores que desenvolvem atividades de ensino (ministram disciplinas), pesquisa (desenvolvem projetos de ensino ou pesquisa), extensão (coordenam cursos de extensão ou seminários) e administração (administram os colegiados de cursos, departamentos, unidades, pró-reitorias, entre outras).

A UFPel incentiva a produção individual do quadro docente através da avaliação dessa produtividade. Para isso, a cada ano, os professores em exercício preenchem o Relatório Anual de Atividades Docente (RAAD). Como incentivo, a UFPel oferece a Gratificação de Estímulo à Docência (GED). A GED é uma gratificação no salário calculada em função da pontuação acumulada pelo professor nas atividades descritas anualmente no RAAD.

5.1.2 Análise e entendimento dos dados

De acordo com levantamento feito junto ao Centro de Informática da UFPel, a Instituição dispõe de diversos sistemas para registro do seu dia-a-dia operacional. Tais dados são armazenados em bancos de dados distintos.

Para este estudo de caso, foi concebida a autorização para análise e utilização dos dados do RAAD e do Sistema de Gestão Acadêmica, com restrições quanto aos dados pessoais de identificação de alunos e docentes. Os dados foram disponibilizados em formatos distintos: parte em banco de dados Microsoft Access, parte em arquivo texto migrado do SGBD IBM DB2 e parte em planilhas e documentos. Todos os dados selecionados foram migrados para o banco de dados livre MySQL, para acesso e manipulação, esta migração será melhor explicada na etapa de seleção dos dados.

- **Sistema de Gestão Acadêmica (SGA):** Este sistema controla toda a vida acadêmica dos alunos da UFPel. Ele está em processo de reengenharia desde o segundo semestre de 2002, sendo assim, ainda possui muitos dados inconsistentes, com modelagem para banco de dados do antigo mainframe onde funcionava. Atualmente, estes dados foram migrados para um servidor com SGBD DB2 da IBM. Em virtude de sua origem, os dados ainda estão sem garantia da qualidade. Ainda existem muitos dados nulos e inconsistentes. Além disso, o sistema antigo não possuía documentação suficiente para facilitar o entendimento dos dados. O entendimento foi feito consultando um especialista no novo sistema e no dia-a-dia operacional do Departamento de Registros Acadêmicos.

- **Relatório Anual de Atividade Docente (RAAD):** Este sistema armazena informações anuais das atividades dos docentes, armazenando as atividades de ensino, extensão, administração, produção intelectual e pesquisa. Para este estudo foram disponibilizados os bancos de dados separados por ano: de 1997 à 2002. Existem algumas diferenças nas estruturas de algumas tabelas, em virtude das atualizações no programa RAAD.

Ambos os sistemas possuem características temporais nos dados armazenados. Esta característica é determinada pelo período de validade dos dados, que determina o intervalo de tempo no qual a informação é verdadeira. Um exemplo típico de aplicação temporal é encontrado na tabela de cadastro de docentes da UFPel. Por exemplo, no ano de 1998, o docente declarou no RAAD que sua titulação era em nível de "Mestrado", já no ano de 2002 sua titulação mudou para "Doutorado".

A partir do conhecimento destes sistemas, foi realizada uma análise nas tabelas em busca de padrões nos dados.

Esta análise foi realizada com auxílio de uma ferramenta desenvolvida em PHP (*Personal Home Page*) e executada em um servidor WEB utilizando o acesso ao banco de dados MySQL e a qualquer outro banco de dados acessada via o padrão ODBC (*Open DataBase Connectivity*) do Windows. Foi escolhida a linguagem PHP por ser uma linguagem independente de plataforma e o banco de dados MySQL v.3.23.47 por ser um banco de dados não comercial.

O módulo que faz a análise, verifica a estrutura das tabelas e dos dados que serão manipuladas no decorrer do processo de DCBD. Além disso, disponibiliza um procedimento para padronização dos dados que foi utilizada na etapa de preparação dos dados.

A análise da estrutura da tabela retorna o nome, tipo de dados e tamanho das colunas da tabela analisada, conforme a Fig.5.2. Quando selecionada uma coluna desta tabela, a ferramenta faz a análise dos dados, retornando o número de registros existentes e o número de ocorrências distintas deste registro, conforme mostrado na Fig.5.3.

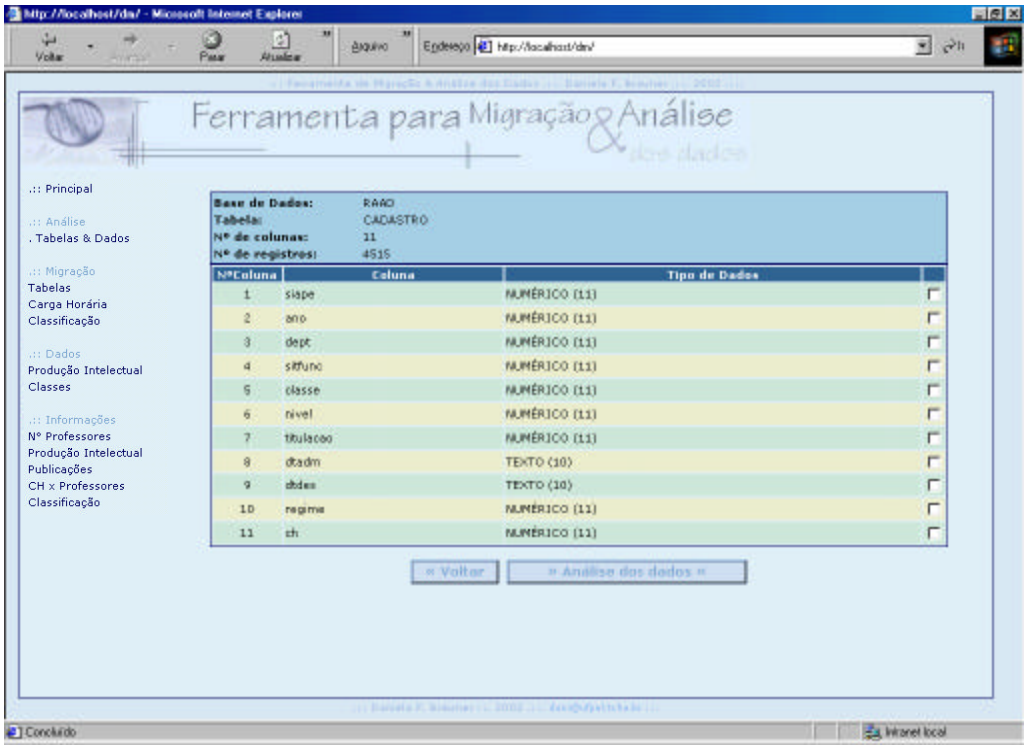

Figura 5.2: Análise da estrutura da tabela "Cadastro"

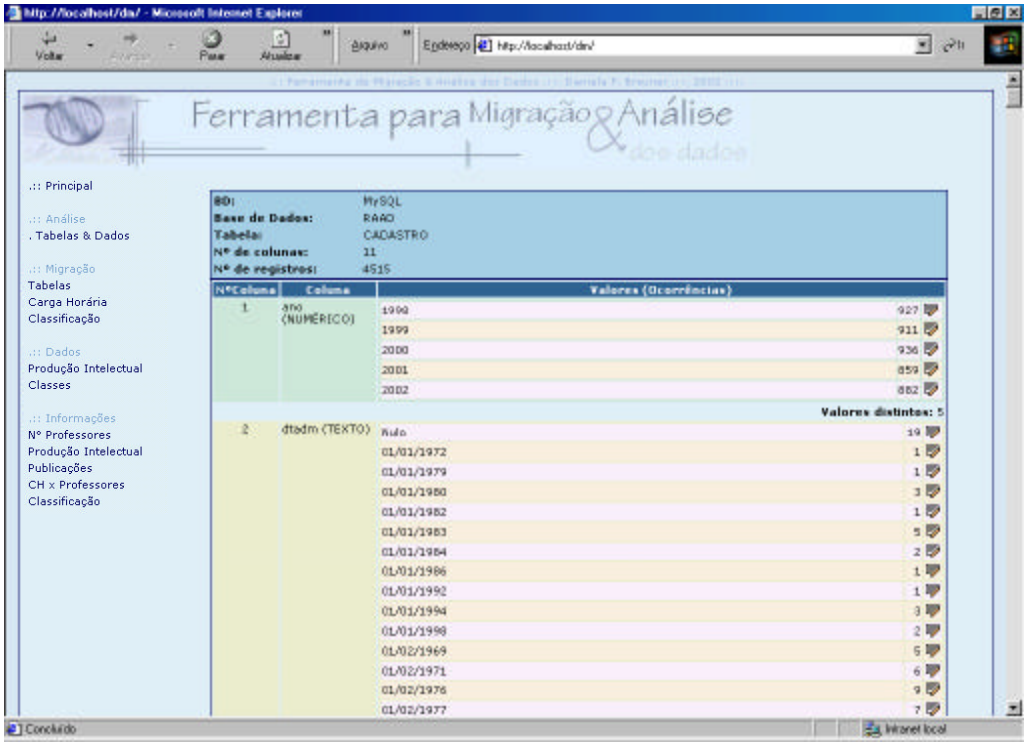

Figura 5.3: Análise dos dados

Os dados que apareceram mais freqüentemente no banco de dados determinaram algum tipo de comportamento. De acordo com esta análise foram utilizadas técnicas heurísticas e foi possível definir algumas alternativas de estudo de padrões para posterior modelagem e aplicação no restante dos dados disponíveis:

1. A relação entre a idade do aluno que prestou vestibular e a sua classificação: para a realização deste estudo, seria utilizada a tabela com registros de alunos do Sistema de Gestão Acadêmica. Esta tabela possui uma coluna que contém a nota no vestibular de cada aluno, além de informações pessoais, tais como: nome, data de nascimento, endereço, cidade natal, curso matriculado, entre outras. Para isto foi necessária a investigação da forma de avaliação do vestibular. Constatou-se que, o vestibular contou com um tipo de avaliação diferente para cada ano, tornando-se inviável a classificação destes alunos, visto que não existe documentação para determinar uma classificação padronizada para as notas cadastradas no sistema.

2. A produção intelectual nos departamentos e unidades da UFPel: este estudo tem a finalidade de detectar o perfil dos departamentos e unidades da UFPel com relação à produção intelectual anual dos docentes. Para a realização deste estudo, seria utilizada a tabela de "Produção", agrupada pelo número de identificação dos professores contendo a soma da pontuação intelectual individual. Para isso seria necessária a tabela com a pontuação referente à GED, para que os valores da produção intelectual pudessem ser disponibilizados de acordo com sua pontuação oficial. Assim seria possível o cruzamento desta tabela resultante com a tabela de docentes, com a finalidade de selecionar os departamentos e unidades dos professores, resultando em análises da produção por departamento e unidade.

3. O perfil do docente da UFPel em relação à sua carga horária de atividades de pesquisa, ensino, extensão e administração: este estudo faz um somatório das cargas horárias dedicadas pelos docentes da UFPel para cada tipo de atividade. A finalidade principal é detectar o perfil dos docentes voltados à pesquisa, ensino, extensão ou administração da Instituição. Para este estudo de caso seriam utilizadas praticamente todas as tabelas disponíveis com informações do RAAD sobre as atividades dos docentes.

5.2 Definição do objetivo

Uma etapa fundamental para todo o processo de DCBD é a definição do objetivo, ou seja, descobrir a necessidade de se implementar um sistema para DCBD. O objetivo é definido a partir do entendimento do problema que será solucionado. Existem diversos tipos de problemas que podem ser solucionados com esse processo, tais como: otimização de campanhas de marketing, prevenção de fraudes em cartões de crédito, previsões futuras, detecção de perfil, entre outros.

Para definição do objetivo, é necessário o entendimento do domínio e dos dados. Para o processo de DCBD ser capaz de resolver o problema eficientemente, devem ser descobertos os fatores importantes que influenciam os resultados, ou seja, as variáveis relevantes para o processo.

Embora os dados institucionais ofereçam diversas possibilidades de estudo, este trabalho focaliza-se na análise das atividades do corpo docente, na tentativa de detecção do perfil dos docentes da UFPel. Especificamente, na tentativa de descoberta de padrões de comportamento nas atividades docentes da UFPel. Este estudo foi limitado a este objetivo em virtude das diversas possibilidades de aplicação que surgiram.

Segundo Fayyad [FAY 96a], deve ser verificado se o processo de DCBD tem chances de ser bem-sucedido. Para isso, o autor indica a utilização de alguns critérios, que, quando atendidos, determinam uma maior possibilidade de sucesso. Estes critérios são divididos em dois grupos: práticos e técnicos. Os práticos abrangem as condições externas envolvidas ao processo e os técnicos tratam de características relacionadas à base de dados disponível.

Os critérios práticos utilizados neste estudo de caso abrangem:

- **Impacto potencial na aplicação:** Este critério é medido de acordo com a aplicação, pois as informações extraídas devem ter um valor significativo para a Organização. O resultado deste estudo pode servir como suporte aos dirigentes da UFPel quanto à distribuição de bolsas, investimentos na capacitação e contratação de docentes.

- **Falta de alternativa:** Este critério define quando é recomendada a utilização do processo de DCBD. Por exemplo, quando existe uma grande quantidade de dados a serem analisados. Neste estudo o processo é recomendado, visto o volume de dados acumulados durante os anos no RAAD.

- **Suporte Organizacional:** Quando há o interesse da Organização nos resultados do processo. Este trabalho possui suporte Organizacional das pessoas diretamente ligadas ao dia-a-dia operacional dos Sistemas analisados. O interesse nos resultados será crescente, visto que nenhum estudo desta natureza havia sido realizado nos dados da UFPel.

- **Problemas legais:** Este critério abrange a privacidade dos dados, garantindo a legalidade na utilização dos dados que contém informações privadas. Para este estudo, foi autorizada a utilização de todos os dados dos sistemas, mantendo em sigilo as informações pessoais dos docentes. Para tanto, foi utilizada uma função de criptografia no código de identificação dos docentes quando for necessária a divulgação dos resultados individuais.

Os critérios técnicos utilizados neste estudo de caso abrangem:

- **Disponibilidade dos dados:** Devem existir dados suficientes para realizar as análises de padrões. Neste estudo de caso, a quantidade de dados disponíveis, conforme Tab.5.2, mostrou-se suficiente para uma primeira análise de padrões no sistema RAAD.

- **Atributos relevantes:** Mesmo que exista uma grande quantidade de dados estes devem possuir alguma relação com o objetivo do processo, senão os resultados não gerarão informações úteis. Neste trabalho, existe uma quantidade suficiente de atributos relacionados ao objetivo do processo.

- **Baixo nível de ruído:** O nível de ruído é medido quando existe a ocorrência de valores incompletos, dados redundantes, dinâmicos ou volumosos demais. Os dados analisados neste trabalho apresentaram um nível de ruído

bastante elevado. Porém, este problema é resolvido durante a fase de limpeza desenvolvida na etapa de preparação dos dados.

- **Conhecimento prévio:** Baseado no conhecimento do sistema, pode-se determinar quais atributos da base de dados são mais importantes para o processo e quais padrões já são conhecidos. Este critério foi contemplado durante a fase de entendimento, no início do processo, onde foi realizado o entendimento do domínio da aplicação e dos dados.

5.3 Preparação dos Dados

A preparação dos dados é uma etapa de grande importância para todo o processo de DCBD que engloba a seleção dos dados, limpeza e transformação.

Apenas um bom gerenciamento do armazenamento dos dados, garantido pelos SGBDs, não é suficiente. Para o sucesso do processo de DCBD é necessário que os dados tenham sido corretamente selecionados, corrigidos e transformados. Embora os SGBDs amenizem estes problemas a maioria dos bancos de dados apresentam problemas com os dados que devem ser corrigidos nesta etapa. Assim, elimina-se o processamento desnecessário do algoritmo de mineração de dados.

Nesta etapa devem ser estudadas e aplicadas as estratégias para tratamento de dados incorretos, além de alternativas para tratar os registros nulos.

5.3.1 Seleção dos dados

Durante a fase de entendimento dos dados, foram conhecidos os bancos de dados disponíveis para este estudo de caso e os dados armazenados. Para o sucesso da aplicação, nesta fase deve-se levar em consideração os critérios técnicos definidos na seção 5.2, para que o conjunto de dados resultante esteja apropriado ao restante do processo.

A **quantidade de dados** é um fator de extrema importância. Para uma solução satisfatória da descoberta de conhecimento é necessária uma amostragem de casos do problema determinado, ou seja, devem existir exemplos que validem o modelo para o efetivo funcionamento do processo.

Além do volume de dados, o conjunto de dados selecionado deve conter apenas os **dados relevantes** ao processo de DCBD, ou seja, aqueles que determinam o padrão de comportamento. Neste trabalho, os dados relevantes são aqueles referentes aos docentes e suas atividades anuais. Para isso foram utilizados os dados do sistema do Relatório Anual de Atividades Docentes (RAAD).

O RAAD armazena suas tabelas em bancos de dados Interbase (Borland). Para este estudo, foram disponibilizados os dados do RAAD migrados para o formato do banco de dados Microsoft Access e em planilhas Excel, além de outros documentos com dados relevantes ao processo.

Contudo, quando se utilizam bases de dados distintas é necessário que os dados sejam migrados para um mesmo tipo de banco de dados ou que os algoritmos de extração estejam preparados para trabalhar com bancos de dados heterogêneos.

Além disso, em virtude do preenchimento do RAAD ser anual, os dados encontravam-se em tabelas separadas por ano de competência. Foram disponibilizados dados de 1997 à 2002. Porém, os dados de 1997 foram descartados do processo, em virtude de não possuírem todas as informações necessárias para o desenvolvimento do processo.

Com a finalidade de padronizar o armazenamento e manipulação dos dados foram utilizados todos os dados armazenados, primeiramente, no banco de dados Microsoft Access e depois migrados definitivamente para uma base de dados criada no MySQL. Para isso foi desenvolvido um módulo da ferramenta de Análise e Migração dos Dados com o propósito de transferir as tabelas do banco de dados de origem, acessado por ODBC para o banco de dados destino, denominado “RAAD”, criado no MySQL.

Esta ferramenta já se encarrega de algumas tarefas de limpeza e enriquecimento dos dados no momento da migração. A migração das tabelas ocorre seguindo a configuração para o procedimento. Um exemplo das etapas de migração é mostrado nas figuras Fig. 5.4, Fig. 5.5, Fig. 5.6 e Fig. 5.7. O exemplo mostra a migração dos dados da tabela de “Ensino” do Microsoft Access para o MySQL. A Fig. 5.4, mostra a tela inicial do módulo de migração, onde são definidos o nome do banco de dados de origem e as tabelas de origem e destino.

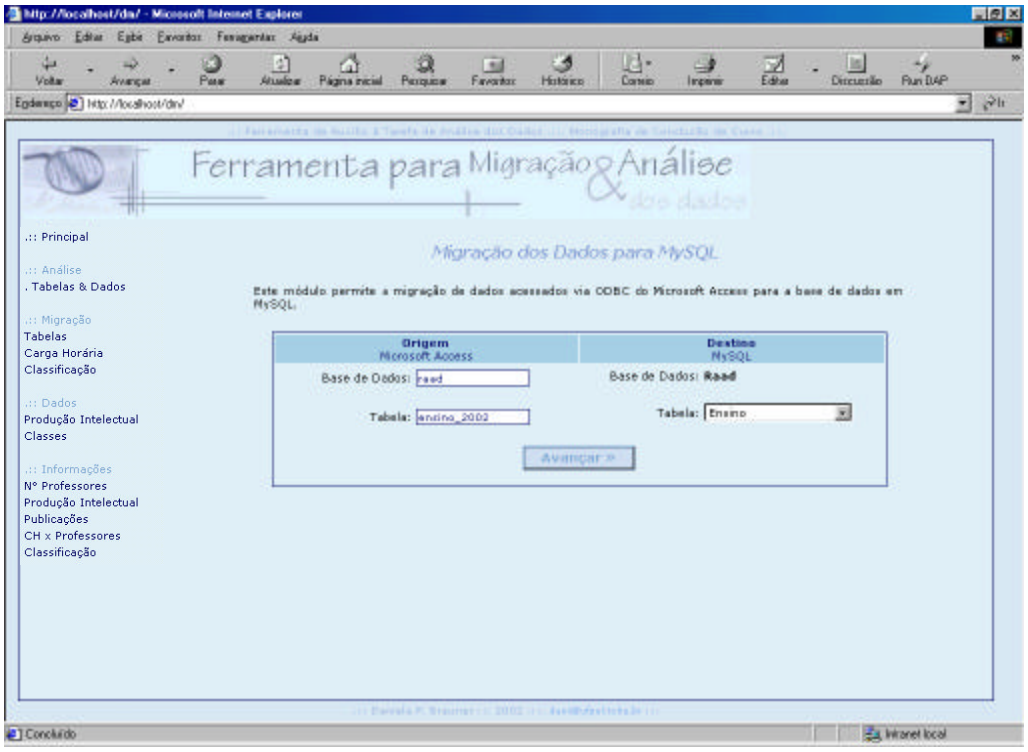

Figura 5.4: Tela inicial do módulo de migração

A Fig.5.5, mostra a tela de análise realizada sobre a estrutura da tabela de origem. A ferramenta analisa o nome da coluna e os tipos de dados associados.

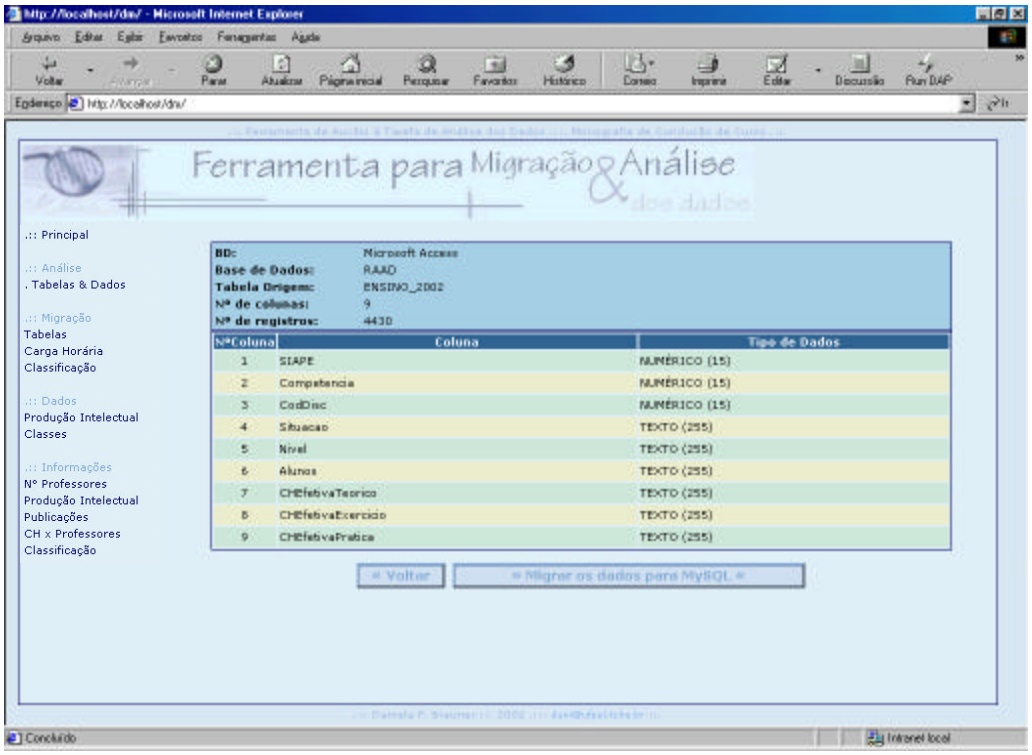

Figura 5.5: Verificação da estrutura da tabela de origem dos dados

A Fig.5.6 apresenta a tela onde é realizada a análise da tabela de destino, mostrando o nome das colunas e os tipos de dados. Nesta etapa são definidas as colunas da tabela de origem que estarão vinculadas às colunas da tabela de destino, que será melhor explicada na etapa de enriquecimento do processo de DCBD.

Por exemplo, na coluna onde é armazenada a carga horária do docente, já são somadas as variáveis relevantes da antiga tabela e armazenado o valor total desta soma. Portanto, para cada coluna de carga horária das novas tabelas armazenadas no MySQL existe um somatório das colunas das tabelas antigas associado. Isto não significa dizer que este somatório já é a carga horária total disponibilizada pelo docente para a atividade, pois existem mais dados espalhados em outras tabelas do sistema que guardam esta informação. Um exemplo disso é a própria tabela de ensino, que armazena diversas entradas para o mesmo professor, pois armazena o código da disciplina ministrada. A carga horária total de ensino do professor, é dada pelo somatório de todas as cargas horárias das disciplinas distintas ministradas naquele determinado ano com o somatório das cargas horárias disponíveis na tabela de cadastro do docente.

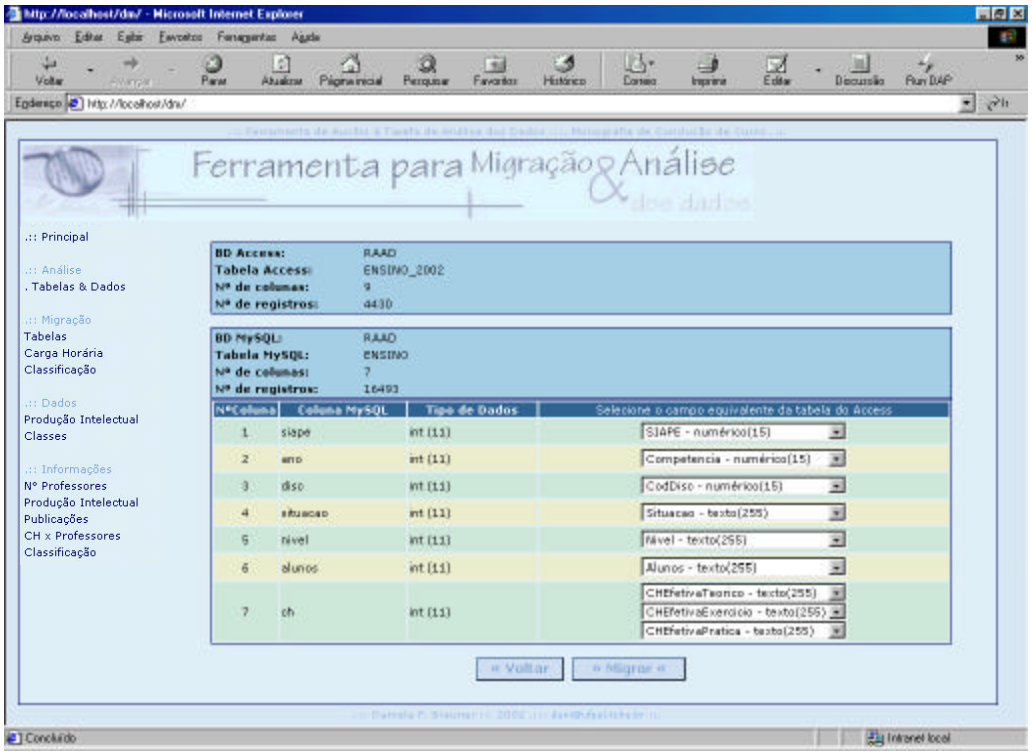

Figura 5.6: Definição das colunas da tabela de destino para migração

Após esta etapa, a ferramenta de migração possui uma rotina de limpeza e transformação para cada tipo de tabela. Estes procedimentos têm a finalidade de limpar e padronizar os dados (tarefas realizadas na etapa de preparação dos dados).

Finalizando o módulo de migração, a Fig.5.7, mostra a confirmação da migração realizada. Esta tela final ainda trás informações sobre o número de linhas transferidas (registros da tabela de origem) e o número de linhas total da tabela de destino.

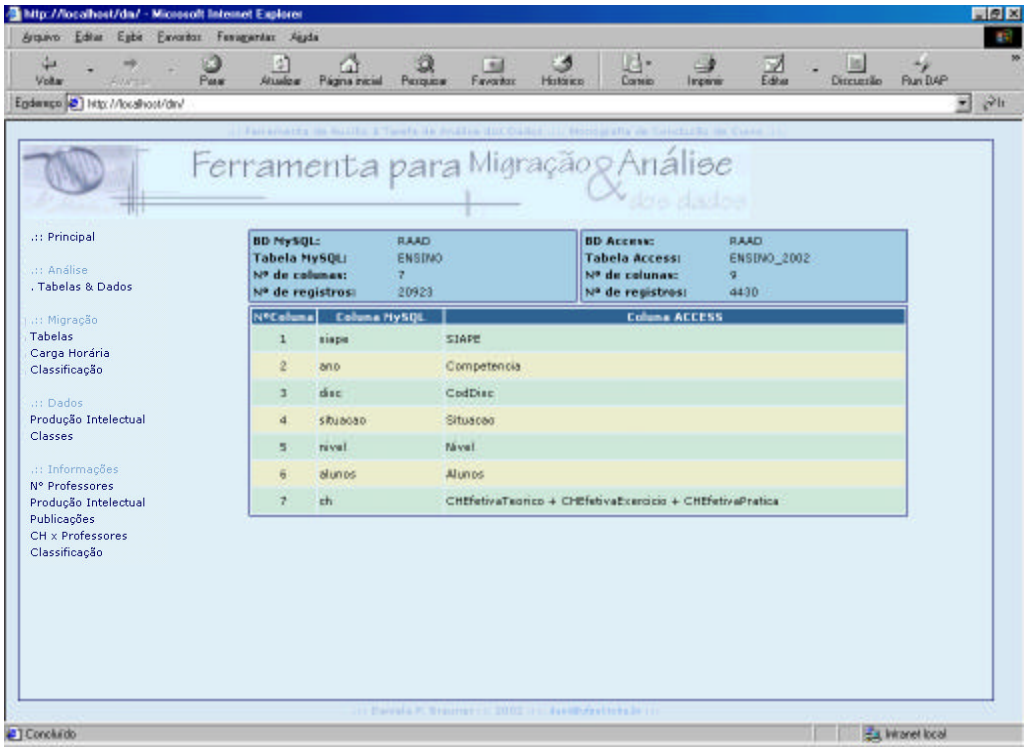

Figura 5.7: Tela para conferência da migração realizada

Portanto, utilizando-se essa ferramenta, foram selecionadas as tabelas anuais de 1998 à 2002 armazenadas no Microsoft Access, tais como: produção intelectual (“Producao”), atividades de ensino (“Ensino”), extensão (“Extensao”), pesquisa (“Pesquisa”), atividades administrativas (“Admin”) e cadastro dos docentes (“Cadastro”). Estas foram migradas para a base de dados “RAAD” do MySQL, gerando uma tabela para cada atividade, identificadas inclusive, pelo ano de competência do RAAD.

Durante a migração, algumas colunas foram somadas, outras eliminadas formando uma única tabela para cada tipo de atividade. Estas tabelas mantiveram a

coluna do ano como chave primária, mantendo os dados históricos, garantindo assim, a característica temporal do banco de dados. A Tab. 5.2 descreve o número de registros das tabelas anuais da base de dados do Access e o número total, referente à quantidade de dados resultante nas tabelas da base de dados criada no MySQL.

TABELA 5.2: Quantidade de dados anuais

Ano	Tabela "Producao"	Tabela "Ensino"	Tabela "Extensao"	Tabela "Pesquisa"	Tabela "Admin"	Tabela "Cadastro"
1998	2396	3953	1186	841	1191	927
1999	3233	3998	1286	839	1168	911
2000	3991	4276	1327	883	1095	936
2001	3837	4266	1316	994	1122	859
2002	3974	4430	1286	1178	1269	882
Total	**17431**	**20923**	**6401**	**4735**	**5845**	**4515**

A estrutura das tabelas criadas no MySQL estão definidas no dicionário de dados apresentado na Tab. 5.3.

TABELA 5.3: Tabelas resultantes na etapa de seleção dos dados

Nome da tabela	Campos	Tipos de dados	Descrição
Producao	Siape	Integer	Chave primária e estrangeira. Referência ao código do professor da tabela de cadastro de docentes
	Ano	Integer	Chave primária e estrangeira. Ano de competência do RAAD, referencia o ano da tabela Cadastro
	Codproducao	Integer	Chave primária. Código da produção intelectual. Número sequencial gerado pelo sistema para cadastramento da produção
	Codgrupo	Integer	Chave estrangeira. Código do grupo de produção intelectual. Referência à tabela Grupo
Cadastro	Siape	Integer	Chave primária. Código de identificação do professor
	Ano	Integer	Chave primária. Ano de competência do RAAD
	Dept	Integer	Chave estrangeira. Código do Departamento da UFPel de lotação do professor, referencia a tabela Departamento
	Sitfunc	Integer	Chave estrangeira. Código da situação funcional do professor, referencia a tabela Cadastro_SitFuncional
	Classe	Integer	Chave estrangeira. Código da classe do professor, referencia a tabela Cadastro_Classe
	Nível	Integer	Nível do professor

Nome da tabela	Campos	Tipos de dados	Descrição
Cadastro (cont.)	Titulação	Integer	Chave estrangeira. Código da titulação do professor, referencia a tabela Cadastro_Titulacao
	DataAdmissao	Varchar(10)	Data de admissão do professor
	DataDesligamento	Varchar(10)	Data de desligamento do professor
	Regime	Integer	Chave estrangeira. Código do regime de trabalho do professor, referencia a tabela Cadastro_Regime
	CH		Parte da carga horária disponibilizada pelo professor para atividades de ensino
Ensino	Siape	Integer	Chave primária e estrangeira. Referência ao código do professor da tabela de cadastro de docentes
	Ano	Integer	Chave primária e estrangeira. Ano de competência do RAAD, referencia o ano da tabela Cadastro
	Disc	Integer	Chave primária. Código da disciplina ministrada pelo professor
	Situacao	Integer	Chave primária e estrangeira Código do período em que é ministrada a disciplina, referencia a tabela Ensino_Situacao
	Nivel	Integer	Chave primária e estrangeira Código do nível do curso em que é ministrada a disciplina, referencia a tabela Ensino_Nivel
	Alunos	Integer	Número de alunos atingidos por essa atividade
	CH	Integer	Carga horária disponibilizada pelo professor para a atividade de ensino
Pesquisa	Siape	Integer	Chave primária e estrangeira. Referência ao código do professor da tabela de cadastro de docentes
	Ano	Integer	Chave primária e estrangeira. Ano de competência do RAAD, referencia o ano da tabela Cadastro
	CodCocepe	Integer	Chave primária. Código da pesquisa
	Projeto	Integer	Chave primária e estrangeira Código do tipo de projeto desenvolvido, referencia a tabela Pesquisa_Projeto
	Orgao	Varchar(255)	Órgão Financiador da pesquisa
	Valor	Integer	Valor financiado
	CH	Integer	Carga horária disponibilizada pelo professor para esta atividade de pesquisa
Extensao	Siape	Integer	Chave primária e estrangeira. Referência ao código do professor da tabela de cadastro de docentes
	Ano	Integer	Chave primária e estrangeira. Ano de competência do RAAD, referencia o ano da tabela Cadastro
	CodExt	Integer	Chave primária. Código da atividade de extensão
	Atividade	Integer	Chave primária e estrangeira Código do tipo de projeto desenvolvido, referencia a tabela Extensao_atividade
	Tipo	Integer	Chave primária e estrangeira Código do tipo de atividade de extensão desenvolvida, referencia a tabela Extensao_tipo
	Alunos	Integer	Número de alunos atingidos por essa atividade

Nome da tabela	Campos	Tipos de dados	Descrição
Extensao (cont.)	CH	Integer	Carga horária disponibilizada pelo professor para esta atividade de extensão
Admin	Codigo	Integer	Chave primária Código de auto-incremento
	Siape	Integer	Chave estrangeira. Referência ao código do professor da tabela de cadastro de docentes
	Ano	Integer	Chave estrangeira. Ano de competência do RAAD, referencia o ano da tabela Cadastro
	Tipo	Integer	Chave estrangeira Código do tipo de administração, referencia a tabela Admin_tipo
	Natureza	Varchar(50)	Descrição da atividade
	CH	Integer	Carga horária disponibilizada pelo professor para esta atividade de administração

5.3.2 Enriquecimento

Durante esta etapa foi observada a **disponibilidade dos dados.** Nem toda a informação relevante pode estar armazenada no banco de dados selecionado. Estas informações devem ser coletadas e convertidas para um banco de dados para que possam ser utilizadas no processo.

Estes dados podem ser encontrados em outras fontes, tais como: documentos, planilhas e até mesmo no conhecimento dos funcionários que trabalham no dia-a-dia operacional da Instituição. Alguns podem ser gerados a partir de outras colunas da tabela de origem. Por exemplo, pode-se gerar a carga horária de ensino somando algumas colunas da tabela de cadastro de docente.

Neste estudo de caso, foram utilizados alguns dados provenientes de documentos, planilhas e de consultas realizadas utilizando-se rotinas em PHP para armazenamento no BD. Por exemplo, a pontuação das atividades docentes de produção intelectual. Estes dados foram disponibilizados através de uma tabela de um documento no formato do Microsoft Word. Nesta tabela constava a pontuação referente à Gratificação de Estímulo à Docência (GED) referente a uma atividade docente. Estes dados foram convertidos para uma coluna na tabela "Grupo" do banco de dados, onde para cada grupo de atividades esta associada uma determinada pontuação na GED.

Outro método de enriquecimento dos dados utilizado neste trabalho, foi a soma de colunas da tabela de origem para alimentação de uma coluna da tabela de destino. Esta tarefa foi realizada durante a migração das tabelas anuais do Access para as tabelas do MySQL. A ferramenta desenvolvida para migração dessas tabelas detecta

o tipo de tabela que está sendo transferida e define a quantidade de colunas que a nova coluna da tabela do MySQL requer.

Um exemplo desse enriquecimento dos dados é a coluna que contém a carga horária na tabela "Ensino". Esta coluna foi gerada a partir da soma das colunas da tabela proveniente do Access: *CHEfetivaTeorica, CHEfetivaPratica* e *CHEfetivaExercicio*.

Além disso, foram armazenados alguns resultados de consultas realizadas pelo módulo de informações da ferramenta desenvolvida.

5.3.3 Limpeza

Qualquer banco de dados pode conter vários problemas de qualidade nos dados (poluição). Por isso, para o perfeito funcionamento do processo de DCBD, é necessário assegurar que os dados utilizados no processo estejam corretos. Existem diversos tipos de poluição, alguns dos encontrados neste estudo de caso são descritos a seguir.

- **Dados dinâmicos:** São aqueles dados que possuem seus conteúdos alterados com muita freqüência. Para o processo de descoberta do conhecimento esta característica pode interferir no processo, por isso é necessário criar uma visão ou replicação dos dados que fazem parte dos sistemas do dia-a-dia da organização, para que não ocorram atualizações indesejáveis durante o processo. Por isso, a utilização de um *data warehouse* é bastante recomendada para o desenvolvimento do processo de DCBD. No *data warehouse*, os dados não sofrem atualizações, eles são apenas carregados para o banco de dados e ficam disponíveis apenas para consultas.

  Neste trabalho, os dados foram migrados para uma tabela única, armazenando informações históricas que não sofrem atualizações. Esta característica é garantida em virtude da chave primária ser constituída do ano de competência do RAAD. O que permite que novas entradas do mesmo professor não sobrescrevam as antigas, pois a cada ano, o mesmo professor terá um comportamento diferente dos demais.

- **Dados ruidosos:** São os dados armazenados de forma errada na base de dados. Quando se trabalha com grandes volumes de dados procura-se

minimizar ao máximo a taxa de erros encontradas em pesquisas em uma base de dados.

Os dados do RAAD apresentavam um nível de ruído bastante elevado. A coluna de titulação, por exemplo, apresentava duas formas de representação da mesma informação dentro do mesmo campo, ou seja, a titulação do professor estava armazenada da seguinte forma: "14 - Doutor". Este tipo de ruído também foi detectado nas colunas de situação funcional e classe da tabela de "Cadastro" e na coluna de tipo de produção intelectual na tabela de "Produção".

Para eliminar este problema, a ferramenta de migração foi desenvolvida com um filtro para cada coluna. Ao migrar os dados a ferramenta armazena os dados codificados, de acordo com um código determinado na tabela "Cadastro_SitFuncional". Esta tarefa será melhor explicada na etapa de transformação na seção 5.3.4.

- **Dados inconsistentes:** Este tipo de poluição dos dados é detectado quando ocorrem dados incompletos (a falta de valores numa coluna) ou valores incorretos (ocasionados pela falha na verificação da entrada dos dados). Um fator que contribui para a ocorrência de dados incompletos é o uso de banco de dados relacional onde não foram utilizadas adequadamente as técnicas do modelo entidade-relacionamento ou as restrições de integridade não foram corretamente construídas.

  Neste estudo de caso, foram detectadas diversas ocorrências de dados incompletos. Para a limpeza desse tipo de poluição, durante a migração, os dados que apresentavam esta característica foram armazenados com o valor "NULL" ou "0" dependendo do tipo de dados da coluna.

  Para os dados incorretos, que também foram detectados, foi realizado um procedimento de limpeza. Por exemplo, nas colunas que armazenavam datas foi detectada a falta de padronização no formato.

Para a limpeza dos dados, no módulo de análise da ferramenta desenvolvida, foi adicionada uma função para a correção dos dados. Na tela de análise da ocorrência dos dados, mostrada na Fig.5.3, ao lado de cada ocorrência distinta, aparece a opção para correção. A Fig. 5.8 mostra a correção de um registro incorreto. Esta alteração afeta o número de ocorrências daquele registro.

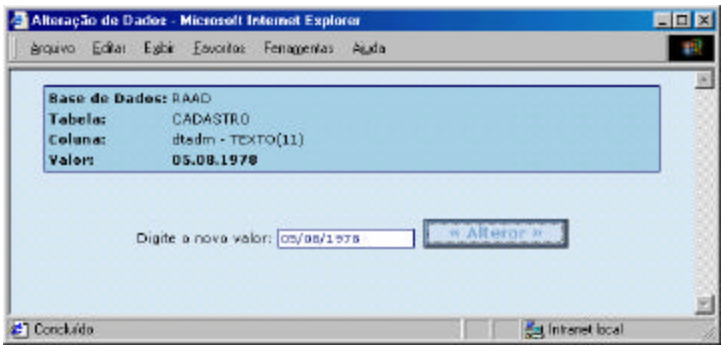

Figura 5.8: Alteração dos dados

A etapa de seleção resulta num conjunto dos dados relevantes para a aplicação das técnicas de mineração de dados. Porém, como estes dados são manipulados no dia-a-dia operacional da Instituição, eles podem apresentar problemas que poderão acarretar em atrasos, retornos de etapas e até mesmo a falha do processo de DCBD. Por isso, torna-se necessária a realização desta etapa de limpeza e correção dos dados durante todo o processo.

5.3.4 Transformação

Nesta etapa todos os dados possíveis são codificados, ou seja, todos aqueles dados que aparecem com freqüência nas colunas do banco de dados, são codificados e armazenados em uma tabela distinta.

Neste estudo de caso, a partir da utilização da ferramenta desenvolvida para auxiliar na análise da estrutura das tabelas e dos dados, foram analisadas as colunas das tabelas e detectada a padronização nos dados de algumas colunas.

Para as colunas onde ocorriam valores padronizados, foi criada uma tabela adicional. Esta granularidade do banco de dados possui a finalidade de gerar uma modelagem relacional ideal, aumentando assim a performance no acesso aos dados, acelerando as consultas.

A transformação dos valores dos dados em códigos foi realizada durante a etapa de migração dos dados, realizada com a ferramenta de migração. Quando detectada a ocorrência de um valor a ser codificado, a ferramenta armazena na coluna apenas o código do dado.

Um exemplo desta transformação dos dados aconteceu na tabela de cadastro de docentes. A coluna com o código de departamento possuía oito dígitos na modelagem do sistema: os dois primeiros dígitos referentes ao código da unidade e os dois seguintes ao código do departamento nesta unidade. Os quatro últimos dígitos eram zeros. No momento da migração, foram cadastrados apenas os primeiros quatro dígitos do código identificador de departamento, visto que os últimos dígitos eram compostos somente por zeros. Por conseguinte, foram criadas as tabelas com departamentos e unidades com seus respectivos códigos. Nas colunas: "sitfunc", "classe" e "titulacao" da tabela do MySQL, foram cadastrados apenas os códigos e criadas tabelas com as respectivas descrições. Para cada coluna foi gerada uma tabela individual.

A lista das tabelas criadas está na Tab. 5.4, com suas respectivas estruturas.

TABELA 5.4: Tabelas criadas para padronização dos dados

Tabela	Colunas	Tipos de dados	Descrição
Cadastro_Titulacao	Codigo	Integer	Chave primária. Código da titulação
	Descricao	Varchar (15)	Descrição da titulação
Cadastro_Sit Funcional	Codigo	Integer	Chave primária. Código da situação funcional
	Descricao	Varchar(15)	Descrição da situação funcional
Cadastro_Classe	Codigo	Integer	Chave primária. Código da classe
	Descricao	Varchar(15)	Descrição da classe
Cadastro_Regime	Codigo	Integer	Chave primária. Código do regime de trabalho
	Descricao	Varchar(15)	Descrição do regime de trabalho
Grupo	Codigo	Integer	Chave primária. Código do tipo de produção intelectual
	Descricao	Varchar(255)	Descrição do tipo de produção intelectual
	Pontos	Integer	Pontuação da produção intelectual para GED
Departamento	Codigo	Integer	Chave primária. Código do departamento
	Descricao	Varchar(100)	Nome do departamento
	Sigla	Varchar(10)	Sigla do departamento
	Unidade	Integer	Chave estrangeira. Referencia a tabela de Unidade
Unidade	Codigo	Integer	Chave primária. Código da Unidade
	Descricao	Varchar(50)	Nome da Unidade
	Sigla	Varchar(10)	Sigla da Unidade
Pesquisa_Projeto	Codigo	Integer	Chave primária. Código do tipo de projeto
	Descricao	Varchar(15)	Descrição do tipo de projeto

Tabela	Colunas	Tipos de dados	Descrição
Extensao_atividade	Codigo	Integer	Chave primária. Código do tipo de atividade
	Descricao	Varchar(30)	Descrição da atividade
Extensao_tipo	Codigo	Integer	Chave primária. Código do tipo de atividade de extensao
	Descricao	Varchar(35)	Descrição do tipo de atividade de extensao
Ensino_situacao	Codigo	Integer	Chave primária. Código do período ministrado da disciplina
	Descricao	Varchar(15)	Descrição do período ministrado da disciplina
Ensino_nivel	Codigo	Integer	Chave primária. Código do nível da disciplina ministrada
	Descricao	Varchar(15)	Descrição do nível da disciplina ministrada
Admin_tipo	Codigo	Integer	Chave primária. Código do tipo de administração
	Descricao	Varchar(15)	Descrição do tipo de administração

5.4 Mineração de dados

A mineração de dados é a etapa onde efetivamente são utilizados os métodos para a descoberta de conhecimento. Os métodos são definidos de acordo com o objetivo do estudo. Neste estudo de caso, o objetivo do processo focaliza a análise das atividades do corpo docente, na tentativa de detecção do perfil dos docentes. Portanto, as técnicas de mineração de dados escolhidas foram: agrupamento, padrões seqüenciais e classificação.

Este sistema híbrido possui a finalidade de primeiramente detectar um padrão nos dados, utilizando técnicas de agrupamento, para depois, classificar o restante dos dados utilizando técnicas de classificação e apresentar os resultados de acordo com os padrões seqüenciais descobertos.

Para isso, foram formuladas consultas SQL aos dados dos docentes da UFPel somando as cargas horárias disponibilizadas anualmente para administração, ensino, extensão e pesquisa. Estes somatórios foram coletados das tabelas do RAAD armazenadas no MySQL. Para cada professor, foram aplicadas técnicas de agrupamento através do comando “GROUP BY” via consultas SQL e outros procedimentos em PHP somando as cargas horárias referentes a cada ano de competência do RAAD.

Estas consultas estão disponibilizadas no módulo de informações desenvolvido junto à ferramenta de análise e migração dos dados.

Além dessa funcionalidade, este módulo da ferramenta ainda possui uma rotina para gerar a tabela para mineração. Esta tabela consiste numa consulta avançada

em SQL que, juntamente com procedimentos em PHP, geram dados relevantes para inserção no algoritmo minerador. Os resultados são armazenados tanto na tabela "Mining" no banco de dados "RAAD" como também em um arquivo texto para transportar estes dados para o programa de execução do algoritmo de classificação.

A tabela "Mining" possui a classificação dos professores de acordo com a carga horária disponibilizada para determinada atividade. As classes criadas estão definidas na Tab. 5.5.

TABELA 5.5: Classes de docentes detectadas

	Classe	Descrição
1	Administração	Docente que se dedica à administração
2	Ensino	Docente que se dedica ao ensino
3	Extensão	Docente que se dedica à extensão
4	Pesquisa	Docente que se dedica à pesquisa
5	Ensino/Administração	Docente que se dedica ao ensino e administração
6	Extensão/Administração	Docente que se dedica à extensão e administração
7	Pesquisa/Administração	Docente que se dedica à pesquisa e administração
8	Ensino/Extensão	Docente que se dedica ao ensino e extensão
9	Ensino/Pesquisa	Docente que se dedica ao ensino e pesquisa
10	Pesquisa/Extensão	Docente que se dedica à pesquisa e extensão
11	Ensino/Administração/Extensão	Docente que se dedica ao ensino, administração e extensão
12	Ensino/Administração/Pesquisa	Docente que se dedica ao ensino, administração e pesquisa
13	Pesquisa/Administração/Extensão	Docente que se dedica à pesquisa, administração e extensão
14	Ensino/Pesquisa/Extensão	Docente que se dedica ao ensino, pesquisa e extensão
15	Ensino/Pesquisa/Extensão/Administração	Docente que se dedica a todas as atividades
16	N.I.	Docente não classificado pela rede neural

Após utilizar a técnica de agrupamento para identificação dos padrões, alguns dados foram classificados para servirem como conjunto de treinamento do algoritmo de classificação. O conjunto de treinamento gerado possuía 136 registros com exemplos das classes com diferentes combinações entre administração, ensino, extensão e pesquisa.

O algoritmo de classificação utilizado foi uma rede neural RBF (*Radial Base Function*) [HAI 2001]. Este algoritmo faz a classificação dos dados de acordo com as classes de saída definidas através do treinamento da rede neural sobre os dados do conjunto de treinamento. O algoritmo foi treinado a partir do conjunto selecionado. A seguir, foi testado sobre o próprio conjunto de treinamento, para verificar a sua eficiência. A rede RBF apresentou 91% de acertos sobre estes dados.

O modelo gerado pela rede neural foi aplicado ao restante dos dados, retornando a classificação dos demais docentes da instituição nos respectivos anos do RAAD. Estes resultados são disponibilizados em formato de um arquivo texto.

5.5 Resultados e interpretação

Dependendo do algoritmo utilizado para mineração dos dados, é gerada uma saída diferente. Esta saída, geralmente, não é interpretável pelos usuários do sistema. Para possibilitar a interpretação dos resultados, foi desenvolvida uma ferramenta para migrar os dados classificados para o banco de dados.

A Fig. 5.9 mostra a tela inicial, onde é localizado o arquivo texto (resultante do algoritmo minerador) e definido o ano dos dados classificados. A seguir a ferramenta lê o arquivo, valida o padrão detectado e armazena o código da classe na coluna "Classe" da tabela "Mining".

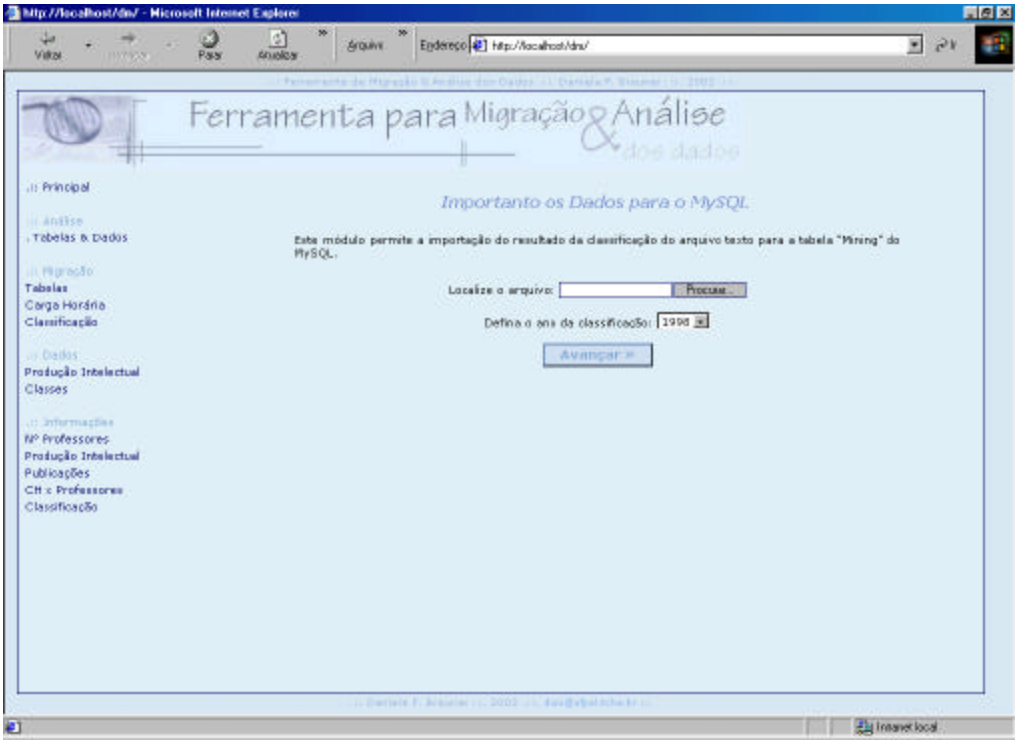

Figura 5.9: Migração do arquivo texto com os dados classificados

Assim, resulta a classificação anual dos professores, realizada de acordo com a carga horária dedicada para cada tipo de atividade.

Além disso, o módulo de informações disponibiliza estatísticas, agrupamentos e informações seqüenciais dos dados armazenados no MySQL. Abaixo segue a lista de informações disponibilizadas neste módulo do sistema.

- Número de professores substitutos e permanentes por departamento
- Número total de professores por departamento
- Proporção entre professores substitutos e permanentes por departamento
- Número de professores substitutos e permanentes por unidade
- Número total de professores por unidade
- Proporção entre professores substitutos e permanentes por unidade
- Pontuação referente à produção intelectual por professor
- Pontuação referente à produção intelectual por Departamento
- Pontuação referente à produção intelectual por Unidade
- Carga horária total de cada atividade docente

A partir das informações disponibilizadas neste módulo, foram gerados gráficos para que estas informações fossem visualizadas. Utilizando as informações da Tab. 5.6, que mostram o número de professores classificados no Departamento de Matemática, Estatística e Computação (DMEC), foi gerado o gráfico da Fig. 5.10.

No gráfico (Fig. 5.10), foram detectadas alterações no comportamento dos docentes deste departamento da UFPel. Por exemplo, nota-se uma forte tendência dos professores do DMEC dedicarem-se exclusivamente para o ensino. Além disso, foi detectado um aumento significativo na dedicação dos professores para o ensino e pesquisa.

A Fig.5.11 mostra as informações do ano de 2002 da Tab. 5.7. Esta tabela mostra a pontuação das Unidades da UFPel de acordo com a sua publicação anual de trabalhos completos em anais de congressos e artigos em periódicos nacionais e internacionais. E o gráfico da Fig.5.12 apresenta o aumento desta produção intelectual dos docentes do Instituto de Física e Matemática (IFM) no decorrer dos últimos cinco anos.

Através destas informações foi detectado que embora algumas unidades possuam cursos de pós-graduação, existem unidades onde existem apenas cursos de graduação que possuem uma pontuação consideravelmente alta em publicações.

TABELA 5.6: Classificação dos docentes do DMEC

Classe	1998	1999	2000	2001	2002
Administração	2	1	1	1	0
Ensino	16	14	19	18	15
Extensão	0	0	0	0	0
Pesquisa	0	0	0	0	0
Administração/Ensino	4	4	5	3	3
Administração/Extensão	0	1	0	0	0
Administração/Pesquisa	0	0	0	0	0
Ensino/Extensão	1	1	0	1	2
Ensino/Pesquisa	1	1	0	5	5
Extensão/Pesquisa	0	0	0	0	0
Administração/Ensino/Extensão	0	1	0	0	0
Administração/Ensino/Pesquisa	2	3	0	4	2
Administração/Extensão/Pesquisa	0	0	0	0	0
Ensino/Extensão/Pesquisa	1	1	1	0	1
Administração/Ensino/Extensão/Pesquisa	0	0	1	0	1
N.I.	1	1	5	1	3

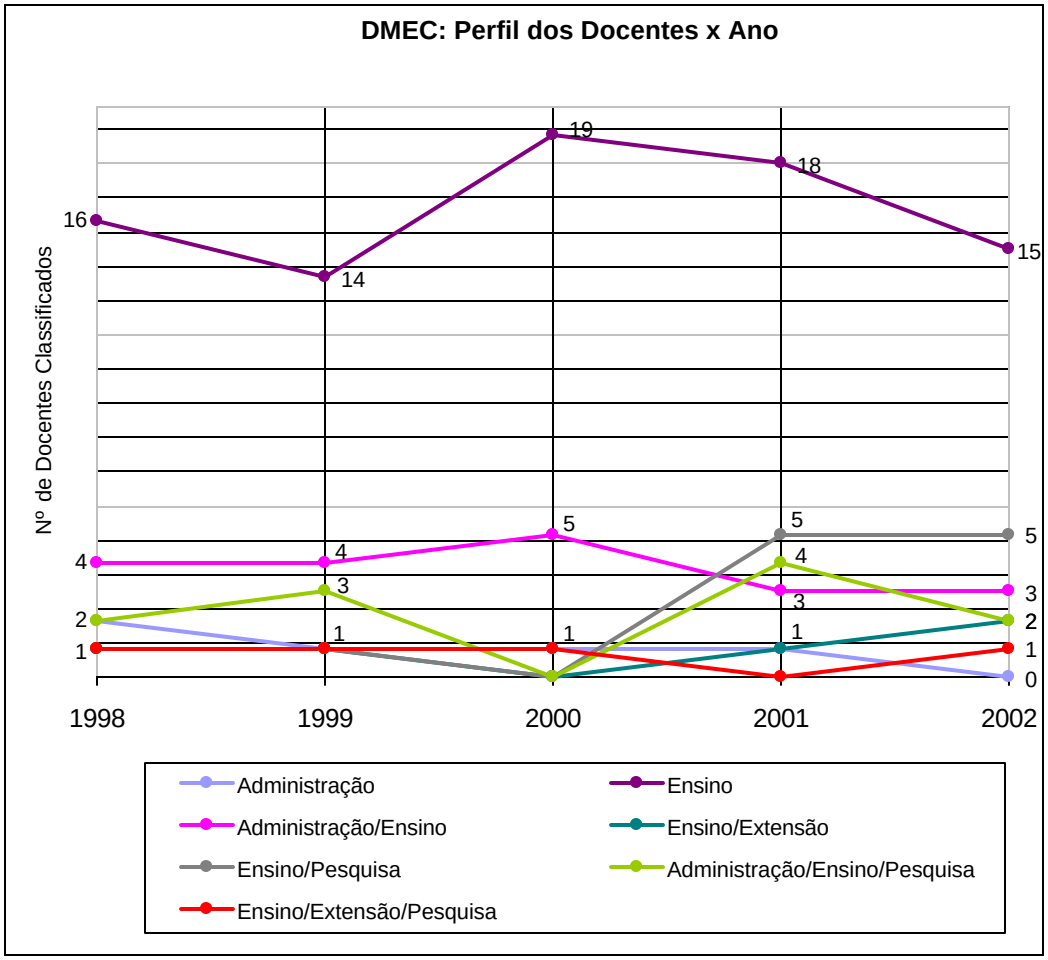

Figura 5.10: Classificação dos docentes do DMEC

TABELA 5.7: Pontuação em publicações das Unidades da UFPel

Unidade	1998	1999	2000	2001	2002
Faculdade de Agronomia Eliseu Maciel	1182	1404	1168	1334	1550
Faculdade de Ciências Domésticas	96	48	6	112	180
Direito	6	34	54	42	76
Veterinária	224	108	300	276	264
Odontologia	106	226	266	444	376
Faculdade de Educação	202	102	184	154	168
Medicina	390	164	410	338	452
Faculdade de Arquitetura e Urbanismo	18	14	66	4	16
Nutrição	44	30	76	70	40
Enfermagem	36	92	198	106	96
Engenharia Agrícola	32	24	30	28	26
Meteorologia	114	0	150	164	190
Escola Superior de Educação Física	12	82	62	100	76
Biologia	306	284	306	236	384
Instituto de Ciências Humanas	32	30	64	64	102
Instituto de Física e Matemática	198	134	152	158	296
Instituto de Letras e Artes	22	62	48	56	62
Instituto de Sociologia e Política	6	4	20	22	12
Instituto de Química e Geociências	70	138	100	64	154
Conservatório de Música	12	0	40	18	6

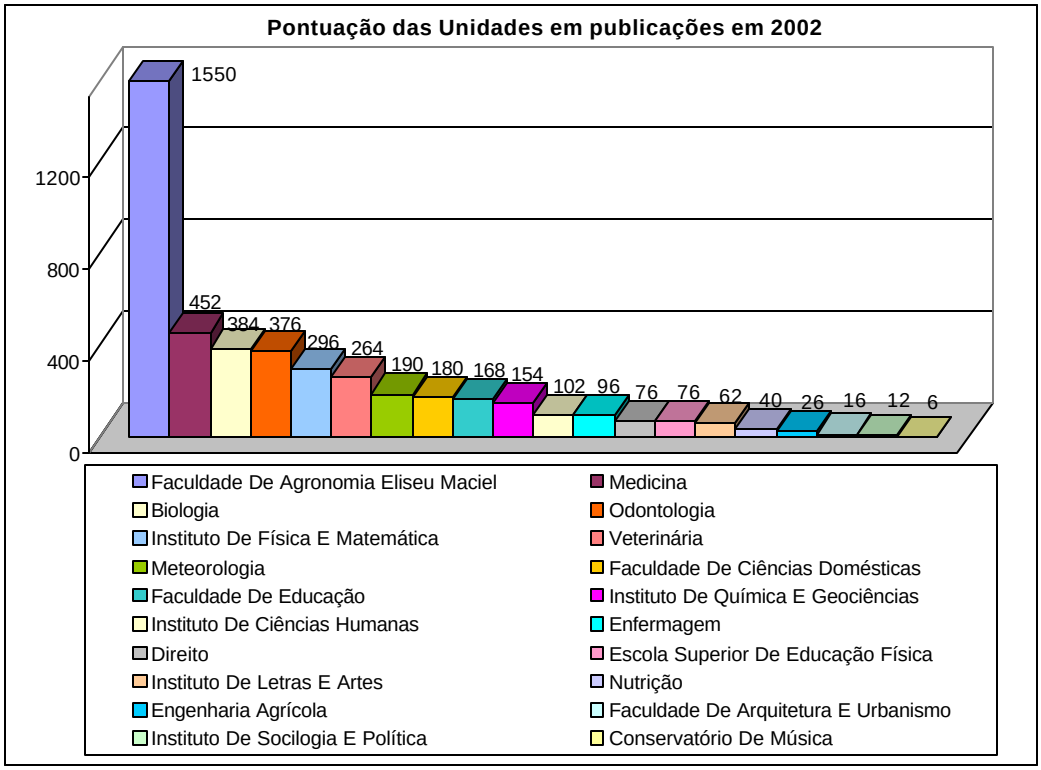

Figura 5.11: Pontuação em publicações das Unidades da UFPel em 2002

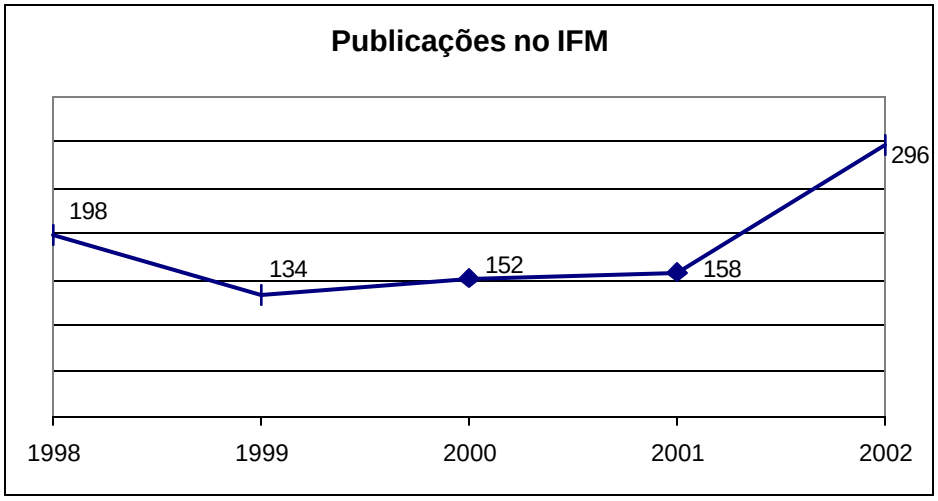

Figura 5.12: Publicações no IFM

O gráfico da Fig.5.13 mostra a porcentagem total dos docentes da UFPel de acordo com a sua situação funcional. Neste gráfico, em contraste com os números disponíveis na Tab. 5.8., é visível a crescente contratação de substitutos.

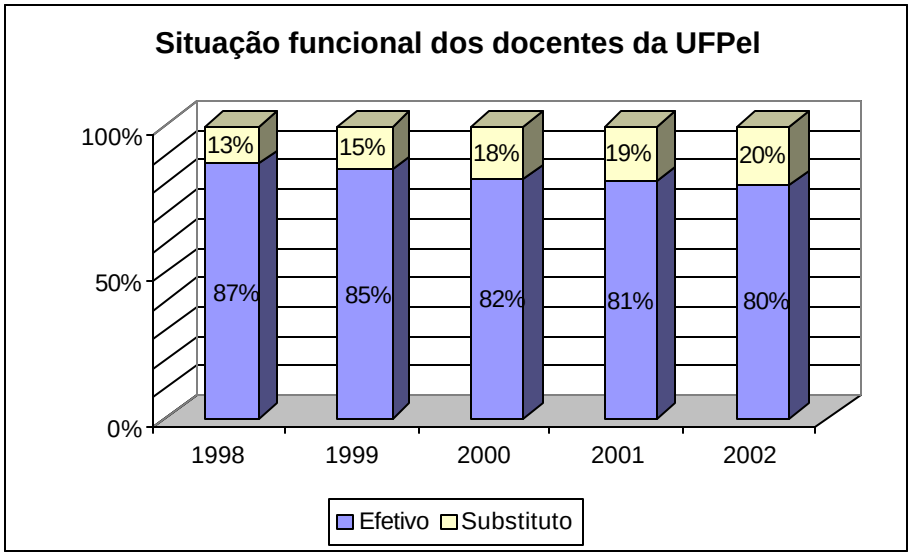

Figura 5.13: Porcentagem de docentes efetivos e substitutos da UFPel

TABELA 5.8: Informações aproximadas do número de docentes da UFPel

Sit.Funcional	1998	1999	2000	2001	2002
Permanente	800	673	766	690	701
Substituto	118	116	164	164	179
TOTAL	**918**	**789**	**930**	**854**	**880**

Os resultados acima comprovam que a ferramenta desenvolvida para análise e migração dos dados limpou e adaptou adequadamente os dados do RAAD, garantindo o funcionamento das técnicas utilizadas para mineração.

As técnicas escolhidas na etapa de mineração dos dados, do modelo de processo de DCBD definido neste estudo de caso, mostraram-se eficientes para o volume de dados disponível. Disponibilizando informações relevantes para a conclusão deste estudo de caso. Estes resultados motivam a continuidade da pesquisa neste estudo de caso, possibilitando a descoberta de novas informações relevantes.

6 CONCLUSÕES E RECOMENDAÇÕES

6.1 Conclusões

A primeira preocupação deste trabalho foi explicar as diferentes abordagens do processo de descoberta de conhecimento em banco de dados encontradas na bibliografia da área. A partir desta análise, foi constatado que nenhum dos processos foi adequado ao estudo de caso proposto, a partir daí foi sugerido um modelo do processo para aplicação neste estudo de caso.

Com a ferramenta desenvolvida, foi possível constatar diversos problemas com os dados coletados anualmente dos professores da UFPel. Estes problemas foram gerados em virtude do programa do RAAD possibilitar a inserção desses valores inconsistentes e também não armazenar corretamente estes dados, conforme os conceitos fundamentais de banco de dados relacionais. Na etapa de análise dos dados, realizada no início do processo já seria possível gerar resultados estatísticos relevantes sobre os dados, não fosse a poluição existente nestes dados.

Com isso, foi detectado que a etapa mais demorada do processo foi a de preparação dos dados. Concluímos que, incomparavelmente, esta é a etapa mais importante do processo de DCBD. Se os dados fossem passados aos algoritmos mineradores sem passarem por todo o tratamento e transformação efetuada, o conhecimento gerado não teria sido confiável.

Quando se trabalha com grandes quantidades de dados que apresentam problemas de padronização, tornam-se necessárias exaustivas análises para criar amostras consistentes destes dados, conforme foi realizado neste trabalho.

Este estudo de caso, além de proporcionar descobertas relevantes sobre os dados dos docentes da UFPel, apresentou um estudo das técnicas comumente utilizadas para extração de conhecimento em bases de dados. Contudo, um início de aplicação real de um sistema de descoberta de conhecimento foi apresentado.

6.2 Recomendações

Sugere-se para trabalhos futuros, um estudo para uma nova modelagem para o armazenamento dos dados da Instituição, como por exemplo, a criação de um *data warehouse*, com isso seria eliminada a maioria dos problemas detectados nos dados, acelerando o processo e possibilitando a utilização imediata de programas para mineração de dados.

Além disso, dando continuidade ao que já foi feito, é indicado o desenvolvimento do algoritmo de mineração de dados integrado à ferramenta de migração e análise. Assim, tornar-se-á possível a mineração dos dados sem a necessidade de geração dos arquivos textos para transporte.

Ainda neste contexto, é aconselhado o aperfeiçoamento da ferramenta para acesso aos arquivos migrados de bancos de dados ao invés de puramente acesso ao banco de dados. Assim, esta ferramenta poderia ser disponibilizada na Internet para análise de qualquer volume de dados exportados para um arquivo e lido através da rede.

Recomenda-se ainda, a aplicação experimental de outros algoritmos para mineração dos dados selecionados para avaliação de desempenho e compatibilidade com o objetivo da aplicação, dando continuidade a esta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

[ADR 97] ADRIAANS, Pieter e ZANTINGE, Dolf. **Data Mining.** London, UK: Addison-Wesley, 1997.

[AUR 99] AURÉLIO, Marco. et al. **Apostila de Descoberta de Conhecimento em Banco de Dados e Mineração de Dados.** Rio de Janeiro: ICA - Laboratório de Inteligência Computacional Aplicada da PUC-RJ, 1999.

[AVI 98] ÁVILA, Bráulio. Data Mining. In: VI ESCOLA DE INFORMÁTICA DA SBC REGIONAL SUL, 1998, Blumenau, SC. **Anais...**Curitiba: PUC-PR, 1998. 194p. p.87-106.

[BER 97] BERRY, Michael J. A.; LINOFF, Gordon. **Data mining techniques: for marketing, sales and customer support.** New York: Wiley Computer Publishing, 1997.

[BER 99] BERSON, Alex et al. **Building Data Mining Applications for CRM.** London,UK: McGraw-Hill, 1999.

[BER 97] BERSON, Alex; SMITH, Stephen J. **Data warehousing, data mining & OLAP.** London, UK: McGraw-Hill, 1997**.**

[CHA 2000] CHAPMAN, Pete et al. **CRIPS-DM 1.0 Step-by-step data mining guide.** 2000. Disponível por WWW em http://www.crisp-dm.org (12/11/2002).

[CLE 99] CLEMENTINE. **Data mining: an introduction.** 1999. Disponível por WWW em http://www.sinter.com.tw/read/DataMiningIntroduction.pdf (12/11/2002).

[CRA 2002] CRAVEN, Mark W.; SHAVLIK, Jude W. **Using neural networks for data mining.** Disponível por WWW em http://www.cs.cmu.edu/~craven/craven.fgcs97.ps (17/09/2002).

[DAT 91] DATE, C.J. **Introdução a sistemas de banco de dados.** Rio de Janeiro: Campus, 1991.

[FAY 96a] FAYYAD, Usama M. et al. **Advances in knoledge discovery and data mining**. Menlo Park, Califórnia EUA: AAAI Press, 1996.

[FAY 96b] FAYYAD, Usama M. et al. **KDD for science data analysis: issues and examples**. Second International Conference on Knowledge

Discovery and Data Mining, Portland, Oregon, Ago.1996, AAAI Press, 1996.

[FOR 97] FORSMAN, Sarah. **OLAP council white paper.** 1997. Disponível por WWW em http://www.olapcouncil.org (30/12/2002).

[GRO 97] GROTH, R. **Data mining: a hands-on approach for business professionals**. New Jersey, EUA: Prentice Hall, 1997.

[HAI 2001] HAIKIN, Simon. **Redes neurais: princípios e prática.** 2ª edição. Porto Alegre, RS: Bookman, 2001.

[HEU 2000] HEUSER, C. A. **Projeto de banco de dados.** Série de Livros Didáticos, número 4. Porto Alegre, RS: Sagra Luzzato, 2000.

[INM 97] INMON, W. H. **Como construir o data warehouse**. São Paulo, SP: Editora Campus, 1997.

[KIM 96] KIMBALL, Ralph. **The data warehouse toolkit. Practical techniques for building dimensional data warehouses**. Union City, CA: John Wiley & Sons, 1996.

[KOK 2000] KOK, J. et al. **Natural data mining techniques**.2000. Disponível por WWW em http://www.wi.leidenuniv.nl/home/kosters/datam.ps (18/11/2002).

[KOR 99] KORTH, H. et al. **Sistema de Banco de Dados**. São Paulo: Makron Books, 1999.

[LUA 2002] LUAN, Jing. **Data mining and knowledge management in higher education – potential applications**. Disponível por WWW em http://www.cabrillo.cc.ca.us/oir/oir_reports/DM_KM2002AIR.pdf (10/08/2002).

[QUO 2001] QUONIAM, Luc. et al. **Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil**. 2001. Disponível por WWW em http://www.ibict.br/cionline/300201/3020104.pdf (02/09/2002).

[THE 97] THEARLING, K. **Understanding data mining: it´s all in the interaction**, 1997. Disponível por WWW em http://www.thearling.com (10/09/2002).

[WEI 98] WEISS, Sholom; INDURKHYA, Nitin. **Predictive Data Mining: a practical guide.** San Francisco, CA, USA: Morgan Kaufmann Publishers, Inc. 1998.

[WUJ 2000] WU, J. **Business Intelligence: What is Data Mining?** DMReview, 2000. Disponível por WWW em http://www.dmreview.com (10/09/2002).