PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Engenharia Elétrica

AVALIAÇÃO DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM UM AMBIENTE DE DATA WAREHOUSE

Tadeu dos Reis Faria

Dissertação de Mestrado PPGEE-33/2002

Orientador: Prof. Dr. Pyramo Pires da Costa Júnior

Maio de 2002

Agradecimentos

Encerro essa fase de meus estudos alegre e consciente de ser uma pessoa privilegiada por ter tido tantas oportunidades de melhorar meu instrumental para a vida. Aprendi um pouco da ciência e muito sobre as pessoas, percebendo o quão é maravilhosa e enriquecedora a convivência, a amizade e a interação com aquelas que conheci nesse período.

O término deste trabalho deve-se em muito à minha mãe, a meu pai (*in memorian*), aos meus irmãos, pelo carinho, pela amizade e por terem me propiciado oportunidades que nunca tiveram.

À Maria Helena, minha companheira, à Débora e ao Gustavo, meus filhos, os quais sempre estiveram presentes com seu amor, carinho e compreensão apesar momentos ausentes.

Ao meu orientador, Profº. Pyramo, que, além dos conhecimentos transmitidos com presteza, me ajudou a entender as diferenças que definem a identidade de cada pessoa.

Aos meus amigos e professores, Palhares e Humberto, minha eterna gratidão, visto que sempre me motivaram, e, com paciência, fizeram despertar em mim a vontade, a coragem e o gosto pelo exercício de docência.

Ao Flavinho, grande amigo, pela sua colaboração na revisão de partes desta dissertação, pelas inúmeras conversas que tivemos antes de tudo começar e pelas discussões decorrentes do confronto de idéias.

À SUDECAP, nas pessoas de Murilo Valadares, Wellington Martins Barros e João Carlos Zamagna, que não mediram esforços para disponibilizar recursos necessários para a pesquisa.

Ao Paulo Bicalho e Nivaldo Fernandes, pelo apoio dado para o encaminhamento deste estudo.

Aos demais professores do Programa de Pós-Graduação em Engenharia Elétrica da PUC-MG, em especial o prof°. Pietro, pela troca de idéias e pela colaboração na organização do trabalho.

Agradeço aos meus colegas de mestrado, sobretudo Fabiano, Claudete e Cynthia, por terem sido muito especiais e me ajudado muito durante o curso.

À Profª. Marlene, pela paciência e cuidado com a correção do texto.

Agradecimentos, sem dúvida, devem ser estendidos a todas aquelas pessoas que não foram citadas e participaram comigo nessa etapa da minha vida, apoiando-me e incentivando-me. Pode ser que caminhos ou projetos nos afastem no decorrer do tempo, mas, com certeza, lembrarei de todos, com saudades.

Por fim, agradeço a Deus, por estar sempre me ajudando a viver a vida atento ao interesse pelo aprendizado, pelo novo, trazendo daí a possibilidade de mais alegria e crescimento.

Sumário

Lista de Abreviaturas e Siglas

ACCESS	Ambiente de Banco de Dados da Empresa MICROSOFT
ADABAS	Sistema Gerenciador de Banco de Dados da Empresa SOFTWARE AG
API	Aplication Program Interface
BD	Banco de Dados
CBR	Case-Based Reasoning
CPS	Complete Process Solution
CRM	Customer Relationship Management
CSCW	Computer Suported Cooperative Work
DBminer	Ferramenta de Apoio ao Processo KDD
DM	Data Mining
DMQL	Data Mining Query Language
DW	Data Warehouse
GUI	Guidance User Interface
EXCEL	Programa da Empresa MICROSOFT para Gerenciar Planilhas Eletrônicas
HAMB	Heuristic Autonomous Model Builder
KBS	Knowledge-Based System
KDD	Knowledge Discovery in Database
MDDB	Mutidimensional Database
NT	Sistema Operacional da Empresa MICROSOFT
OLAM	On-line Analytical Mining
OLAP	On-line Analytical Processing
OLE DB	Object Linking e Embedding Database
OMIS	Organizational Memory Information System

ORACLE	Sistema Gerenciador de Banco de Dados da Empresa ORACLE
PA	Planning Component
PBH	Prefeitura Municipal de Belo Horizonte
PAC	Problem Analysis Component
RPS	Reusable Process Units
SGBD	Sistema Gerenciador de Banco de Dados
SPSS	Ferramenta para Análises Estatísticas de Dados da Empresa SPSS
SQL	Structured Query Language
SQL server	Sistema Gerenciador de Banco de Dados da Empresa MICROSOFT
Sudecap	Superintendência de Desenvolvimento da Capital (Órgão da Prefeitura de Belo Horizonte)
UGM	User Guidance Module
UML	Unified Module Language

Lista de Tabelas

Lista de Figuras

Resumo

O grande número de sistemas informatizados tem coletado e armazenado um enorme volume de informações em banco de dados, criando boas oportunidades para a aplicação de técnicas na descoberta de padrões de comportamento nos dados armazenados. Este trabalho tem como objetivo principal a elaboração de uma análise do processo de descoberta de conhecimento em banco de dados , abordando as suas características, bem como os fatores e dificuldades envolvidos em sua aplicação, principalmente os métodos de trabalho e os ambientes que lhe dão apoio. Ele também relata uma experiência de aplicação desse processo em um ambiente de *data warehouse*. Essa experiência foi orientada para a verificação do vínculo existente entre o que se apresenta como base teórica de aplicabilidade da tecnologia e sua implementação em um caso real. São analisadas as razões dos principais problemas encontrados bem como as lições aprendidas. De forma complementar, foi relacionado um conjunto inicial de requisitos que devem ser suportados por um ambiente de apoio à condução do processo KDD.

1- Introdução

A percepção de que o conhecimento permeia o nosso viver vem desde a época de Platão e Aristóteles quando eram feitas reflexões sobre a dicotomia sujeito e objeto, conhecedor e conhecido. No entanto, a idéia do conhecimento – como um recurso estratégico para manter a competitividade – ainda é recente.

O grande interesse pelo conhecimento pode ser explicado, segundo Davenport [DAV98a], por se tratar de um diferencial competitivo sustentável. Uma empresa rica em conhecimentos e que os gere de forma eficiente terá passado para um outro nível de qualidade, criatividade e/ou eficiência. Essa mesma empresa, porém, nem sempre tem disponíveis todos os conhecimentos de que necessita. É importante ressaltar, também, que as fontes para extração de conhecimento são as mais diversificadas possíveis.

Nesse contexto, uma das ações que tem despertado interesse é a descoberta de conhecimento em banco de dados. Esse interesse é justificado, principalmente, pelo aumento de nossa capacidade de gerar e coletar dados, a qual se explica, sobretudo, pelo domínio e conhecimento da tecnologia de banco de dados, que já apresenta um alto grau de maturidade; pelo aumento do processo de informatização nas empresas e, principalmente, pela crescente demanda das organizações por mais informações para que o seu potencial competitivo esteja sempre adequado ao posicionamento de mercado.

Freqüentemente, as informações extraídas dos bancos de dados são usadas para suprir as demandas operacionais das organizações. Contudo, ao longo do tempo, elas guardam uma história, que espelha a forma como o negócio que representam vem evoluindo. O poder de utilizar o conhecimento implícito nos dados e a capacidade de tomar decisões baseadas nesse conhecimento é que, cada vez mais, estão se transformando em um diferencial gerador de vantagem competitiva.

As pesquisas relacionadas a esse assunto vêm sendo impulsionadas pelo aumento acelerado de informações produzidas pelos sistemas automatizados e armazenadas nos bancos de dados. Isso ocorre porque esse crescente volume de informações parece ultrapassar a capacidade humana de digerir e interpretar as relações existentes entre elas.

A área de descoberta de conhecimento em bancos de dados é recente. Seus conceitos ainda estão se consolidando e, apesar de diversas pesquisas já realizadas sobre o assunto, existem inúmeras questões em aberto [CHE00a]. Os maiores esforços foram concentrados na construção, implementação e otimização de técnicas de extração de padrões em grandes volumes de dados, conhecidas como algoritmos de mineração [HAN01a]. Vemos agora que essas técnicas estão se tornando *commodities*, já com um alto nível de confiabilidade nos resultados de sua aplicação.

A partir de experiências com aplicações de algoritmos de mineração, surgiram questionamentos da seguinte ordem: como selecionar e preparar os dados para os algoritmos; como selecionar e divulgar os padrões encontrados; que pessoas devem estar envolvidas nas atividades. Esses questionamentos – que influenciavam diretamente na qualidade dos resultados obtidos, indicando que a tarefa não se restringia à aplicação dos algoritmos – foram transformados em etapas e aninhados em um processo de descoberta de conhecimento, denominado por Fayyad et al [FAY96a] como *Knowledge Discovery in Database* – KDD, ou seja, descoberta de conhecimento em banco de dados.

Segundo Fayyad et al [FAY96a] e Brachman et al [BRA96a], KDD é um processo de descoberta de conhecimento novo, interessante e útil em grandes volumes de dados. Na verdade, é um processo intensivo envolvendo inúmeras iterações e interações entre usuários de um banco de dados, por meio de diversos itens de tecnologia, objetivando extrair padrões de relacionamento e comportamento entre esses dados.

As características de iteração são evidenciadas pela constante necessidade de acertar os dados, após cada execução de um algoritmo, eliminando as inconsistências apresentadas, para aumentar a representação do domínio do problema investigado. As interações estão associadas à efetiva participação dos usuários nas fases do processo, que fazem inferências e tomam decisões no sentido de maximizar os resultados das atividades.

Em relação às necessidades de apoio ao processo KDD, podemos identificar duas consideradas básicas. A primeira é a evidência de que devem existir sistemas informatizados e integrados que auxiliem nas fases do processo. Os sistemas atuais, que, segundo Dunkel et al [DUN97a], se enquadram nesse perfil não trabalham de forma integrada, uma vez que a área de KDD possui características multidisciplinares relacionadas com outros campos de pesquisas, como: banco de dados, estatística, inteligência artificial, gestão de conhecimento, *Computer Suported Cooperative Work* - CSCW. A segunda necessidade é que as metodologias de trabalho que suportam o processo ainda não estão disponíveis e não há sinalizações para o estabelecimento de um padrão único. Apesar de haver trabalhos como o de Engels et al [ENG99a] [BAR00a], faltam estudos que promovam uma maior integração entre os usuários e as ferramentas, de forma a atingir mais rapidamente o objetivo, que é a produção do conhecimento. Essa deficiência reforça observações de que KDD é um processo não-trivial [FAY96a] [BRA96a2]. Para Dunkel et al [DUN97a], a definição do processo e a interação entre pessoas e sistema são tão importantes que, dependendo de cada problema, KDD pode ser visto como '*Human-Assisted Computer Discovery*` ou como `*Computer-assisted Human Discovery*`.

Embora as técnicas de mineração tenham um papel central no processo KDD, as atividades de entendimento do domínio do problema e da preparação adequada dos dados, para representar mais fielmente o que se deseja, consomem aproximadamente 80% tempo usado no processo. Mesmo assim, atenções especiais são voltadas para os algoritmos de mineração, que já estão com alto nível de consolidação [ENG99a].

Essa realidade mostra que ainda existem desafios a serem percorridos, e o maior deles é o de propiciar uma integração eficiente entre usuários e ferramentas, para que os resultados produzidos no processo KDD sejam propulsores de um novo estágio de nível de conhecimento.

Existem diversas ferramentas disponíveis que atendem os requisitos de um processo KDD, mas elas não são integradas e nem são direcionadas para a organização e divulgação do conhecimento produzido [GOE99a] [HAN01a]. Há trabalhos nesse sentido [BAR00a] [MOU99a] [ENG99a] [CHE00a] [LIN99a], porém eles falham por não documentarem e divulgarem as experiências vividas, nesse processo, pelas pessoas envolvidas. Isso é importante na geração de um novo conhecimento, visto que, para que se gere conhecimento, é necessário que os conceitos criados por uma pessoa ou grupo de pessoas sejam compartilhados por outras pessoas [NON97a].

Na literatura científica, existem trabalhos que relatam resultados obtidos por meio do processo KDD ou de algoritmos de mineração de dados; trabalhos que descrevem as etapas do processo [FAY96a] [BRA96a] e, também, trabalhos orientados para o apoio às atividades desse processo. Todavia, poucos relatos o validam observando as necessidades de tecnologia para maximizar a produção dos resultados e o reaproveitamento das experiências vividas no processo.

Este trabalho descreve uma experiência de descoberta de conhecimento em banco de dados em um ambiente de *data warehouse.* Tem como objetivo principal a elaboração de uma análise do processo de KDD, abordando as suas características, bem como os fatores e dificuldades envolvidos em sua aplicação, principalmente os métodos de trabalho e os ambientes que lhe dão apoio. Os requisitos para análise foram levantados com estudo aprofundado do processo KDD, a partir de alguns trabalhos relacionados com a atividade de apoio nas tarefas que compõem o processo KDD e dos resultados de um estudo de caso orientado para a verificação de como se comporta o processo KDD ao utilizar dados de um DW. Esse estudo de caso foi orientado, também, para a verificação do vínculo existente entre o que se apresenta

como base teórica de aplicabilidade da tecnologia e sua implementação em um caso real: o controle de Obras e Gerenciamento de Projetos da Prefeitura de Municipal de Belo Horizonte, especificamente os de responsabilidade da SUDECAP. Nesse estudo, será feita uma investigação para verificar se os critérios estabelecidos para determinar os preços das obras controladas pela SUDECAP podem influenciar negativamente na condução dos projetos. Como escopo para esse estudo foi estabelecido que as análises seriam centradas no comportamento do fator K, como descrito no início do capítulo 2. Os dados a serem analisados compõem um DW com informações dos assuntos orçamento e empreendimentos.

A estrutura da dissertação é organizada como segue. Após a introdução, o Capítulo 2 descreve o processo de KDD, suas características e os principais problemas na sua aplicação. No Capítulo 3, são apresentados alguns trabalhos relacionados ao processo KDD, que têm tido um papel importante como alternativas para solução de deficiências apresentadas na tarefa de descoberta de conhecimento em banco de dados. Ele mostra alguns projetos que objetivam a produção de ferramentas de apoio na condução do processo e, também, uma visão conceitual de ontologias aplicada à representação e reutilização do conhecimento. No Capítulo 4, é apresentado um estudo de caso de aplicação do processo KDD envolvendo um ambiente de *data warehouse* para gestão de obras e controle orçamentário. Nele é feita uma descrição da aplicação, um relato dos resultados de cada etapa do processo e, também, uma análise da aplicação do processo no ambiente da Sudecap. Para a tarefa de mineração dos dados, é usada uma ferramenta para trabalhar com tabelas multidimensionais, que incorpora regras de associação, classificação e *clusterização*. Por último, no Capítulo 5, são apresentadas as conclusões e sugestões para trabalhos futuros.

2- Processo KDD

A atividade de descoberta de conhecimento é abordada na literatura sob diferentes pontos de vistas. Alguns autores a tratam como aplicação de um conjunto de algoritmos sobre o conjunto de dados com o objetivo de extrair padrões. Esse ponto de vista é denominado de *Data Mining* (Mineração de Dados). Já para outros autores a atividade de descoberta de conhecimento é vista como um conjunto de tarefas que envolvem uma necessidade, o entendimento do domínio da aplicação, a preparação e análise dos dados e a disseminação dos conhecimentos extraídos das análises. Essa visão é fundamentada na organização das atividades em etapas, com um seqüenciamento organizado, constituindo um processo, que será chamado Processo KDD. A atualidade do processo KDD se evidencia pelos inúmeros estudos e pesquisas a ele relacionado. Estudos relacionados a esse processo têm aumentado consideravelmente e, com isso, vários trabalhos têm sido produzidos.

O objetivo central deste Capítulo é apresentar o processo KDD, determinando as suas principais características e indicando as dificuldades na sua aplicação nas organizações. Nele são apresentadas duas abordagens de autores diferentes, para o estudo do processo KDD. Por se tratar de assunto referente à descoberta e geração de conhecimento, a seção 2.1 apresenta um estudo sobre como se dá a criação do conhecimento na sociedade [NON97a]. Esse estudo é importante para o entendimento de alguns conceitos referentes ao processo KDD. A seção 2.2 apresenta a abordagem de Fayyad et al [FAY96a] [FAY96b] [FAY97c], que é mais voltada para o entendimento das atividades que compõem o processo KDD, sem se preocupar muito com a iteração que pode ocorrer na execução das atividades e com a interação das pessoas na sua execução. Outra abordagem – a apresentada por Brachman et al [BRA96a] – é mostrada na seção 2.3. Ela ressalta a necessidade da participação das pessoas no processo, além de entrar em mais detalhes sobre sua estrutura. Completando as abordagens vistas nas seções anteriores, a seção 2.4 mostra outras características do

processo e algumas questões importantes que podem determinar o bom andamento de um trabalho que trata de descoberta de conhecimento em banco de dados.

Embora estejam publicadas há algum tempo, as abordagens apresentadas por Fayyad et al [FAY96a] e Brachman et al [BRA96a] são atuais, sendo sempre referenciadas quando o assunto abordado é KDD.

2.1- Processo de Criação de Conhecimento Segundo Nonaka

A preocupação atual com o conhecimento pressupõe a necessidade de sua gestão na organização. A criação de um ambiente que propicie a gestão ampla do conhecimento torna necessária uma profunda e permanente reflexão dos membros da organização sobre os conceitos e valores envolvidos no processo de geração do conhecimento [MUS00a]. A existência de uma forte sinergia entre os componentes pessoa e tecnologia de uma organização é condição básica para que a gestão do conhecimento tenha sucesso.

O produto final do processo KDD é um novo conhecimento. É importante, portanto, considerar a contribuição de autores com uma visão mais teórica que tenham fundamentado as formas de como o conhecimento é gerado e transmitido entre os membros da organização.

Para Nonaka [NON97a] tanto a informação quanto o conhecimento são ligados ao contexto, dependem da situação e são criados de forma dinâmica na interação social entre as pessoas. A informação é um fluxo de mensagens, enquanto o conhecimento é criado por esse próprio fluxo de informação, apoiado nas crenças e compromissos de quem o detém, reforçando, assim, a convicção de que o conhecimento está essencialmente ligado com a interação das pessoas.

Nonaka [NON97a] classifica o conhecimento em dois tipos: o primeiro – chamado de Tácito (implícito, aceito) – é mais subjetivo, inclui elementos cognitivos e é centrado

em experiências e práticas. O segundo – denominado de Explícito – é mais objetivo e gira em torno de um conhecimento mapeado e estruturado.

O pressuposto apresentado por Nonaka [NON97a] é que o conhecimento humano é criado e expandido através da interação social entre o conhecimento tácito e o conhecimento explícito. Essa interação, chamada de 'Conversão do Conhecimento', é estudada sob o ponto de vista de quatro métodos de conversão, conforme ilustra a Figura 2.1.

CONVERSÃO DO CONHECIMENTO

	Conhecimento tácito	**Conhecimento explícito**
Conhecimento tácito	Socialização (conhecimento compartilhado)	Externalização (conhecimento conceitual)
Conhecimento explícito	Internalização (conhecimento operacional)	Combinação (conhecimento sistêmico)

Figura 2.1- Modos de Conversão do Conhecimento

A Socialização implica a troca de conhecimento tácito entre as pessoas. É um processo de compartilhamento de experiência que propicia a criação e a ampliação do conhecimento. No entanto, o conhecimento não se torna explícito, não podendo ser utilizado na organização como um todo.

A Externalização é um processo de conversão do conhecimento tácito – o qual é difícil de ser formulado e comunicado – em conhecimento explicito, articulável e estruturado. É vista como o meio mais eficaz de criação de conhecimento [MUS00a].

Na Internalização, as pessoas usam o conhecimento explícito para ampliar o conhecimento tácito. Observa-se, também, o processo de compartilhamento do

conhecimento, pois, para internalizar, é necessário que o conhecimento esteja compartilhado.

A Combinação é um processo de sistematização de diferentes conhecimentos explícitos em um sistema de conhecimento, ou seja, cria-se, a partir dela, um novo conhecimento. Esse modo de conversão envolve diferentes conhecimentos explícitos para criar um novo conhecimento. Nesse processo, tal qual no processo KDD, é importante a reutilização do conhecimento já existente.

A criação do conhecimento na organização é uma contínua e dinâmica interação entre o conhecimento tácito e explícito. O tipo de conhecimento criado por cada modo de conversão é certamente diferente. A Socialização gera o 'Conhecimento Compartilhado'. A Externalização gera o que pode ser nomeado de 'Conhecimento Conceitual'. A Combinação faz nascer o 'Conhecimento Sistêmico'. A Internalização produz o 'Conhecimento Operacional'. Esses conhecimentos interagem entre si gerando uma espiral de conhecimento [NON97a], na qual o conhecimento tácito e o explícito crescem à medida que sobem os níveis ontológicos.

Observa-se uma profunda interação pessoal na geração do conhecimento. A ampliação ou a criação de um conhecimento têm forte dependência de um outro já existente. É inegável, nesse aspecto, a necessidade de se ter mecanismos que propiciem o trabalho de transmissão e disseminação desse conhecimento. Para isso a formalização e a estruturação do conhecimento são determinantes. Isso, conforme mostra as abordagens apresentadas nas próximas seções, é fundamental, também, no processo KDD.

2.2- O Processo KDD Segundo Fayyad et al

Fayyad et al [FAY96a] apresentam uma definição de KDD que é uma das mais referenciadas na literatura: "Processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e finalmente compreensíveis a partir dos dados".

Essa definição está relacionada com o pressuposto de que o processo é complexo, pois envolve pesquisa de estruturas, padrões, modelos ou parâmetros com um grau de autonomia nem sempre bem definidos. Isso o caracteriza como não-trivial. É necessário observar a relevância dos resultados extraídos em forma de padrões; esses padrões devem ser válidos para novos dados e ter algum grau de certeza. Devem ser novos e válidos para que as pessoas possam promover algum diferencial com a sua utilização e, por fim, devem ser compreensíveis para que as pessoas interessadas ou ligadas ao assunto possam entendê-los e aplicá-los.

O termo 'processo' indica que há vários passos iterativos e interativos envolvendo as atividades de conhecimento do domínio da aplicação, preparação de dados até a definição e refinamento dos padrões estabelecidos. Tais passos são vistos como interativos porque permitem ao usuário refinar os resultados alcançados, mudando dinamicamente o foco de dados e obtendo resultados com diferentes níveis de abstração e sob diferentes ângulos de visão. Eles são ditos iterativos porque, em qualquer ponto do processo, pode-se retornar a etapas anteriores, até que o objetivo seja atingido.

O processo KDD apresentado por Fayyad et al [FAY96a] está ilustrado na Figura 2.2 e consiste da seqüência iterativa de etapas mostradas abaixo:

- Compreensão do domínio da aplicação
  Nessa etapa é feita uma análise do ambiente da aplicação, entendendo os objetivos do usuário e delimitando o escopo da aplicação. É importante formalizar os conhecimentos já existentes sobre o assunto Fayyad et al [FAY96a] não é apresenta essa etapa na figura que ilustra a sua abordagem sobre o processo (Figura 2.2).
- Criação de um conjunto de dados alvo

Nessa etapa seleciona-se um conjunto de dados, ou foca-se um subconjunto de variáveis ou uma amostragem dos dados sobre os quais o processo será executado.

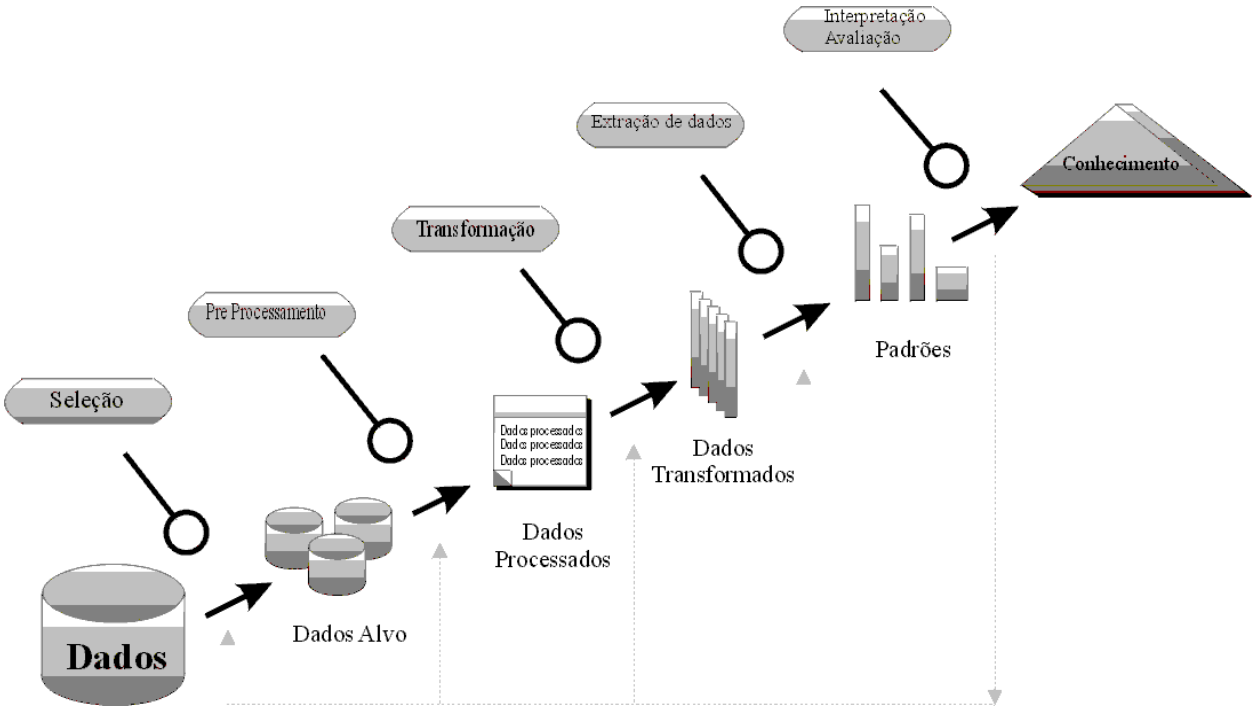

Figura 2.2- Processo KDD Segundo Fayyad et al

Fonte - [FAY96a]

- Limpeza e pré-processamento dos dados

  Nessa etapa são realizadas operações básicas sobre os dados, tais como: remoção de ruídos, valores desconhecidos ou inconsistentes; definição de estratégias para tratar campos sem valores; obtenção de informação necessária para modelar ou tratar os dados com ruídos. O objetivo central é consolidar informações relevantes para o algoritmo de mineração de dados.

- Transformação e redução dos dados

  Nessa etapa é feito o ajustamento de dados para atender melhor ao objetivo. Consiste da escolha de um subconjunto de atributos disponíveis para o algoritmo que seja relevante para o objetivo da tarefa.

Técnicas[1] para redução da dimensão e transformação podem garantir maior eficiência ao processo, mantendo as mesmas características dos dados originais.

- Mineração dos dados

  Nessa etapa ocorre o conjunto das seguintes atividades:

  - Definição da tarefa de mineração de dados: define se o objetivo do processo é classificação, associação, *clusterização* etc. Os diversos tipos de tarefas de mineração serão discutidos na seção 2.4;
  - Definição do algoritmo de mineração de dados: selecionam-se um ou mais algoritmos que serão usados para estabelecer novos padrões por meio dos dados. Cada tarefa de mineração pode envolver mais de um tipo de algoritmo;
  - Mineração: aplicam-se os algoritmos selecionados nos dados preparados para buscar padrões interessantes. Essa etapa pode envolver constantes ajustes em parâmetros para refinamentos dos métodos usados pelos algoritmos para a tarefa de mineração .

- Interpretação e avaliação

  Nessa etapa será feita a visualização, a interpretação e avaliação dos padrões extraídos. Inclui-se, nessa etapa, a remoção de padrões redundantes ou com baixo nível de acerto [JAN01]. O seu resultado pode indicar a necessidade de retornar a qualquer outra atividade para outra iteração. Os padrões dela extraídos só terão validade se o grau de interesse apresentado (e.g. novidade, validade, utilidade) for verificado [FRE99a]. Quando isto ocorrer, esses padrões passam a ser considerados 'conhecimento'.

- Consolidação do conhecimento descoberto

  Nessa etapa incorpora-se esse conhecimento à execução do sistema, documentando-o e divulgando-o às áreas interessadas. Inclui, também, a verificação de possíveis conflitos com antigas crenças. Aqui aplica-se a internalização – modo de conversão do conhecimento descrito na seção anterior.

---

[1] Técnicas para transformação e redução de dados podem ser encontradas em Han et al [HAN01a].

Fayyat et al [FAY96a] definem de maneira clara o processo KDD e nomeiam as etapas de forma seqüencial de modo a caracterizá-lo. Percebe-se, porém, um distanciamento entre a forma de execução das etapas e o modo como elas estão se relacionando. Nesse contexto, essa abordagem deixa de explorar questões como:

- A não-indicação de métodos de trabalho para conduzir o processo, principalmente no que se refere ao redirecionamento e às decisões sobre novas iterações;
- A importância da visão do processo como um fator crítico de sucesso para a organização;
- A importância do pleno conhecimento do domínio dos dados como um fator determinante para o encaminhamento do processo;
- A incorporação de resultados extraídos em alguma iteração em novas etapas;
- A importância e a forma de incorporação dos novos conhecimentos na empresa;
- A utilização de um conjunto integrado de ferramentas para apoiar o processo, sendo que essas ferramentas ainda apresentam pouca integração [DUN97a] [ENG01a];
- A importância das pessoas envolvidas no processo.

2.3- O Processo KDD segundo Brachman et al

O processo KDD proposto por Fayyad et al [FAY96a] é baseado em uma hierarquia de procedimentos que geram, quando bem sucedidos, um conhecimento novo. Brachman et al [BRA96a] ampliam a visão de Fayyad et al, destacando não só a necessidade e importância das pessoas que interagem no processo, mas também a idéia de que o próprio processo deve auxiliar quem o está trabalhando na tarefa de descoberta.

Esse processo é considerado um empreendimento complexo. Existe um fluxo básico para a execução de cada uma de suas etapas, mas, apesar dessa ordem, ele se

caracteriza como iterativo, podendo o usuário, a qualquer momento, mover-se de uma etapa para outra, sem seguir uma ordem definida, quantas vezes for necessário.

O processo KDD é estabelecido por três etapas principais: seleção e evolução do modelo, análise dos dados e apresentação dos resultados. Há também, incorporados ao processo, quatros pontos que Brachman et al [BRA96a] consideram complexos: descoberta da tarefa, descoberta dos dados, limpeza dos dados e conhecimento prévio do domínio. Essas etapas estão ilustradas na Figura 2.3 e descritas abaixo:

- Descoberta da tarefa

  Normalmente o usuário tem uma necessidade que pode incluir um problema a ser respondido ou um objetivo a ser atingido. Porém, ele nem sempre tem claro ou não sabe realmente o modo de determinar essa necessidade sob o ponto de vista de análise. O analista deve despender um tempo inicial, aprofundando o conhecimento sobre o problema, para determinar que tarefas serão executadas a fim de que os objetivos globais do processo fiquem bem definidos e claros.

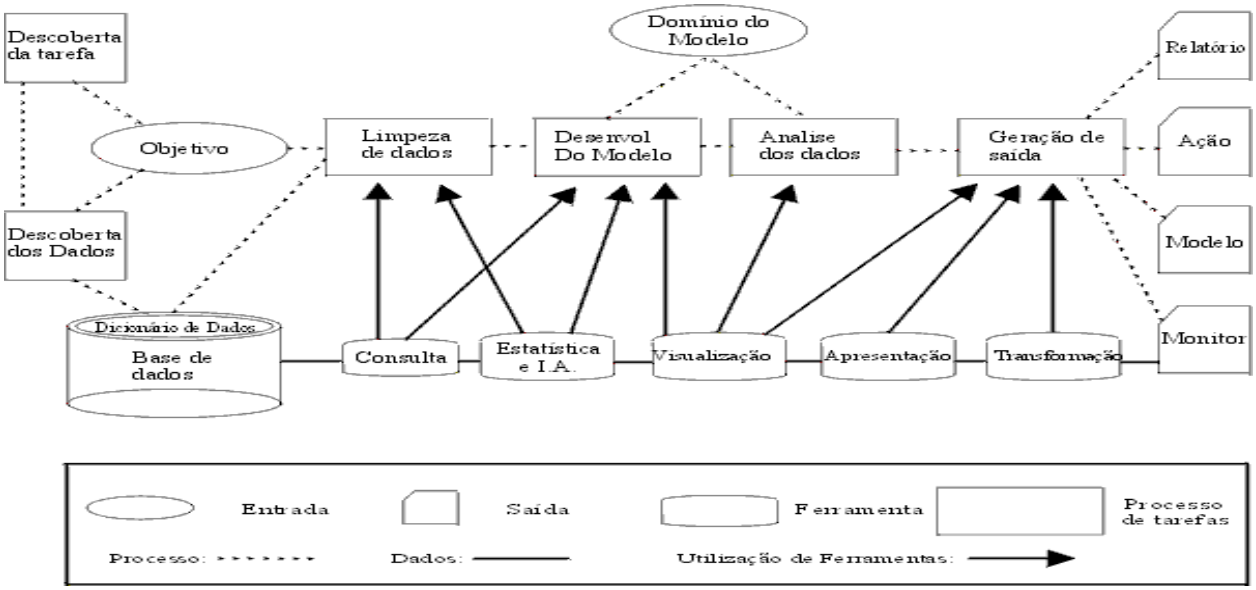

Figura 2.3- Processo KDD Segundo Brachman et al

Fonte – Brachman et al [BRA96a]

- Descoberta dos dados

  É imperativo conhecer a estrutura, o conteúdo e a qualidade dos dados e também verificar se os dados que serão disponibilizados comportam o desenvolvimento do processo de forma a atingir os resultados esperados pelo usuário. Essa etapa está intimamente relacionada com a etapa de descoberta da tarefa, podendo ser executada simultaneamente, uma completando a outra, pois a definição dos objetivos da aplicação e as tarefas que devem ser realizadas para atingi-los estão fortemente ligados aos dados e ao seu domínio [BOO99a] [CHE00a]. Essas duas etapas podem ser executadas em diversos momentos. Em um primeiro momento, nas suas próprias definições; outras tantas vezes nas iterações do processo quando elas podem ser novamente analisadas para aprofundar o entendimento do domínio.

- Limpeza dos dados

  Os dados devem apresentar uma boa qualidade para que o processo possa ter bons resultados. Eles podem apresentar ruídos, estar inconsistentes ou não ter valores. As fontes dos dados podem, também, estar com valores diferentes, mas com mesmos significados. O processo de limpeza deve ser cuidadoso para não excluir dados que tenham significado especial no domínio da aplicação, como, por exemplo, os que representam desvios de comportamento padrão.

- Desenvolvimento do modelo

  São raros os casos em que o analista simplesmente começa o desenvolvimento do modelo com uma hipótese formalizada. O trabalho nessa fase é encontrar um subconjunto da população do banco de dados que represente, de forma mais fiel possível, todo o conjunto de dados. Essa etapa complementa a etapa de análise de dados e com ela tem uma contínua iteração. O analista pode caminhar entre elas diversas vezes. Nessa etapa são realizadas três atividades:

  - Segmentação dos dados – usa, de forma geral, técnicas de aprendizado não supervisionadas [WES98a];
  - Definição do modelo a ser utilizado – a existência de uma série de técnicas para análises de modelos, como regressão, árvores de decisão,

redes neurais [JAN01a], que podem ser utilizadas em grandes volumes de dados para sua melhor definição;

- Seleção dos parâmetros – ajustes e seleção de parâmetros necessários para o modelo.

Nessa etapa o conhecimento prévio do domínio e a experiência do analista ajudam na definição do modelo.

- Análise dos dados

  Essa etapa é o núcleo do processo, da qual, efetivamente, os resultados serão obtidos. O analista tem uma hipótese sobre os dados e vários tipos de ferramentas para construir o modelo. Nessa etapa é realizada uma especificação formal desse modelo, sua avaliação e possíveis refinamentos baseados na verificação dos resultados apresentados. Para essa tarefa existem diversas ferramentas disponíveis para mineração e visualização de dados, das quais tem surgido grande número de pesquisas e trabalhos acadêmicos [HAN01a].

- Apresentação de resultados

  As saídas podem ser geradas sob formas de relatórios, descrições textuais, gráficos. Elas podem ser produzidas automaticamente por ferramentas ou formatadas pelo analista. O processo de comunicação na organização e a formalização do conhecimento são fatores determinantes na criação do conhecimento [NON97a], que, em última instância, é o objetivo maior do processo.

Brachman et al [BRA96a] ressaltam, também, a importância de incluir o conhecimento prévio do analista a respeito do domínio dos dados e da aplicação no processo de KDD. Muitos desses conhecimentos podem não estar formalizados ou documentados, mas existem diversas técnicas que ajudam nesse sentido, como, por exemplo, os diagramas da UML [BOO99a]. Contribui, também, para isso a documentação produzida quando é feita a definição dos dados no SGBD (dicionário de dados, integridade de dados).

A abordagem apresentada por Brachman et al [BRA96a] traz, com muito mais detalhes, uma descrição para o processo KDD. Se comparada com a formalização apresentada por Fayyad et al [FAY96a], essa abordagem amplia as etapas do entendimento do domínio da aplicação, a participação do usuário e resolve algumas deficiências apontadas, como:

- Inclusão das etapas de descoberta dos dados e descoberta da tarefa as quais ajudam a definir com mais segurança a necessidade do processo na empresa e o pleno conhecimento do domínio dos dados. Ao considerar o processo como interativo e iterativo, a ampliação ou redefinição do domínio dos dados podem acontecer em qualquer outra etapa, ocorrendo uma reestruturação do processo, bem como o ajuste de seus objetivos;
- Não há uma indicação formal de como fazer e formalizar a reutilização do conhecimento. No entanto, já apontam alguns caminhos. É importante a observação de Nonaka [NON97a] que afirma ser a criação de um conhecimento dependente de um outro já existente;
- A deficiência de ferramentas ou técnicas de apoio às atividades do usuário no processo KDD;
- A deficiência de apoio à tarefa do analista, no que se refere a técnicas ou ferramentas de descoberta do conhecimento. Fica patente a necessidade de focar mais o processo no analista e nas suas tarefas, pois o processo de descoberta de conhecimento se apresenta muito mais complexo que simplesmente a descoberta de padrões interessantes [CHE00a].

Neste trabalho, a etapa de limpeza dos dados poderá ser tratada como pré-processamento dos dados e inclui as atividades de limpeza, integração, transformação e redução dos dados. A mineração dos dados pode, em alguns casos, substituir as etapas de desenvolvimento do modelo e análise dos dados. Até então os termos ‘usuário’ e ‘analista’ representavam pessoas que se envolviam com o processo e que nele utilizavam as técnicas e ferramentas disponíveis. Doravante, usuário será a pessoa de uma área de negócio, com conhecimentos do mercado em que a empresa está inserida e que precisa dos resultados de um processo KDD. Sua participação no

mesmo é fundamental. Analista será a pessoa que tem conhecimentos das técnicas e ferramentas que suportam o processo KDD, tendo, necessariamente, fortes conhecimentos da área de informática. Atividades e tarefas são usadas como sinônimos no decorrer do texto.

2.4- Data Warehouse

Pela crescente demanda de informações gerenciais nas organizações, a tecnologia de *data warehouse* tem, cada vez mais, sido assunto nas pesquisas relacionadas ao processo KDD [HAN01a]. *Data warehouse* pode ser definido como um repositório de informações coletadas de diversas fontes, organizadas por um assunto principal, armazenadas de acordo com um esquema único e que, normalmente, reside em um único local [KIM98a] [TAM98a]. As dados são armazenados no DW de forma a prover informações sobre a perspectiva histórica, e, usualmente, são sumarizados. O termo *data warehousing* refere-se ao processo de construção e uso *do data warehouse.*

As informações necessária para construção do DW podem originar de fontes heterogêneas, podendo ter codificações e formatos diferentes. Por isso, na sua construção, é necessário um processo de integração, limpeza e consolidação do dados. Isso pode ser visto como um fase de pré-processamento dos dados para o processo KDD.

O DW é, quase sempre, modelado considerando a estrutura de banco de dados multidimensional. As Tabelas de Dimensões contêm a descrição das dimensões do negócio. Cada tabela é uma visão diferente do negócio que ela representa. As Tabelas de Fatos representam os relacionamentos entre as dimensões. Elas armazenam as medições numéricas do negócio, as quais são obtidas na interseção das dimensões [KIM98a]. Essas medidas são valores indicadores de performance do negócio. Exemplo:

venda do produto Y por R$50,00 em 17/10/01
na loja A, com custo de R$35,00.

Tabela de fato - venda

Dimensões - produto, tempo, loja

Medidas - valor venda , valor custo

Essa estrutura pode ser representada conforme a Figura 2.4.

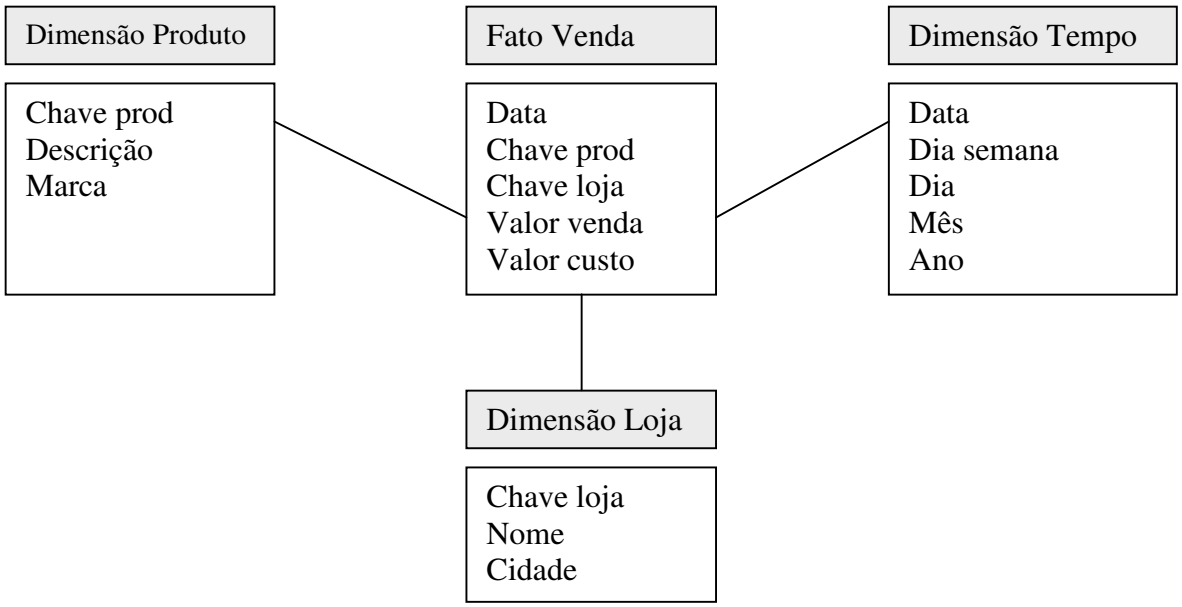

Figura 2.4- Modelo Dimensional [KIM98a]

Para analisar e visualizar os dados do DW de forma eficiente e eficaz, as ferramentas que usam a técnica de *On-line Analytical Processing* – OLAP, que tem tido importante papel no processo de consolidação do DW nas organizações. OLAP é uma tecnologia que usa uma visão multidimensional dos dados para prover acessos rápidos a informações estratégicas para a empresa.

Data warehouse já está com alto grau de consolidação nas empresas [kIM98a] [INM97a], sendo, portanto, um local rico em informações e deve ser considerada nas aplicações KDD. O DW, pelo trabalho de limpeza, integração e pré-processamento dos dados e por toda a infra-estrutura implantada para a sua construção – tratamento de informações de múltiplos bancos de dados, conexões ODBC/ OLE DB, ferramenta

OLAP – poderá ser uma das mais promissoras plataformas para fazer mineração de dados [HAN99a].

2.5- Tarefas do Processo KDD

As abordagens apresentadas nas seções anteriores mostraram a atividade de definição de que tarefa o processo KDD deve executar. Neste item são mostradas algumas dessas tarefas, observando que a utilização do processo KDD está relacionada com diversos domínios de aplicações: medicina, biologia, geoprocessamento, *marketing*, entre outros. Desse modo, podem-se identificar diversas tarefas de KDD, que são dependentes do domínio da aplicação e, principalmente, do interesse e necessidade do usuário. De forma geral, cada tarefa extrai um conhecimento ou regra diferente de um banco de dados. Cada tarefa requer um algoritmo diferente para a extração de conhecimento. Algumas das principais tarefas de KDD são:

- Regras de Associação

  Descoberta de associações interessantes ou correlações entre grandes conjuntos de dados. Foram introduzidas por Agrawal et al [AGR93a]. O exemplo clássico de regra de associação é o processo chamado *market basket analysis*. Esse processo analisa os hábitos de compras de clientes, encontrando associações entre os diferentes itens que eles colocam nas "cestas de compras". Uma tupla de dados, para esse tipo de análise, consiste num conjunto de atributos binários chamados de itens. Cada tupla corresponde a uma transação, e um item pode assumir um valor verdadeiro ou falso, dependendo se ele está ou não presente na transação. Essas regras são freqüentemente expressas em forma de $X \Rightarrow Y$, isto é, " $A_1 \wedge ... \wedge A_m \Rightarrow B_1 \wedge ... \wedge B_n$ " onde A_i (para $i \in \{1, ..., m\}$) e B_j (para $j \in \{1,...,n\}$) são pares de valores para o atributo [HAN01a]. Interpreta-se $X \Rightarrow Y$ como uma "tupla do banco de dados que satisfaz uma condição em X e que tende, também, a satisfazer uma condição em Y" . De outra forma, podemos considerar que em um banco de dados com D transações – onde cada transação $T \in D$ é um conjunto de

itens – $X \Rightarrow Y$ expressa que se a transação T contém X então T provavelmente conterá Y [AGRA93a] [FRE00a] [HAN01a] [HIP00a] [POS01a]. Cada regra é associada a um fator suporte superior, denominado Fsup, e a um fator de confiança, Fconf. Fsup é definido como a razão do número de tuplas satisfazendo ambos X e Y sobre o número total de tuplas, isto é, $Fsup = |X \cup Y| / N$, onde N é o número total de tuplas. Fconf é definido como a razão do número de tuplas que satisfazem ambos X e Y sobre o número de tuplas que satisfazem X, isto é, $Fconf = |X \cup Y| / |X|$. A tarefa de descobrir regras de associação consiste em extrair do banco de dados todas as regras com Fsup e Fconf maiores ou iguais a um Fsup e Fconf especificado pelo usuário. Recentemente a descoberta de regras de associação tem sido estendida a outros tipos de atributos que não sejam estritamente binários , incluindo, também, regras baseadas nos valores dos atributos (regras quantitativas) [POS01a] [ZHU98a].

Um dos algoritmos mais comentados na literatura para tratar as regras de associação é o *Apriori* [AGR93a] [AGR96a]. Ele é, também, uns dos mais implementados nos sistemas de apoio ao processo KDD [FRE00a].

- Regras de Classificação

Métodos usados para extrair modelos que descrevem importantes classes de dados e também para predizer tendências dos dados [FRE00a]. É a derivação de um modelo ou função que distingue objetos de uma classe de objetos de todas outras classes [HAN01a]. O modelo resultante é baseado em análises em um conjunto de dados preparados de acordo com o conhecimento do domínio da aplicação e que é usado para determinar classes para os dados não conhecidos. A classificação é feita em dois passos: no primeiro passo, um modelo é construído descrevendo uma pré-determinada classe de dados ou conceitos. Esse modelo é construído analisando tuplas do banco de dados. Tal processo é chamado de *supervised learning*. No segundo passo, o banco de dados é classificado de acordo com o modelo treinado no passo anterior.

Existem diversos métodos usados para proceder à classificação:

- Indução por árvore de decisão

  Trata-se de uma estrutura semelhante a uma árvore, representada como um gráfico, onde cada nó representa um atributo de teste e cada deslocamento representa um resultado desse mesmo teste. Nos nós folhas representam-se as classes. Uma árvore de decisão é geralmente construída de maneira *top-down.* Inicialmente, todas as tuplas que estão sendo analisadas são colocadas no nó raiz da árvore. O algoritmo, então, seleciona uma partição de atributos e divide o conjunto de tuplas no nó raiz de acordo com o valor do atributo selecionado. O objetivo desse processo é separar as classes para que tuplas de classes distintas tendam a ser associadas a diferentes partições, produzindo subconjuntos de dados cada vez menores, até que um critério de parada seja encontrado [HAN01a] [DZE96a]. As principais vantagens dos algoritmos baseados em árvores de decisão são sua inteligência computacional e sua simplicidade [AUR99a].

- Algoritmos genéticos

  Trata-se de modelos inspirados na evolução natural e na genética, aplicados a problemas complexos de otimização. Baseiam-se no conceito evolucionário da sobrevivência do mais apto – a seleção natural. A menor unidade de um algoritmo genético, chamada gene, representa uma unidade de informação do domínio do problema. Uma série de genes, ou um cromossomo, representa uma possível solução completa para o problema, ou seja, uma regra candidata. Para que um cromossomo seja avaliado, é necessário convertê-lo numa solução para o problema, decodificando-o. Uma vez que o cromossomo foi decodificado, o módulo de avaliação determina o quanto as soluções são boas ou ruins [AUR98a] [CAR00a]. Esses dois módulos – decodificador e avaliador – são as únicas partes do algoritmo genético responsáveis por entender o domínio do problema.

- Classificação Bayesian

  Trata-se de um método baseado em classificador estatístico. Trabalha com probabilidades definindo cada classe para cada registro [CEE96a].

- Classificação por redes neurais

  Trata-se de métodos cujos primeiros conceitos surgiram no início da década de 40. Em 1943, Warren McCulloch e Walter Pitts apresentaram a primeira discussão sofisticada sobre *neuro-logical network*. De forma resumida, uma rede neural é um conjunto de unidades conectadas por *inputs* e *outputs* onde cada conexão tem um peso associado. Durante a fase de aprendizado, a rede aprende pelo ajuste dos pesos bem como por sua capacidade de predizer sobre a classe correta para o atributo que está sendo tratado. Uma rede neural envolve um longo tempo de treinamento, sendo, dessa forma, mais adequada a aplicações onde isso é aceitável. Uma das grandes vantagens da rede neural é a sua tolerância com dados não muito bem trabalhados (*noisy data*). Vários algoritmos têm sido desenvolvidos para extração de regras de redes neurais treinadas. Um dos algoritmos mais populares para tratar redes neurais é conhecido como *backpropagation*, proposto na década de 80. Essas redes neurais têm sido amplamente estudadas na área de inteligência computacional [BRG00a] [HAN01a].

- Case-Based Reasoning - CBR

  Trata-se de classificadores baseados em instâncias, ao contrário dos outros classificadores que se baseiam em distância euclidiana. Os exemplos ou "casos" tratados pelo CBR são descrições simbólicas complexas. CBR têm sido usados em aplicações que trabalham com diagnósticos de problemas e interpretações de casos. Quando é dado um novo caso para ser classificado, o algoritmo, em um primeiro instante, verifica se já não existe um caso semelhante em uma base de casos. Se existe, a solução do caso é retornada. Se não existe, o algoritmo, então, pesquisa, em uma base de casos treinados, componentes que são similares àquele novo caso. A resposta deve ter informações suficientes para classificar esse novo caso. O algoritmo pode usar o conhecimento adquirido anteriormente e as estratégias para resolver problemas objetivando a proposta de uma solução adequada. Os desafios para os algoritmos CBR incluem a definição

eficiente de métricas, o desenvolvimento de técnicas para indexação dos casos treinados e métodos para combinar soluções [RIC98a] [BAR00a]. CBR têm um forte relação com os projetos relacionados a memória organizacional.

- Lógica *fuzzy*

  Trata-se de sistemas que têm facilidades para tratar de fronteiras de dados ou de classes [HAN01a]. Sistemas para classificação que se baseiam em regras têm a desvantagem da rígida exigência por atributos contínuos. Suponhamos que uma empresa apresente a seguinte regra para liberação de empréstimos: SE (**anos empregado** > 2) E (**rendimento** > 100,00) ENTÃO crédito = "aprovado". Por essa regra, clientes que têm **anos empregado** = 2 e **rendimento** = 99,00 não conseguiriam o empréstimo. A lógica fuzzy pode ser introduzida no sistema para permitir que limites *fuzzy* ou fronteiras possam ser definidos. Por conseguinte, com a lógica fuzzy, pode-se ter a idéia de que um rendimento de 99,00, em alguns casos, é tão bom quanto um de 100,00.

- Regras de *Clusterização*

  É o processo de agrupamento de um conjunto de objetos físicos ou abstratos em classe de objetos similares. Um *cluster* é uma coleção de dados ou objetos que têm características similares dentro de um mesmo *cluster* e têm pouca similaridade com dados ou objetos de outros *clusters* [BER99a] [HAN01a]. O algoritmo para essa tarefa deve criar as classes por meio da produção partições do banco de dados em conjuntos de tuplas. Essas partições são feitas de modo que tuplas com valores de atributos semelhantes sejam agrupadas dentro de uma única classe. Um bom agrupamento caracteriza-se pela produção de segmentos de alta qualidade, cuja similaridade intraclasse é alta e a interclasse é baixa. A qualidade do resultado da *clusterização* depende, também, da medida utilizada para medir a similaridade usada pelo método e de sua implementação, além de sua habilidade de descobrir algum ou todos os padrões escondidos [AUR99a]. O processo de *clusterização* tem sido usado em aplicações como

reconhecimento de padrões, análises de dados, processamento de imagens, pesquisa de mercado. Pode ser usado, igualmente, como um pré-processamento para outros métodos (classificação, caracterização).

- Sumarização

  São técnicas de criação de descrições compactas para um conjunto de dados. Por exemplo, definição das médias, desvio padrão, estabelecimento de correlações, cálculo da regressão. O conhecimento que pode ser extraído por essas técnicas tem um aspecto exploratório, sendo que um conjunto de dados é reduzido a uma descrição para posterior análise. A sumarização tem um aspecto descritivo, e sua representação compacta visa a uma melhor compreensão do conjunto de dados. Uma análise mais genérica deve ser feita para maior refinamento.

2.6- Questões Complementares sobre o Processo KDD

As seções anteriores mostraram as abordagens de Brachman et al [BRA96a] e Fayyad et al [FAY96a] que descrevem o processo KDD. Esta seção complementa os assuntos apresentados anteriormente. Tem como objetivo detalhar algumas características ainda não apresentadas e enumerar possíveis dificuldades que podem ocorrer durante o processo. Tais dificuldades servem como avisos a serem observados para que o projeto possa ser visto como um empreendimento possível de ser executado com sucesso na organização. Assim, na seção 2.6.1, são apresentadas algumas das possíveis fontes de extração de dados. Na seção 2.6.2, são levantadas questões relevantes com relação à tarefa de mineração de dados. As dificuldades relacionadas aos critérios que devam ser adotados para verificar a validade dos padrões descobertos são apresentadas na seção 2.6.3, e o processo KDD, visto como um projeto, é estudado na seção 2.6.4.

2.6.1- Tipos de Dados para Mineração de Dados

Em princípio, a mineração de dados pode ser feita em qualquer tipo de arquivo de dados, desde que os algoritmos estejam preparados para tal. O amadurecimento da tecnologia de banco de dados e sua enorme utilização são fatores que devem ser observados quando da definição de ferramentas para esse ambiente.
Encontram-se relacionados abaixo alguns tipos de repositórios de dados que podem ser usados na tarefa de mineração de dados :

- Banco de dados relacional

  Esse tipo de repositório é um dos mais populares e ricos em volume de informações. Por isso, vem se tornando o principal foco dos trabalhos que estão sendo produzidos atualmente.
- *Data warehouse*
- Banco de dados orientado a objetos.
- Banco de dados objeto-relacional.
- Banco de dados espacial.
- Banco de dados temporal.
- Banco de dados multimídia.
- Dados de sistema legados.

  São sistemas antigos que dão apoio aos processos operacionais das empresas.

Informações sobre esses tipos de bancos de dados podem ser encontradas em Korth et al [KOR99a] e Navate et al [NAV00a].

As pesquisas e os algoritmos para cada tipo de dado estão em diferentes níveis de profundidade. Essa percepção é importante no processo de definição de qual ferramenta deve ser definida para uso em um projeto. Ela deve ser compatível com os dados que serão tratados.

2.6.2- Questões que Influenciam a Atividade de Mineração de Dados

Assuntos como diversidade de tipos de dados; performance; interação com usuário; apresentação, representação e visualização do conhecimento mostrado abaixo são determinantes para o bom encaminhamento do processo.

- Mineração de diferentes tipos de conhecimentos no banco de dados

  Diferentes usuários podem ter interesses em diferentes tipos de conhecimentos. A atividade de mineração de dados deve ser feita com ferramentas que permitam diversos tipos de análises, para diferentes tarefas, incluindo algumas como associação, *clusterização*, classificação, caracterização de dados [HIP00a] [FRE00a] [HAN01a]. Essas tarefas podem usar os mesmos bancos de dados de forma diferente e requerem o desenvolvimento de diversas técnicas de mineração de dados.

- Mineração interativa de conhecimento em múltiplos níveis de abstração

  Como é difícil saber exatamente o conhecimento que pode ser descoberto em um banco de dados, o processo de mineração de dados deve ser interativo. Podem ser aplicadas técnicas como agregação e generalização para facilitar a exploração dos dados em enormes volumes de dados. O usuário pode interagir com as ferramentas de mineração de dados para vê-los em diversos níveis de granularidade e ver os padrões descobertos em diferentes ângulos de visão.

- Incorporação de conhecimento prévio

  O conhecimento prévio ou informações referentes ao domínio da aplicação em estudo pode ser usado para guiar o processo KDD e para ajudar na melhor definição dos padrões descobertos. Esse conhecimento prévio pode ser representado em diversos níveis de abstração. As definições relacionadas ao ambiente de banco de dados, como regras de integridade, podem ajudar a diminuir o tempo despendido no processo de descoberta, e ajudar a definir o interesse pelos padrões descobertos.

- Apresentação e visualização dos resultados

  Os conhecimentos descobertos devem ser expressos em linguagens de alto nível, como representação visual ou qualquer outra forma em que o

conhecimento seja de fácil entendimento para as pessoas. Isso é especialmente importante se o processo de mineração for interativo, que requer que os sistemas de mineração sejam dotados de diferentes técnicas de representação de conhecimentos como tabelas, árvores, regras, gráficos, matrizes.

- Tratamento de dados defeituosos ou incompletos

  Os dados armazenados no banco de dados podem estar incompletos, redundantes, inconsistentes ou defeituosos. Os resultados da mineração sobre esses dados podem provocar confusões ou gerar dúvidas. Como conseqüência, os métodos de avaliação de interesse pelo padrão podem apresentar resultados pobres. São necessários métodos para fazer limpeza dos dados e para tratar dados incompletos [HAN01a]. Os algoritmos de mineração podem ser usados para análises e tratamento dos dados que representam casos especiais.

- Dados dinâmicos

  Se as bases de dados usadas para mineração de dados forem atualizadas freqüentemente, os conhecimentos extraídos anteriormente podem divergir dos atuais. Para assegurar que essas alterações não levem a descobertas de regras conflitantes, podem ser tomadas precauções, como separação de dados históricos ou análises de séries temporais.

- Tratamento de diversos tipos de dados

  Uma vez que dados de banco de dados relacionais e *data warehouse* são muito usados, é importante o desenvolvimento de sistemas eficientes de mineração para esses dados. Outros bancos de dados, porém, podem conter tipos de dados complexos, como multimídia, dados espaciais, hipertexto. Sistemas específicos de mineração devem ser construídos para tratar o maior número de tipos de dados.

- Mineração em banco de dados heterogêneos e em rede

  As redes de computadores conectam diversas fontes de dados distribuídas e heterogêneas. A descoberta de conhecimento de diferentes fontes de dados estruturados, semi-estruturados e não-estruturados constitui desafio para a construção de ferramentas de mineração. A mineração de dados na Web é uma das áreas que mais tem despertado atenção [DOR00b] [KAD00a].

Problemas relacionados a esses assuntos podem aparecer durante as diferentes fases do processo, provocando re-execuções de atividades gerando novos resultados. Esse ciclo pode se repetir diversas vezes, onerando o processo.

2.6.3- Avaliação das Regras e Padrões Descobertos

As ferramentas de mineração de dados têm potencial para gerar um grande número de regras ou padrões. Um desafio do processo KDD é a avaliação do grau de interesse dos padrões extraídos, de forma a produzir somente aquilo que interessa ao usuário. Três questões básicas podem ser analisadas: todos os padrões descobertos são interessantes? As ferramentas de mineração podem gerar todos os padrões interessantes? As ferramentas de mineração podem produzir somente padrões interessantes?

Para a primeira questão Fayyat et al [FAY96a] definem como padrão interessante aquele que é facilmente entendido pelas pessoas, válido em dados novos, com algum grau de certeza, potencialmente útil, e ser novidade. Os padrões que atendem a esses requisitos são denominados de conhecimento. Existem diversas medidas objetivas para medir o interesse das regras. Essas medidas são baseadas na estrutura das regras descobertas e em medidas estatísticas que as envolvem. Em geral, cada medida de avaliação é baseada em *threshold,* que pode ser controlado pelos usuários. Para cada tipo de tarefa de mineração há uma medida específica. Embora as medidas objetivas possam identificar o quanto o padrão é interessante, elas são insuficientes, a menos que combinem com a avaliação subjetiva dos usuários. As medidas subjetivas são baseadas em crenças que cada usuário tem sobre os dados e sobre o comportamento a que esses dados se referem. As medidas objetivas buscam padrões que não são esperados, isto é, contradizem a crença do usuário e aqueles que oferecem informações estratégicas sobre as quais o usuário pode agir. Um padrão que era esperado é valido se ele vem confirmar uma hipótese elaborada pelo usuário.

A questão que se coloca em seguida – se ferramentas de mineração podem gerar todos os padrões interessantes – refere-se à completeza dos algoritmos. É difícil elaborar algoritmos que possam encontrar todos os padrões possíveis. De acordo com Freitas [FRE00a], a maioria das tarefas de descoberta é considerada bem definida, mas não determinística. De forma geral, usando somente os dados de treinamento, não se tem garantias de que as medidas possam ser verdadeiras, visto que os dados treinados formam um subconjunto de dados que representam toda a população do banco de dados. Ao contrário, as regras de associação são bem definidas e determinísticas [FRE00a] . Qualquer algoritmo para regras de associação deve produzir o mesmo resultado, que se baseia nas medidas de suporte e confiança definidas pelo usuário.

Finalmente o questionamento – se as ferramentas de mineração podem produzir somente padrões interessantes – refere-se a um problema que vem sendo aprimorado no processo KDD. É fortemente desejável que os sistemas gerem somente padrões interessantes, pois isso elimina o tempo necessário para que o usuário possa avaliar o quanto o padrão é válido. Progressos nesse sentido estão sendo alcançados [Han01a], porém cada otimização é um grande desafio.

As medidas para determinar o interesse dos padrões são essenciais para guiar o processo de descoberta e para evitar que regras irrelevantes, triviais ou falsas sejam consideradas corretas ou novas, com influência direta na qualidade dos resultados obtidos.

2.6.4- Gerência do Processo KDD

Os assuntos tratados anteriormente ilustram a complexidade do processo KDD e mostram o quanto a interatividade e a iteratividade estão presentes durante seu decorrer. O processo não é usualmente conduzido como uma lista de tarefas

previamente ordenadas. Pode haver uma seqüência de repetição na execução dessas tarefas, outras podem omitidas ou serem executadas em diferentes ordens. As tarefas interagem e se complementam em qualquer fase. É necessário que o analista e o usuário estejam sempre avaliando o resultado de cada fase, redirecionando o processo baseado nessas avaliações, em suas experiências e conhecimentos sobre o domínio da aplicação.

Outro ponto a ser destacado é o grande número de ferramentas que podem ser utilizadas. A definição de uma ferramenta no processo é influenciada por diversos fatores, incluindo a extensão do projeto, a quantidade de usuários envolvidos, o número de tarefas que podem ser executadas e o tipo e forma como os resultados serão apresentados. Nesse sentido, a decisão pode ser por uma ferramenta mais genérica que apóia aplicações de diversas áreas e um número variado de tarefas de mineração de dados ou por uma ferramenta desenhada para uma aplicação específica com um objetivo único.

Do ponto de vista de gerenciamento de projetos, as observações anteriores são dificultadoras para o controle e acompanhamento do processo. Elas dificultam principalmente a reutilização das experiências adquiridas quando da execução de um outro projeto. Isso mostra a necessidade de documentação da execução de cada fase de maneira cuidadosa. Portanto, a documentação deve ser entendida como um fator determinante para o sucesso de outros projetos.

Conforme já citado nas seções anteriores, o “gargalo” do processo não está na aplicação dos algoritmos para análises de dados, mas sim na utilização, de forma eficiente, de cada técnica e na combinação das ferramentas disponíveis de forma tal que o usuário possa ter os resultados desejados e que esses propiciem um diferencial competitivo para a organização. Por ser um processo centrado em pessoas [BRA96a], o projeto tem um perfil multidisciplinar e deve ser gerenciado como tal.

2.7- Conclusões

A adoção do processo KDD em uma organização não é uma tarefa fácil. Os sistemas informatizados ainda são construídos de forma não-estruturada dificultando a recuperação dos dados; as ferramentas de apoio ainda não são adequadas. Isso pode conduzir a aplicação para problemas de maior ou menor importância, o que pode onerar o custo do projeto.

O processo KDD deve ser entendido como parte de várias estratégias adotadas pela organização para manter seu negócio bem posicionado junto ao mercado. Portanto, para que o processo KDD tenha relevância no negócio, ele deve estar fortemente ligado a uma necessidade ou estratégia da empresa e deve produzir resultados, mesmo considerando que o retorno sobre os investimentos feitos para o projeto ainda é difícil de ser mensurado.

As pesquisas nessa área indicam a dificuldade de criar um ambiente integrado para o processo KDD e para propiciar maior apoio ao usuário. Isso é aceitável se se considerar a complexidade e multidisciplinaridade da área. Apesar disso, diversos trabalhos têm sido desenvolvidos para ajudar na condução do processo. Na próxima seção serão mostrados alguns trabalhos nesse sentido.

3- Trabalhos Relacionados ao Apoio ao Processo KDD

O desenvolvimento de sistemas para auxiliar no processo KDD é uma atividade cujo nível de complexidade é definido pelas características que o sistema deverá ter, como: quais são os tipos de tarefas; qual o nível de integração das suas funções na execução das suas diversas etapas; qual o nível de integração com outras ferramentas; qual será o nível de participação do usuário no controle do processo de descoberta [ENG99a].

O Capítulo anterior objetivou, principalmente, mostrar KDD na visão processo. Nele, foi observada a importância de se ter um ambiente de apoio para facilitar o andamento desse processo, sem, no entanto, citar exemplos ou descrever esses ambientes. O objetivo deste Capítulo é mostrar um posicionamento a respeito de trabalhos que estão sendo desenvolvidos para facilitar a condução do processo KDD. A seção 3.1 descreve um sistema genérico que pode ser usado em diversas aplicações. Ele pode trabalhar com um grande número de tarefas de mineração de dados. A condução do processo KDD, usando esse sistema, é totalmente feita pelo usuário. A seção 3.2 descreve o resultado de um projeto que orienta o usuário na definição de quais tarefas devem ser executadas no processo. A proposta desse projeto é otimizar o processo KDD ao demandar uma pequena participação do usuário. Na seção 3.3 é descrito o projeto de um sistema que pretende ser totalmente autônomo para a execução das atividades do processo KDD. A participação do usuário no processo KDD, ao usar esse sistema, é mínima. Por último, na seção 3.4, é descrita a tecnologia de memória organizacional e a maneira como os conceitos de ontologias podem contribuir para o seu desenvolvimento.

3.1- Projeto DBminer

O DBminer é um sistema para auxiliar o processo KDD, resultado de anos de trabalho em pesquisa na área de descoberta de conhecimento. Esse sistema foi

desenvolvido na Simon Fraser University, Canadá. As suas principais características são:

- Possuir forte integração com ambientes para On-line Analytical Processing – OLAP, conjunto de operações feitas por meio de uma ferramenta com o objetivo de manipular os dados de um DW de acordo com nível de abstração desejado por cada usuário. As operações básicas de uma ferramenta OLAP são:
  - *roll-up* - incrementa o nível de agregação;
  - *drill-down* - decrementa o nível de agregação;
  - *slice-and-dice* - seleção e projeção entre dimensões do DW.

  As ferramentas para OLAP demandam um rápido processamento para grandes volumes de dados contido no DW.
- Integrar, de forma interativa, diversas regras para mineração de dados: associação, *clusterização*, classificação, sumariação. Essa integração conduz a uma promissora metodologia para mineração de dados chamada On-line Analytical Mining – OLAM. Os sistemas que trabalham com essa metodologia provêm uma visão multidimensional dos dados e permitem que usuários selecionem, dinamicamente, funções para *data mining* e, também, diversas operações de um ambiente OLAP;
- Possuir uma forte integração com SGBD's relacionais e com sistemas de DW;
- Possuir uma interface interativa e amigável para quem o usa.

A arquitetura do DBminer, ilustrada na Figura 3.1, possibilita a integração e a transformação de dados de um banco de dados relacional ou de um DW em um banco de dados multidimensional. Usando esse banco multidimensional, os usuários fazem solicitações empregando os recursos de uma ferramenta para OLAP ou uma das regras de mineração de dados disponíveis [HAN01a].

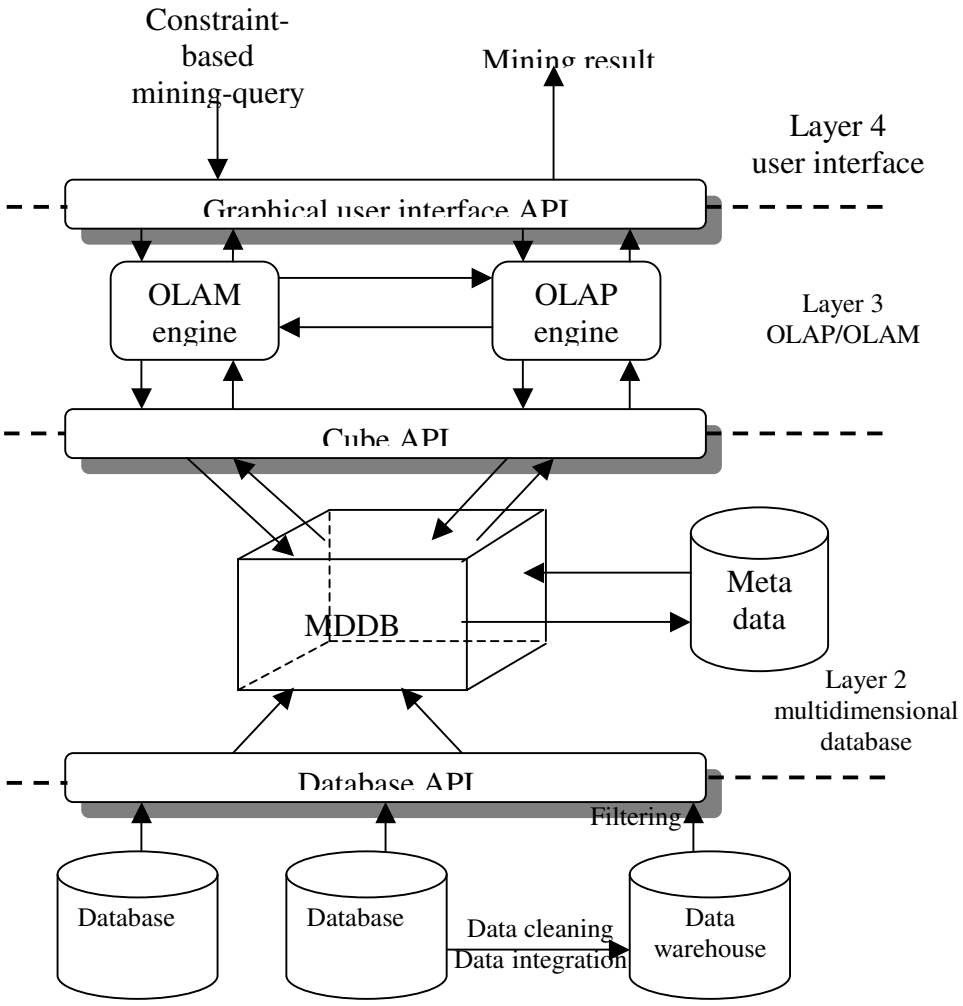

Figura 3.1- Arquitetura Integrada de OLAM e OLAP

Fonte- [HAN01a]

Principais componentes da arquitetura DBminer:

- OLAM *engine*

  efetua múltiplas tarefas de mineração de dados (associação, *clusterização* classificação etc.) de forma similar à execução de uma consulta com uma ferramenta OLAP.

- Interface de apoio ao processo e orientação ao usuário

  orienta o usuário na seleção das técnicas disponíveis, na interpretação, na avaliação e no armazenamento dos resultados finais. A escolha da regra a ser executada é feita por meio de um guia, o qual, dependendo da tarefa, solicita

parâmetros necessários para a sua execução. Os passos a serem executados são mostrados ao usuário em uma linguagem similar à da SQL chamada Data Mining Query Language – DMQL,que é um padrão para linguagem de consulta (*query*) de *data mining* [HAN99b], de fácil utilização e altamente interativa. A principal justificativa para essa padronização é o sucesso e aceitação pelos padrões já estabelecidos para a SQL. As análises feitas pelo usuário são traduzidas para o padrão DMQL, podendo ser alteradas pelo usuário, caso seja necessário, e, executadas.

O DBminer suporta operações para:

- Análises OLAP

  Apresentam o conteúdo de um cubo em vários níveis de abstração, agregação e generalização. Os resultados podem ser mostrados em forma gráfica ou textual [kIM98a] [INM97a].

- Análises de séries temporais

  Incluem funções[1] para análise de similaridades, tendências, análises de periodicidade etc.

- Regras de Associação.
- Regras de Classificação.
- Regras de *Clusterização* .
- Análise tridimensional de cubos

  Apresentam o conteúdo de três cubos combinados em gráficos tridimensionais.

Considerando-se o apoio ao processo KDD, podem-se fazer as seguintes observações sobre o DBminer:

---

[1] Essas funções são descritas em Berndt el al [BET96A] e Han et al [HAN01a].

- É sabido que as atividades de preparação de dados são complexas e consomem a maior parte do tempo no processo. O DBminer não tem função para auxiliar tais atividades, que incluem a limpeza, integração e consolidação (agrupamento) dos dados. Pressupõe-se que elas tenham sido executadas no processo de definição do DW. A atividade de seleção dos dados que serão utilizados para a mineração dos dados é executada pelo DBminer como parte da preparação da *query* de *data mining*.
- É importante que os resultados interessantes para o usuário sejam armazenados, evitando, assim, que re-execuções desnecessárias de tarefas ocorram no futuro. O Dbminer permite o armazenamento das *queries* DMQL e dos resultados, entretanto, ele não oferece ao usuário funções para uma documentação adequada, incluindo comentários ou emitindo pareceres sobre os resultados, o que pode causar dificuldades no entendimento desses resultados em futuras leituras.
- O pós-processamento das regras encontradas é integrado com o processo de mineração. Isso porque, ao escolher o tipo de tarefa a ser executada e os dados a serem analisados, são, também, avaliados e definidos os valores das medidas de interesse para a regra. Com isso, ao apresentar os resultados, o Dbminer já terá observado os valores das medidas estabelecidas *a priori*.
- A visualização dos conhecimentos extraídos pode ocorrer de diversas formas, dependendo do tipo de tarefa que está sendo executada e da preferência do usuário. Um mesmo resultado pode ser visto de diferentes formas.

É inegável que os objetivos do projeto DBminer vão ao encontro da solução da boa parte dos problemas apresentados no capítulo 2. Contudo, as seguintes questões ainda estão em aberto:

- A sua integração com outras ferramentas. Essa integração é pequena, ficando o usuário limitado aos recursos do próprio sistema;
- A forma como são registradas as experiências bem sucedidas, as tentativas descartadas durante o processo e também o raciocínio que levou a elas. A história do processo pode ser útil na tomada de decisões futuras;

- Qual o apoio efetivo ao usuário para resolver questões não genéricas, com alto nível de complexidade, que envolvam uma análise mais aprofundada nos dados.

Freqüentemente, encontram-se artigos científicos e referências na *internet* sobre evoluções no sistema. Isso demonstra os esforços para aprimorá-lo no sentido de facilitar o trabalho do usuário no desenvolvimento de uma aplicação do processo KDD.

O Quadro 3.1 mostra, de forma resumida, em quais das principais etapas do processo KDD apresentado nas abordagens de Fayyad et al [FAY96a] e Brachman et al [BRA96A] o Dbminer mais colabora.

Quadro 3.1- Resumo dos principais apoios do Dbminer nas etapas do processo KDD

ETAPA	**NÍVEL DE COLABORAÇÃO**
Compreensão do domínio	Não possui funções para apoio no entendimento do domínio
Preparação dos dados	Não tem função para auxiliar essa etapa, que incluem as atividades de limpeza, integração e consolidação (agrupamento) dos dados. No entanto, a sua integração ferramentas OLAP, pode, em alguns, casos ajudar no entendimento dos dados.
Mineração dos dados	É nessa etapa que o Dbminer mais colabora. Ele efetua múltiplas tarefas de mineração de dados (associação, clusterização classificação etc.) de forma similar à execução de uma consulta com uma ferramenta OLAP.
Interpretação / avaliação	As facilidades disponíveis para visualização dos resultados ajudam o usuário no entendimento desses resultados, colaborando, assim, com a redução do tempo despendido no processo.
Incorporação/divulgação	Não possui funções para essa etapa.

3.2- Projeto UGM

O projeto User Guidance Module - UGM, ou seja, Módulo de Orientação ao Usuário, foi desenvolvido por pesquisadores da Universidade Karlusruhe, Alemanha,

motivados pelos diversos problemas enfrentados pelos usuários no desenvolvimento de uma aplicação KDD. Engel et al [ENG99a] reforçam a idéia na necessidade de um apoio ao usuário com algumas considerações:

- O problema central do processo KDD é a análise dos dados, mas o fator crítico de sucesso está na definição do problema e na análise e pré-processamento dos dados. Juntas, essas fases consomem aproximadamente 80% do tempo no processo;
- A documentação de aplicações do processo, juntamente com os seus objetivos e resultados, é importante para a organização, pois além de propiciar sua reutilização, contém informações referentes ao projeto e as experiências corporativas.
- Considerando que o processo KDD consome tempo e diversos recursos, é importante pensar em algum tipo de reutilização do processo.
- A avaliação e definição de formas de iteração estão relacionadas com a definição de tarefas apropriadas para o pré-processamento de dados e a inicialização das técnicas e parâmetros para a indução.

Segundo Engels et al[ENG99a], o usuário deve ter apoio adequado na descrição do problema, no tratamento de sua complexidade, quando da definição da solução, na seleção e utilização de técnicas adequadas, e, ainda, na documentação e no armazenamento das experiências adquiridas no desenvolvimento da aplicação. Essas observações foram consideradas no desenvolvimento do UGM, que, ao ser projetado, preocupou-se com todas as atividades básicas do processo KDD descritas no Capítulo 2, observando, porém, duas diferenças significativas. A primeira é que existem dois momentos em que as mesmas etapas devem estar presentes. O primeiro, na definição do processo quando as tarefas são planejadas e definidas; o segundo momento na aplicação desse processo quando as tarefas são executadas. A segunda diferença, é a inclusão da tarefa de documentação das atividades presentes em todo o processo. Engel et al [ENG99a] ressaltam que as abordagens para o processo KDD não tratam do problema de documentar as experiências vividas na aplicação do processo KDD.

O principal objetivo do UGM é disponibilizar uma estrutura que sirva como orientação para os usuários nas aplicações de descoberta de conhecimento em banco de dados, visando diminuir o tempo de desenvolvimento do processo KDD; melhorar o seu resultado; facilitar a descrição do contexto em que o processo está envolvido (característica dos dados, requerimentos da solução etc.); reutilizar técnicas e algoritmos usados em um projeto; e, também, elaborar protótipos para aplicação de KDD. Esse sistema provê um guia que dá suporte a usuários que não têm experiência em encaminhar um processo KDD bem como àqueles que já têm alguma experiência e que desejam ajuda na escolha do um algoritmo mais apropriado para uma aplicação.

Na Figura 3.2, é ilustrada a estrutura do projeto UGM, desenvolvido para:

- Apoiar a descrição do problema que deverá ser analisado no processo;
- Combinar as tarefas para o problema KDD corrente. Permite, também, fazer uma decomposição dessas tarefas;
- Propor um plano inicial para que o usuário possa executá-lo ou tomá-lo como base para um futuro refinamento do problema;
- Gerenciar os componentes da solução com o objetivo de sua reutilização.

O UGM está estruturado em três principais componentes:

- Componente de Análise do Problema ou Problem Analysis Component - PAC
  Conduz um diálogo com o usuário para a definição e análise do problema. Nesse momento, é iniciado um processo de busca nos repositórios existentes com o objetivo de reutilizar partes de processos já definidos.
- Componente de planejamento ou Planning Component – PA
  Usa os repositórios existentes para planejar uma seqüência de tarefas a serem executadas. As tarefas são refinadas por decomposição até se chegar a tarefas simples, que, quando executadas de forma planejada, resolvem o problema definido.

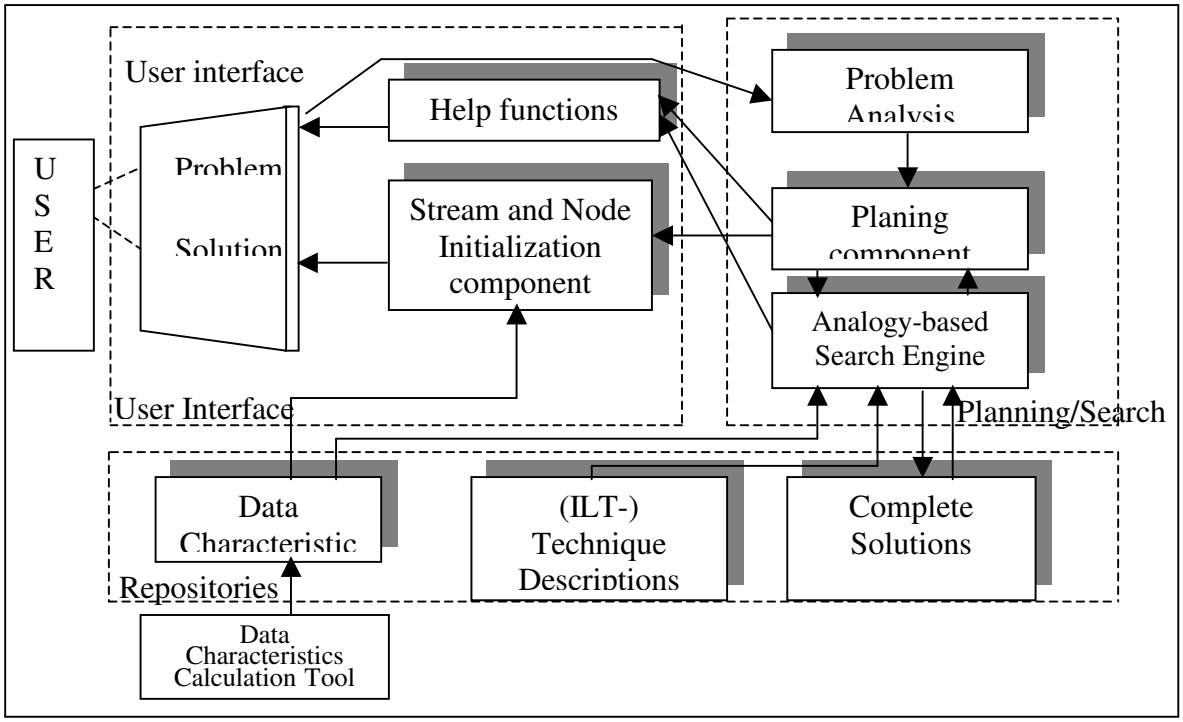

Figura 3.2- Estrutura do UGM

Fonte- Engels et al [ENG99a]

- Repositórios

  Existem três tipos de repositórios que auxiliam na definição do domínio do problema, na decomposição das tarefas e na reutilização dos componentes:

  - Complete Process Solution - CPS, ou seja, Soluções Completas de Processo e *Reusable Process Units* – RPS, ou seja, Unidades Reutilizáveis do Processo. Ambos contêm descrições das soluções já extraídas de aplicações anteriores do processo KDD. A descrição dos componentes é feita usando uma abordagem de descrição lógica, que permite a utilização de algoritmos, para verificar o quanto o componente pode ser bom para a aplicação. Basicamente, essa verificação é estabelecida por um conjunto de características que descreve a similaridade e a distinção do corrente RPS em relação à definição do problema;
  - Descrição de técnicas: contém descrições das técnicas disponíveis, tais como inicialização e adaptação de regras, efeitos dos resultados, restrições de

entrada. Mostra como é feita a inicialização e a adaptação de regras e os efeitos dos seus resultados;

- Características dos dados: contêm informações sobre os dados em termos de dimensões, tipos de dados e domínio. O gerenciamento desse repositório é feito por uma ferramenta chamada *Data Characteristics Calculation Tool* que foi incluída no UGM.

A construção de uma aplicação é conduzida por diálogos com o usuário e começa da definição do problema, quando seus objetivos são traçados juntamente com o contexto no qual ele será resolvido, até a documentação dos resultados obtidos. O usuário define plano do processo usando o PAC*,* que, a partir das informações definidas por ele, pode, de forma inteligente, responder questões relevantes ao problema, como entradas sem valores; saídas criadas, mas não utilizadas. Os requerimentos para a solução são definidos à medida que a descrição do problema é detalhada. Como é baseado em reutilização, o PAC pode realizar pesquisas nos repositórios por projetos similares e dar *feedback* para certas decisões; pode fazer propostas para pré-processamento; e, também, ver a aplicabilidade de certos algoritmos sobre os dados. Nesse último caso, o PAC é auxiliado por uma ferramenta que pesquisa as características dos dados do domínio em questão. O trabalho desse componente gera uma decomposição de tarefas que está fortemente ligada às RPS. A decomposição é interrompida quando é encontrada uma unidade nos repositórios (CPS ou RPS) que possa ser utilizada, ou até chegar a tarefas simples que possam ser mapeadas para técnicas apropriadas. O PAC passa o controle ao componente planejamento, que tenta identificar componentes do processo em um repositório e projeta uma seqüência de passos que podem resolver a funcionalidade requerida. Esse trabalho é iterativo. No início de uma aplicação o plano é basicamente uma decomposição hierárquica de tarefas, as quais são refinadas até serem mapeadas para as técnicas existentes. A documentação é produzida automaticamente durante o processo

A metodologia UGM baseia-se no conceito de decomposição de tarefas, nos recursos que o usuário pode usar para refinar o problema e no conjunto de componentes, descritos pelas suas pré e pós condições e armazenadas no repositórios. O UGM possibilita, na definição do plano, definir qual a melhor técnica a ser aplicada. Tem, igualmente, a habilidade para usar a base de conhecimento para analisar a descrição do problema em relação às características dos dados, propondo algoritmos para o usuário executar certas tarefas .

A implementação do UGM baseou-se no sistema Clementine desenvolvido pela Integral Solution Ltd (ISL), que possui recursos para acesso, importação e exportação dos dados e técnicas e métodos para análise dos dados. Ele é orientado a objetos. Aceita que componentes e aplicações desenvolvidas por sejam incorporados no seu ambiente.

A estrutura apresentada por Engels et al [ENG99a] é um grande passo para a criação de um ambiente mais apropriado para apoiar o trabalho do usuário nas atividades do processo KDD. Ela deixa claro que, com a já existência de um grande número de algoritmos para análises de dados e descoberta de conhecimento – algoritmos que possuem um certo nível de complexidade e uma definida funcionalidade para resolver um problema específico – o desafio é fazer com que esses algoritmos trabalhem juntos. O UGM caminha nesse sentido, promovendo a reutilização e provendo ao usuário um ambiente que o apóia numa mais precisa definição do problema, na formalização da descrição do problema, na decomposição das tarefas e na pesquisa e armazenamento dos dados. Contudo, ele ainda não resolve alguns problemas destacados no capítulo 2. Entre eles podemos citar:

- Falta de apoio para avaliação dos resultados

  A experiência, o conhecimento prévio do domínio dos dados e da aplicação e o domínio das técnicas, por parte do usuário, são determinantes para a avaliação do processo. O ambiente é baseado em casos de sucesso. Ele não explora as experiências dos usuários em outros projetos. Não foram abordadas, também, as

formas de utilização e parametrização das tarefas de descoberta do conhecimento (associação, *clusterização*, etc).

- Integração de ferramentas

  O ambiente está restrito aos recursos oferecidos pelo Clementine. Os problemas de ambientes heterogêneos não foram considerados.

- Deficiência no apoio ao entendimento do domínio da aplicação e na organização do projeto

  O ambiente está mais voltado para o lado operacional do processo. As dificuldades para entender e mapear o domínio dos dados e fazer a preparação dos mesmos não foram tratadas.

- Visualização dos resultados

  Questão importante para geração de conhecimento que também não foi enfatizada como parte relevante.

O UGM tem papel fundamental no processo KDD, pois é um caminho para dar mais autonomia de execução a algumas etapas do processo. Pode ser, também, um motivador para novas pesquisas, principalmente em relação aos sistemas autônomos para executar todas as fases do processo KDD.

O Quadro 3.2 mostra, de forma resumida, em quais das principais etapas do processo KDD apresentado nas abordagens de Fayyad et al [FAY96a] e Brachman et al [BRA96A] o UGM mais colabora.

Quadro 3.2- Resumo dos principais apoios do UGM nas etapas do processo KDD

ETAPA	NÍVEL DE COLABORAÇÃO
Compreensão do domínio	Pouca funcionalidade de apoio a essa etapa.
Preparação dos dados	Apoia mais na definição das características dos dados. Não foi disponibilizadas funções ajudar na preparação dos mesmos.
Mineração dos dados	De forma geral pode-se dizer que é onde ele mais colabora para o processo KDD, visto sua preocupação em ajudar o usuário nas atividades de descrição do processo, na escolha das tarefas, na definição de que algoritmos utilizar e na reutilização de conhecimento.
Interpretação / avaliação	Pouca funcionalidade para essa etapa.
Incorporação/divulgação	Não possui funções para essa etapa.

3.3- Projeto HAMB

O rápido crescimento da disponibilidade de dados para exercitar a tarefa de descoberta de conhecimento e a cobrança por resultados imediatos trazem junto uma demanda por sistemas inteiramente autônomos – sistemas que têm a capacidade de atuar sem a necessidade de um estímulo humano ou de outro *software*, além de manter controle sobre o seu comportamento [JEN98a] – para suporte ao processo KDD. Por serem autônomos, eles reduzem o tempo utilizado pelos usuários para executar manualmente as tarefas eliminando, também, os erros produzidos pelo cansaço mental ou distração dos usuários [LIV01a]. Esses sistemas têm capacidade para verificar um número muito maior de hipóteses do que as exploradas manualmente.

Uma funcionalidade necessária para um sistema de descobrimento autônomo é a sua habilidade para selecionar adequadamente as tarefas a serem executadas e a seqüência de sua execução. Nesse sentido, foi criado o projeto *Heuristic Autonomous Model Builder* - HAMB – um sistema autônomo que implementa uma estrutura baseada em agenda e justificação de tarefas. Essa estrutura consiste de uma agenda de tarefas priorizadas pela sua plausibilidade. A plausibilidade de cada tarefa é

calculada em função do peso do argumento dado para a execução da tarefa e o valor do interesse estimado do item na operação. O Quadro 3.3 mostra a visão de uma agenda com suas tarefas e seus pesos.

As tarefas são colocadas na agenda e executadas com o uso de métodos heurísticos. Ao propor uma tarefa, o método deve, também, propor um ou mais argumentos com seus respectivos pesos para a execução da mesma. Nos casos em

Quadro 3.3- Agenda de Tarefas

Fonte- Livingston [LIV01a]

Agenda de Tarefas					
Tarefa	**Operação**	**Itens**	**Argumento**	**Peso**	**Plausibilidade**
Tarefa-1	**Operação-1**	**Item-11** **Item-12** **...**	**Argumento-11** **Argumento-12** **...**	**Peso-11** **Peso-12** **...**	**Plausibilidade-1**
Tarefa-2	**Operação-2**	**Item-21** **Item-22** **...**	**Argumento-21** **Argumento-22** **...**	**Peso-21** **Peso-22** **...**	**Plausibilidade-2**
Tarefa-3	**Operação-3**	**Item-31** **Item-32** **...**	**Argumento-31** **Argumento-32** **...**	**Peso-31** **Peso-32** **...**	**Plausibilidade-3**
Tarefa-n	**Operação-n**	**Item-n1** **Item-n2** **...**	**Argumento-n1** **Argumento-n2** **...**	**Peso-n1** **Peso-n2** **...**	**Plausibilidade-n**

que o método tenta recolocar uma tarefa já agendada, somente o(s) novo(s) argumento(s) para a sua execução é (são) aceito(s). Segundo Livingston [LIV01a], esse mecanismo é suficiente para selecionar a próxima tarefa a ser executada no processo KDD.

Os dois fatores usados para estimar a plausibilidade – peso do argumento dado para a tarefa proposta e o interesse do item envolvido na tarefa – provêm duas importantes propriedades na função de plausibilidade que podem torná-la suficiente para selecionar uma tarefa de descoberta [LIV01a]. Esses dois fatores sustentam três teses:

- A escolha de tarefas com argumentos mais fortes permitirá que o sistema de descoberta selecione aquelas que são mais apropriadas para o problema em questão;

- A escolha de tarefas que operam sobre itens mais interessantes antes daquelas que operam em itens menos interessante, possibilitará que resultados mais interessantes sejam encontrados mais rapidamente no processo de descoberta;
- O uso de métodos heurísticos para executar e propor tarefas permitirá que sistemas autônomos de descobrimento sejam guiados pelo conhecimento já extraído, isto é, conhecimento geral sobre o tipo de descoberta que está sendo executado e conhecimento sobre o domínio específico.

O HAMB utiliza regras de indução como o método básico para identificar os padrões de conhecimentos.

A Figura 3.3 ilustra, em nível macro, o funcionamento da estrutura apoiada em agenda e justificação.

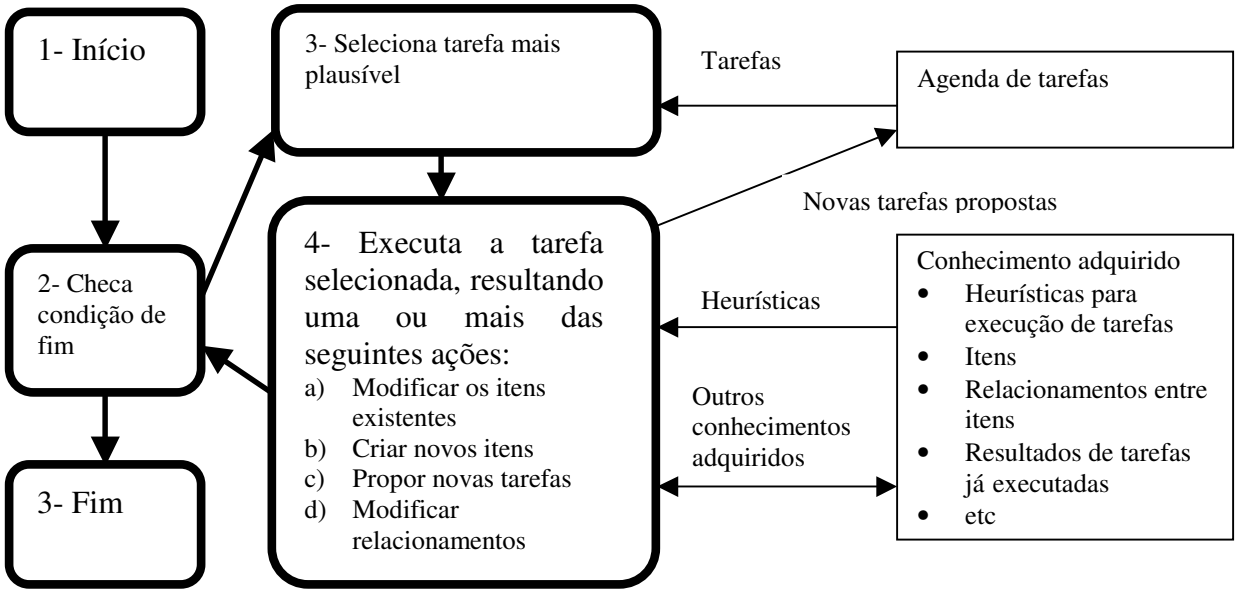

Figura 3.3- Controles e Componentes da Estrutura Baseada em Agenda
Fonte- Livingston [LIV01a]

Os principais objetos da estrutura são:

Itens

São elementos do escopo do problema, pesquisados pelo programa de descoberta para encontrar um conhecimento. São instâncias e/ou conjuntos de instâncias de

conceitos usados no problema de descoberta. Os itens são representados por *frames* tendo os seguintes atributos:

Nome – identificador único;

Descrição – breve descrição que pode ser usada em algum diálogo ou relatório;

Interesse – grau de interesse que o usuário tem por ele;

Propriedades – aspectos do item que podem ser usados para identificar o interesse;

Relacionamentos – relacionamentos entre um item e outros;

Comentários – anotações feitas pelos programas sobre os itens.

Tarefas

São operações feitas sobre os itens ou conjuntos de itens, que o programa de descoberta, usando heurísticas, tenha identificado como potencial para produzir conhecimento. As tarefas são representadas usando quatro tuplas (tipo-operação, itens, argumentos-adicionais, raciocínio). Tipo-operação é o tipo de operação a ser executada; *itens* compõem uma lista sobre os quais a tarefa será executada; argumentos adicionais são parâmetros necessários para a tarefa executar, como threshold; raciocínio são justificativas para a execução da tarefas.

A seleção de uma tarefa para a execução baseia-se na plausibilidade descrita na agenda. A que possui o maior valor será selecionada.

As tarefas são executadas usando heurísticas, regras de ações e condições que especificam procedimentos para a sua execução. Os resultados de uma execução podem ser a proposta de uma nova tarefa, a criação de um novo item, a modificação de um item existente ou a mudança do valor de um relacionamento entre itens. As tarefas são propostas, isto é, colocadas na agenda, na inicialização ou na execução de uma heurística para uma outra tarefa. A proposta de uma tarefa, durante a inicialização, é automática. Para cada item é proposta uma tarefa.

O programa autônomo de descobrimento se encerra quando a razão custo-benefício em relação aos resultados deixa de ser atrativa, isto é, quando os recursos usados para continuar os processos não são compensados por uma descoberta interessante. Duas condições básicas podem finalizar o processo: a primeira, no momento em que o número de ciclos definidos pelo usuário por meio de um *threshold* é alcançado; a segunda, no momento em que o valor da plausibilidade de todas as tarefas fica abaixo de outro *threshold* definido pelo usuário.

O HAMB emite periodicamente um relatório sobre o comportamento de uma execução. Esse relatório deverá servir de apoio ao usuário para definir sobre paradas não-programadas do processo.

O HAMB pode ser usado para implementar um completo sistema autônomo de descobrimento, provendo um mecanismo de seleção de tarefas que é sensível ao interesse e à adequação da tarefa para o problema de descoberta. O HAMB reutiliza o conhecimento para filtrar resultados pouco interessantes.

Sob o ponto de vista do processo KDD, a proposta anterior mostra novos horizontes, principalmente em relação à otimização do tempo despendido pelo usuário no processo. Contudo, alguns itens podem ser considerados como não abordados ou ainda devem ser aprimorados em novas versões do produto. Entre os itens não abordados, incluem-se:

- Seleção de conhecimentos similares

  A produção de regras é enorme sendo várias delas similares. O processo de eliminação das regras que produzem resultados semelhantes não foi abordado.
- Restrição do domínio / trivial

  O HAMB foi testado e validado em um domínio específico. Cada aplicação tem suas características. Ele deve ser testado em outras aplicações e ampliado, se for o caso, para tornar-se genérico.
- Pouca interação com o usuário

No capítulo 2 foram mostradas as características interativas e iterativas do processo KDD. As avaliações dos usuários para definir redirecionamentos são freqüentes e importantes.

- Dificuldade para redirecionar o processo

  O usuário tem pouca opção para mudar o curso do processo.

- Visualização dos resultados

  Os resultados são difíceis de ser entendidos. O projeto deve ser ampliado para dar mais opções ao usuário.

Segundo Livingston [FIN01a], vários sistemas têm sido criados para atender a diferentes fases de um processo de descoberta empírico, porém nenhum deles integra todas as fases. O HAMB tem como meta fazer a seleção dessas tarefas automaticamente. A participação do usuário nesse ambiente é pequena. Apesar de alguns sistemas de apoio ao processo KDD estarem ampliando sua autonomia, eles ao contrário da proposta apresentada no HAMB, ainda não executam de forma autônoma.

O Quadro 3.3 mostra, de forma resumida, em quais das principais etapas do processo KDD apresentado nas abordagens de Fayyad et al [FAY96a] e Brachman et al [BRA96A] o HAMB mais colabora.

Quadro 3.3- Resumo dos principais apoios do HAMP nas etapas do processo KDD

ETAPA	**NÍVEL DE COLABORAÇÃO**
Compreensão do domínio	Não possui funções para essa etapa.
Preparação dos dados	Não possui funções para essa etapa.
Mineração dos dados	É onde ele mais colabora. Com as tarefas agendadas ele trabalha de forma autônoma até produzir os resultados. Ele, por meio de heurísticas, define qual algoritmo, e que dados deve usar.
Interpretação / avaliação	Pouca funcionalidade para essa etapa.
Incorporação/divulgação	Não possui funções para essa etapa.

3.4- Memória Organizacional

No desenvolvimento do processo KDD, a necessidade de coletar, representar e distribuir o conhecimento referente ao domínio da aplicação é constante. Isso converge para os trabalhos relacionados à área de Memória Organizacional.

Memória Organizacional pode ser caracterizada como um sistema flexível de computador que captura o *know-how* acumulado na empresa e outros conhecimentos úteis, tornando-os disponíveis para melhorar a eficiência e eficácia dos processos que exigem níveis elevados de conhecimento [LIA99a]. Essa tecnologia, segundo Abecker et al [ABE99a], pode contribuir para:

- uma melhor utilização dos documentos disponíveis, mas que efetivamente são pouco explorados;
- a formalização das regras de negócios em um sistema de w*orkflow*;
- a melhor utilização das habilidades e conhecimentos humanos;
- o registro das experiências e k*now-how* das melhores práticas para tratar bases de dados;
- o armazenamento dos processos de tomada de decisão nas melhores práticas para tratar bases de dados.

A maioria dessas atividades é apoiada total ou parcialmente pela tecnologia de informação, mas ainda falta uma visão mais aprofundada que caracterize as propriedades específicas de um Sistema de Informação de Memória Organizacional ou *Organizational Memory Information System* - OMIS.

Segundo Abecker et al [ABE99a], a aplicação de sistemas baseados em conhecimento ou *Knowledge-Based System* - KBS apresenta, na prática, sérias deficiências e a análise desses problemas tem levado ao desenvolvimento de abordagens de OMIS. Studer et al [STU99a] apresentam alguns critérios para diferenciar um sistema baseado em conhecimento de um OMIS:

- KBS está focado na solução de uma simples tarefa, enquanto um OMIS apóia uma coleção de diferentes processos do negócio e também diferentes tarefas;
- O conhecimento gerenciado por um KBS tem um alto nível de formalização, enquanto o OMIS consiste em conhecimento em diferentes níveis de formalização (documentos, hipertextos e bases formais de conhecimento);
- Um OMIS integra diferentes tipos de conhecimentos (melhores práticas, experiências, processo de conhecimento, padrões) em diferentes níveis de representação. Como um KBS é direcionado para resolver tarefas únicas, o nível de conhecimento requerido é pequeno e homogêneo;
- *Groupware* e técnicas de disseminação de conhecimento normalmente não fazem parte de um KBS, mas são essenciais para um OMIS porque o conhecimento gerenciado tem que ser comunicado entre as pessoas da organização. O OMIS deve ser integrado com aplicações legadas e componentes de sistemas já existentes que são utilizados para uma específica estratégia de gerenciamento de conhecimento.

Segundo Liao et al [LIA99a], um OMIS armazena dados, informações e conhecimento de diferentes fontes de uma empresa, que estão representados em diferentes formas como banco de dados, documentos e base formal de conhecimento . Ele deve ser permanentemente estendido, atualizado e acessível na empresa. Na Figura 3.4.1, está ilustrado o modelo proposto por Liao et al [LIA99a] como uma possível organização de um OMIS.

O nível objeto consiste em diversas fontes de informação e conhecimento com representação formal, entendida pelo computador, e com representação informal, tratável pelas pessoas. A decisão para formalizar um conhecimento depende da relação custo e benefício na sua utilização.

O nível de descrição do conhecimento permite um uniforme e inteligente acesso às fontes do nível objetos.

O nível de aplicação faz a conexão entre o modelo de informação e uma situação que apresenta uma necessidade concreta de aplicação.

Quando necessita de uma informação ou conhecimento, o usuário submete uma consulta ao OMIS. Da mesma forma, o OMIS pode armazenar um conhecimento criado na ocasião de uso da base de conhecimento, enriquecendo, assim, essa base para subseqüente consulta.

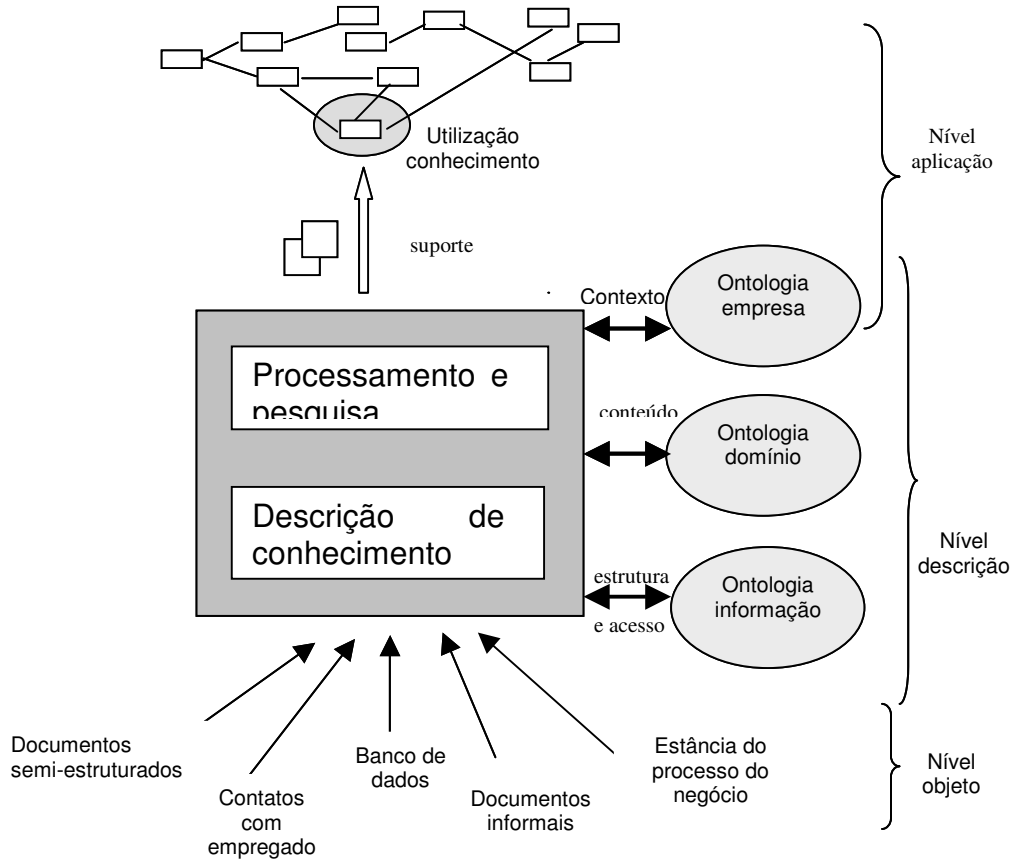

Figura 3.4.1- Modelo de Memória Organizacional

Fonte- Liao et al [LIA99a]

O principal desafio no projeto de um OMIS é a construção de um repositório de informação ou conhecimento que pode ser capturado ou criado na sua utilização [DOR00b]. Todo conhecimento ou informação devem ser descritos por um número de atributos representando o metamodelo, o conteúdo de informação e o contexto de sua criação e de sua aplicação.

Os conceitos de Ontologia têm-se apresentado como uma tecnologia que poderá dar uma grande contribuição no apoio à construção de um OMIS.

Ontologia é um entendimento compartilhado e comum de algum domínio que pode ser comunicado através das pessoas e sistemas informatizados [BEN98a] e que fornece uma descrição concisa, uniforme e declarativa do conhecimento, independentemente do local e da forma como os dados estão armazenados. Ela pode, dessa forma, ser compartilhada e reutilizada em diferentes aplicações. Dentro da Ciência da Computação, o termo Ontologia originou-se na comunidade de Inteligência Artificial (IA), como um modelo semântico para ser utilizado publicamente [DOR00b]. Uma definição amplamente utilizada é a de Gruber [GRU93a] *´an ontology can be defined as a formal, explicit specification of a shared conceptualization'*. Dessa definição o termo formal refere-se ao fato de que a ontologia deve ser tratável pelo computador; o explícito refere-se ao tipo de conceito usado sendo que as restrições no seu uso são explicitamente definidas; e o compartilhado reflete a idéia de que ontologia não é privada, mas pública. No contexto de organização de conhecimento, conceitualização compreende não só objetos, padrões, conceitos e entidades que existem em alguma área de interesse, mas também os relacionamentos que existem entre eles.

No modelo ilustrado na Figura 3.4.1, é feita referência a três tipos de ontologias, que estão representadas, em conjunto, na Figura 3.4.2.

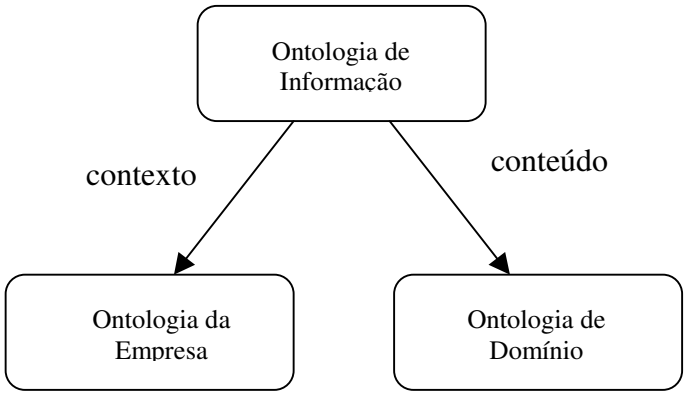

Figura 3.4.2- Três Dimensões de Descrição do Conhecimento [LIA99a]
Fonte- Liao et al [LIA99a]

Ontologia de informação descreve o vocabulário do metamodelo de informação, que caracteriza os diferentes tipos de fontes de informação com suas respectivas estruturas, acessos e formatos. Os conceitos nessa ontologia são estáveis e independentes de domínio.

Ontologia de empresa é usada na modelagem do processo. Ela descreve o contexto da informação. É expressa como uma estrutura organizacional e como um modelo do processo. Seu propósito é modelar a necessidade de conhecimento no processo do negócio bem como descrever o contexto do processo. A expectativa na construção dessa ontologia é que ela seja o mais independente possível de uma empresa real. Ela deve servir a mais de uma empresa.

Ontologia do domínio é usada para modelar o conteúdo das fontes de informação. Tipicamente os conceitos nessa ontologia são muito específicos para uma aplicação em especial. Esforços em pesquisas têm sido feitos no sentido de construí-la da forma mais independente possível de uma única situação. Normalmente os conceitos nessa ontologia são muito específicos para uma aplicação em particular. Cada base de conhecimento, utilizada em cada contexto, pode possuir sua própria ontologia descrevendo o significado dos conceitos usados de acordo com sua visão particular do mundo.

Segundo Jurisica et al [JUR99a], um dos principais usos de ontologia é a documentação e reutilização das informações ou conhecimento que são trocados entre as aplicações. Os sistemas podem se conectar por meio dela, visto que é compartilhada, aumentando assim o seu potencial de uso. No caso de um OMIS, ela permite o uso do conhecimento de forma mais sofisticada.

No contexto de um processo KDD, um OMIS, principalmente se ele usar a abordagem de ontologia, é importante, pois permite representar um modelo de documentação de forma flexível e adaptável ao problema em questão. Essa flexibilidade permite tratar alguns pontos do processo KDD que não foram explorados nas abordagens anteriores. Entre eles a documentação, de forma histórica, do processo

KDD – contemplando as experiências adquiridas pelo usuário; a facilidade de disseminação do conhecimento – visto que a ontologia é pública; e a formalização do conhecimento.

Uma das dificuldades nessa abordagem é a falta de critérios consolidados para a definição, construção e utilização de um OMIS. Além da evolução necessária na tecnologia, existe uma demanda para diminuir os custos, principalmente na recuperação da informação.

O Quadro 3.4 mostra, de forma resumida, em quais das principais etapas do processo KDD apresentado nas abordagens de Fayyad et al [FAY96a] e Brachman et al [BRA96A] o HAMB mais colabora.

Quadro 3.4- Resumo dos principais apoios da tecnologia MO nas etapas do processo KDD

ETAPA	**NÍVEL DE COLABORAÇÃO**
Compreensão do domínio	Um OMIS, por princípio, deve dispor de funções para colaborar nessa etapa.
Preparação dos dados	Deve disponibilizar informações que ajudem a identificar e selecionar os dados a serem preparados.
Mineração dos dados	Deve possuir funções para integrar suas informações com uma ferramenta ou algoritmo de mineração de dados.
Interpretação / avaliação	Sua preocupação em reutilização e documentação do conhecimento ajudam nessa etapa.
Incorporação/divulgação	Os conceitos de antologias ajudam no compartilhamento e divulgação dos resultados.

3.5- Conclusões

Nas seções anteriores, foram mostradas diferentes linhas de pesquisa com o propósito de facilitar o processo de gerenciamento do conhecimento para atender aos mais diversos interesses. Nessa seção, é feito um resumo, mostrado na Quadro 3.5, sobre os projetos analisados, identificando algumas características consideradas positivas, bem como aqueles pontos que ainda se apresentam como deficientes. Esse resumo não tem um caráter comparativo, visto que os objetivos dos projetos têm propostas diferentes.

Quadro 3.5- Resumo sobre Projetos/Tecnologias de Apoio ao Processo KDD

Pontos positivos			
DBminer	**UGM**	**HAMB**	**MO**
Integração com ambientes DW e ferramentas OLAP	Liberação do usuário de algumas atividades	Automicidade	Formalização do conhecimento
Forte integração com banco de dados relacional	Reutilização do conhecimento e de algoritmos	Reutilização do conhecimento e de algoritmos	Reutilização do conhecimento
Integração de um número variado de regras de mineração	Facilidade para descrição do problema	Formalização do conhecimento	Utilização do seu produto por diversos sistemas
Facilidades para visualização dos resultados	Automatização de partes do processo	Diminuição de custos com pessoas	Documentação das experiências
Interatividade com o usuário	Orientação para o usuário construir a aplicação	Velocidade na obtenção dos resultados	Representação flexível dos modelos
Utilização de uma linguagem padronizada para OLAM	Padronização do fluxo do processo		Compartilhamento do conhecimento
Pontos negativos			
DBminer	**UGM**	**HAMB**	**MO**
Falta de apoio para documentação dos domínios dos dados e dos padrões descobertos	Baixa interatividade com os usuários	Dificuldade para selecionar o grande número de padrões produzidos	Dificuldade para recuperação dos conhecimentos
Forte dependência do usuário	Dificuldade no processo de visualizar os resultados	Baixíssima interatividade com o usuário	Falta de critérios bem definidos para a construção
Falta apoio para documentar o problema	Falta de apoio para avaliar os resultados	Dificuldade para redirecionar o processo	Custos altos para a sua contrução
Falta apoio para a preparação/exploração dos dados	Pouca integração com outras ferramentas	Falta apoio para a documentação das experiências	Dificuldade para integração com outras ferramentas
Nenhuma reutilização do conhecimento	Falta apoio para o entendimento do domínio da aplicação	Dificuldade no processo de visualizar os resultados	Total dependência das pessoas
Pouca integração com outras ferramentas	--	--	--

4- Estudo de Caso: Aplicação do Processo KDD

Nos capítulos anteriores, foram apresentadas visões distintas, mas interdependentes sobre KDD. A primeira, é uma abordagem teórica do processo KDD, confirmando que o mesmo é complexo. A segunda, os trabalhos e pesquisas que estão sendo feitos auxiliarem a condução do processo KDD. Este capítulo, considerando que relatos sobre a convergência dessas duas visões em casos reais ainda são escassos, tem dois objetivos básicos. O primeiro é apresentar uma análise dos resultados obtidos em um caso real de utilização do processo KDD. Esse estudo é relevante por duas razões principais. A primeira: verificar como as fases do processo descrito por Fayyad et al [FAY96a] e Brachman et al [BRA96a] se adaptam a uma aplicação do mundo real. A segunda: verificar como as etapas do processo se comportam ao usar dados de um DW. O segundo objetivo é verificar o comportamento de uma ferramenta integrada de DM em um caso real, considerando principalmente os aspectos de interdisciplinaridade desse campo de pesquisa.

O Capítulo é organizado da seguinte forma: na seção 4.1, é apresentado o problema a ser analisado pelo processo KDD. O domínio dos dados da aplicação é mostrado na seção 4.2. Na seção 4.3, justifica-se a escolha da ferramenta para análise dos dados. A seção 4.4 apresenta uma análise dos resultados obtidos com a utilização da ferramenta. A avaliação do processo KDD é feita na seção 4.5, visto a existência de iterações entre as etapas até o final do projeto. Por último, na seção 4.6, são relacionados elementos fundamentais a serem observados na construção de um ambiente de apoio para o processo KDD.

4.1- Definição do Problema

Fundada em 9 de dezembro de 1969, a Sudecap – Superintendência de Desenvolvimento da Capital - é uma autarquia da Prefeitura de Belo Horizonte e sua função é cuidar da reestruturação urbana da cidade, projetando, executando e mantendo obras públicas.

A Sudecap responde pela construção, manutenção e recuperação de ruas, avenidas, edificações e espaços públicos de Belo Horizonte. Seu principal desafio é conciliar o crescimento da cidade com a qualidade de vida, preservando o meio ambiente e evitando expor a população a transtornos. Para isso, a Sudecap emprega mais de mil funcionários.

Para estabelecer suas metas, a Sudecap conta também com uma importante aliada: a própria comunidade. Através do Orçamento Participativo, um programa da PBH, a população tem a oportunidade de expressar suas necessidades e definir as prioridades que a Sudecap ajuda a realizar.

Além disso, a Sudecap busca estar sempre em contacto direto com as pessoas, por meio de uma política de acompanhamento social das obras e de outros canais de comunicação, tais como a distribuição de material informativo e o suporte de uma central telefônica , à qual o cidadão pode solicitar serviços de tapa-buracos, limpeza de boca-de-lobo e colocação de tampas e grelhas.

Dentre outras, são funções da Sudecap:

- A elaboração de projetos, construção e manutenção de vias públicas pavimentadas, viadutos municipais, túneis, trincheiras e passarelas;
- A construção e manutenção de escolas, creches, centros de saúde, bem como de todos os prédios públicos de Belo Horizonte;
- A elaboração do Plano Diretor de Drenagem de Belo Horizonte;
- A execução de obras de drenagem e o gerenciamento de programas de regularização de bueiros e tapa-buracos, assim como a limpeza e manutenção de aproximadamente 30.000 bocas-de-lobo.

Os recursos financeiros necessários para a execução das obras são previstos no orçamento da Prefeitura. A execução de qualquer empreendimento pode ser feita com recursos próprios ou com a contratação de empresas prestadoras de serviços. Tal empreendimento é entendido como sendo a intervenção que resulte em implantação,

modificação, criação ou reparação de elementos da infra-estrutura ou de bem imóvel, mediante construção, ampliação, reforma ou restauração.

Caso ocorra a contratação, ela deve ser feita por um processo de licitação, que consiste na escolha mais vantajosa de pessoa física ou jurídica que execute o projeto para a Sudecap, tendo sempre em conta a estrita conformidade do mesmo com os princípios reguladores estabelecidos em legislação específica. A modalidade de licitação é definida pelo porte do empreendimento e pela urgência de sua demanda, sendo consideradas mais urgentes as obras que se processam para correção de defeitos originários principalmente da ação das chuvas, defeitos que surgem sem que deles houvesse previsão.

Após a definição pela execução de um empreendimento, a Sudecap deve gerir todas as suas fases para que o projeto tenha sucesso. Nessa gerência, inclui-se o planejamento e projeto da obra, o controle da qualidade dos serviços executados, as medições, o controle de custos e a liberação da obra para a comunidade. No acompanhamento da obra, o foco incide principalmente sobre os custos, uma vez que, considerando o próprio conceito de orçamento participativo, ela passa a ser uma parceria entre a administração (prefeitura) e a comunidade, que, na condição de parceira, tem, naturalmente, direito à prestação de contas.

Na formação do preço final da obra e na gestão e acompanhamento dos custos, a Sudecap define e mantém políticas de preços para os projetos por ela administrados. Essa atribuição é controlada por órgãos fiscalizadores e regida pelas leis e princípios da administração pública, subordinando-se também ao desejo da sociedade de que os impostos arrecadados tenham uma destinação correta e justa, o que torna obrigatoriamente de domínio público a elaboração e execução dessas políticas.

O valor previsto para um projeto é determinado observando-se todos os componentes formadores do preço final: custos diretos, indiretos e margens de lucro. A

estimativa de preço é estabelecida após a identificação do conjunto de atividades e recursos, tendo em vista também os prazos para a execução da obra.

O valor do Custo Direto (CD) é obtido pelo custo dos insumos e da mão-de-obra que serão gastos na execução específica de um determinado serviço. Mensalmente é feita uma pesquisa de mercado para compor o custo de cada serviço. Também integram os custos finais as Despesas Indiretas e a Bonificação ou lucro (BDI). O BDI obtido em unidades monetárias gera outro valor, chamado de BDI em percentual, que corresponde à divisão do BDI em valor monetário pelo CD. O preço total da atividade será a soma do CD mais BDI percentual aplicado no CD, isto é, CD * (1 + BDI percentual); o preço total do empreendimento será a soma do preço das atividades. Esses valores são levados ao conhecimento do público, através do edital de licitação. As empresas interessadas em apresentar propostas para a realização da obra devem apresentar os valores dos serviços considerando que eles não podem ultrapassar em 20% os preços unitários da Sudecap, quando considerados por atividade e nem exceder o valor total obtido por ela.

Definida a melhor proposta, dentre as apresentadas pelas empresas, é calculado um fator chamado de 'Fator K', que é estabelecido pela divisão do preço total da contratada pelo CD calculado pela Sudecap.

Durante a execução desse projeto, as seguintes ocorrências podem surgir:

Aditivo no contrato que é alteração das condições iniciais, constatada a insuficiência de alguma delas para o cumprimento do mesmo, por exemplo, valor, prazo, etc;

Serviço extra que ocorre quando há necessidade de execução de alguma atividade/serviço não planejada na contratação do projeto;

Ressarcimento que é feito à empresa prestadora de serviço quando o projeto teve atrasos provocados pela Sudecap;

Realinhamento de preços que ocorre quando algum acontecimento na economia provoca alterações substanciais nos preços de algum serviço/atividade;

Reajustamento do Contrato que se baseia em cláusulas já relacionadas no contrato.

Esse fator é utilizado durante todo o andamento do projeto. Ele será utilizado para remuneração dos serviços não previstos na planilha da licitação por meio dos preços da tabela de custo direto da Sudecap, bem como em todas situações que impliquem acompanhamento dos custos. A Sudecap considera que a forma de calcular o Fator K, por ser baseada numa combinação de valores definidos por ela e pela empresa contratada, dá transparência à contratação de serviços, inibindo possíveis manipulações, durante a execução do projeto, que objetivem vantagens adicionais, que não as originalmente contratadas, para as partes.

Embora a Sudecap considere justa a política de definição de preços, diversos questionamentos ainda são feitos:

- Um Fator K alto indica a possibilidade de mais aditivos no futuro?
- O valor do Fator K influencia no número de realinhamentos?
- Como o Fator K influencia nas alterações ocorridas no contrato durante a sua vigência?

Essas questões foram levantadas pelos administradores da Sudecap devido ao número de projetos gerenciados, ao alto valor financeiro envolvido e à responsabilidade delegada ao governo pelo "Belohorizontino" para gerir o que, em última análise, são os seus recursos.

O processo KDD foi visto como uma alternativa bastante interessante para ajudar a determinar as questões envolvidas no problema descrito. Foi estabelecido como objetivo para o trabalho investigar os dados disponíveis em banco de dados sobre os empreendimentos já realizados, para identificar possíveis correlações entre atributos que caracterizam o empreendimento, tais como: tipo de empreendimento, empresas prestadoras de serviços, tipos de contratos, advogados responsáveis pelo empreendimento, medições e o valor do empreendimento, determinado, fundamentalmente, pelo fator K obtido. Isso estabelecido, foi iniciado o processo KDD,

considerando tanto a seqüência de etapas proposta por Fayyad et al [FAY96a] como a descrição usada na abordagem de Brachman et al [BRA96a].

4.2- Preparação dos Dados

Os dados utilizados no processo foram extraídos de um DW, oriundo de um projeto, também patrocinado pela Sudecap, que se iniciou no final de 1999. A disponibilização desse DW para os usuários, devido a problemas técnicos e administrativos internos à Sudecap, deu-se no momento da definição do processo KDD.

As fontes de dados para preencher o DW eram basicamente de três sistemas legados implementados em plataformas diferentes. O Sistema de Controle Orçamentário - SF11, desenvolvido para plataforma *mainframe* usando o SGBD ADABAS [SAG02a], apresenta dados bastante confiáveis. O Sistema SIMEC, responsável pelo registro de medições dos contratos, desenvolvido em Clipper, tem boa qualidade de dados. O último sistema, o INFORMA, desenvolvido internamente pela Diretoria de Planejamento usando o ACCESS, tenta suprir uma lacuna no gerenciamento de empreendimentos. Esse sistema, como acontece nos casos de sistemas que não tiveram um projeto inicial bem definido, foi expandido na medida das necessidades, tendo que sofrer várias mudanças estruturais no decorrer do processo KDD. Por isso, ele utiliza mecanismos de projeto não adequados, principalmente na organização dos arquivos, o que acarreta necessidade de manutenção constante.

É fundamental, nesse ambiente, a integração das informações entre os sistemas, que são necessárias para compor dimensões que interligam assuntos do DW. Esses sistemas, na Sudecap, não têm nenhuma integração e são gerenciados por pessoas de áreas diferentes. A integração das informações é mantida pelo usuário do sistema INFORMA. Esse usuário, na prática, é encarregado da tarefa de manter as ligações entre os contratos, medições e empreendimentos.

O DW é composto por um conjunto de 27 Tabelas de Fatos e 90 Tabelas de Dimensões. O primeiro trabalho realizado foi o estudo de todas tabelas para verificar quais poderiam conter informações que relacionassem com o problema colocado. Como destacado por Fayyad et al [FAY96a], o processo é sempre interativo e iterativo, o que foi confirmado na experiência, já que os dados (dimensões e fatos) a serem analisados não foram todos identificados no início do processo. Eles foram alterados e acrescidos à medida que o entendimento do domínio dos dados dos sistemas foi aumentando.

Esse trabalho foi lento, pois as consultas realizadas com a ferramenta OLAP levaram a resultados inesperados, desviando o trabalho dos seus objetivos iniciais e concentrando no entendimento dos sistemas legados. Na verdade, conforme mostrado em Han et al [HAN01a], a tarefa de preparação de dados para a atividade de *data mining*, quando se usa um ambiente de DW, fica concentrada, principalmente, na sua preparação para o DW.

O trabalho dessa etapa resultou da relação de tabelas de fatos e dimensões mostradas no Quadro 4.2 (Os diagramas com os relacionamentos entre as Dimensões e os Fatos estão no Anexo A). A seqüência do trabalho deu-se com a definição dessas tabelas no ambiente DW do SQL server, de onde o DBminer seleciona os dados para processar. Nesse momento, apareceu a necessidade de transformar alguns tipos de dados devido às diferenças apresentadas entre as definições do ambiente ORACLE e o SQL server.

Algumas dimensões não foram tratadas nas análises, no entanto foram mantidas para ter a compatibilidade com as informações do DW e para confirmar os resultados mostrados pela função OLAP do DBminer, confrontado-os com os resultados mostrados pela ferramenta que já existia para consultar o DW. Durante o processo de análise, algumas tabelas foram desmembradas, gerando planilhas EXCEL e tabelas no ACCESS para serem usadas pelo SPSS [SPS02a].

Quadro 4.2- Relação das Tabelas de Fatos Usadas na Avaliação

Tabela de Fatos	Descrição	Dimensões	Medidas
Contratos	Contém informações relativas aos contratos	Contrato	Valor
		Empresa contratada	Prazo
		Advogado responsável	Fator K
		Situação contrato	Quantidade de contratos
		Data contrato	
		Data fim vigência	
		Data início vigência	
		Regional	
		Finalidade	
Aditivos contratos	Contém ocorrências referentes aos aditivos.	Aditivo contrato	Valor aditivado
		Contrato	Prazo contrato
		Empresa contratada	
		Advogado responsável	
		Data aditivo	
Medições	Contém informações quantitativas referentes à percentuais e indicadores da evolução da obra ou serviço.	Contrato	Quantidade medida
		Empreendimento	Valor medido
		Itens de medição	Fator K
		Medição	Quantidade de contratos
		Planilha	
		Regional	
		Subemprendimento	
		Supervisor	
		Data atualização	
		Data fim período	
		Data GLM	
		Data início período	
Planilhas de obra	Contém informações do conjunto de atividades listadas na seqüência lógica de sua ocorrência, determinando recursos e prazos para sua execução.	Tipo de obra	Prazo da planilha
		Empreendimento	Valor da planilha
		Situação planilha	
		Subemprendimento	
		Data atualização	
		Data início	

4.3- Seleção da Ferramenta de Análise

A definição do algoritmo ou ferramenta a serem usados no processo KDD faz parte das atividades apresentadas por Fayyad et al [FAY96a] e Brachman et al [BRA96a]. A escolha adequada é um dos fatores críticos de sucesso do empreendimento. Quando ela ocorre sem critérios relevantes, pode trazer ao projeto

conseqüências do tipo: aumento do número de pessoas envolvidas no processo, tempo para obter resultados, resultados incorretos, custos desnecessários com equipamentos.

O número de ferramentas disponíveis no mercado tem aumentado devido ao crescente número de pesquisas e estudos sobre técnicas de mineração de dados, e, também, ao aumento da demanda do mercado por tecnologias de tratamento de informações gerenciais. Já existem centenas delas disponíveis – algumas saindo dos laboratórios, outras já com um bom nível de maturidade.

A avaliação que deve ser feita não é uma simples tarefa de escolher a melhor ferramenta para atender todos os casos possíveis. Devem-se considerar as particularidades de cada empresa, observando principalmente as características e necessidades de seu negócio e os tipos de dados a serem tratados.

No caso em questão, a ferramenta escolhida foi o DBminer. A decisão deu-se, sobretudo, por dois motivos. O primeiro foi o fator custo e disponibilidade. O DBminer poderia ser usado para testes e por um período limitado sem custos. Isso foi importante pois a Sudecap não tinha orçado valores necessários para sua aquisição. Além disso, ele estava disponível para *download* da *internet*. O segundo motivo – e o principal deles – é que essa ferramenta estava em consonância com os principais critérios, encontrados em Han01 et al [HAN01a], Berson et al [BER99a] e Goebel et al [GOE99a], a serem observados na seleção de uma ferramenta:

- **Maturidade do produto**

  O DBminer já se apresenta como um produto comercial e com freqüência encontram-se referências, na *internet* ou em artigos, relatos sobre sua utilização em casos de testes.
- **Credibilidade dos fornecedores**

  O DBminer é resultado de anos de pesquisa em universidade (Simon Fraser University, Canadá) que tem profundo interesse pelo assunto KDD.
- **Compatibilidade com Sistemas Operacionais**

O DBminer trabalha com diversos sistemas, dentre eles, o NT, que é o utilizado pela Sudecap. Essa compatibilidade facilita a organização do ambiente e não provoca aumento de custos com *hardware* ou *software.*

- **Comunicação com SGBD e DW**

  Os dados a serem tratados estão todos gerenciados pelo SGBD ORACLE e organizados em um DW. O DBminer trabalha com o SQL server. Estes ambientes têm fácil integração.

- **Tipos de dados usados na aplicação**

  O DBminer tem algoritmos específicos para tratar dados multidimensionais, que é o tipo de dados a ser tratado nessa aplicação.

- **Fonte dos dados**

  A comunicação com os SGBD´s é feita via OLE DB, que facilita o trabalho de acesso aos dados.

- **Manipulação dos dados**

  O DBminer trata com facilidade dados do tipo hora, data, variáveis contínuas, dados incompletos, etc.

- **Visualização dos dados**

  O DBminer permite visualizar os resultados produzidos de diversas formas, o que agiliza a interpretação dos resultados.

- **Escalabilidade**

  Não há limites para quantidade de linhas ou colunas do banco de dados no processo de análise.

- **Interface com o Usuário**

  O DBminer possui uma interface que orienta o usuário em todas tarefas a serem realizadas.

- **Interface com DMQL**

  Todos procedimentos a serem executados podem ser vistos por meio de uma DMQL. Isso facilita o entendimento pelo usuário da relação existente entre o conhecimento extraído e o Banco de Dados.

- **Quantidade de tarefas disponíveis**
  O DBminer tem implementado as tarefas mais populares: classificação, associação *clusterização* e descrição dos dados, além de uma forte integração com o ambiente OLAP.

A maioria das características anteriores é comum entre diversas ferramentas. O tratamento de dados mutidimensionais e a interface com DMQL são itens que mais distinguem o Dbminer.

Os recursos de Software e Hardware usados no projeto estão relacionados no Anexo B.

4.4- Mineração dos Dados (Análise dos Dados)

Não havia, por parte da Sudecap, nenhum fato concreto ou indício de manipulação do fator K que servisse como norte para orientar o início das análises dos dados. Nem após a execução das tarefas de entendimento do domínio e preparação dos dados isso foi evidenciado. Assim, não foi estabelecida nenhuma hipótese inicial para conduzir as análises. Foi definido que elas teriam um caráter exploratório, objetivando a identificação de alguma situação que pudesse se relacionar com o objetivo do trabalho.

Inicialmente, decidiu-se pela identificação de segmentos significativos dos dados usando algoritmos de classificação, de *clusterização* e técnicas de exploração de dados, com o objetivo de encontrar atributos ou valores de atributos que pudessem direcionar mais as análises e, também, ser considerados nas análises das regras de associação. Alguns resultados encontrados foram validados com recursos do SPSS.

Definidas quais técnicas seriam utilizadas, isto é, quais tarefas de mineração seriam usadas, foi estabelecida uma ordem inicial para as atividades a serem executadas, que estão relacionadas abaixo:

1- Aplicar técnicas de prospecção geral de dados;
2- Levantar potenciais atributos para classificação;
3- Aplicar as regras de classificação;
4- Aplicar as regras de associação;
5- Fazer validação dos resultados encontrados.

Feito esse planejamento, começaram as análises dos dados do DW, e, a princípio, o que parecia ser somente a execução das tarefas, envolveu iterações entre elas. Isso provocou a criação de um ciclo, que se repetiu algumas vezes no decorrer do processo, como será mostrado mais adiante.

As primeiras análises produziram resultados que, avaliados pelos usuários, não mostravam exatamente o comportamento dos dados. Verificando os dados, percebeu-se que algumas tabelas do DW tinham linhas com informações de agregações. Foi, então, necessário criar uma outra estrutura de tabelas eliminando essas agregações.

Após a eliminação das agregações, a primeira análise foi realizada na tabela de contratos. A intenção era observar como era a distribuição dos valores do fator K. Foram identificados, conforme Gráfico 4.4.1, um grande número de contratos com valores do fator K igual a 1 (um). Essa informação causou surpresa à equipe de gerência de obras. Levada ao conhecimento do especialista em elaborar contratos, foi explicado que contratos que não se tratavam de obras, isto é, aqueles referentes a projetos e consultorias, tinham fator k menor ou igual a 1 (um), pois a composição do preço do contrato só tinha o custo direto.

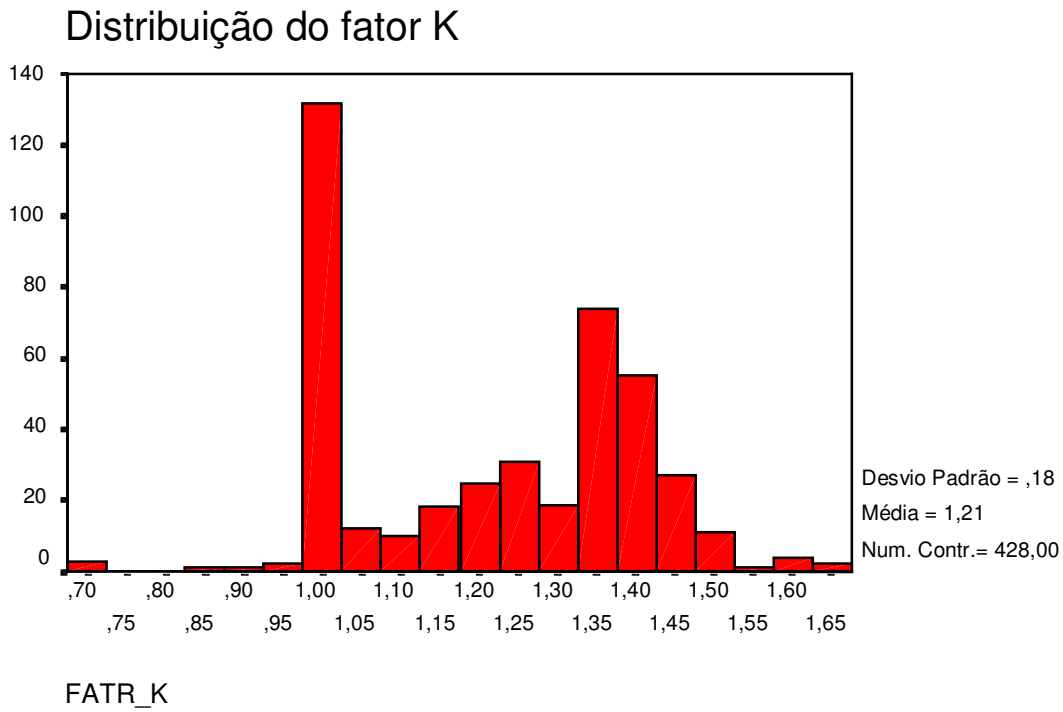

Gráfico 4.4.1- Distribuição do Fator K nos Contratos

A própria conclusão citada e a aparência da distribuição já sugerem uma necessidade de fracionar os dados em classes para tipos de contrato, o que, de fato, será feito mais adiante.

A prospecção, ainda feita na tabela de contratos, mostrada no Gráfico[1] 4.4.2, identifica bem o comportamento dos empreendimentos no decorrer do tempo. O número de empreendimentos concluídos e em andamento está coerente com a época em que eles foram contratados. O resultado de uma outra análise mostra que a quase totalidade desses empreendimentos é de contratos com pessoas jurídicas. Os tipos de empreendimentos apresentados mostram a existência de números consideráveis de empreendimentos oriundos de propostas elaboradas pelo programa de orçamento participativo.

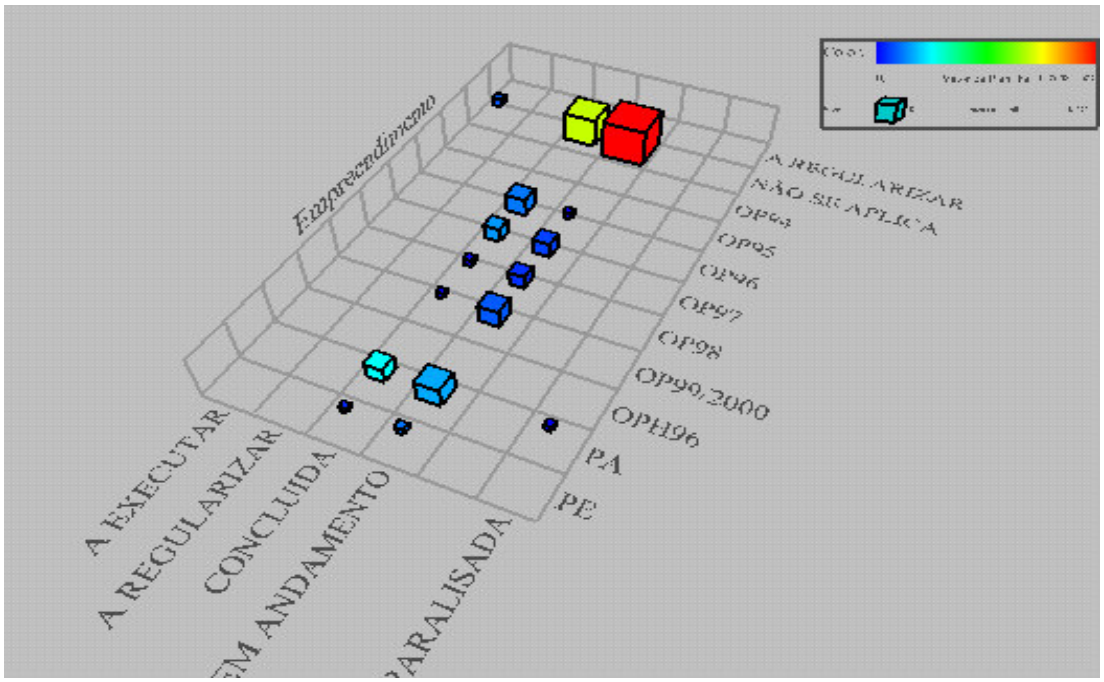

Gráfico 4.4.2- Relação Empreendimento por Situação

[1] A legenda no canto superior esquerdo é útil para visualização *online.* Ela mostra os valores mínimos e máximos para as medidas envolvidas e as dimensões de cada cubo. Essa legenda tem o mesmo significado para os próximos gráficos.

O Gráfico 4.4.3, produzido a partir da tabela de aditivos, trouxe informações relevantes sobre o comportamento desses aditivos. A maioria deles foi feita para estender o prazo inicial de contratação. Essa informação estava coerente com a crença da equipe que define as políticas de preços, pois de acordo com essa equipe, o valor do contrato não era o que mais provocava aditivos. Consideram-se aditivos do tipo prazo aqueles que alteram o prazo inicial para execução do contrato; aditivos do tipo valor aqueles que alteram o valor do preço inicial do contrato.

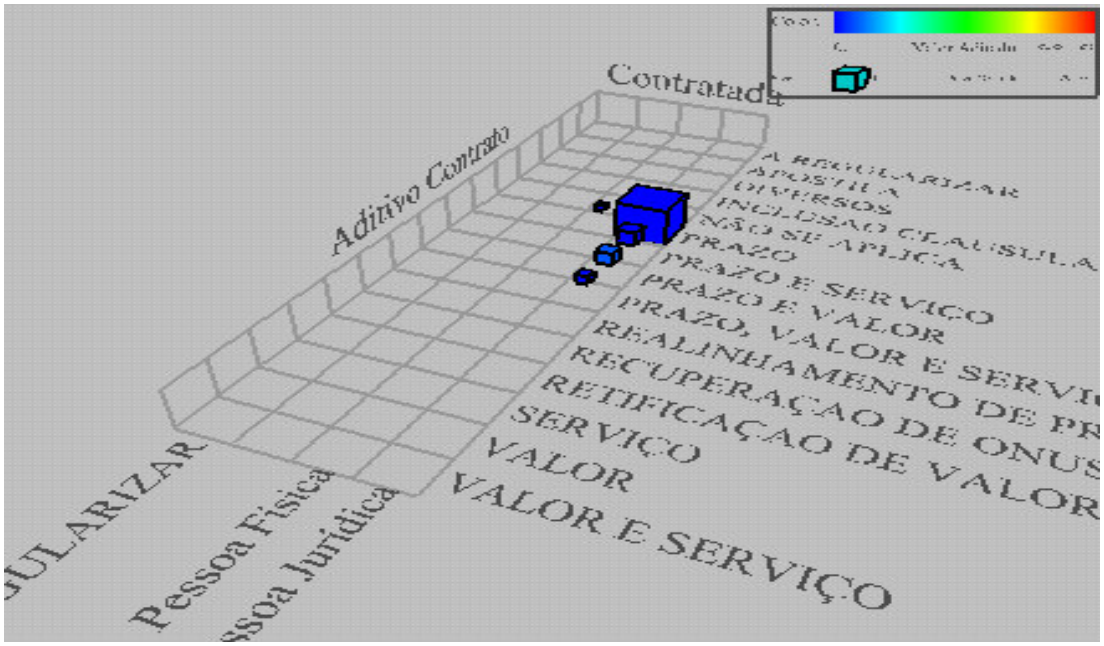

Gráfico 4.4.3- Relação de Tipos Aditivos por tipo de Contratada

O Gráfico 4.4.4, mostrado a seguir, feito a partir da tabela de medições, tem um comportamento similar ao Gráfico 4.4.2, com as medições distribuídas em projetos do orçamento participativo. Mais adiante, foi feita, ainda nessa tabela, uma análise para os empreendimentos concluídos e, da mesma forma, a aplicação de regras de associação objetivando a busca de correlações.

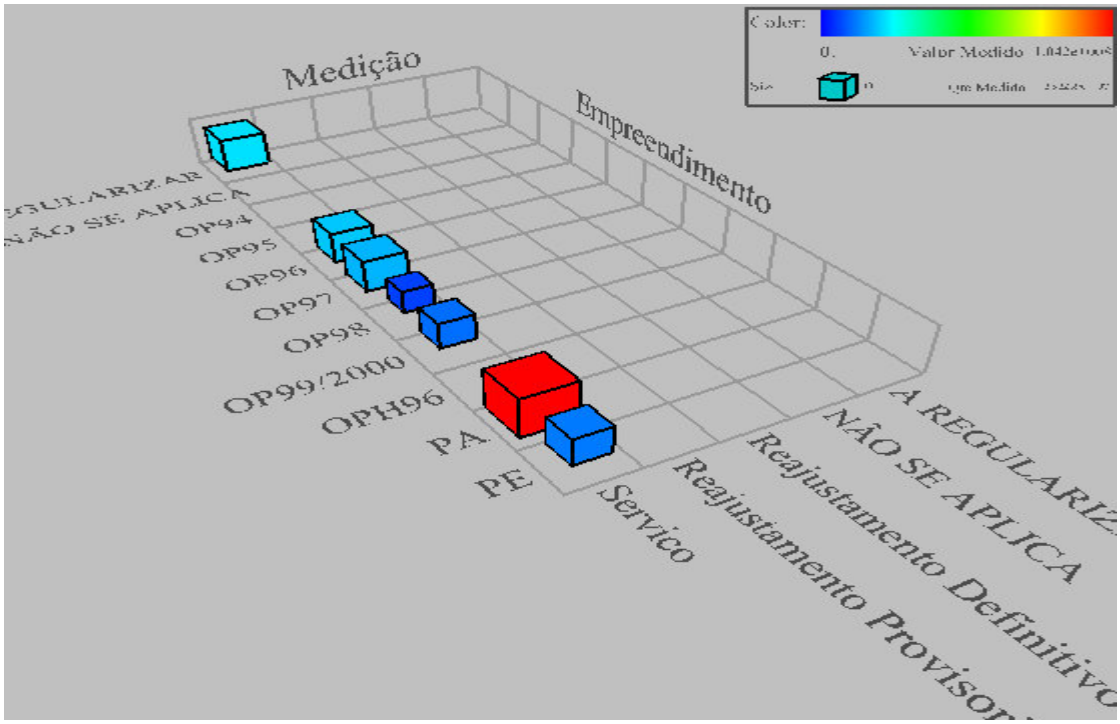

Gráfico 4.4.4- Relação de Tipos de Medição por Tipo de Empreendimento

A análise feita sobre a tabela de planilhas produziu o Gráfico a seguir, que mostra a existência de um número relativo de contratos de consultoria, confirmando as informações extraídas do Gráfico 4.4.1, que indicava a existência de certa quantidade desses contratos (fator k menor ou igual a 1).

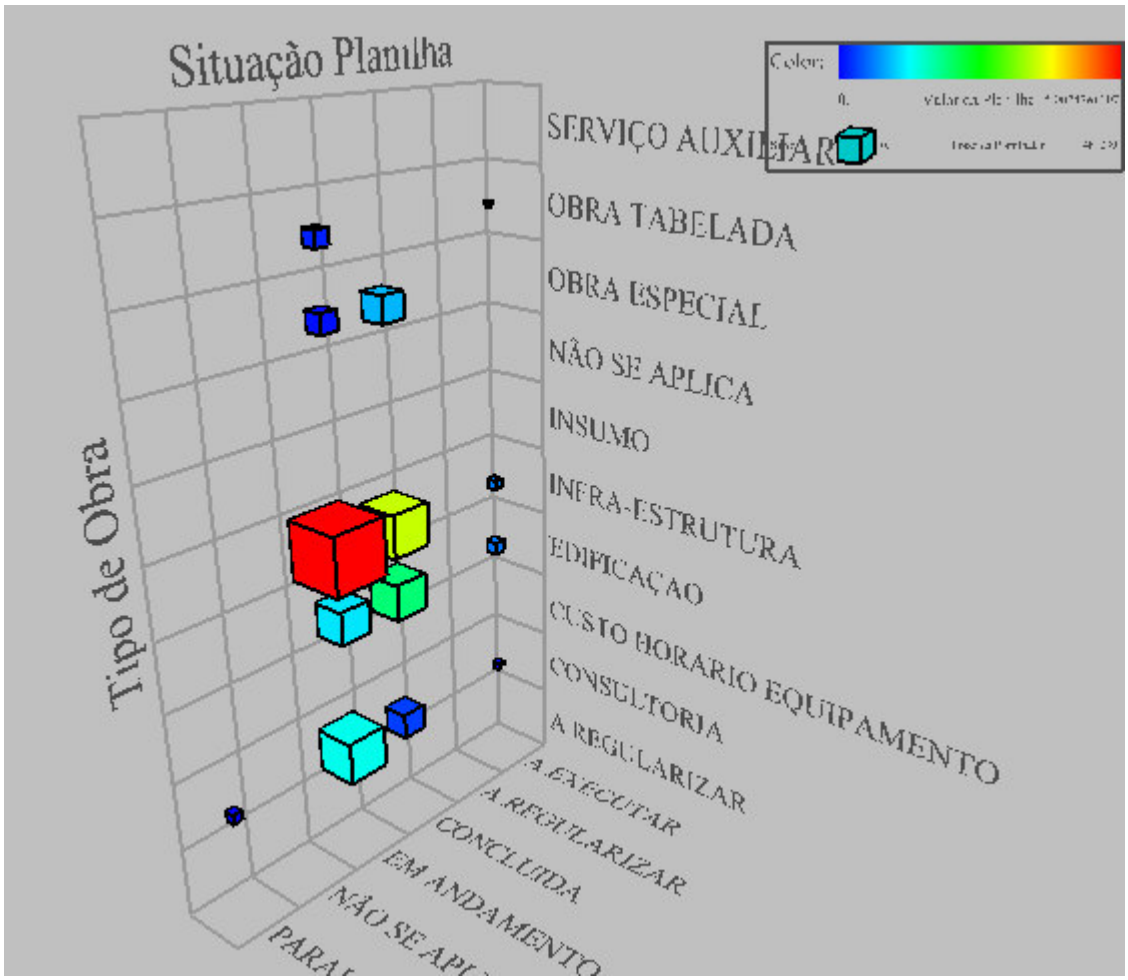

Gráfico 4.4.5- Relação Situação da Planilha por Tipo de Obra

Os resultados anteriores propiciaram um melhor entendimento dos dados e ajudaram, igualmente, na escolha de quais atributos seriam mais relevantes para a continuidade das análises e, em especial, para a aplicação das regras de classificação. Essas regras seriam utilizadas para um melhor detalhamento de alguns relacionamentos existentes entre as tabelas de fatos e, também, para identificar conjuntos de dados para aplicar as regras de associação.

Um dos experimentos feitos com o algoritmo de regras de classificação foi a produção de uma árvore de decisão, a partir da tabela de contratos, que mostra a situação dos contratos de acordo com a data de fim de vigência do mesmo. O algoritmo produziu quatro padrões que representam a realidade em nível da situação do contrato por data de vencimento, como a seguir:

Quadro 4.4.1- Regras de Indução Usando a Variável por Fim de Vigência

IF Data Fim Vigência = 1998, **THEN Situação Contrato IS OR CONCLUIDO (81.82%) OR EM ANDAMENTO (18.18%) .**
IF Data Fim Vigência = 1999, **THEN Situação Contrato IS OR CONCLUIDO (85.82%) OR EM ANDAMENTO (14.18%) .**
IF Data Fim Vigência = 2000, **THEN Situação Contrato IS OR CONCLUIDO (15.00%) OR EM ANDAMENTO (84.55%) OR PARALISADO (0.45%) .**
IF Data Fim Vigência = 2001, **THEN Situação Contrato IS OR CONCLUIDO (2.78%) OR EM ANDAMENTO (97.22%) .**

Outro experimento realizado com regra de classificação foi a análise da tabela de aditivos com objetivos de criar classes de tipos de aditivos. As regras produzidas assemelharam-se às informações mostradas no Gráfico 4.4.2 e, também, as confirmaram.

Os experimentos feitos nas tabelas de medições e planilhas não trouxeram novidades relacionadas ao objetivo do trabalho, além das mostradas anteriormente.

Os resultados produzidos até então serviram para definir os seguintes padrões de comportamento dos dados, julgados interessantes pelos usuários:

- O comportamento do fator K : fator K maior que 1 (um) é para contratos de obras ou similares; fator K menor ou igual a 1 (um) é para contratos de projetos e consultoria.
- Comportamento dos aditivos : a maioria dos aditivos feitas nos contratos – cerca de 97% – é de prazos.

Com base nos dois padrões anteriores, foram criados alguns conjuntos de dados para dar prosseguimento às análises. Esses conjuntos foram produzidos da seguinte forma:

- A tabela de contratos foi dividida em duas. Uma para os contratos com valores de fator K maior que 1(um), referente aos contratos de obras e similares; a outra para os contratos com valores de fator K menores ou iguais a 1 (um), referentes aos contratos de consultoria.
- As tabelas de medições e planilhas foram divididas seguindo o mesmo critério adotado na tabela de contrato.
- A tabela de aditivos também foi dividida em duas, observando o mesmo critério adotado para a tabela de contratos. Posteriormente, ela foi redividida observando, também, o critério do tipo de aditivos. Foi criada uma para os aditivos de prazo e outra para os de valor.

Esses conjuntos de dados foram criados, de forma temporária, no DW, para serem preparados para o Dbminer; no ACCESS e em planilhas EXCEL para serem tratados pelo SPSS.

O prosseguimento das análises deu-se com o uso da regras de associação para tentar identificar alguma situação interessante. As análises aconteceram nos dois conjuntos de dados produzidos para a tabela de contratos, mas para ambos os resultados foram semelhantes. Na avaliação, foram identificadas associações de informações da área jurídica, e, após avaliações por parte do usuário, verificou-se que elas foram produzidas porque os dados da área jurídica estavam incompletos. O que mais chamou atenção foram as regras produzidas, mostradas no Quadro 4.4.2, envolvendo data de assinatura de contrato em dezembro, regras, essas, com valores de suporte e confiança acima do esperado pelo usuário, que não a explicou com clareza. A justificativa mais plausível foi que os contratos sempre eram assinados em final de ano fiscal.

Quadro 4.4.2- Regras Associação na Tabela de Contrato

descrição	implica	situação	suporte	Conf.
Situação Contrato = [CONCLUIDO]	**==>**	**Advogado Responsável = [A REGULARIZAR] AND Data Contrato = [December]**	**26,397**	**63,134**
Data Contrato = [December]	**==>**	**Advogado Responsável = [A REGULARIZAR] AND Situação Contrato = [EM ANDAMENTO]**	**26,012**	**49,27**

A avaliação de regras de associação na tabela de aditivos, mostrada no Quadro 4.4.3, produziu uma regra com data do aditivo em agosto. Essa regra teve valor de confiança muito alto. Não houve, no entanto, uma explicação, por parte do usuário, para esse resultado. Por isso, ela não foi considerada no esclarecimento do problema proposto. Os resultados foram semelhantes para todos os conjuntos de dados.

Quadro 4.4.3- Regra Associação na Tabela de Aditivos

descrição	implica	situação	suporte	Conf.
Data Aditivo = [August]	**==>**	**Advogado Responsável = [A REGULARIZAR]**	**15,909**	**100**

As regras produzidas a partir da tabela de planilha não produziram resultados que ajudassem a elucidar os objetivos do trabalho.

Ao aplicar os algoritmos de regras de associação na tabela de medições, só se encontrou a regra mostrada no Quadro 4.4.4, que não ajudou no esclarecimento do objetivo do trabalho. Ela foi considerada por aparecer a informação de data de fim de medição para o mês de fevereiro, e foi analisada, mas os usuários não conseguiram identificar fatos que a justificassem. Os resultados foram semelhantes para todos os conjuntos de dados.

Quadro 4.4.4- Regra Associação na Tabela de Medições

descrição	implica	situação	suporte	Conf.
Data Início Período = [February]	**==>**	**Data Fim Período = [February] AND Empreendimento = [A REGULARIZAR] AND Itens de Medições = [ACERTO DE VALORES ACUMULADOS] AND Medição = [32] AND Sub Empreendimento = [NÃO SE APLICA]**	**13,588**	**99,399**

Uma vez encontrados os resultados anteriores, teve início o processo de validação dos mesmos. Para isso, foi utilizado o SPSS para verificar a possibilidade de existência de correlações entre algumas variáveis.

A primeira verificação, por sugestão de um dos usuários, foi descrever o comportamento do fator K para os empreendimentos já concluídos. A média do fator K para esses empreendimentos ficou um pouco acima da média encontrada para todos os empreendimentos. A comparação dos empreendimentos já concluídos com outros do mesmo grupo, também concluídos, mas que receberam aditivos de valor, mostrou que a média do fator k passou de 1,26 para 1,29. Para os 10 (dez) empreendimentos concluídos que receberam os 10 (dez) maiores aditivos, a média do fator K aumentou de 1,29 para 1,39. As obras concluídas que não receberam aditivos tinham média do fator K igual a 1,22. Esses valores, na opinião do usuário, indicavam uma possível correlação entre o fator K e a produção de aditivos. Para confirmar a análise anterior, foram feitos os seguintes testes usando regressão linear[1]

[1] Regressão linear é uma técnica destinada a estimar o relacionamento entre duas variáveis. A equação, gerada por meio do método dos mínimos quadrados, pode ser usada para predição de valores de uma variável dependente face aos movimentos da outra variável, chamada independente.

1) Verificar a correlação da variável valor do contrato e valor do fator K para os contratos com valores de fator K maior que 1 (um)

O valor[1] do r^2 (0,042), representado no Quadro 4.4.6, indica que o fator k explica pouco o valor do contrato, apesar da curva estimada de correlação, ao lado, ter apresentado , para algumas faixas, uma possível correlação entre as variáveis.

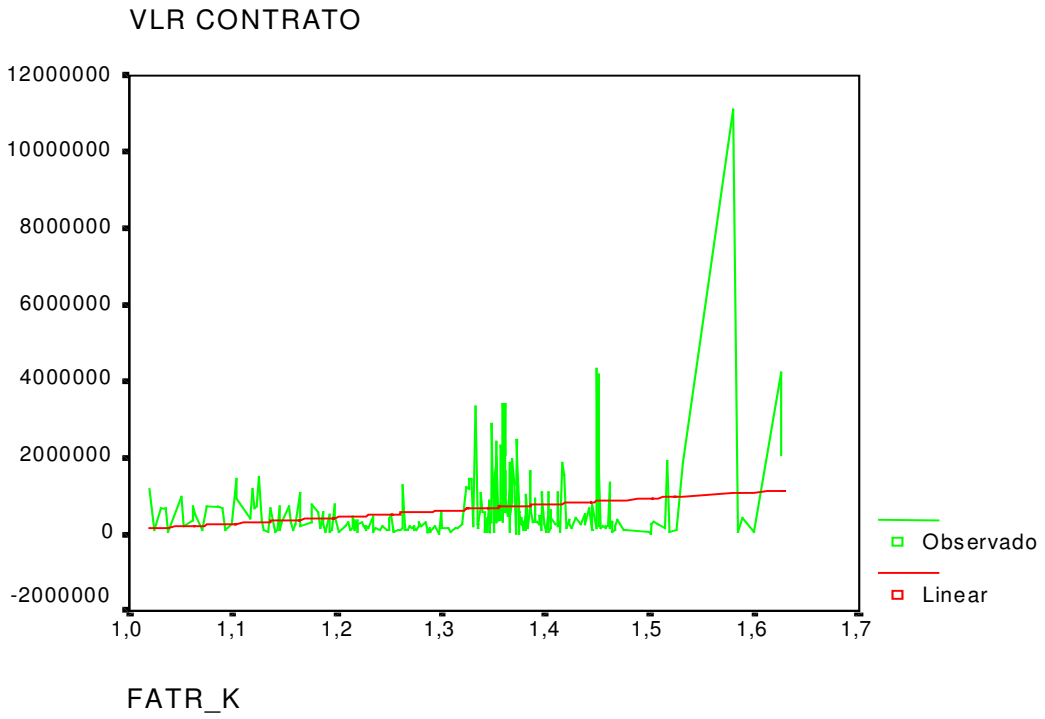

Gráfico 4.4.6- Correlação na Tabela de Contratos

[1] r^2 também chamado de coeficiente de determinação, aponta a qualidade do modelo preditivo.

Com o intuito de entender melhor o comportamento dos dados de cada faixa da curva anterior, foram criados, usando as variáveis valor do contrato e fator K, três *clusters* para o conjunto de dados usado no item anterior. Para cada um dos *clusters* foi analisada correlação entre as mesmas variáveis. Todas as correlações tiveram resultados semelhantes ao indicado no Gráfico a seguir, produzidos a partir de um dos *clusters*, onde o r^2 (0,025) ficou ainda menor do que a correlação para o conjunto com todos os dados dados.

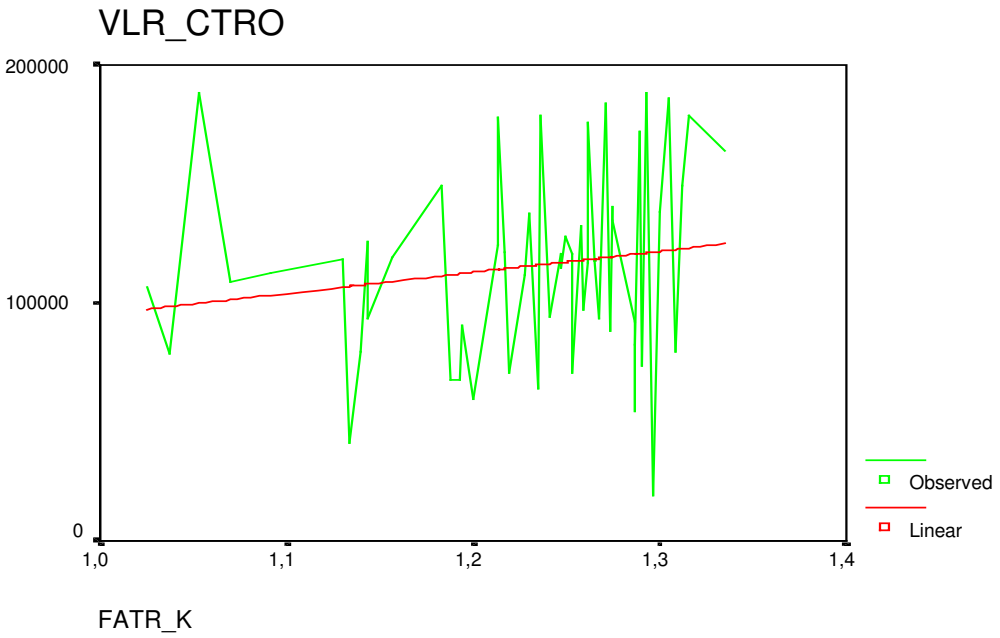

Gráfico 4.4.7- Correlação entre a Variável Fator K e o Valor do Contrato

2) Verificar a correlação da variável prazo do contrato e valor do fator K para os contratos com valores de fator K maior que 1 (um).

O resultado, apresentado no Gráfico abaixo, não indicou correlação entre as variáveis do item 2. O r^2 encontrado foi de 0,032, confirmando a baixa correlação.

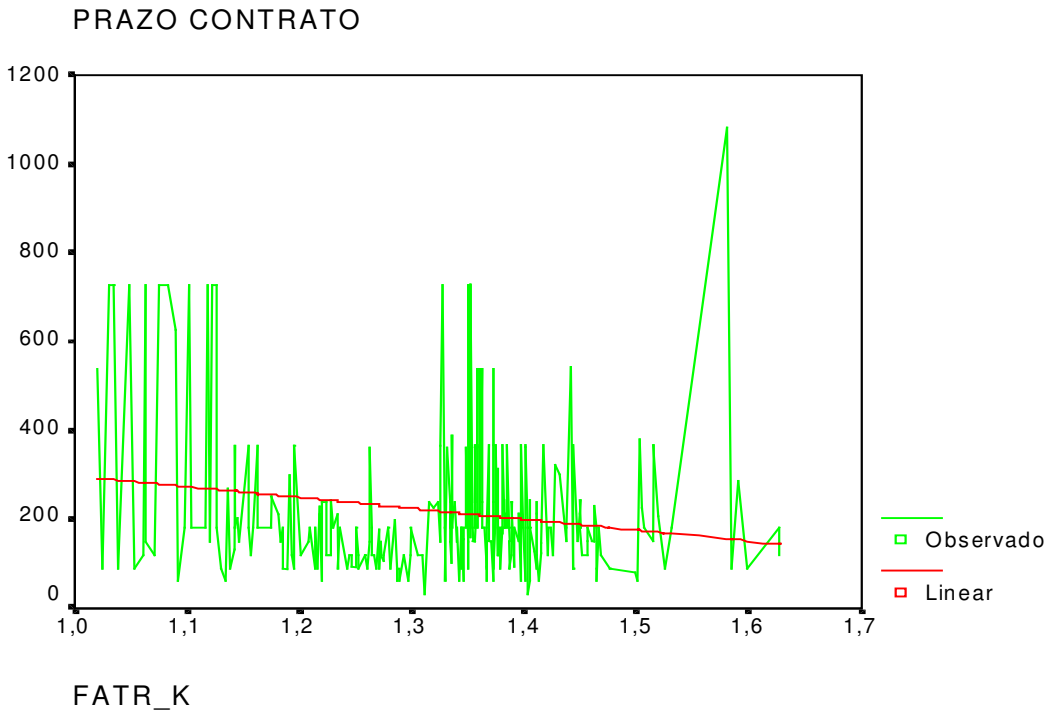

Gráfico 4.4.8- Correlação entre a Variável Fator K maior que um e Prazo do Contrato

Da mesma forma que no item 1, foram criados, para os contratos com Fator K maior que 1 (um), três *clusters* baseados nas variáveis prazo do contrato e valor do fator K. No entanto, em nenhum deles a correlação trouxe valores do r^2 muito diferente de 0,032, confirmando a não-correlação entre o fator k e o prazo do contrato.

3) Analisar a correlação existente entre a variável valor do contrato e valor do fator K para os contratos com valores de fator K menor ou igual a 1 (um).

Também, conforme explica o R2 (0,001) e o Gráfico abaixo, a correlação não existiu. Ao mudar a variável valor do contrato para prazo do contrato, o r^2 praticamente não teve alteração. Vistos esses resultados, esse conjunto de dados não foi segmentado.

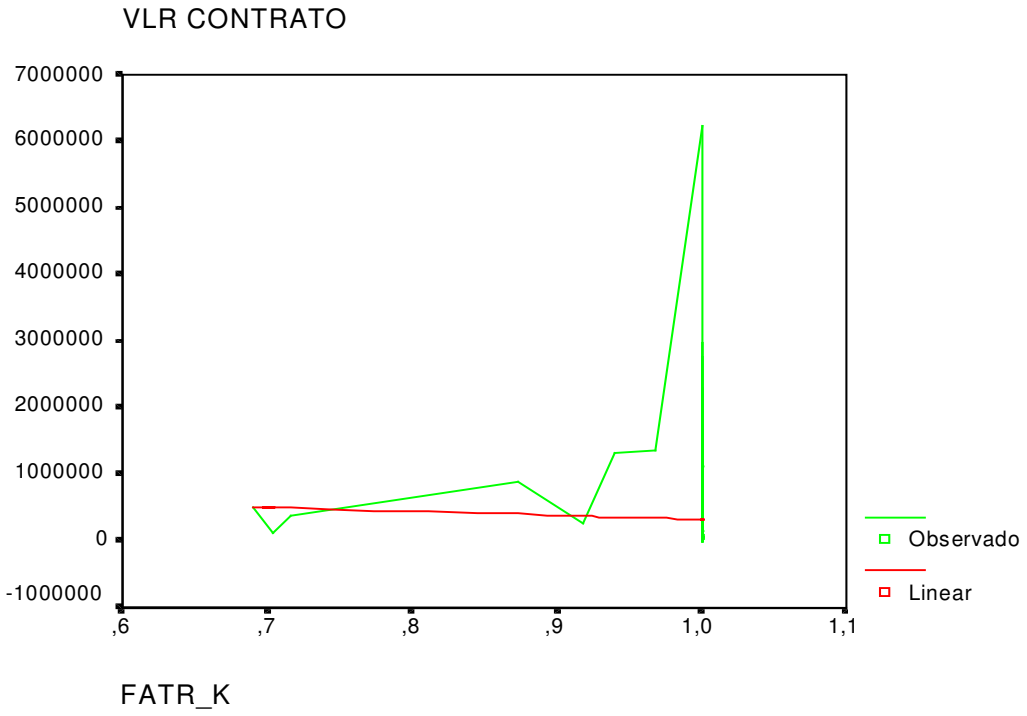

Gráfico 4.4.9- Correlação entre a Variável Fator K menor que um e Valor do Contrato

4) **Analisar, na tabela de aditivos, a correlação existente entre a variável valor do aditivo e valor do fator K para os contratos com valores de fator K maior que 1 (um).**

Esse experimento teve, como resultado o Gráfico abaixo, o r^2 0,079 indicando baixa correlação entre as variáveis. Esses dados não foram segmentados pois os resultados não indicaram o relacionamento entre as variáveis.

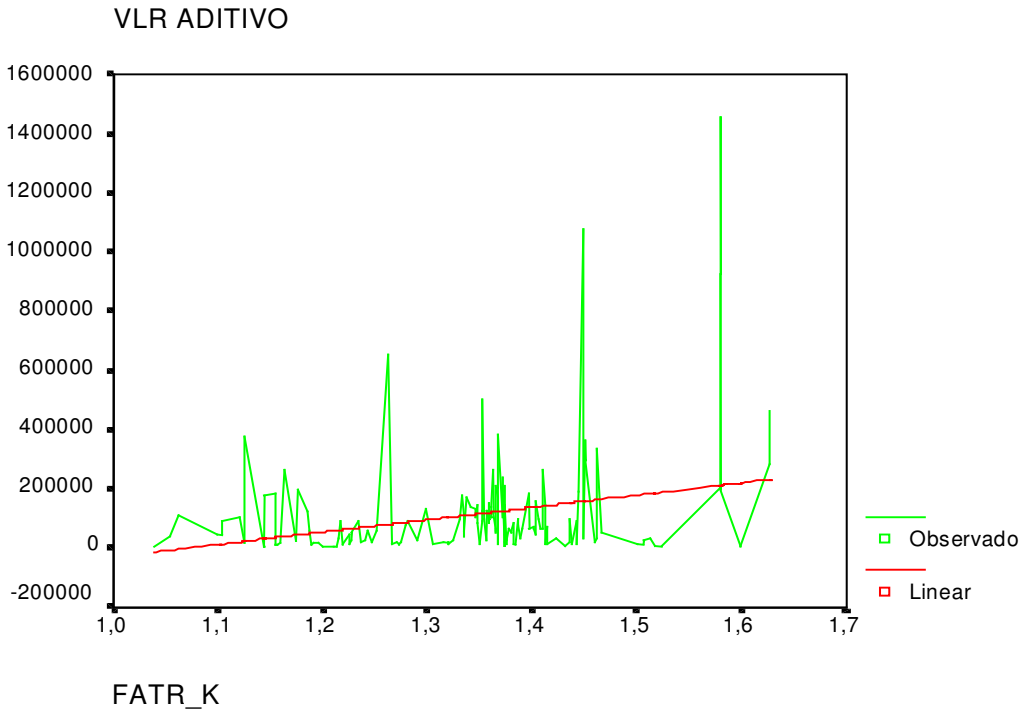

Gráfico 4.4.10- Correlação entre a Variável fator K maior que um e Valor Aditivo

5) Analisar, na tabela de aditivos, a correlação existente entre a variável valor do fator K e a valor do aditivo para os contratos com valores de fator K menor ou igual a 1 (um).

Esse experimento teve, como resultado o Gráfico abaixo, com r^2 0,003 indicando baixa correlação entre as variáveis. Esses dados não foram segmentados pois os resultados não indicaram o relacionamento entre as variáveis.

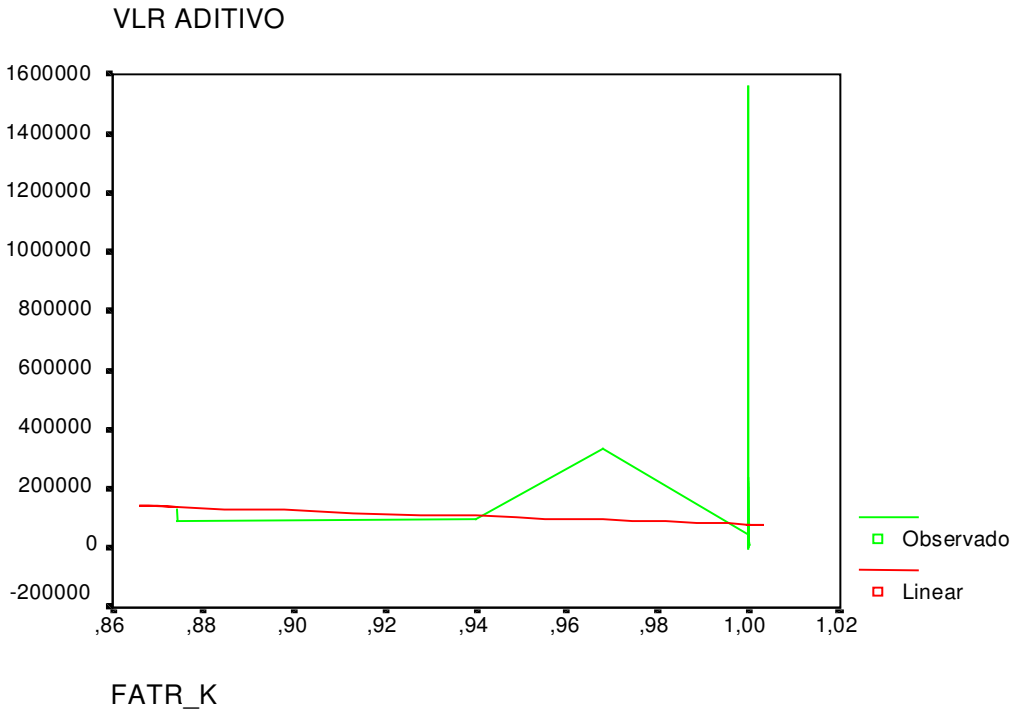

Gráfico 4.4.11- Correlação entre a Variável Fator K menor que um e Valor Aditivo

6) **Analisar, na tabela de aditivos, a correlação existente entre a variável prazo do aditivo e valor do fator K para os contratos com valores de fator K maior que 1 (um).**

Esse experimento teve, como resultado o Gráfico abaixo, o r^2 0,068 indicando baixa correlação. Também, esses dados não foram segmentados pois os resultados não indicaram o relacionamento entre as variáveis.

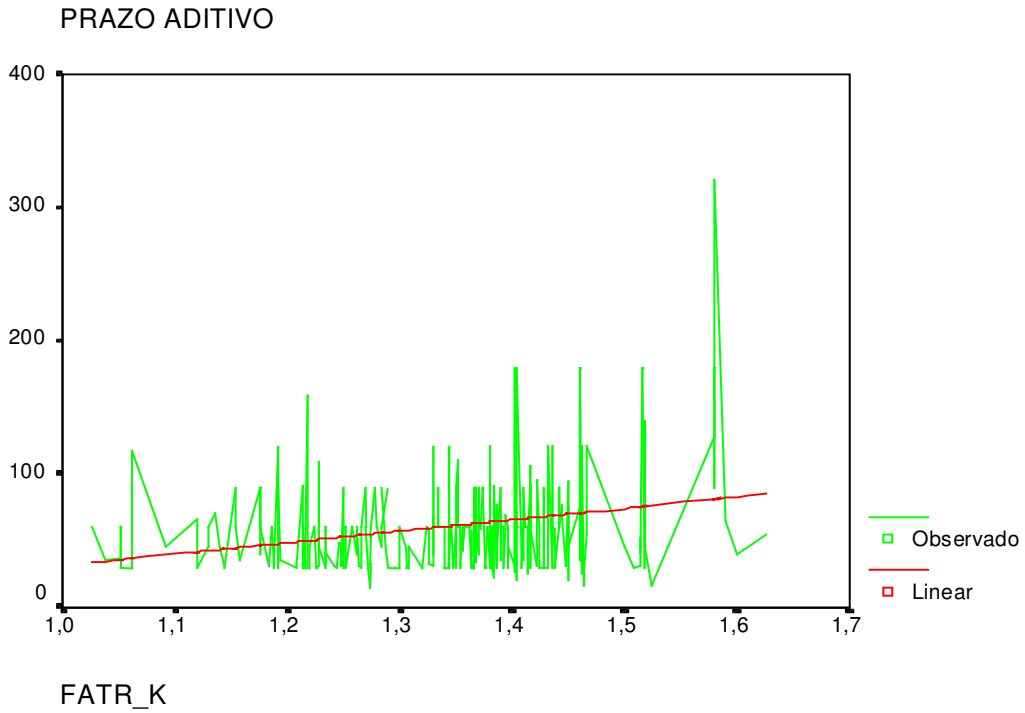

Gráfico 4.4.12- Correlação entre a Variável Fator K maior que um e Prazo Aditivo

7) **Analisar, na tabela de aditivos, a correlação existente entre a variável prazo do aditivo e valor do fator K para os contratos com valores de fator K menor ou igual a 1 (um).**

Esse experimento teve, como resultado o Gráfico abaixo, o r^2 próximo de zero indicando baixa correlação. Esses dados não foram segmentados pois os resultados não indicaram o relacionamento entre as variáveis.

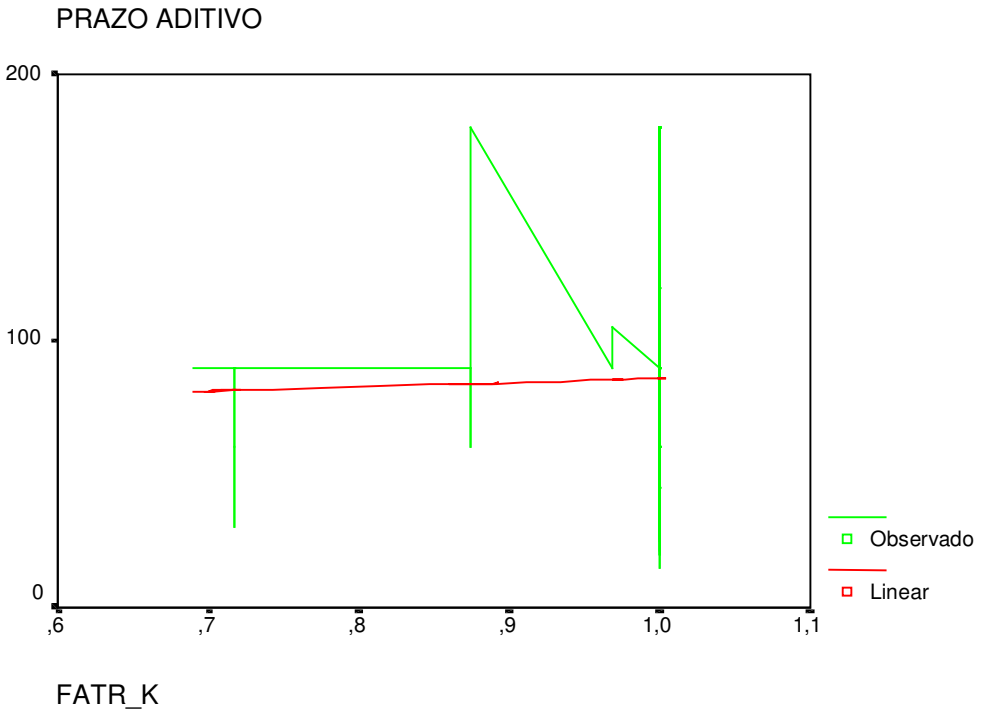

Gráfico 4.4.13- Correlação entre a variável Fator K menor que um e Prazo Aditivo

Outras duas análises foram feitas, ambas na tabela de contratos. A primeira, para verificar a correlação entre as variáveis valor do contrato e valor do aditivo. Essa análise apresentou o r^2 igual a 0,47, mostrando uma possível correlação entre as variáveis. O usuário explicou que contratos com valores maiores teoricamente são mais difíceis de serem planejados, provocando assim aditivos no futuro. A segunda análise, foi para verificar a correlação entre valor do contrato e prazo de aditivo. O resultado dessa análise apresentou uma correlação muito pequena, r^2 igual a 0,057.

Observação geral: O r^2, nesse caso, indica ausência de correlação na relações lineares. Estatisticamente falando poderiam existir outras correlações (não-lineares, logarítmicas, etc).

Quando começaram as análises, existiam, somente, dúvidas por parte de algumas pessoas, com relação a uma possível influência do fator K no valor final do preço dos contratos. Existia, também, a crença, por parte da equipe de controle de política de preços, de que o fator K era uma forma justa de controlar a remuneração das novas atividades do contrato e que ele não era usado para manipular os preços.

Com as análises feitas, usando os dados selecionados do DW, não se chegou a um resultado conclusivo sobre as questões anteriores. Houve indícios, principalmente após a aplicação dos métodos estatísticos, de que o valor do fator K não tinha correlação com a produção de aditivos, nem com os de prazos nem com os de valores. Essas análises também não apresentaram correlação entre o fator K e os valores dos contratos. Isso, segundo a equipe de gerência de preços, parece lógico se se considerar que o prestador de serviço, na composição do preço do empreendimento, não tem muita margem para compô-lo. Essa composição não pode ter variação acima de 20% sobre o valor estipulado pela Sudecap de cada atividade e nem ser maior do que o valor total do preço do empreendimento. Além do mais, o fator K só terá uso se houver aditivos. Só haverá aditivos se a obra for mal planejada ou se houver uma situação especial, como emergência ou questão de segurança. Não existe, então, *a priori*, a previsão de que mudança no contrato poderia ocorrer no futuro, dificultando, assim, a manipulação pelo prestador de serviço. Se existisse essa previsão, ela seria planejada no início.

Podemos considerar que do ponto de vista do problema levantado – a influência do fator K no valor final dos preços dos contratos – as análises, usando os dados selecionados do DW e o método de regressão linear para verificar a correlação entre algumas variáveis, não identificaram nenhuma influência. Mas é prudente que as análises sejam aprofundadas, tratando e complementando os dados com um novo

escopo, e usando, também, outros tipos de correlações. Talvez, assim, os resultados ficassem mais definitivos.

4.5- Análise do Processo KDD da Sudecap

Esta seção tem como objetivo apresentar considerações a respeito da aplicação do processo KDD relatado nas seções anteriores. São feitas análises sobre a execução de cada etapa dessa experiência.

4.5.1- Compreensão do Domínio

A definição clara dos objetivos do processo é importante para o desencadeamento dos trabalhos de entendimento e preparação dos dados. Um problema mal definido poderá ou prolongar o tempo do projeto com a preparação de novos dados ou produzir resultados pouco interessantes.

Nessa experiência, a expectativa era que a documentação do processo de construção do DW fosse fundamental e essencial para a realização dessa etapa. Isso foi verificado, mas não foi o suficiente. Não havia documentação adequada de quais eram as fontes de dados, quais os procedimentos e periodicidade das cargas no DW, o objetivo de cada dimensão, etc. Para entender o problema, foi necessário interagir com pessoas que usavam com freqüência as informações do DW. O processo mostrou-se muito dependente dos conhecimentos informais que essas pessoas possuem. Essa dependência agravou o andamento do projeto, pois a disponibilidade delas era pequena, visto que ele ocorreu em paralelo a uma reforma administrativa na Prefeitura. Foi importante o envolvimento de pessoas com uma visão gerencial do problema e de outras com uma visão mais operacional dos dados. A participação da equipe de tecnologia de informação, que era responsável pelos sistemas informatizados e pela administração do DW, foi, também, fundamental.

Os resultados dessa etapa surgiram após várias consultas ao DW e diversas entrevistas. Essa característica interativa está mais próxima da visão de Brachman et al

[BRA96a] do que da de Fayyad et al [FAY96a]. Na visão de Brachman et al, a hipótese inicial é apenas um ponto de partida; é por meio da interação com o banco de dados e com as pessoas que se chega ao real objetivo.

Á medida que os resultados das análises começaram a aparecer, o entendimento do problema foi ampliando, o que indicou que o processo é incremental no aspecto de aquisição de conhecimento.

O fato de os dados estarem no DW não foi determinante. A complexidade da estrutura organizacional, a grande quantidade de pessoas envolvidas e a diversidade de assuntos no DW foram os fatores que definiram o seu andamento.

4.5.2- Preparação dos Dados

O objetivo dessa etapa é gerar um conjunto de dados a serem utilizados nas etapas seguintes. De acordo com a abordagem de Fayyad et al [FAY96a], ela segue a compreensão do domínio, devendo gerar um conjunto de dados que seja suficiente para a solução do problema. No entanto, nessa experiência, tomou a forma de processo de aquisição de conhecimento, já que, por várias vezes, o conjunto de dados, formado, no caso, por tabelas de dimensões e fatos, revelou-se insuficiente e foi alterado durante o projeto.

Além de tratar o caráter incremental do conjunto de dados, essa etapa serviu, também, a procedimentos de validação que identificava a aderência entre o conjunto de dados e o objetivo do processo.

A etapa de preparação de dados seguiu a seqüência de atividades resumidas na Figura 5.2 e descritas como se segue:

- Localização dos dados

  O primeiro passo nessa etapa foi identificar as fontes de informação que seriam usadas. A princípio, a interação seria somente com o DW, mas, com a percepção de

que algumas informações do DW estavam incompletas, foram identificadas todas as fontes de dados para serem usadas para a carga do DW.

- Exploração inicial

  Foi realizada uma série de consultas ao DW, com a ferramenta OLAP, com o intuito de aprender mais sobre ele. Durante o processo, foram extraídos alguns arquivos do DW para serem examinados por uma ferramenta com funções estatísticas.

- Preparação básica

  Essa atividade teve como objetivo criar alguns conjuntos de dados, a partir das informações obtidas com a atividade de exploração inicial. Esses conjuntos de dados foram utilizados como *input* para o SPSS.

- Verificação de relacionamentos

  As tabelas de fatos incluíam relacionamentos com algumas dimensões que não tinham nenhuma relação com o problema em questão. Nessa atividade, esses relacionamentos foram excluídos.

- Documentação

  Foi feita uma descrição não-estruturada das ocorrências das atividades dessa etapa.

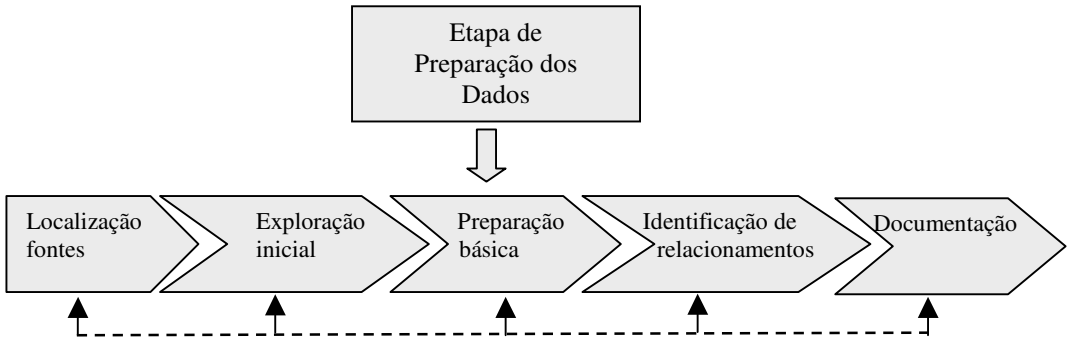

Figura 4.5.2- Fluxo da Etapa de Preparação dos Dados

Nessa experiência específica, os itens abaixo foram determinantes na execução da etapa:

- Tratamento de valores faltosos

  As fontes dos dados para o DW são sistemas legados desenvolvidos em plataformas distintas e com baixo nível de integração. Na carga do DW, foi identificado um grande número de dados que não possuíam valores. Eles foram preenchidos de

acordo com o método que usa uma constante para o preenchimento. Com isso, nas análises, surgiram resultados representativos para valores como 'não se aplica', 'pendente'.

- Falta de informação para algumas tabelas

  Algumas informações, principalmente as referentes à área jurídica, não foram carregadas no DW. Os sistemas de informação estavam sendo alterados para tratá-las , quando, então, seriam carregadas.

- Descontinuidade do processo de carga no DW

  O início da utilização do DW na Sudecap aconteceu junto com a reforma administrativa na Prefeitura. Isso trouxe diversas modificações para o projeto e, entre elas, o desligamento do responsável pela administração do DW e conseqüentemente a paralisação dos processos de carga. Os objetivos em relação ao DW foram alterados: decidiu-se priorizar a substituição dos sistemas legados com o desenvolvimento de novos sistemas. Essa decisão provocou um *gap* entre os objetivos que levaram à construção do DW e à sua efetivação. Além de dificultar o processo, ela colaborou com a pouca expressividade de alguns resultados.

- Ampliação dos conjuntos de dados

  O processo teve um caráter incremental. A cada análise percebia-se a necessidade de ampliação dos dados. O conjunto proposto inicialmente pelo administrador do DW foi alterado e ampliado diversas vezes.

- Formatação de dados em diferentes tipos de arquivos

  As tabelas de fatos quando foram segmentadas tiveram que ser reformatadas em planilhas EXCEL ou ACCESS para serem acessados pelo SPSS. O uso do SPSS tinha como objetivo aumentar a efetividade dos resultados.

- Ampliação dos objetivos

  O processo de seleção foi um processo de aquisição de conhecimentos o que motivou, diversas vezes, a possibilidade de alterar os objetivos do trabalho.

Conforme apontado nos capítulos anteriores, a maior parte dos esforços humanos no processo é concentrada na formulação adequada do problema e na preparação dos dados e não na aplicação dos algoritmos de mineração em particular. A experiência da Sudecap confirmou essa observação, contrariando a expectativa inicial de que o uso do DW facilitaria essas etapas. Ela ocorre durante todo o caminhar do processo, sendo os dados alterados à medida que os resultados vão surgindo. Essa etapa demonstrou ser a mais crítica do processo, já que a qualidade de informação na entrada determina a qualidade dos resultados obtidos com o processo. No caso da Sudecap, podem-se enumerar algumas razões para essa criticidade:

- Dependência excessiva de pessoas com conhecimento dos sistemas de informação e do DW. Como não havia uma documentação adequada, sem elas o trabalho ficaria comprometido;
- O tempo gasto com as análises e a preparação dos dados;
- Utilização de ferramentas auxiliares para manipulação das fontes dos dados.

Duas questões merecem ter uma análise ampliada em outros trabalhos. A primeira: será que estabelecer um objetivo para o processo KDD e restringir o uso dos dados a esse objetivo, como mostra a abordagem de Fayyad et al [FAY96a], não vai contra as práticas de DM? A pessoa que está trabalhando no processo, no seu início, não tem ainda, pleno conhecimento de quais dados serão necessários e nem quais relacionamentos irão surgir entre eles. A segunda questão: o que fazer com os dados selecionados após serem usados nas análises? Esses dados poderão ser úteis em novos projetos. Ainda são poucos os relatos de trabalhos abordando esse assunto.

4.5.3- Mineração dos Dados (Análise dos Dados)

Essa tarefa, segundo a abordagem de Fayyad et al [FAY96a], é composta pelas atividades de seleção das tarefas a serem executadas, a escolha do algoritmo e as

análises propriamente ditas. O seu encaminhamento deve ser ligado ao objetivo do projeto. As etapas anteriores – compreensão do domínio e preparação de dados – são executadas com o objetivo bem definido de construir um conjunto de dados adequado a um algoritmo para a tarefa específica. Por conseguinte, a definição das tarefas a serem executadas não é feita na etapa de mineração de dados, mas sim no início do processo. Nesse caso, é mais apropriado que a escolha do algoritmo se realize após a seleção dos dados, pois , nesse momento, já se conhecem todos os tipos de dados a serem tratados.

Essa seqüência está em conformidade com aquela mencionada na abordagem de Brachman et al [BRA96a], que coloca a seleção das tarefas a serem executadas como ponto inicial do processo, e, também, com a da abordagem de Fayyad et al [Fay96a], que, mesmo não colocando tal seleção no início do fluxo, considera-a condição básica para a execução das tarefas antecessoras. Todavia, nessa experiência de caráter exploratório, que poderia levantar diversas hipóteses para serem confirmadas ou refutadas, esse aspecto – a definição a *priori* de uma tarefa especifica – não foi tão relevante, pois os seus resultados poderiam não ser suficientes para um julgamento da hipótese. Com isso, a atividade de preparação dos dados não foi centrada na tarefa a ser executada, mas sim na relação que eles tinham com o objetivo a ser alcançado. Considerando isso, em vez de um algoritmo específico para uma tarefa, escolheu-se uma ferramenta que integrava diversas tarefas com seus respectivos algoritmos preparados para os dados do domínio da aplicação.

A aplicação dos algoritmos, como era de se esperar, foi, em linhas gerais, fácil. Percebeu-se que a grande dificuldade seria organizar, para cada análise, o motivo da parametrização; quais entradas e/ou variáveis consideradas e sob que circunstâncias e estratégias eles foram alcançados. Outra dificuldade foi como proceder para não utilizar muito tempo analisando regras com resultados similares, já que processo de análise gerou um grande número delas. Essas dificuldades foram agravadas ao se considerar o número de interações e o conjunto distinto de ferramentas usadas durante o processo.

4.5.4- Interpretação/Avaliação dos Resultados

A etapa de mineração, em si, não apresenta dificuldades. Os dados, se bem preparados, são facilmente tratados pelos algoritmos escolhidos. A escolha do algoritmo requer uma interferência humana. Essa interferência, embora pequena, tem grande importância, pois cada algoritmo tem parametrizações a serem feitas e, se elas não forem adequadas, as análises podem resultar em resultados pouco interessantes [CHE00a].

Diversas regras foram produzidas no processo de análise. A maioria delas não foi considerada pelas seguintes razões:

- A regra não correspondia a um conhecimento anterior ou não atendia a uma expectativa;
- A regra se referia a um conjunto de atributos ou combinação de atributos que não tinham significado para o objetivo do projeto;
- As regras eram similares ou redundantes.

A interpretação e a avaliação são etapas críticas no processo, sendo influenciadas pela forma como o algoritmo apresenta os resultados[1]. Essa forma de apresentação dos resultados deve ser flexível, de modo a permitir ao usuário escolher a melhor visualização (gráficos, textos) e a melhor maneira de interagir com os resultados.

O conhecimento do domínio dos dados pelo usuário e a forma de vizualização dos resultados são fatores que, juntos, podem propiciar maior independência do usuário em relação ao analista – que conhece bem o funcionamento do algoritmo – no entendimento dos resultados.

---

[1]A importância de diferentes formas de visualização de resultados podem ser encontradas em [GRA02a] [FOO02a].

O processo de avaliação e interpretação depende do conhecimento do domínio dos dados e de resultados de análises feitas em outros momentos. Por isso é importante que esse processo seja auxiliado por um mecanismo que possibilite interpretação dos resultados com base em resultados de outras análises. No caso em questão, a extração dos resultados com o DBminer foi simples. A dificuldade foi maior na sua comparação com resultados extraídos anteriormente. Quando foi usado o SPSS, a avaliação foi mais difícil, visto que o processo de visualização não é tão flexível.

A forma de visualização dos resultados tem grande importância nessa fase. Para isso é importante integrar os conceitos da visualização[1] com os de DM. Nessa experiência, ao usar o DBminer, percebeu-se essa integração nos seguintes pontos:

- Visualização dos dados
  Os dados do DW puderam ser vistos em diferentes níveis de granularidade e abstração, ou como combinações de diferentes atributos ou dimensões.
- Visualização dos resultados da mineração
  Os resultados das análises foram vistos em diferentes formas (gráficos, textos).
- Alteração visual interativa
  A visualização pode ser usada no processo para ajudar o usuário a tomar decisões. Os resultados, em forma de gráficos ou textos, podem ser manuseados de diversas maneiras para facilitar o entendimento e a tomada de decisão.

4.5.5- Incorporação dos Resultados

A divulgação dos resultados restringiu-se às pessoas envolvidas no processo. Ainda não havia cultura de divulgação de novos conhecimentos no âmbito geral da empresa.

---

[1]Conceitos de visualização podem ser encontrados em [GRI02b] [HIN02a] [HOF02a] .

Essa etapa é a que efetivamente traz à organização os resultados obtidos com o projeto. Verificou-se, por meio dessa experiência, que a divulgação, mesmo sendo restrita a grupo pequeno de pessoas, é facilitada se usar ferramentas especializadas para comunicação.

4.5.6- Conclusões sobre a Aplicação do Processo KDD

Este capítulo estava comprometido com o objetivo principal do trabalho que era a elaboração de uma análise do processo de KDD, abordando as suas características, bem como os fatores e dificuldades envolvidos em sua aplicação, principalmente os métodos de trabalho e os ambientes que lhe dão apoio. Eram dois seus objetivos. O primeiro apresentar uma análise dos resultados obtidos em um caso real de utilização do processo KDD. Esse estudo é relevante por duas razões principais. A primeira: verificar como as fases do processo descrito por Fayyad et al [FAY96a] e Brachman et al [BRA96a] se comportam diante uma aplicação do mundo real. O segundo: verificar como as etapas do processo se comportam ao usar dados de um DW. O segundo objetivo é verificar o comportamento de uma ferramenta integrada de DM em um caso real. Essa seção apresenta uma análise dos resultados encontrados em relação aos objetivos relacionados.

Sobre o comportamento das etapas do processo

Percebeu-se que a seqüência em que elas foram realizadas teve algumas variações em relação à apresentada por Fayyad et al [FAY96a] e Brachaman et al [BRA96a], duas delas com significativas diferenças. A primeira referente à etapa de preparação dos dados. As abordagens estudadas no capítulo 2 mostraram que ela era um pouco dependente do tipo de tarefa de mineração a ser realizada. Nesse processo, não houve essa dependência. A preparação dos dados foi encaminhada observando a relação que os dados tinham com o objetivo do projeto. Houve uma seleção inicial que foi ampliada durante o trabalho. A segunda diferença é que não houve uma definição prévia da tarefa de mineração a ser executada. A definição fez parte de um processo interativo, entre o usuário e o DBminer, à medida que os resultados eram produzidos. Na prática, pelas facilidades do DBminer, no momento da execução da etapa de mineração dos

dados, já eram, também, realizadas as etapas de seleção da tarefa, a escolha do algoritmo e a avaliação dos resultados.

A proposta de Brachman et al [BRA96a], com relação às interações entre as pessoas, mostrou-se mais concreta do que aquela do fluxo apresentado por Fayyad et al [FAY96a].

Na Figura 4.5.6, estão representadas as principais etapas realizadas no processo e, também, os principais níveis de iteração ocorridos entre elas. Esse ciclo foi evidenciado nessa experiência, tendo como razões principais as características dos dados do DW e a utilização do DBminer, pelas suas facilidades de interação com o usuário.

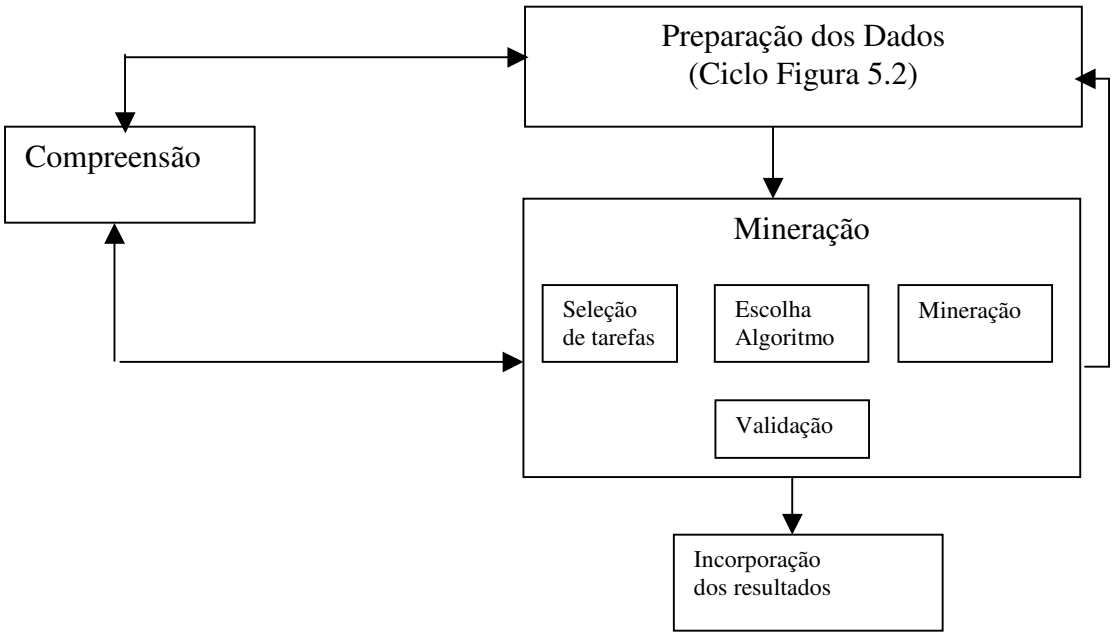

Figura 4.5.6- Ciclos de Atividades Realizados no Processo

Sobre o comportamento das atividades do processo ao utilizar um ambiente de DW

percebeu-se que há uma integração entre os objetivos de um *data warehousing* e o KDD. Eles possuem atividades comuns que podem colaborar entre si. No entanto,

KDD, ao objetivar a etapa de mineração de dados, pode requerer diferentes estruturas de dados, processos computacionais e, possivelmente, diferentes grupos de usuários. O objetivo das análises feitas no processo KDD difere, também, do objetivo das feitas no DW. No DW, a organização usa os dados selecionados e preparados para fazer análises objetivando a obtenção de informações que possam ajudar no processo de tomada de decisões. As análises sobre os dados do DW aumentam à medida que a tecnologia e a cultura de tratamento de informações gerenciais vão se consolidando na empresa. De acordo com Han et al [HAN01a], inicialmente o DW é usado para a geração de relatórios e para responder a consultas pré-definidas. Progressivamente, ele é usado para fazer análises em dados sumarizados, que são apresentados em gráficos e relatórios. Mais adiante, pode ser usado para análises multidimensionais mais sofisticadas, empregando recursos como as operações *slice-and-dice*. Por fim, o DW pode ser usado para mineração de dados. Os resultados obtidos com as funcionalidades das ferramentas OLAP, porém, são diferentes dos conseguidos com as técnicas de mineração de dados. OLAP permite a sumarização e agregação de dados, as quais ajudam a simplificar as análises desses dados, enquanto a mineração de dados permite a descoberta de padrões e conhecimentos interessantes em grandes volumes de dados.

No início deste trabalho, havia a expectativa de que, por usar os dados de um DW, algumas etapas do processo KDD seriam realizadas com mais facilidade. Isso não se concretizou como foi imaginado. Em cada etapa do processo, constatou-se o seguinte:

- Entendimento do domínio

  O fato de usar os dados do DW não correspondeu à expectativa, talvez pelo fato de o DW ainda ser incipiente na Sudecap e pelas equipes usuárias não estarem integradas.

- Preparação dos dados

  A alta qualidade dos dados do DW efetivamente colabora com a preparação dos dados para a mineração. Mesmo assim, nessa experiência foi necessário tratar algumas dimensões, pois nem todas eram relacionadas ao objetivo. Houve a

necessidade de uma fase de seleção. Foi necessário eliminar as informações sobre agregações que existiam em algumas tabelas do DW. As estruturas de tabelas do DW atenderam os requisitos do Dbminer, porém, para o uso do SPSS foi necessário preparar os dados em uma outra estrutura, no caso foi usado planilhas EXCEL. As facilidades de ferramenta OLAP foram fundamentais para a análise exploratória, que foi feita com o intuito de entender os dados.

- Mineração dos dados e avaliação dos resultados
  A utilização da ferramenta OLAP contribuiu para algumas pesquisas feitas para verificar e entender os resultados do DM.

Sobre o comportamento do processo ao utilizar uma ferramenta integrada de mineração

percebeu-se que ela foi fundamental para agilizar o processo. O uso de uma ferramenta integrada promove ao usuário um conhecimento maior sobre os dados preparados. Na etapa de mineração ficou evidenciado o caráter interdisciplinar do processo KDD, principalmente, ao usar as facilidades de visualização, onde se percebeu uma forte integração com a disciplina comunicação; e ao usar o SPSS, para interpretar e descrever alguns dados, verificou-se a integração do processo KDD com o campo de pesquisa de estatística . Na seção 4.7

Sobre a complexidade do processo KDD, de forma geral, os principais problemas encontrados nas etapas do processo estão resumidos no Quadro a seguir.

Quadro 4.5.7- Resumo dos Problemas Ocorridos Durante o Projeto

ETAPA	PROBLEMA	SOLUÇÃO
Compreensão do domínio	Documentação inadequada do ambiente DW	Entrevistas com pessoas envolvidas e consultas exploratórias no DW
	Grande número de pessoas envolvidas com os assuntos do DW	Entrevistas com pessoas envolvidas e apoio da equipe de tecnologia de informação
	Reforma administrativa na prefeitura	Prolongamento do tempo de execução da etapa
Preparação dos dados	Documentação inadequada do ambiente DW	Entrevistas com pessoas envolvidas e apoio da equipe de tecnologia de informação
	Falta de conhecimento das fontes de dados	Entrevistas com pessoas responsáveis pelos sistemas de informação
	Dados faltosos	Não consideração dos resultados das análises
	Paralisação da carga do DW	Definição de uma data até a qual os dados estavam íntegros
	Tendência a ampliar objetivos	Definição do escopo
Mineração dos dados	Sequenciamento das atividades	Utilização de uma seqüência inicial modificada durante o processo
	Documentação dos resultados	Feita usando recursos de documentação disponíveis na Sudecap
	Reutilização de resultados existentes pelas ferramentas	Não houve
	Tratamento de resultados semelhantes	Eliminação feita por meio de observações visuais
Interpretação / avaliação	Produção de grande número de resultados	Eliminação prévia de resultados supostamente pouco interessantes com base no conhecimento do domínio da aplicação
	Grande número de pessoas a serem consultadas sobre os resultados	Discussão com cada equipe separadamente
	Resultado pouco visual do SPSS	Inclusão de textos explicativos
Incorporação/divulgação	Não preocupação da empresa com esta atividade	Elaboração da atividade somente entre as pessoas envolvidas
Geral	Documentação dos resultados e experiências do processo	Utilização de recursos da Sudecap

Sobre os resultados da aplicação do processo KDD para esclarecimentos relativos ao Fator K

A expectativa, por parte da Sudecap, era que o processo KDD ajudasse a entender mais o comportamento do fator K e, por conseqüência, esclarecer as dúvidas

envolvidas no problema descrito no início do capítulo. Apesar dos problemas descritos nas seções anteriores, como paralisação das carga do DW e ausência de algumas informações, os resultados, segundo avaliação das pessoas envolvidas, foram satisfatórios, principalmente se considerar o nível de informação gerencial disponível na Sudecap. As análises feitas nos dados propiciaram, além de um entendimento maior sobre o comportamento do fator K, alguns critérios a mais para definir a influência desse fator no estabelecimento dos preços dos empreendimentos. Propiciaram, também, resultados que colaboraram com questões além das estabelecidas no contexto deste trabalho. As análises ajudaram, também, num conhecimento maior dos dados disponíveis no DW potencializando mais o seu uso na empresa. O sucesso da utilização da aplicação do processo KDD na Sudecap será maior se as análises continuarem após essa fase, definindo um novo escopo, resolvendo, de alguma forma as deficiências encontradas até então.

4.6- Ambiente de Apoio ao Processo KDD

Conforme apontado nas seções anteriores, a ausência de um ambiente para auxiliar na organização dos trabalhos e documentação das experiências foi sentida no projeto da Sudecap. Considerando isso, incluímos esta seção com o objetivo de apresentar sugestões, em um nível macro, para a construção de um ambiente que servisse de apoio nas atividades de um processo KDD.

Nesse trabalho, os registros de todos procedimentos e experiências importantes foram feitos de maneira pouca organizada, de forma tal que, ao longo do projeto, um grande volume de informação foi acumulando, tornando o seu gerenciamento complicado e dificultando possibilidade de reutilização. Essa complexidade foi ressaltada também por Brachman et al [BRA96a], Engels [ENG99a] e Bartlmae et al [BAR00a].

Percebeu-se que o KDD, no seu decorrer, envolve grupos de pessoas com conhecimentos, interesses e visões distintas, mas com um único objetivo. Para se chegar a esse objetivo – que é a criação de um novo conhecimento –, ao explorar as experiências das pessoas envolvidas, na forma como elas se procedem, no grande número interações entre elas no decorrer do projeto e no relato das iterações das etapas, nota-se toda uma filosofia de atividades cooperativas[3].

Pelo o observado nessa experiência, a execução do processo KDD pode ser prolongada e, dependendo da complexidade do domínio da aplicação, pode ser, também, complicada. Normalmente, as equipes que conduzem o processo têm um caráter temporal e se desfazem após o projeto ser completado, e é comum, conforme apontado por Bartlmae [BAR00a], que as experiências adquiridas por essas equipes não sejam documentadas e conseqüentemente não sejam reaproveitadas. Na experiência da Sudecap isso foi verificado: com a reforma administrativa ocorrida na prefeitura, várias pessoas que estavam envolvidas no projeto passaram a trabalhar em outras atividades. É importante que as pessoas que participam do processo tenham um método de trabalho, e que as experiências adquiridas não sejam exclusivas delas, mas que sejam organizadas, compartilhadas e mantidas na instituição para serem reutilizadas. A idéia de usar um ambiente para manter a história de todo o processo na organização deve ser considerada.

A abordagem apresentada por Brachman et al [BRA96a] ressalta a importância da criação de um ambiente de trabalho para auxílio à execução do processo. Ressalta, ainda, que nesse ambiente seriam documentados os passos executados no processo, com detalhes que possibilitassem futuras comprovações, bem como aqueles das regras descobertas para uso futuro. Segundo Chedini [CHE00a], esse ambiente de apoio deve ser construído para facilitar o acesso, o compartilhamento e a reutilização do conhecimento produzido no processo. Alguns trabalhos [BAR00a] [MOU99a] [ENG99a]

---

[3] As características de atividades cooperativas podem ser encontradas em [ARA97a] [ARA97b] [BOC95a].

[CHE00a] [LIN99a] já tratam desse assunto. Alguns deles adotam a abordagem de apoio ativo, em que os sistemas dão recomendações de como escolher um algoritmo ou a tarefa a ser executada. Outros poucos tratam da documentação do processo, mas não se preocupam com a possibilidade de reutilização das experiências e nem com a integração com ferramentas de mineração de dados. Todos eles apresentam avanços na questão de apoio ao processo, porém neles ainda se percebem limitações no que se refere ao caráter cooperativo das interações das pessoas; no tratamento de grande número de iterações e apoio à gerência da aplicação.

Essa experiência e outros trabalhos [BAR00a] [MOU99a] [ENG99a] [CHE00a] [LIN99a] sobre processo KDD serviram para entender pontos que devem ser observados na construção de um ambiente de apoio, entendendo, principalmente que

- O ponto de partida de uma aplicação é o entendimento do problema com vistas às necessidades do negócio e ao planejamento inicial das tarefas a serem executadas;
- Esse planejamento deve ser flexível em relação à ordem de execução das tarefas, à criação de novas tarefas e suas decomposições;
- Quanto maior for o número de pessoas com conhecimentos sobre o domínio da aplicação a participar do projeto, mais ricos serão os resultados produzidos.
- Os requerimentos iniciais não são completamente especificados;
- Existe a inserção de novas tecnologias durante o processo;
- Existem diversas informações que caracterizam cada tarefa, e elas devem ser organizadas;
- Os resultados encontrados devem ser padronizados e incluídos no sistema. Eles devem, também, ser úteis para o prosseguimento do projeto e para outras aplicações;
- Quanto maior for o número de pessoas com conhecimentos sobre o domínio da aplicação a participar do projeto, mais ricos serão os resultados produzidos.

Para propor uma estrutura, partimos das percepções e necessidades observadas durante o processo e do conceito de que uma aplicação KDD é composta por diversas tarefas distribuídas nas etapas que compõem o processo. O ponto de vista aqui

apresentado tem a intenção de ampliar o aspecto meramente organizacional e documentacional de experiências. Considera que a condução do processo tem características de trabalho cooperativo e que o ambiente de apoio deve incluir facilidades para isso e para a reutilização das experiências. Ele deve servir de suporte a todas as etapas do processo e ser construído de forma que as ferramentas que apóiam cada etapa do processo possam integrar com os seus objetos e usar as informações por ele organizadas. Ele deve, também, gerenciar as informações com um nível de abstração e detalhes suficientes para auxiliar cada etapa do processo, com perspectiva de atender um número diverso de aplicações. Para isso ele deve auxiliar o usuário

- Na definição dos objetivos da aplicação;
- No relacionar as hipóteses levantadas acerca do problema;
- No gerenciamento e controle das tarefas a serem executadas, incluindo informações sobre as decomposições feitas e os resultados obtidos nas execuções;
- Na documentação dos recursos utilizados para a execução das tarefas, como as ferramentas e os conjuntos de dados para os quais se deve ter a descrição dos atributos e tipos, os conceitos, objetos, as relações existentes entre os dados e as regras existentes entre eles;
- No gerenciamento dos resultados gerados no processo, entendendo que a idéia básica é que esses resultados devem ser utilizados para a solução de outro problema;
- No gerenciamento dos redirecionamentos do processo com justificativas para tal.

A construção desse ambiente, pensando no perfil cooperativo das atividades de documentação e de integração dos resultados do processo, poderia valer-se dos conceitos e resultados obtidos com pesquisas que envolvem a área de Computer Suported Cooperative Work – CSCW [ARA97a] [BOC95a] e incluir mecanismos que

- Promovam a estruturação dos pontos de entendimentos entre o grupo;
- Promovam a percepção do grupo em relação às atividades produzidas por outra pessoa, para que cada membro possa ter a noção do contexto onde está inserido o seu trabalho;

- Dêem suporte adequado à comunicação das pessoas, para que os participantes do processo possam trocar informações sobre os resultados encontrados;
- Dêem suporte para um constante gerenciamento e acompanhamento das atividades que estão sendo realizadas pelos participantes como um todo e daquelas realizadas individualmente por cada um deles.

A Figura 4.6 apresenta uma sugestão de uma possível estrutura conceitual para uma ferramenta de apoio ao ambiente mencionado anteriormente. O seu processamento seria um misto entre a operação manual feita pelo usuário e a gravação e a verificação de conhecimentos mais interessantes e similares feita de forma automática por agentes.

As características de implementação para facilitar o trabalho em equipe poderiam seguir as abordagens apresentadas por Araújo et al [ARA97a] [ARA97b].

Esse ambiente, no seu funcionamento para gerenciar a documentação das experiências, deve ser capaz de ler as informações fornecidas pelo usuário acerca de um resultado, analisá-las e armazená-las de forma estruturada, para serem acessadas em outro momento. Ele deve, também, ter funções básicas para a descrição do problema e a descrição da solução.

O processo de documentação começaria com a descrição das tarefas realizadas, das suas decomposições e das hipóteses levantadas inicialmente. Prosseguiria com a descrição das características dos dados usados, incluindo os atributos e tipos que mostram as características de cada domínio; com os conceitos e objetos que descrevem as entidades; com os relacionamentos entre os objetos e com as regras que descrevem as relações existentes entre os objetos. A estrutura de ontologia pode ser usada para apoiar essa descrição.

Os resultados obtidos devem ser formalizados na base de casos, incluindo as medições que indiquem a sua similaridade com o problema resolvido. Essa base de

casos deve ser acessada quando o conhecimento estiver sendo documentado para verificar se ele é mais ou menos interessante do que um já existente. O processo de armazenamento de casos deve ser responsável pela garantia da sua qualidade, isto é, ele deve resguardar que as medidas, a sua estruturação e sua avaliação estejam adequadas. Os conceitos e as técnicas relacionadas a CBR podem ser úteis para gerenciar esses conhecimentos.

As informações armazenadas por essa estrutura poderão ser usadas por ela mesma no processo de novas documentações. Nesse caso, o processo tem uma intervenção manual grande. Essas informações poderão, também, ser acessadas e manipuladas por meio da integração com uma ferramenta de mineração. O processo de documentação, nesse caso, seria feito automaticamente.

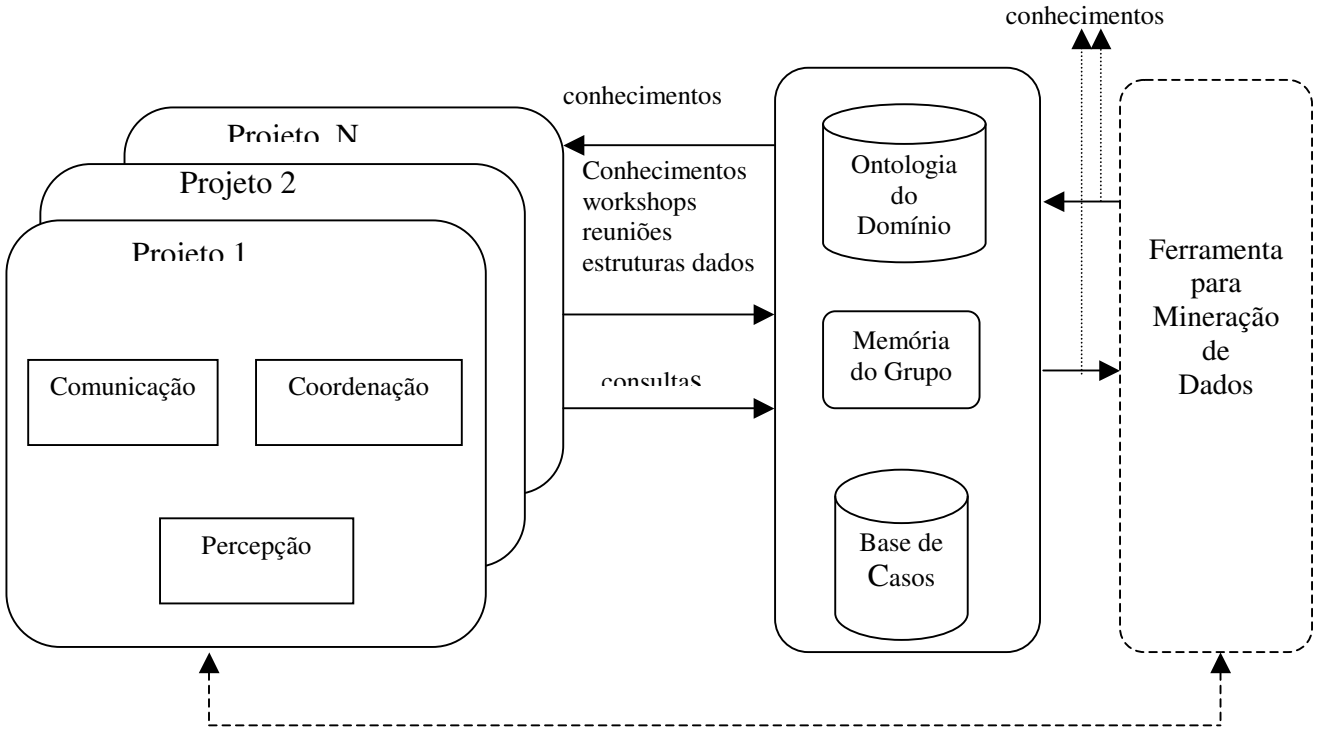

Figura 4.6- Arquitetura de um Ambiente para Apoio ao Processo KDD

5- Conclusão

Neste trabalho foi realizado um estudo de duas das principais abordagens sobre o processo KDD, colocando-se em evidência as suas características, bem como os fatores e dificuldades envolvidos nas suas aplicações. Os conceitos apresentados nas abordagens discutidas no trabalho foram consolidados e validados através de uma aplicação prática desse processo KDD em um ambiente de DW com informações de controle de obras da Prefeitura de Belo Horizonte.

As principais contribuições que este trabalho pretende oferecer são:

- Expor o processo KDD sob diferentes abordagens: a de Brachman et al [BRA96a] e a de Fayyad et al [FAY96a].
- Expor trabalhos que têm sido desenvolvidos para ajudar na condução do processo KDD.
- Posicionar processo de *data warehousing* no contexto KDD, mostrando que os dados preparados para o DW, se estiverem íntegros e completos, atendem, quase na sua totalidade, os requisitos necessários para a etapa de mineração de dados. Mostrar, também, que as ferramentas OLAP têm grande importância na fase de exploração e entendimento dos dados.
- Indicar requisitos importantes a serem considerados na construção de um ambiente de apoio às atividades do processo KDD. Indicar, também, uma proposta de estrutura para esse ambiente.
- Enriquecer, com um estudo de caso, a literatura científica sobre KDD, ainda pequena em relatos de experiências e verificação empírica de resultados do processo.

A avaliação prática do processo KDD possibilitou um entendimento mais abrangente dos problemas ressaltados na literatura, e trouxe, como resultados, pontos que podem ser explorados em trabalhos futuros para trazer maiores benefícios na sua

adoção. De forma geral, pôde-se perceber

- A importância da visão do processo KDD como uma questão de estratégia de negócios. Este estudo de caso indicou que a adoção do processo de descoberta de conhecimento em uma empresa é lenta.
- A importância de a empresa ter políticas para gerenciamento de conhecimento ou entender essa questão como uma estratégia de manter o seu negócio com nível de competitividade compatível com o mercado em que atua;
- A importância de os objetivos da aplicação do processo KDD estarem alinhados com as estratégias da empresa; caso contrário, considerando o quanto o processo é trabalhoso e complexo, poderá diminuir o apoio das pessoas que devem estar envolvidas nas diversas etapas. O valor dos resultados obtidos e o tempo para consegui-los dependem da motivação e esforços despendidos pelas pessoas no processo;
- A imprevisibilidade do processo. As abordagens propostas por Fayyad et al [FAY96a] e Brachman et al [BRA96a] sugerem uma disposição seqüencial das etapas, existindo a possibilidade de iterações entre elas. A quantidade de iteração foi mais alta do que a imaginada inicialmente, tornando a execução do processo complexa e comprometendo as expectativas sobre prazo para produção de resultados e a qualidade dos mesmos;
- O grande número de fatores que influenciam os resultados. O acompanhamento do projeto deve ser constante para evitar que fatores como tempo, falta de envolvimento das pessoas, o grande número de informações manipuladas (a definição da aplicação, as fontes de dados, as técnicas e parametrização utilizadas, os procedimentos utilizados, os resultados obtidos, etc) e a repetição de atividades influenciem na produção de resultados úteis. Uma das maiores preocupações, nesse sentido, é com o fator qualidade dos dados e com a pré-disposição da organização em melhorá-la caso não esteja adequada.
- A importância de um escopo bem definido para o objetivo do trabalho. Um empreendimento nessa área, pelas suas características multidisciplinares e pelas interações entre diferentes pessoas, tem grande possibilidade de identificar nichos

de informações que ainda não estão organizadas em banco de dados. Se isso acontece, os usuários podem julgar os resultados como duvidosos ou incompletos pelo fato de o processo não ter tratado tais informações. É importante que a empresa tenha motivação para estar sempre trabalhando as informações necessárias para atender o objetivo definido no início; caso contrário, como o processo leva a outras indagações e necessidades, o projeto pode se prolongar por muito tempo;

- A importância do especialista do domínio da aplicação. Ele é quem ajuda a validar os resultados, a organizar e disseminar os conhecimentos adquiridos pelo processo, e a confrontar o que foi descoberto com as crenças ou verdades aceitas até então;
- A importância da pessoa que realiza a tarefa de preparação dos dados. Sua familiaridade com as técnicas de preparação de dados e com o uso das ferramentas de apoio à manipulação dos dados dá mais agilidade ao processo.
- A necessidade de diversas ferramentas, geralmente não integradas. Ao longo do processo, para o cumprimento dos objetivos de cada etapa, são usados diversos tipos de ferramentas. Dentre outras, podemos citar ferramentas manipulação de textos, gráficos e de visualização, EXCEL, SQL. È importante que os sistemas para KDD sejam desenvolvidos estruturados em forma de componentes, de tal modo que um novo componente seja facilmente adicionado. É importante, também, que esses sistemas sejam integrados com outras ferramentas (SGBD, OLAP, CASE), pois facilita o andamento do processo;
- A importância de um suporte para interpretação dos resultados. O processo de interpretação e avaliação dos resultados depende da intuição e do conhecimento do domínio por parte dos usuários. Por essa razão, a visualização dos resultados tem um papel importante. As ferramentas de apoio ao processo devem permitir que o usuário trabalhe com a forma mais adequada de visualizar os resultados produzidos. O desenvolvimento de algoritmos com essas características promovem ainda mais o caráter interdisciplinar do KDD, utilizando, nesse caso, os estudos e pesquisas na área de comunicação;
- A importância de uma etapa de exploração dos dados disponíveis. Esse caráter exploratório é um pré-processo de aquisição de conhecimento que aprofunda a

compreensão do domínio, e pode, em conseqüência, prover outros critérios para definir como o processo pode ser encaminhado;
- O quanto o conhecimento interdisciplinar e experiência das pessoas sobre o domínio da aplicação são importantes para o bom andamento do processo;
- O quanto a aplicação dos métodos estatísticos contribuem para os resultados do processo KDD.

Vários trabalhos podem dar continuidade a este estudo já que existe ainda um longo caminho a ser percorrido para o KDD tornar-se mais facilmente tratável pela comunidade de usuários, mormente na ampliação do seu uso pelas organizações.

Do ponto de vista do problema levantado – relação do fator K na política de preços –, as análises, com os dados carregados no DW, não identificaram nenhuma relação. Para a continuidade desta experiência pode-se prever um trabalho que encaminhe os seguintes itens:
- Completar os dados faltosos e refazer as análises. Isso inclui a atualização do DW com informações da obras dos últimos meses;
- Incluir nos dados preparados, informações sobre o orçamento e verificar a existência de correlações entre as variáveis do orçamento com o fator K. Isso, além de ampliar o conhecimento do comportamento do fator K nas questões levantadas, poderá produzir outras informações úteis para a administração da prefeitura e, também, despertar maior interesse pelo assunto KDD ampliando a sua utilização para outros domínios de aplicação;
- Fazer uma coleta de informações externa à prefeitura sobre obras similares às administradas pela Sudecap, realizadas por outras instituições. Com essas informações e as do DW, fazer uma investigação comparativa, tendo em vista a identificação de possíveis desvios na política de preços estabelecidos pela Sudecap. A realização desse item deve ser motivada, mesmo sabendo das dificuldades e a falta de métodos padronizados para a coleta dessas informações.

Esses itens não foram realizados nesse trabalho pelo grande esforço para se chegar a um conjunto de dados estáveis e pelo tempo que iria despender.

Do ponto de vista do processo várias questões podem ser abordadas em trabalhos futuros. Dentre elas podemos citar:

- Definição de um ambiente para documentar, organizar e representar formalmente o conhecimento produzido no processo KDD. Tanto o projeto UGM como o HAMB fazem referência a bases de conhecimento, porém, eles não falam de ferramentas para apoiar essa tarefa. Os conceitos de OM e ontologia poderão ser úteis. Essa organização poderá ser proveitosa na construção de aplicações que são bastante dependentes de experiência e conhecimento. Podem ser úteis, também, para construir agentes para escolher tipos de algoritmos mais adequados a determinados conjuntos de dados. Esse ambiente deve definido e criado para ser usado cooperativamente, atendendo e ampliando os requisitos listados na seção 5.8. Ele poderá aumentar o nível de produtividade do processo na obtenção de resultados e, igualmente, aumentar qualidade desses resultados;
- Definição de uma sistemática ou metodologia de trabalho para tratar a questão KDD na empresa. Apesar deste trabalho ser baseado em apenas uma experiência, as observações feitas no seu decorrer sugerem que o sucesso de um empreendimento KDD depende do engajamento da empresa por meio de políticas de gerenciamento do conhecimento. É prematuro tornar essa afirmação definitiva. São necessárias avaliações em outras experiências e é preciso propor métodos de trabalhos para o processo KDD dentro de um contexto maior: as estratégias da empresa e as atividades de gerenciamento de conhecimento;
- Definição de metodologias e critérios para avaliação de ferramentas de apoio ao processo KDD. Para este trabalho, a escolha do DBminer seguiu um número reduzido de critérios, ligados principalmente às características dos dados. Até poucos anos atrás, as ferramentas de apoio ao processo KDD estavam sendo usadas em caráter experimental ou em ambiente de pesquisas. Percebemos agora uma nova realidade, na qual um número cada vez maior de ferramentas sofisticadas é disponibilizado no mercado. Vemos, também, que KDD é um assunto cada vez

mais freqüente nas organizações, que, em última estância, esperam retorno sobre os investimentos feitos com as tecnologias. É importante, portanto, que sejam propostos critérios e enumeradas características mais importantes que uma ferramenta necessita ter para apoiar efetivamente os usuários, atendendo-os em suas expectativas.

Uma questão merece ser estudada com mais profundidade: que decisões devem ser tomadas em relação aos dados trabalhados após o fim do processo? Eles podem ser úteis para futuras consultas. É importante fazer um estudo propondo alternativas para a estruturação e manutenção desses dados.

Este trabalho aspira ser mais um motivo para despertar ou aumentar o interesse sobre o assunto KDD, entendendo que o desenvolvimento da tecnologia junto com a participação das pessoas propicia inovações nas instituições. Segundo Nonaka [NON97a], organizações que inovam, não só processam informações, mas também criam constantemente novos conhecimentos, a fim de redefinir tanto os problemas quanto as soluções e, desse processo, recriam o seu próprio meio.

6- Abstract

The great number of computing that systems has been collecting and storing an enormous volume of information in databases, creating good opportunities for the application of techniques to discover patterns of behavior in the stored data. This work main objective is the elaboration of an analysis of the process of discovery knowledge in databases, approaching its characteristics, as well as the factors and difficulties involved in its application, mainly its methods of working and the environment that gives support to it. It also describes an experience of application of the process in a data warehouse. That experience was guided for the verification of the existent link between what it is presented as a theoretical basis of the technology applicability and its implementation in a real case. The reasons of the main problems and the learned lessons are analyzed. In a complementary way, it was related an initial set of requirements that they should be supported by a support environment to the conduction of the KDD process.

7- Referências Bibliográficas

[ABE99a] ABECKER, Andreas; DECKER, Stefan. ***Organizational Memory: Knowledge Acquisition, Integration, and Retrieval Issues***. In: Frank Puppe (ed.) *XPS-99 / 5. Deutsche Tagung Wissensbasierte Systeme*, Würzburg: Springer Verlag, March 1999. Disponível em: http://www.dfki.uni-kl.de/~aabecker/Publications.html Acesso em 15/06/2001.

[AGR93a] AGRAWAL, T. Imielinski; SWAMI A. **Mining Association rules between sets of itens in large databases**. In: Proc. Int. Conf. Management of Data (SIGMOD), May 1993, 207-216.

[AGR96a] AGRAWAL, Rakesk; MANNILA, Heikki; SRIKANT, Ramakrishnan; TOIVONEN, Hannu; VERCKAMO, A. Inkeri. **Fast Discovery of Association Rules**. In: Advances in Knowledge and Data Mining in Databases: FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. Califórnia, USA: AAAI/MIT Press, 1996. p 307-328.

[ARA97a] ARAUJO, Renata M.; DIAS, Márcio S.; BORGES, Marcos R. **A Framework for the Classification of Computer Supported Collaborative Design Approaches**. In : CRIWG, 1997, San Lorenzo de El Escorial, Madrid, Espanha, 1997.

[ARA97b] ARAUJO, Renata M.; DIAS, Márcio S.; BORGES, Marcos R. **Suporte por Computador ao Desenvolvimento Cooperativo de Software: Classificação e Propostas**. XI Congresso de Engenharia de Software, Brasil, Outubro 1997.

[AUR99a] AURELIO, Marco; VELLASCO, Marley; LOPES, Carlos Enrique. **Descoberta de Conhecimento e Mineração de Dados**. 101 p. Apostila. Laboratório de Inteligência Computacional Aplicada, Departamento de Engenharia Elétrica, PUC-Rio, 1999.

[BAR00a] BARTLMAE, Kai; RIEMENSCHNEIDER, Michael. **Case Based Reasoning for Knowledge Management in KDD-Projects**. In: Proc. Of Third Int. Conf. On Practical Aspects of Knowledge Management, Basel, Switzerland, Out 2000. Disponível em: http://sunsite.informatik.rwth-aachen.de/publications/CEUR-WS/vol-34/ Acesso em: 03/06/2001

[BEN98a] BENJAMINS, Richard V., FENSEL, Dieter, PÉREZ, Asuncíon Gómez, **Knowledge Management througt Ontologies**, Universitat Karlsruhe, Institut AIFB. Disponível em: http://citeseer.nj.nec.com/benjamins98knowledge.html Acesso em: 15/06/2001.

[**BER99a**] BERSON, Alex; SMITH, Stephen; THEARLING, Kurt. **Building Data Mining Applications for CRM**. New York,USA: McGraw-Hill,1999. 510p.

[BET96a] BERNDT, Donald j.; CLIFFORD James. **Finding Patterns in Time Series: A dynamic Programming Approach.** In: Advances in Knowledge and Data Mining in Databases: FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. Califórnia, USA: AAAI/MIT Press, 1996. p 229-248.

[BOC95a] BOCK, Geoffrey E; MARCA, David A. **Designing Groupware; a guidebook for designers, implementors, and users**. McGraw-Hill, 1995. 231p.

[BOO99a] BOOCH, Grady; RUMBAUGH, James; JACOBSON, Ivar. **The Unified Modeling Language User Guide.** MA: Addison-Wesley,1999. 482p.

[BRA96a] BRACHMAN, Ronald; ANAND, Tej. **The Process of Knowledge Discovery in Databases**. In: Advances in Knowledge and Data Mining in Databases: FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. Califórnia, USA: AAAI/MIT Press, 1996. p 37-57.

[**BRG00a**] BRAGA, Antônio Pádua; LUDEMIR Tereza B.; CARVALHO, André Carlos. **Redes Neurais Artificiais Teoria e Aplicações**. Rio de Janeiro: LTC, 2000. 262p.

[**BRU00a**] BRUHA, Ivan. **Data Mining, KDD, and Knowledge Integration: Methodology and A Case Study**. 2000. Disponível em: http://www.ssgrr.it/em/ssgrr2000/papers Acesso em: 15/10/2001.

[CAR00a] CARVALHO, D. R.; FREITAS, Alex A. **A genetic algorithm-based solution for the problem of small disjuncts. Principles of Data Mining and Knowledge Discovery**. In: Proc. 4th European Conf. PKDD. 2000. Lyon, France. Lecture Notes in Artificial Intelligence. p345-352.

[CHN99a] CHEN, Qing. **Mining Exceptions and Quantitative Association Rules in OLAP Data Cuben Rules**. Jul 1999,103p. Thesis of Master of Science in the Departament of Computer Science, Simon Frase University, Canada.

[CHE00a] CHEDINI, Cinara G. **Um Modelo de Apoio à Documentação de Aplicações de Descoberta de Conhecimento em Base de Dados**, 2000. 116p. Dissertação de mestrado. Porto Alegre, RS, Pontifícia Universidade Católica do Rio Grande do Sul,.

[CEE96a] CHEESEMAN, Peter; STUTZ, John. **Bayesian Classification autoclass): Theory and Results**. In: Advances in Knowledge and Data Mining in Databases: FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. Califórnia, USA: AAAI/MIT Press, 1996. P 154-180.

[DAV98a] DAVENPORT, Tomas H.; PRUSAK, Laurence. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual**. Rio de Janeiro: Campus, 1998. 237p.

[DZE96a] DZEROSKI, Saso**. Inductive Logic Programming and Knowledge Discovery in Databases.** In Advances in Knowledge and Data Mining in Databases: FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. Califórnia, USA: AAAI/MIT Press, 1996. p 117-152.

[DOR00a] DORNELES, Carina. F.; HEUSER, Carlos A. **Extração de Dados com base em uma Ontologia**. In: International Symposium on Knowledge Management/Document management, Curitiba, Novembro 2000. p 407-418.

[DOR00b] CHEDINI, Cinara G. **Extração de Dados Semi-Estruturados com Base em uma Ontologia**, 2000. 88p. Dissertação de mestrado. Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[DUN97a] DUNKEL, Brian; SOPARKAR, Nandit; SZARO John; UTHURUSAMY, Ramasamy. **Systems for KDD: From concepts to practice**. In: Future Generation Computer Systems, vol. 13, 1997. p 231-242.

[ENG99a] ENGELS, Robert; LINDNER, Guido; STUDER, Rudi. **A Methodology for Provinding User Support for Developing Knowledge Discovery Applications**, Universitat Karlsruhe, Institut AIFB, disponível em: http://www.aifb.uni-karlsruhe.de/WBS/publications/ Acesso em 15/10/2000.

[FAY96a] FAYYAD, Usama.; SHAPIRO, Gregory. P.; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery: an Overview**. In: Advances in Knowledge and Data Mining in Databases: FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. Califórnia,USA: AAAI/MIT Press, 1996. p.1-34.

[FAY96b] FAYYAD, Usama.; SHAPIRO, Gregory. P.; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery**. In: Databases American Association of Artificial Intelligence. 1996. p 37-54.

[FAY97c] FAYYAD, U.; STOLORZ, P. **Data mining and KDD: Promise and challenges**, Elsevier Science, 1997. p 99-115.

[FRA91a] FRAWLEY, W. J.; SHAPIRO, G. P.; MATHEUS, C. J. Knowledge **Discovery in Databases: an overview**. In: Knowledge Discovery in Databases: SHAPIRO, G. P.; MATHEUS, C. J. . Califórnia,USA: AAAI/MIT Press,1991. p 1-27.

[FRE00a] FREITAS, Alex A. **Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A position Paper**. ACM SIGKDD, July, 2000. Disponível em: http://www.acm.org Acesso em: 15/10/2000.

[FRE99a] FREITAS, Alex A. **The principles of Transformation between Efficiency and Effectineness: Towards a fair Evaluation of the Cost-Effectiveness of KDD Techniques**. Disponível em: http://www.ppgia.pucpr.br/~alex Acesso em: 10/09/2000.

[FOO02a] FOONG, David Low Yuh. **A visualization-Driven Approach for Startegic Knowledge Discovery**. In: Information Visualization in Data Mining and Knowledge. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p 181-190.

[GOE99a] GOEBEL, Michael; GRUENWALD, Le. **A Survey of Data Mining and Knowledge Discovery Software Tools**. In: ACM SIGKDD, 1999. p. 20-33.

[GRA02a] GRADY, Nancy; AUVIL, Loreta; BECK, Allan; BONO, Becker; MENEZES, Claudio. **Integrating Data Mining and Visualization Processes. In: Information Visualization in Data Mining and Knowledge**. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p 299-304.

[GRI02a] GRINSTEIN, Georges; HOFFMAN, Patrick; PICKETT, Ronald M.; LASKOWSKI, Sharon J. **Benchmark Development for the Evaliation of Visualization for Data Mining**. In: Information Visualization in Data Mining and Knowledge. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p 129-177.

[GRI02b] GRINSTEIN, Georges; WARD, Matthew O. **Introdution to Data Visualization**. In: Information Visualization in Data Mining and Knowledge. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p 01-20.

[GRO98a] GROTH, Robert. **Data Mining: a hands-on approach for business professionals**; USA: Prentice Hall, 1998. 264p.

[GRU93a] GRUBER, Thomas H. **Toward Principles for the Design of Ontologies Used for Knowledge Sharing**, August, 1993. Disponível em: http://ksl.stanford.edu.com Acesso em: 25/03/2000

[HAN99a] HAN, Jiawei, SONNY, H. S.; CHIANG J. Y. **Issues for ON-line Analytical Mining of Data Wahehouses**, School of Computing Science, Simon Frazer University, British Columbia, Canada. Disponível em: http://db.cs.sfu.ca/sections/publication/kdd/ Acesso em: 10/02/2000.

[HAN99b] HAN, J.; WANG, W.; FU, Y.; KOPERSKI, K.; ZAIANE, O . **DMQL: A Data Mining Query Language for Relational Databases**, School of Computing Science , Simon fraser University, Canada. Disponível em: http://db.cs.sfu.ca/sections/publication/kdd/ Acesso em: 10/02/2000.

[HAN01a] HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. San Diego, USA: Academic Press, 2001. 550 p.

[HIN02a] HINKE, Thomas H.; NEWMAN, Timothy S. **A Taxonomy for Integration Data Mining and Data Visualization**. In: Information Visualization in Data Mining and Knowledge. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p 291-298.

[HIP00a] HIPP, J.; GUNTZER, U.; NAKHAEIZADEH, G. **Algorithms for association Rule Mining – A General Survey and Comparison**. ACM SIGKDD July, 2000. Disponível em: http://www.acm.org Acesso em: 15/10/2000.

[HOF02a] HOFFMAN, Patrick; GRINSTEIN, Georges. **A Survey of Visualization for High-Dimensional Data Mining**. In: Information Visualization in Data Mining and Knowledge. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p. 47-82.

[INM97a] INMON, W.H. **Como Construir o *Data Warehouse***. 2 ed. Rio de Janeiro: Editora Campus, 1997. 388p.

[JAS99a] JASPER, Robert; USCHOLD, Mike. **A Framework for Understanding and Classifying Ontology Applications**, 1998. Disponível em: http://citeseer.nj.nec.com/uschold99framework.html Acesso em: 15/03/2001.

[JEN98a] JENNINGS, Nicholas R.; WOOLDRIDGE, Michael J. **Agent technology: foundations, applications, and markets**. Springer, 1998.

[JUR99a] JURISICA, Igor; MYLOPOULOS John; YU, Eric. **Using Ontologies for management: An InformationSystems Perpective**. In: Annual Conference of American Society for Information Science, Washington, nov 1999. Disponível em: http://www.toronto.edu/pub. Acesso em 20/06/2001.

[KAD00a] KADE, Adrovane M.; HEUSER, Carlos A. **Uma Interface visual para linguagem de consulta a dados semi-estruturados**. In: International Symposium on Knowledge Management/ Document management, Curitiba, Novembro 2000. p 421-438.

[KIM98a] KIMBALL, Ralph. **Data Warehouse Toolkit**. São Paulo: MAKRON Books,1998. 388p.

[KIM00a] KIMBAL, Ralph; MERZ, Richard. **The data webhouse toolkit: building the web-enabled data warehouse**. John Wiley & Songs, Inc., 2000.

[KOR99a] SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN S. **Sistema de bancos de dados**. 3ª ed. São Paulo: MAKRON Books, 1999. 778p.

[LIA99a] LIAO, Minghong; ANDREAS Abecker; ANSGAR Bernardi; KNUT Hinkelmann; SINTEK Michael. **Ontologies for Knowledge Retrieval in Organizational Memories**. Out 1999. Disponível em: http://www.dfkiuni-kl.de/pubblications Acesso em: 15/03/2001

[LIN99a] LINDNER, Guido; STUDER, Rudi. **Algorithm Selection Support for Classification**. Set 1999. Disponível em: http://www.dfkiuni-kl.de/pubblications Acesso em: 10/02/2001

[LIV01a] LIVINGSTON, Gary Ray. **A Framework for Autonomous Knowledge Discovery from Databases**. 2001.184p. Tese de Doutorado – Universidade de Pittsburgh, Pittsburgh.

[MOU99a] MOURA, Ricardo; Garcia, Ana Cristina. **DKAT: A Knowledge Configuration Tool applied to Design**. In: Twelfth Workshop on Knowledge Acquisition, Modeling and Management Voyager Inn, Banff, Alberta, Canada, 1999.

[MUS00a] MUSSI, Clarissa C.; ANGELONI, Maria T. **Mapeamento das fontes do conhecimento organizacional: um suporte ao compartilhamento do conhecimento tácito**. In: International Symposium on Knowledge Management/ Document management, Curitiba, Novembro 2000. pg 1-17.

[NAV00a] NAVATHE, Shamkant B.; ELMASRI, Ramez. **Fundamentals of database systems**. 3 ed. Addinson Wesley Publishing, 2000. 955p.

[NON97a] NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Criação de Conhecimento na Empresa: como as empresas japonesas geram a dinâmica da inovação**. Rio de Janeiro: Campus, 1997. 225p.

[POS01a] PÔSSAS, Bruno V.; VIEIRA, Fabiana; WAGNER, Meira; RESENDE, Rodolfo. **Geração de Regras de Associação Quantitativas**. Universidade Federal de Minas Gerais, Departamento de Ciência da Computação.. Disponível em www.dataminingbr.com.br Acesso em 10/05/2001.

[PYL99a] PYLE, D. **Data Preparation for Data Mining**, USA: Morgam Kaufmann Oublishers, 1999. 540p.

[PRE00a] PRESSMAN, Roger. **Software engineering**. 5th edition. McGraw-Hill, 2000. 852p.

[RIC98a] RICHTER, M. **Introdution do CBR**. In: Case Based Reasoning Technology. From Foundations to Applications: LENS, M.; BARTSCH-SPORL, B.; BURKHARD, H-D.; WESS, S. Lecture Notes in Artificial Intelligence. Springer Verlag, New York. 1998.

[RHO02a] RHODES, Philip J. **Discovering New Relationships: A Brief Overview od Data Mining and Knowledge Discovery**. In: Information Visualization in Data Mining and Knowledge. FAYYAD, U. M.; GRINSTEIN Georges G. WIERSE, Andreas. San Diego,USA: Academic Press, 2002. p 277-290.

[SAG02] **Site Software AG Inc**, 2002. Disponível em www.softwareag.com Acesso em 06/01/2002.

[SPS02] **Site SPSS Inc**, 2002. Disponível em www.spss.com Acesso em 10/01/2002.

[STA01a] STAAB Steffen; STUDER Rudi; SCHNURR, Hans-Peter; SURE, York. **Knowledge Process and Ontologies**. IEEE Intelligent Systems, 2001.

[STU99a] STUDER, Rudi; FENSEL, Dieter, DECKER, Stefan; BENJAMINS, Richard, **Knowledge Engeneering: Survey and Future**. Set 1999. Disponível em: http://www.dfkiuni-kl.de/pubblications Acesso em: 11/02/2001.

[SHA01a] SHAPIRO, Gregory Piatetsky. **A Pré-história do Data Mining – Descoberta de Conhecimento em Bases de Dados**. Dez 2001. Disponível em: http://www.marketingdeprecisão.com.br/artigos_nova.asp: Acesso em: 15/12/2001.

[TAM98a] TAM, Jenny Yin. **Datacube: Its Implementation and Application in OLAP Mining**, 1998, 97p. Thesis of Master of Science in the Departament of Computer Science, Simon Frase University, Canada.

[VAS00a] VASCONCELOS, Flávio C. e CYRINO, Álvaro B. **Vantagem Competitiva: os modelos teóricos atuais e a convergência entre estratégia e teoria organizacional**, Revista de Administração de Empresas, vol 40, num 4, Out 2000. p 20-37.

[WES98a] WESTPHAL, Christopher; BLASXTON, Tereza. **Data Mining Solutions: Methods and Tools for Solving Real-World Problems**. New York,USA: Wiley, 1998. 617p.

[ZHU98a] ZHU, Hua. **On-Line Analytical Mining of Association Rules**. Dez 1998,117p. Thesis of Master of Science in the Departament of Computer Science, Simon Frase University, Canada.

8- Anexos

8.1- Anexo A Modelo de Dados das Tabelas de Fatos usadas no Projeto

1) Tabela de fatos com informações sobre os aditivos de contratos.

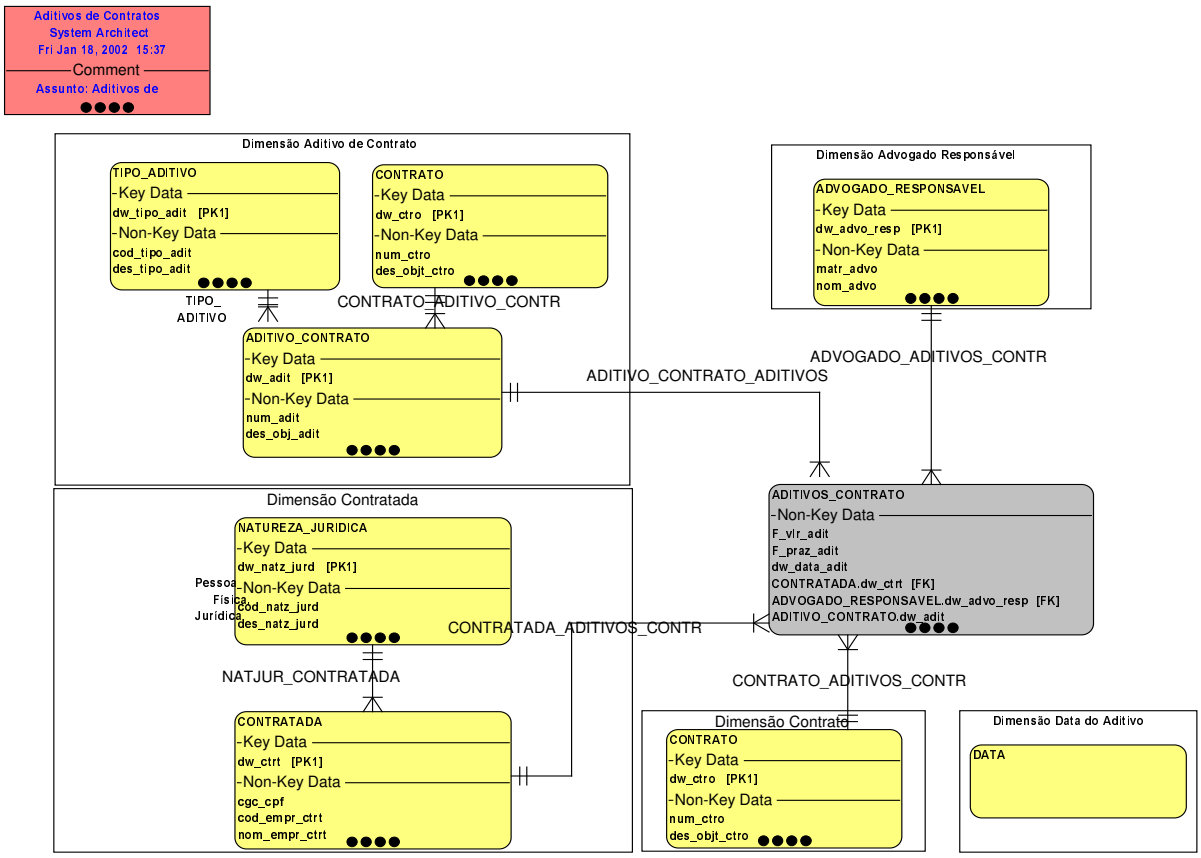

Figura 8.1.1- Modelo de Dados da Tabela de Fatos Aditivos de Contratos

Anexos

2) Tabela de fatos com informações de contratos.

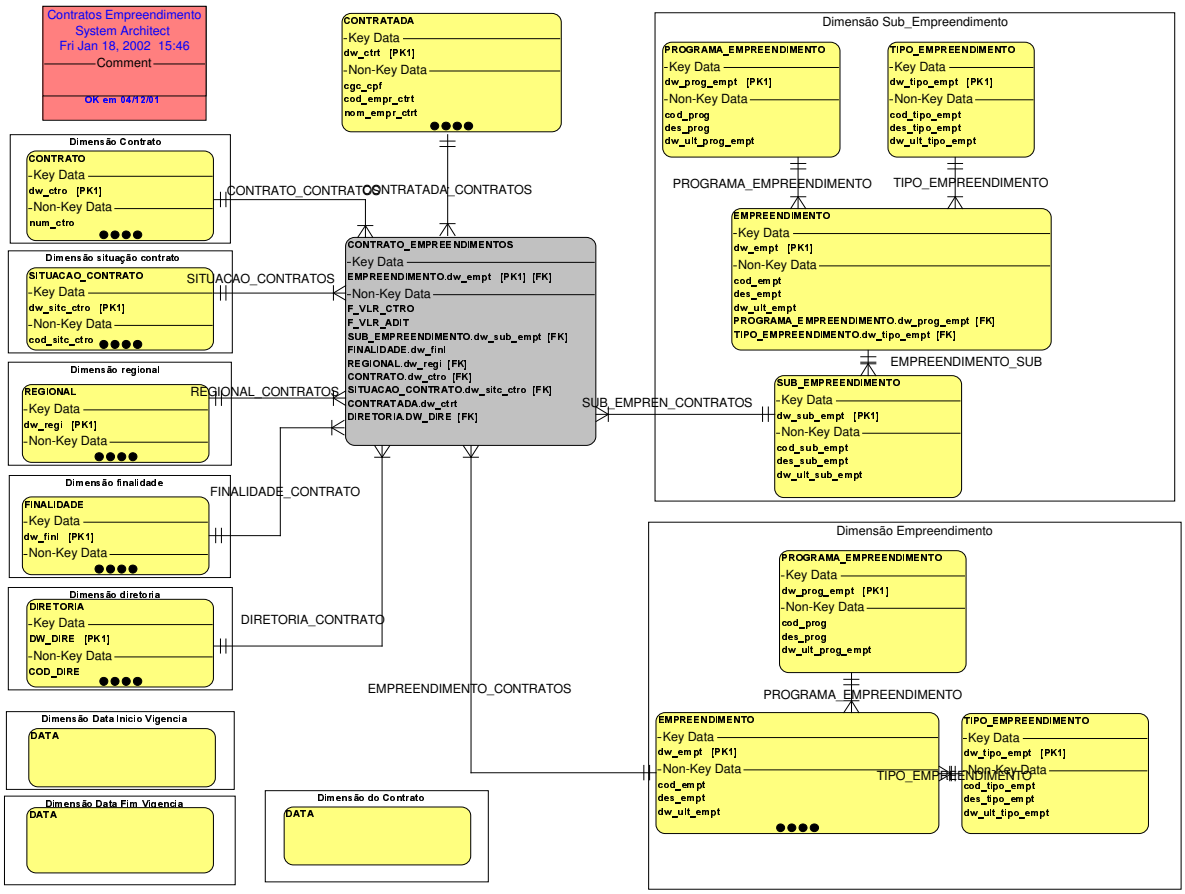

Figura 8.1.2- Modelo de Dados da Tabela de Fatos Contratos

Anexos

3) Tabela de fatos com informações de medições dos empreendimentos .

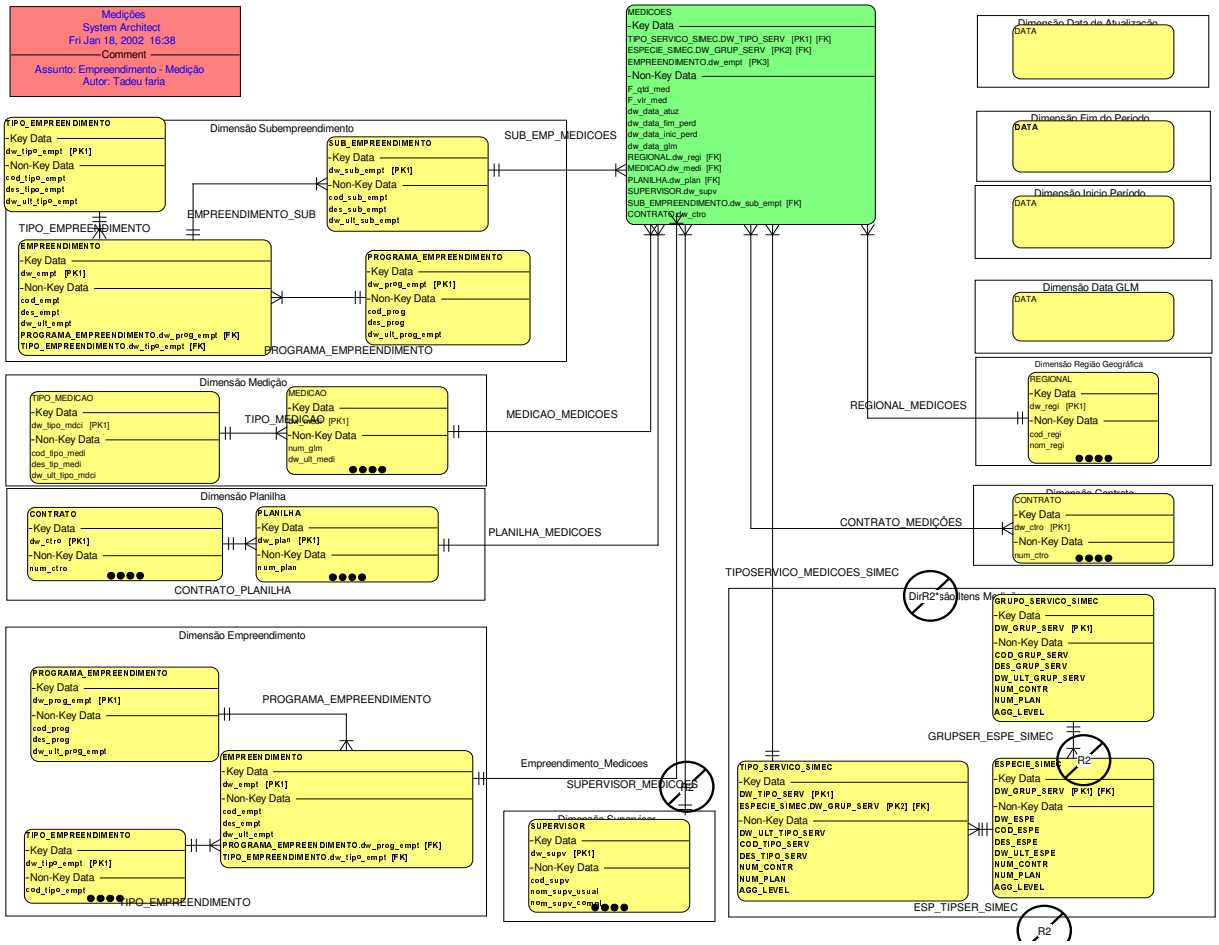

Figura 8.1.3- Modelo de Dados da Tabela de Fatos Medições

Anexos

4) Tabela de fatos com informações das planilhas de execução dos empreendimentos.

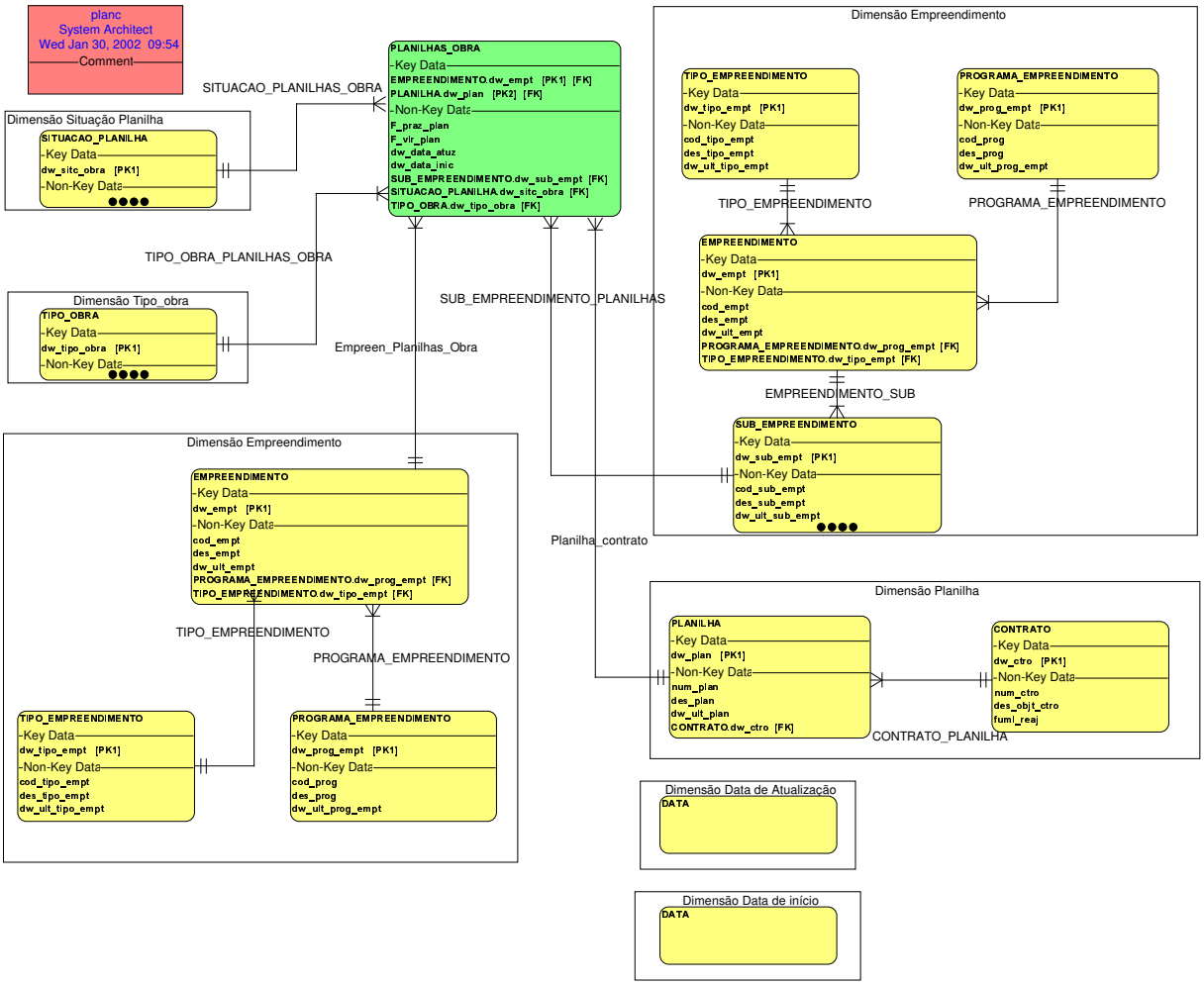

Figura 8.1.4- Modelo de Dados da Tabela de Fatos Planilhas de Obras

Anexos

8.2- Anexo B Recursos de Hardware e Software utilizados no projeto

Os recursos utilizados no projeto foram os seguintes:

Hardware :

- Servidor IBM NETFINITY dual Pentium III de 733 Mhz, 1 Gb de memória RAM, RAID de 4 discos hot-swap de 9 Gb (1 spare disk), utilizado como servidor de banco de dados onde se encontra instalado o ORACLE 8.05 e onde esta DW. Sua identificação é S10-SUDECAP.
- Servidor ALCABYT dual Pentium II 750 MhZ, 250 Mb de RAM, utilizado como servidor de banco de dados onde se encontra instalado o SQL server. As tabelas usadas pelo Dbminer estão neste SQL server. Sua identificação é S8_SUDECAP.
- Estação com processador AMD Athlon com 1.2 Mhz, 250 Mb de RAM, utilizado para o Dbminer e SPSS e Microsoft Office 98.

Software :

- Sistema Operacional : Windows NT 4.0 para o servidor S10-SUDECAP. No servidor S8-SUDECAP e na estação foi usado Windows 2000.
- Banco de dados : ORACLE 8.05 para gerenciar os dados do DW e SQL server para gerenciar as tabelas de fatos e dimensões tratadas pelo Dbminer. Foi usado, também, o ACCESS.
- Ferramenta OLAP: foi usado, no início do projeto, o Metacube 4.2 da INFORMIX. Posteriormente, foi usado o ambiente de análise do SQL server.
- Ferramentas de mineração: foi usado o DBminer para as regras de associação, classificação, clusteriação e análises 3D. Para os métodos estatísticos foi utilizado o SPSS e, em alguns casos, o EXCEL.
- Ferramenta Office: Microsoft Office 98.