MARCELA DOS SANTOS

APLICAÇÃO DA MINERAÇÃO DE DADOS NO SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO: UM ESTUDO DE CASO

Monografia apresentada à disciplina de Projeto de Pesquisa em Informação II do Curso de Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná.

Orientadora: Prof ª. Denise Fukumi Tsunoda

CURITIBA
2004

"A mente humana, uma vez ampliada por uma nova idéia,
nunca mais volta ao seu tamanho original "

Oliver Wendell Holmes

SUMÁRIO

LISTA DE FIGURAS

LISTAS DE TABELAS

LISTAS DE SIGLAS

ABES	- Associação Brasileira de Engenharia Sanitária
AESBE	- Associação das Empresas de Saneamento Básico Estaduais
ASSEMAE	- Associação Nacional dos Serviços Municipais de Saneamento
BNH	- Banco Nacional de Habilitação
CESBs	- Companhias Estaduais de água e Saneamento Básico
DCBD	- Descoberta do Conhecimento em Bases de Dados
DEX	- Despesas Operacionais
KDD	- *Knowledge Discovery in Databases*
SNIS	- Sistema Nacional de Informações sobre Saneamento Básico
WEKA	- *Waikato Enviroment for Knowledge Analysis*

RESUMO

Este trabalho apresenta um estudo sobre a aplicação da técnica mineração de dados, a fim de transformar dados em fonte de informação para tomada de decisão, utilizando para isso a base de dados de domínio público do Sistema Nacional de Informações sobre Saneamento - SNIS, com intuito de descobrir conhecimento útil para aprimorar a formulação de indicativos de desempenho na área de saneamento básico. A pesquisa , de natureza exploratória, buscou descrever as etapas do processo de Descoberta de Conhecimento em Base De Dados (DCBD), as técnicas, modelos e ferramentas que podem ser utilizadas na mineração de dados. O processo de Descoberta de Conhecimento em Bases de Dados neste trabalho mostrou-se eficiente para este estudo de caso, disponibilizando informações relevantes, despertando e motivando a continuidade da pesquisa.

Palavras-Chave: Descoberta do Conhecimento em Bases de Dados, Mineração de Dados, Saneamento Básico.

1 INTRODUÇÃO

O avanço tecnológico tem proporcionado vários benefícios às pessoas que de uma maneira ou de outra fazem uso das variadas informações que se encontram contidas, disponibilizadas ou geradas por sistemas computacionais. Porém o avanço da tecnologia da informação tem trazido facilidade e crescimento na quantidade de dados armazenados acarretando um crescente problema de abundância de dados para as áreas da ciência, negócios, e governo.

Atualmente, os sistemas são implementados com a finalidade de auxiliar as tarefas humanas em qualquer área de atuação. Estes atuam gerando e coletando dados operacionais, ou seja, dados gerados no dia-a-dia. Mas a coleta e o armazenamento de dados por si só não contribuem para alavancar estratégias. É necessário que se façam análises nesta grande quantidade de dados, estabelecendo indicadores para descobrir relações de causa e efeito, pois processar informação correta é um dos requisitos mais essenciais para uma boa tomada de decisão. A análise das informações contidas nas bases passa por diversos processos e com o aumento do volume dos dados fica cada vez mais complexo explorar as informações contidas nos dados. Devido a esse volume e a dificuldade em explorar os dados, as informações acabam ficando escondidas, sendo formadas implicitamente por relações entre os campos, formando padrões imperceptíveis a olho nu.

Ao longo do tempo, percebeu-se que a velocidade de coleta de informações era muito maior que a velocidade de processamento ou análise delas. Num ambiente de extremas mudanças, torna-se necessário à aplicação de técnicas e ferramentas que agilizem o processo de extração de informações relevantes de grandes volumes de dados. Pensando em aumentar o nível do conhecimento armazenado, surgiu uma nova área da tecnologia da informação chamada de Descoberta do Conhecimento em Base de Dados (DCBD), que procura extrair padrões que estejam escondidos em bases de dados. Para realizar a descoberta de informações implícitas, a DCBD utiliza a técnica de mineração de dados, capaz de revelar o conhecimento que está implícito

em grandes quantidades de dados armazenados nos bancos de dados de uma organização, e fazer uma análise antecipada dos eventos, possibilitando prever tendências e comportamentos futuros, permitindo aos gestores, tomarem decisões baseadas em fatos e não em suposições.

A mineração de dados, que é o foco principal do trabalho, surgiu com o intuito de auxiliar na constante busca pela informação, agindo sobre armazéns com grandes quantidades de dados e tentando extrair destes apenas a informação que é relevante à instituição. Essa área tem despertado interesse tanto no meio comercial quanto científico e está sendo amplamente difundida e aperfeiçoada pelo fato de fornecer resultados promissores para qualquer domínio de aplicação.

Mas o que realmente se vê, é que muitas organizações sejam elas empresarias ou governamentais são incapazes de aproveitar totalmente o que está armazenado em seus arquivos. As informações estão escondidas em montanhas de dados, e não podem ser descobertas utilizando sistemas de gerenciamento de banco de dados convencionais.

Neste contexto, de minerar grandes bancos de dados e melhorar a obtenção, tratamento, apresentação e disponibilização de informações é que surgiu a proposta do trabalho.

O trabalho abordado é representado pela necessidade de auxiliar o desenvolvimento de uma pesquisa científica referente à regulação dos serviços de saneamento básico no Brasil. A base de dados disponibilizada pelo Sistema Nacional de Informação sobre saneamento básico (SNIS), traz informações nas esferas estadual e municipal. Os dados contribuem para a regulação e o controle da prestação dos serviços e para a elevação dos níveis de eficiência e eficácia na gestão das entidades prestadoras dos serviços, por meio do conhecimento de sua realidade, orientando investimentos, custos e tarifas, das empresas prestadoras de serviço. Para realizar a pesquisa utilizou a base de dados que contén informações do diagnóstico realizado no ano de 2001. O SNIS está disponível no endereço http://www.snis.gov.br.

Porém existem questões que só serão levantadas com a aplicação da mineração de dados. Sendo assim, a principal problemática do estudo refere-se a identificar a seguinte questão: como transformar os dados da base do Sistema Nacional de Informações sobre Saneamento (SNIS) em informações relevantes para o pesquisador?

Assim o estudo torna-se oportuno, já que os métodos tradicionais de análise de dados, baseados, principalmente, nos procedimentos humanos de tratamento direto com os dados, simplesmente não funcionam para grandes volumes de dados. À medida que uma base de dados cresce, torna-se ineficiente e impraticável a análise manual. Com isso, os dados coletados em grandes bancos de dados têm se tornado arquivo não utilizado em algumas instituições e importantes decisões são freqüentemente tomadas baseando-se, não na informação armazenada, e sim na intuição do tomador de decisão, simplesmente porque ele não dispõe da ferramenta apropriada para efetuar a extração da informação de maneira segura. Passa a ser primordial conhecer os dados que hora estavam ocultos e transformá-los em informação estratégica, seja em um ambiente empresarial, governamental ou científico.

É neste cenário que o presente trabalho aborda a mineração de dados, mostrando sua aplicação em um ambiente científico, auxiliando a descoberta de conhecimento e ressaltando sua importância e relevância. A partir da base de dados disponibilizada pelo SNIS será possível aplicar a descoberta de conhecimento realizada através da mineração de dados, e buscar padrões implícitos na base de dados. Estes dados serão interpretados e servirão de fonte de informações. Cabe ressaltar que o assunto ainda é novo, porém já existem trabalhos que exploraram o tema e apresentam a mineração de dados em uma forma mais técnica, o que se pretende é demonstrar na prática como está sendo aplicada a mineração de dados e como auxiliar o especialista no assunto a interpretar as informações obtidas.

Como conseqüência do estudo do processo de descoberta de conhecimento aplicado aos dados da base de dados do SNIS, espera-se contribuir

significativamente para a pesquisa que está sendo desenvolvida na área de regulação dos serviços de saneamento no Brasil.

Além de servir como referência, contribuirá significativamente com o meio acadêmico por ser uma aplicação real do processo de DCBD.

1.1 OBJETIVOS

Os objetivos deste trabalho estão divididos em objetivo geral e três específicos.

1.1.1 Objetivo Geral

Aplicar a técnica de mineração de dados para extrair conhecimento referente a informação financeira de prestadores de serviços de saneamento básico.

1.1.2 Objetivos Específicos

Como objetivos específicos do presente trabalho de pesquisa têm-se:

a) levantar literatura pertinente sobre mineração de dados;
b) demonstrar na prática as etapas da mineração de dados;
c) transformar dados de um determinado tema, em fonte de informação para tomada de decisão.
d) repassar o conhecimento e resultados obtidos na pesquisa.

1.2 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está estruturado em 6 seções:

a) na primeira seção é apresentada uma introdução ao trabalho desenvolvido, identificando o problema, a justificativa e os objetivos;
b) na segunda seção é apresentada fundamentação teórica mostrando os embasamentos da área de Descoberta de Conhecimento em Bases de Dados, onde serão descritos os conceitos, definições, técnicas, algoritmos e ferramentas, dando ênfase às fases do processo de DCBD e a mineração de dados. Será expostas também as informações sobre saneamento básico e a descrição do domínio que será analisado;
c) a terceira seção se detém na metodologia utilizada para concretizar a pesquisa, desde vistas a bibliotecas universitárias, termos utilizados na busca de material através da internet, descrição das etapas do processo DCBD;
d) a quarta seção é descrição do estudo de caso, onde serão demonstradas todas as etapas para efetuar a mineração de dados;
e) na quinta seção serão expostas as considerações finais, recomendações e contribuições esperadas;
f) por último as referências utilizadas no trabalho.

2 LITERATURA PERTINENTE

Para contextualizar a pesquisa e entender a real importância da mineração de dados, buscou-se na literatura autores que abordam o assunto.

2.1 DADOS, INFORMAÇÃO E CONHECIMENTO

A tríplice dados, informações e conhecimento tem sido importante fator de competitividade para as organizações. Através do gerenciamento desses recursos informacionais pode-se subsidiar atividades para melhorar o desenvolvimento da organização.

Porém a tarefa de distinguir, na prática, o que vem a ser dado, informação e conhecimento não é fácil. Essa distinção fica mais complicada quando se tenta identificar os limites de cada um os conceitos e percebe-se que os três são intimamente interligados. DAVENPORT e PRUSAK (1998 p.18) ilustram da seguinte maneira:

TABELA 2.1 – DADOS, INFORMAÇÃO E CONHECIMENTO

Dados, Informação e Conhecimento		
Dados	Informação	Conhecimento
Simples observação sobre O estado do mundo	Dados dotados de relevância e propósito	Informação valiosa da mente humana Inclui reflexão, síntese e contexto
Facilmente estruturado Facilmente obtido por máquina Freqüentemente quantificado Facilmente transferível	Requer unidade de analise Exige consenso em relação ao significado Exige necessariamente a mediação humana	De difícil estruturação De difícil captura em máquinas Freqüentemente tácito De difícil transferência

FONTE: DAVENPORT, T; PRUSAK, L. O que queremos dizer com conhecimento. In: _____ **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. Rio de Janeiro: Campus, 1998 p. 28.

2.2 DADOS

Segundo JAMIL (2001, p. 160) dado é a "representação convencionada de uma grandeza qualquer. Expresso em unidades padronizadas pode ser obtido por observação, medidores ou processo automáticos". Refere-se a algo que é preciso conforme o tipo de medição feita e corresponde diretamente ao processo em que é coletado.

Já Davenport define dados como "observações sobre o estado do mundo, e sua observação pode ser feita por pessoas ou por tecnologia apropriada" (DAVENPORT, 2000 p. 35).

Os dados podem ser descritos através de representações formais, estruturais, podendo obviamente ser armazenados em um computador e processados por ele. Eles são coletados, por meio de processos organizacionais, nos ambientes interno e externo.

Em resumo, MORESI[1] citado por FELIX (2003 P. 24), afirma que "dados são sinais que não foram processados, correlacionados, integrados, avaliados ou interpretados de qualquer forma".

2.3 INFORMAÇÃO

A informação tornou-se um recurso imprescindível nas organizações, conhece-la de forma efetiva e saber trabalha-la é fator decisivo para que a empresa tenha diferencial competitivo.

REZENDE (2003 p. 5) relata que "o desafio dos anos de 1980 foi migrar os dados para as informações, por meio de desenvolvimento dos sistemas de informações, que tinham por finalidade analisar dados e organizar a informação para melhorar o processo decisório empresarial".

Ao tratar de informação MCGEE & PRUSAK (1994 p.24) esclarece que:

> A informação são dados coletados, organizados, ordenados aos quais são atribuídos significados e contextos. Informação deve informar, enquanto os dados absolutamente não têm essa missão. A informação deve ter limites, enquanto os dados devem ser ilimitados. Para que os dados se tornem úteis como a informação a uma pessoa encarregada do processo decisório é preciso que sejam apresentados de tal forma que essa pessoa possa relaciona-la e atuar sobe eles.

[1] MORESI, E. Delineando o valor da informação de uma organização. **Ciência da informação**, Brasília, v. 29 n. 1, 2000.

Já FELIX (2003, p. 26) aborda que “quando um conjunto de dados é processado, relacionado ou transformado de maneira a possuir um significado, ele se torna uma informação”.

2.4 CONHECIMENTO

Segundo REZENDE (2003 p.51), “cada vez mais as informações e conhecimento vêm impulsionando o desenvolvimento mundial. Se por um lado os avanços tecnológicos ampliam o universo de informações e conhecimento, por outro lado isso vem exigindo profissionais mais capacitados em manipular esse material e empresas mais ágeis em criar, espalhar e manter seu *Know-how”*.

Neste contexto, e com a evolução de novas tecnologias, surgiram ferramentas básicas para realizar o armazenamento e recuperação de grandes volumes de dados com eficiência e segurança. O homem passa a enfrentar um novo tipo de problema que é o acúmulo de dados gerados pelas novas tecnologias. Porém, nossa capacidade de analisá-los, sumariá-los e deles extrair conhecimento, é lenta, exigindo dos novos bancos de dados, ferramentas mais inteligentes que possam nos auxiliar na sua análise, na descoberta e na extração de conhecimento destes conjuntos volumosos de dados (BERNARDES, 2001, p. 24).

Devido a essas variáveis, a importância do conhecimento tornou-se vital para organizações que estão inseridas em uma economia globalizada, pois a globalização exige que a empresa se adapte às exigências do mercado. No entanto para melhor adaptação da organização, ela deve estar sempre monitorando e armazenando as informações de seus clientes, fornecedores, empregados e concorrentes e, destas informações extrair conhecimento que tornem a empresa mais competitiva e mais moldada ás mudanças exigidas pelo mercado.

Neste cenário, DAVENPORT e PRUSSAK (1998 p. 6) conceituam conhecimento como sendo:

> Mistura fluida de experiência condensada, valores, informação contextual e *insight* experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações. Ele tem origem e é aplicado na mente dos conhecedores. Nas organizações, ele costuma estar embutido não só em documentos ou repositórios, mas também em rotinas, processos, práticas e normas organizacionais.

Os autores NONAKA e TAKEUCHI (1997 p. 66), relatam que "a base das empresas japonesas bem sucedidas está na compreensão de conhecimento, que vê o corpo e a mente como um todo". Os autores ainda classificam dois tipos de conhecimento:

a) tácito – é um tipo de conhecimento muito difícil de ser expresso por meio de palavras e é adquirido com a experiência, de maneira prática. É subjetivo, prático e análogo;
b) explícito – é um tipo de conhecimento que pode ser facilmente expresso em palavras, números e pode ser prontamente transmitido entre pessoas, formalmente e sistematicamente. Envolve o conhecimento de fatos. É objetivo, teórico e digital.

TABELA 2.2 – TIPOS DE CONHECIMENTO E SUAS CARACTERÍSTICA

Conhecimento tácito	**Conhecimento explícito**
Subjetivo	Objetivo
Da experiência (corpo)	Da racionalidade (mente)
Simultâneo (aqui e agora)	Seqüencial (lá e então)
Análogo (prática)	Digital (teoria)

FONTE: NONAKA, I.; TAKEUCHI, H.**Criação do conhecimento na empresa –** como as empresas japonesas geram a dinâmica da inovação. RIO DE JANEIRO: CAMPUS, 1997 p.67.

Resumidamente o conhecimento explícito é aquele que pode ser empacotado como informação. Pode ser encontrado em documentos de uma organização, como exemplo: reportagens, artigos, manuais, patentes, pinturas, imagem, vídeo *software* etc. Já o conhecimento tácito é o pessoal, interno ao indivíduo, resultado de suas experiências, compartilhamento e troca através de contatos diretos e face a face com outras pessoas. (NONAKA e TAKEUCHI, 1997, p. 67).

Portanto o conhecimento tácito e o explícito necessitam de gerenciamento, para facilitar a aprendizagem e proporcionar a criação de novos conhecimentos individuais e organizacionais.

2.5 GERENCIAMENTO DO CONHECIMENTO

Os principais objetivos do gerenciamento do conhecimento organizacional são melhorar a produtividade do conhecimento dos trabalhadores da organização, proporcionar meios para a rápida construção e a utilização da coleção do conhecimento da organização.

O gerenciamento do conhecimento dá uma acentuada importância às pessoas, seus trabalhos práticos, culturais e também decide como e quais tecnologias deverão ser empregadas. O mesmo possui uma arquitetura baseada numa capacidade social e tecnológica para produzir repositórios /bibliotecas digitais que possam possibilitar comunicação do conhecimento dos trabalhadores por toda organização, porém para que isso ocorra são necessárias ferramentas, sistemas de navegação e também capacidade de proporcionar a criação de um fluxo de conhecimento. (BERNARDES, 2001, p. 33)

Segundo a proposta de BORGHOFF e PARESCHI[2] citado por BERNARDES (2001 p. 34) a arquitetura do gerenciamento do conhecimento é composta de quatro componentes:

a) fluxo do conhecimento – tido como meta fundamental do gerenciamento do conhecimento, sendo componente central do esquema de gerenciamento e a ponte que junta os outros três componentes. Ele suporta a interação entre o conhecimento tácito com o explícito;
b) cartografia do conhecimento – o conhecimento organizado necessita, para ser descrito, de um número grande de maneiras de pesquisar e agrupar os

[2] BORGHOFF, Uwe; PARESCHI, Remo. *Information technology for knowledgemanagemen*t. Germany : Springer-Verlag, 1998.

conhecimentos diversificados de uma organização, ou seja, é necessária uma ferramenta de mineração de dados, que seja capaz de atender a uma variedade de interesses do usuário. As ferramentas para cartografia do conhecimento serão capazes de mapear e catalogar o conhecimento da organização em todos os seus aspectos, desde a competência essencial das habilidades individuais, à prática da comunidade e de interesse na base de dados de clientes;

c) comunidade dos trabalhadores do conhecimento – o trabalho prático é um grande distribuidor de geração de conhecimento. Ele é levado a vários lugares, através da troca informal do conhecimento tácito entre colegas e companheiros. Para tratar esta situação, a organização pode contar com auxilio da tecnologia da informação proporcionando ajuda aos trabalhadores de forma que possam compartilhar seus arquivos eletrônicos;
d) repositórios e bibliotecas de conhecimento – as corporações têm percebido que documentos na forma eletrônica ou de papel contêm um grande valor explícito do conhecimento que pode ser eficientemente organizado através de mídia eletrônica para acesso e reuso inteligente. As tecnologias da informação, para suportarem esta arquitetura de memória corporativa, estão baseadas no gerenciamento de banco de dados, no gerenciamento de documentos e no processo de suporte de negócio.

Integrar todos os componentes do esquema proposto ainda é desafio para tecnologia da informação, pois as empresas tendem a utilizar repositórios e bibliotecas de conhecimento para gerar o fluxo de conhecimento, deixando de utilizar *software* para disponibiliza-lo e recupera-lo. É neste cenário que as técnicas de mineração de dados ganham importância, pois transforma imensos volumes de informação em conhecimento.

2.6 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS

A dificuldade em obter conhecimento útil de grandes volumes de dados faz com que haja a necessidade de se ter meios para o devido tratamento e extração de informações que possam vir a ter utilidade para uma organização. Essa necessidade vem fazendo com que técnicas e ferramentas sejam criadas e aprimoradas com o intuito de facilitar essa tarefa.

Neste cenário cresce a importância e o interesse pela Descoberta de Conhecimento em Bancos de Dados (DCBD), ou *Knowledge Discovery in Databases* (KDD) e uma de suas etapas em particular, a mineração de dados (*Data Mining*).

O processo de KDD inclui, além da mineração de dados, passos como a preparação inicial da informação, sua seleção, filtragem, introdução de conhecimento já obtido, interpretação e consolidação dos resultados (FAYYAD, et al. 1996, p. 28).

Para facilitar e compreender a mineração de dados é necessário esclarecer a descoberta de conhecimento em bases de dados – KDD. A definição aceita por diversos pesquisadores de mineração de dados foi elaborada por (FAYYAD, et al 1996 p. 30), "é um processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados".

AMARAL (2001 p. 20) facilita a compreensão esclarecendo cada um dos termos relevantes, para o processo de KDD:

a) dados: conjunto de fatos ou casos em um repositório de dados;
b) padrões: denota alguma abstração de um subconjunto dos dados em alguma linguagem descritiva de conceitos;
c) processo: a extração de conhecimento de base de dados envolve etapas como preparação dos dados, busca por padrões e avaliação do conhecimento;

d) válidos: os padrões descobertos devem possuir algum grau de certeza, ou seja, precisam satisfazer funções ou limiares que garantam que os exemplos cobertos e os casos relacionados ao padrão encontrado sejam aceitáveis;
e) novos: um padrão encontrado necessita fornecer novas informações sobre os dados. O grau de novidade serve para determinar quão novo ou inédito é um padrão; Pode ser medido por meio de comparações entre as mudanças ocorridas nos dados ou no conhecimento anterior;
f) úteis: os padrões descobertos devem ser incorporados para serem utilizados;
g) compreensíveis: um dos objetivos de realizar a mineração de dados é encontrar padrões descritos em alguma linguagem que pode ser compreendida pelos usuários permitindo uma análise mais profunda dos dados;
h) conhecimento: o conhecimento é definido em termos dependentes do domínio que estão relacionados fortemente com medidas de utilidade, originalidade e compreensão.

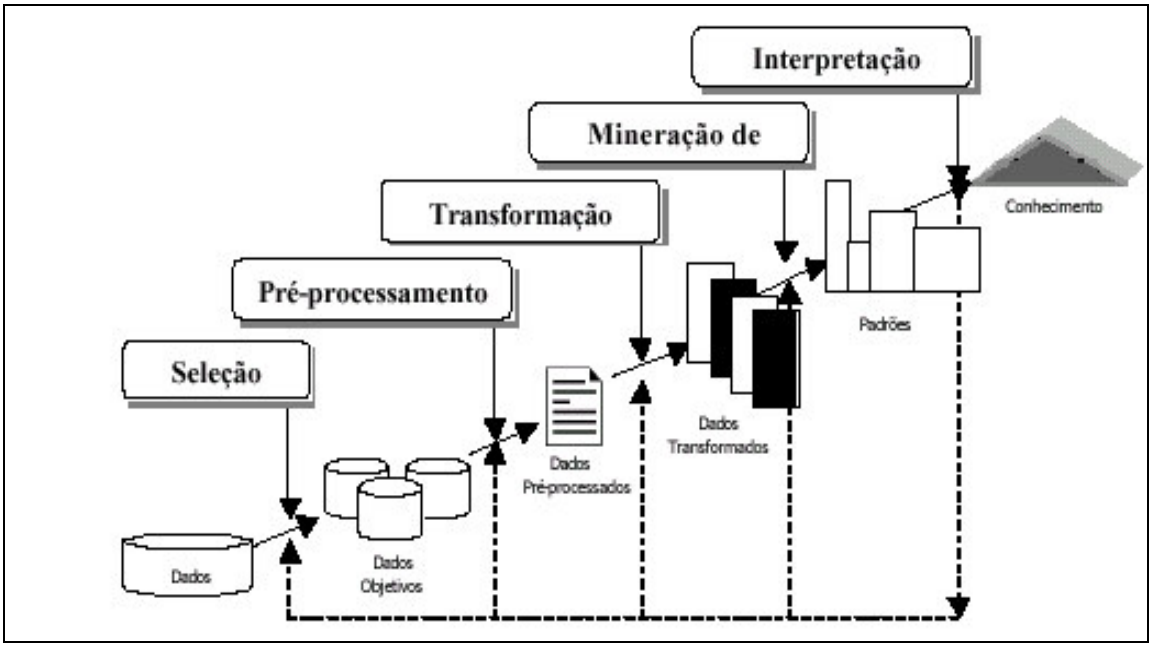
O processo KDD envolve duas grandes áreas de atividade com objetivos metas bem definidas:

a) preparação de dados que diz respeito ao entendimento da área de aplicação e da definição do conjunto de dados a serem submetidos à mineração;
b) mineração dados que é uma área específica do KDD que trata das técnicas e algoritmos utilizados na detecção dos padrões de dados.

Segundo FAYYAD(et al. 1996, p.30), o processo KDD mostra a interatividade das etapas e a interatividade do usuário ao processo. A cada etapa o

usuário analisa as informações geradas e procura incorporar sua experiência e tomar decisões para obter resultados cada vez melhores. O processo é composto de cinco etapas básicas conforme mostra a figura 2.1:

FIGURA 2.1 – ETAPAS DO PROCESSO KDD DEFINIDAS POR FAYYAD



FONTE: AMARAL, F.C.N. Introdução. In: _____ **Data Mining**: técnicas e aplicações para o marketing direto. São Paulo: Berkeley Brasil, 2001.p. 22

a) seleção dos dados: tem por objetivo selecionar um conjunto de dados, pertencentes a um domínio, para que, a partir de um critério definido pelo especialista do domínio, possa ser analisado. Após definido o objetivo, parte-se para a etapa de seleção dos dados, onde é feito um subconjunto de dados selecionados a partir da(s) base(s) de dados disponíveis. Este subconjunto conterá apenas aqueles dados relevantes para a solução do problema. O sucesso do processo depende da escolha correta dos dados que formam o conjunto de dados alvo, pois é neste subconjunto que, mais adiante no processo, serão aplicados o algoritmo para descoberta de conhecimento.

b) pré-processamento: após a seleção dos dados, inicia-se a etapa de pré-processamento dos dados. Nesta etapa serão realizadas tarefas que eliminem ou tratem os ruídos ou registros com dados ausentes. Para que a base de dados tornem-se consistente, faz-se necessário que os ruídos e os dados ausentes sejam eliminados ou, com o auxílio do especialista do domínio, tratados. Ruídos referem-se a situações em que dois ou mais registros, compostos por atributos que possuem os mesmos valores, mas que, no entanto, chegam a classes diferentes. Já dados ausentes, correspondem a registros que não possuem todos os valores dos atributos preenchidos. Em ambas situações, os registros redundantes ou mal formados devem ser eliminados, ou modificados, de tal forma que tenham a mesma classe ou todos seus valores preenchidos, respectivamente. A presença do especialista do domínio nesta etapa é muito relevante.
c) transformação: após serem pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos de aprendizado possam ser aplicados. Os dados devem estar no formato exigido pelo algoritmo escolhido na etapa de mineração.
d) mineração de dados: tem por objetivo construir hipóteses. Nesta etapa, é escolhido o método e são definidos os algoritmos que realizarão a busca pelo conhecimento implícito e útil do banco de dados. É a fase mais importante do processo de KDD onde dados são transformados em informação. Por isso, é importante que seja realizada quando os dados estiverem corretos e a tarefa seja adequada para alcançar o objetivo.
e) interpretação dos resultados: É a última etapa, onde é realizada a interpretação dos resultados obtidos após a aplicação do algoritmo minerador. Os resultados do processo de descoberta do conhecimento podem ser mostrados de forma que possibilite uma análise criteriosa para identificar a necessidade de retornar a qualquer uma das etapas

anteriores do processo de KDD, caso os resultados não sejam satisfatórios. É fundamental que esta etapa seja realizada em conjunto com o(s) especialista(s) do domínio para que a análise seja feita corretamente.

2.6.1 Mineração de Dados

A mineração de dados é a etapa do KDD onde ocorre efetivamente a descoberta de conhecimento. Nesta fase é escolhida a técnica que se deseja utilizar para construir o modelo para a descoberta de padrões. Esta escolha depende fundamentalmente do objetivo da aplicação.

Os objetivos básicos da mineração de dados são a predição e a descrição. A predição utiliza dados existentes na base de dados para previsão de valores desconhecidos ou futuros de outros dados de interesse. A descrição descobre padrões que descrevem os dados e são facilmente interpretáveis pelo usuário (FAYYAD et al. 1996, p 33). Para alcançar estes objetivos são utilizadas técnicas de modelagem.

2.6.2 Técnicas de modelagem

As técnicas de modelagem para mineração de dados utilizam algoritmos para gerar um modelo. Entre os algoritmos utilizados podem ser citados: árvores de decisão, regras de associação, rede neurais, algoritmos genéticos e técnicas estatísticas.

Segundo FAYYAD (et al. 1996, p. 31) “os algoritmos de mineração de dados consistem da composição de três componentes: o modelo, o critério de escolha, e o algoritmo de busca”.

Com relação ao modelo existem dois fatores relevantes: a função do modelo e a forma de representação do modelo. As funções mais comuns são:

a) classificação - a tarefa de classificação consiste em construir um modelo de algum tipo que possa ser aplicado a dados não classificados visando categorizá-los em classes;
b) regressão - mapeia os dados para uma variável de valor real. De modo análogo à classificação, atribui a cada objeto um valor de um atributo especial, chamado atributo-meta. Porém, o atributo-meta é numérico, podendo assumir valores contínuos; ao contrário da classificação, onde o atributo meta é categórico, podendo assumir apenas as classes pré-definidas;
c) agrupamento (*clustering*) - gera um modelo descritivo, que divide a base de dados em subconjuntos de dados com características em comum. Esta tarefa identifica grupos de registros correlatos, que são usados como ponto de partida para futuras explorações;
d) sumarização - fornece uma descrição compacta de um subgrupo da base de dados;
e) modelagem de dependência - descreve dependências significativas entre atributos. Esta tarefa é uma generalização da tarefa de classificação, no sentido que nesta tarefa há vários (mais de um) atributos-meta;
f) associação ou análise de relacionamento entre atributos - determina as relações entre os campos de uma base de dados. A premissa básica para associações é encontrar relevantes entre atributos de uma linha da tabela do banco de dados;
g) análise de sequência - modela padrões sequenciais com o objetivo de obter o estado do processo gerando a sequência ou para extrair e mostrar desvios e tendências.

As formas de representação dos modelos mais comuns incluem árvores de decisão, regras de decisão, modelos lineares, modelos não lineares, modelos baseados em exemplos e modelos de dependência gráfica probabilística.

Geralmente, os modelos mais complexos tendem a representar melhor os dados, mas também criam dificuldades no momento de interpretar os resultados. (FAYYAD, et al.1996, p. 32). É importante ressaltar as características desejáveis para um algoritmo de mineração de dados:

a) descoberta de conhecimento compreensível, normalmente definido como sendo regras expressas em um alto nível de abstração;
b) alto grau de autonomia, necessário para descobrir conhecimento previamente desconhecido pelo usuário;
c) eficiência no processo de descoberta de conhecimento, necessária para lidar com grandes bancos de dados; e
d) integração com bancos de dados genéricos, além da possibilidade de lidar com grandes bancos de dados.

No contexto deste trabalho, a tarefa de mineração de dados foi enfocada por descoberta de regras de associação. Para esse tipo de tarefa o algoritmo mais usado é o *Apriori* (AGRAWAL et al., 1993 p. 208), que faz uma varredura no conjunto de dados procurando por subconjuntos que tenham relacionamentos que sejam freqüentes.

2.6.3 Regras de Associação

A descoberta de regras de associação é uma das técnicas mais utilizadas na etapa de mineração de dados, que procura identificar padrões de dados em bases de dados de grande dimensão, permitindo, após a sua interpretação, adquirir conhecimento específico acerca do problema em análise, permitindo que diferentes decisores possam obter diferentes perspectivas da mesma informação. O estudo da descoberta de regras de associação foi introduzido por Agrawal, Imielinsky e Swami em maio de 1993 (AGRAWAL et al., 1993 p. 207).

A sua aplicabilidade prática às diferentes áreas de negócio das organizações em conjunto com a fácil compreensão que lhe é inerente, até mesmo para não perito em mineração de dados, tem feito das regras de associação um método extremamente popular.

Resumidamente uma regra de associação é definida como: "Se X então Y" ou "X Y", onde X e Y são conjuntos de elementos e X Y = Ø, ou seja, uma regra de associação é um relacionamento da forma $X \Rightarrow Y$, onde X e Y são conjuntos de itens e a interseção deles, $X \cap Y$, é o conjunto vazio. Diz-se que X é antecedente da regra, enquanto que Y é o conseqüente da mesma. Um algoritmo baseado em regras de associação consiste em descobrir esse tipo de regra entre os dados preparados para a mineração. (FREITAS, et al, 2001).

A associação ou afinidade de grupos visa a combinar itens importantes, tal que, a presença de um item em uma determinada transação pressupõe a de outro na mesma transação. (AGRAWAL, 1993 p. 207).

A regra de associação possui dois parâmetros básicos: o suporte e a confiança. Parâmetros que limitam a quantidade de regras a serem extraídas e descrevem a qualidade delas. O objetivo de um algoritmo para a descoberta de regras de associação é identificar todas aquelas que tenham suporte (Sup) e confiança (Conf) maiores do que os valores mínimos estipulados, onde o suporte é um número mínimo de ocorrências e a confiança é o percentual das transações que satisfazem X e Y (FREITAS et al, 2001).

Considerando que os conjuntos de itens X e Y estão sendo analisados, o suporte é definido como a fração de registros que satisfaz a união dos itens no conseqüente (Y) e no antecedente (X), correspondendo à significância estatística da regra, isto é, Sup = $|X \cup Y| / N$, onde N é número total de registros. A confiança é expressa pelo percentual de registros que satisfaz o antecedente (X) e o conseqüente (Y), isto é, $|X \cup Y| / |X|$ (AGRAWAL et al., 1993, p. 208).

AGRAWAL et al., (1996, p. 310) formalizou o problema da mineração de regras de associação que diz o seguinte:

> Seja L = {i1, i2, ..., in} um conjunto de literais chamados itens. Seja D um conjunto de transações, onde cada transação T é um conjunto de itens tal que $T \subseteq$ L, Associado com cada transação está um atributo que a identifica unicamente, chamado *TI*D. Uma transação T contém X, sendo X um conjunto de itens em L, se $X \subseteq T$. Uma regra de associação é uma implicação do tipo $X \rightarrow Y$, onde $X \subset L$, $Y \subset L$ *e* $X \cap Y = \emptyset$. A regra $X \rightarrow Y$ é válida no conjunto de transações D com o grau de confiança c, se c % das transações em D que contêm Y. A regra $X \rightarrow Y$ tem suporte s em D, se s% das transações em D contêm $X \cup Y$. Um conjunto de X contendo k itens é chamado de um conjunto-de-k-itens. O conjunto de itens que aparece à esquerda do operador de implicação, no caso o X, é chamado de *antecedente* (ou *precedente*) da regra; já o conjunto que aparece à direita, no caso o Y, é chamado de *conseqüente.*

2.6.4 Decomposição da tarefa

A tarefa de mineração por todas as regras de associação em um banco de dados pode ser decomposta em dois passos segundo PARK, 1995 e ZAKI 1998 citado por PITONI[3]:

a) gerar todos os conjuntos de itens que tenham um suporte acima do suporte mínimo estabelecido. Estes são chamados de conjuntos de itens freqüentes;
b) gerar as regras de associação utilizando os conjuntos de itens freqüentes. Deve-se selecionar apenas as regras que possuam o grau de confiança mínimo.

A facilidade de interpretação das regras de associação, aliada a uma prática muito forte, incentivou inúmeros investigadores a desenvolverem algoritmos de

[3] PARK, Jong Soo; CHEN, Ming-Syan; Yu, Philip S. An Effective Hash-Based Algorithm for Mining Association Rules. In: ACM SIGMOD, 1995. **Proceedings.** p. 175-186.
ZAKI, Mohammed J.;OGIHARA, Mitsunori. Theorical Foundations of Association Rules. In:3rd SIGMOD'98 WORKSHOP ON RESEARCH ISSUES IN DATA MINING & KNOWLEDGE DISCOVERY, 1998. **Proceedings.** Seattle.

descoberta de regras de associação. O algoritmo padrão atualmente mais utilizado é sem dúvida o *Apriori* sendo o mesmo utilizado para realização do estudo de caso.

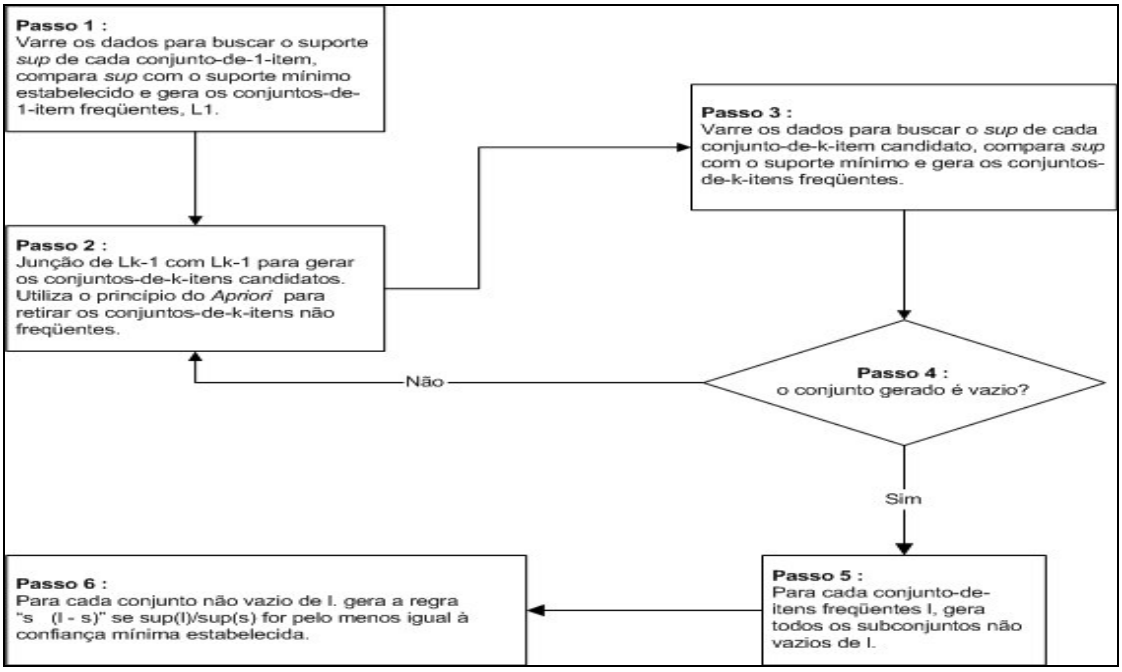
2.6.5 Algoritmo *Apriori*

O algoritmo *Apriori* é um dos mais conhecidos para encontrar grandes conjuntos de itens em bancos de dados de transações. Ele utiliza os conjuntos de itens de tamanho k para gerar os conjuntos de itens de tamanho (k + 1). O primeiro passo do algoritmo é encontrar os conjuntos de itens com 1 item. Este conjunto é denominado L_1. O conjunto L_1 é usado para gerar L_2, que representa os conjuntos de itens com 2 itens e assim por diante até que nenhum conjunto de itens possa ser gerado. PITONI (2002 p. 29) descreve o funcionamento do algoritmo *apriori* da seguinte forma:

> Ele faz diversas passagens sobre a base de transações para encontrar todos os conjuntos de itens freqüentes, sendo que, em cada um destes passos, primeiro gera um conjunto de itens candidatos, e então percorre a base de dados para determinar se os candidatos satisfazem o suporte mínimo estabelecido. Na primeira passagem, o suporte para cada item individual (conjunto-de-1-item) é contado e todos aqueles que satisfazem o suporte mínimo são selecionados. Estes são os conjuntos-de-1-item freqüentes. Na segunda iteração, conjuntos-de-2-itens candidatos são gerados pela junção dos conjuntos-de-1-item freqüentes e seus suportes são determinados pela pesquisa no banco de dados, sendo assim encontrados os conjuntos-de-2-itens freqüentes. O algoritmo prossegue iterativamente, até que o conjunto-de-*k*-itens freqüentes encontrado seja um conjunto vazio.

Para melhor visualização os passos descritos acima são ilustrados na figura 2.2.

FIGURA 2.2 – FLUXOGRAMA DO ALGORITMO *APRIORI*



FONTE: PITONI, M. R. **Mineração de Regras de Associação nos Canais de Informação do Direto.** Porto alegre, 2002. 59 f. Monografia – Curso de Ciência da Computação, Universidade federal do Rio Grande do Sul.

O algoritmo *Apriori* usa o princípio de que cada subconjunto de um conjunto de itens freqüente também deve ser freqüente. O objetivo é reduzir o número de candidatos a serem comparados com cada transação no banco de dados. Todos os candidatos gerados que contenham algum subconjunto que não seja freqüente são eliminados.

A tabela 2.3 a seguir mostra um exemplo, com dados transacionais e uma lista de cinco transações de um supermercado.

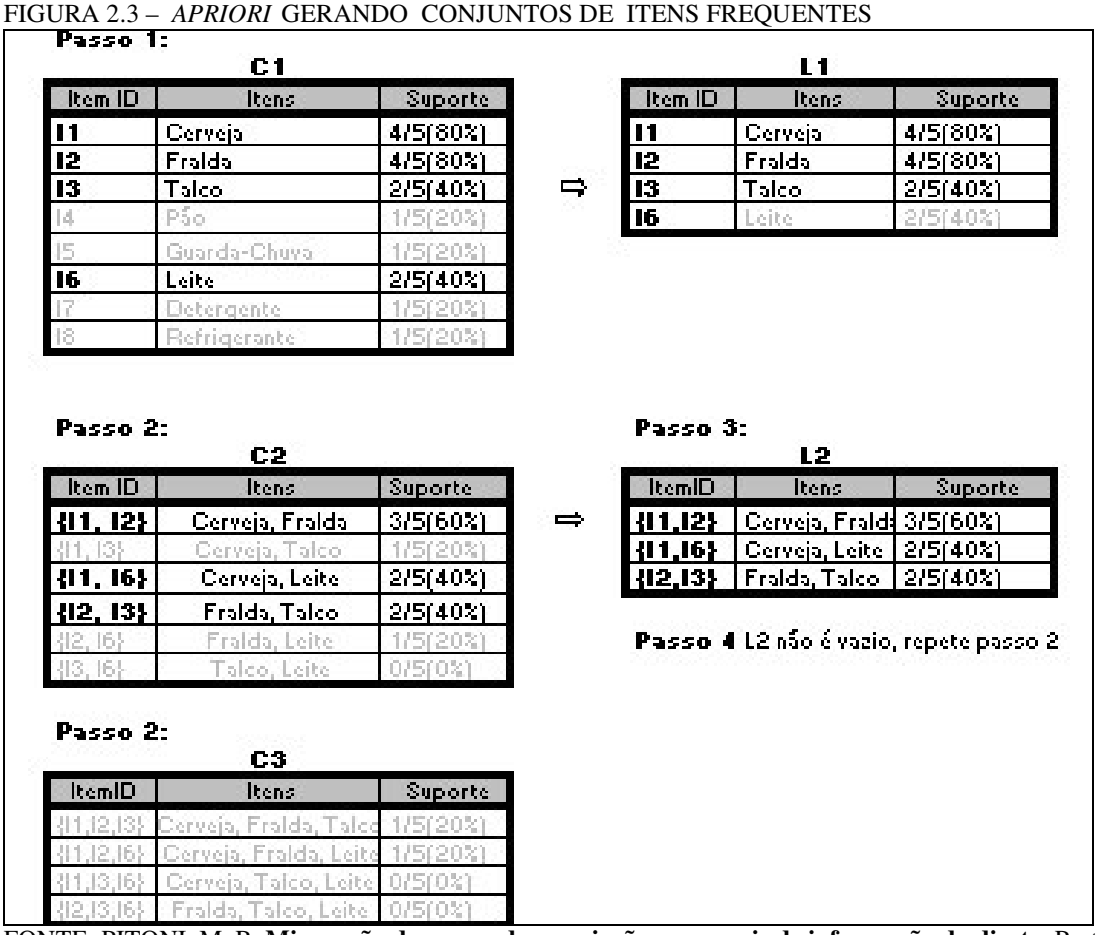
TABELA 2.3 – DADOS DE COMPRA EM UM SUPERMERCADO

TID	Lista de Itens
1	Cerveja, Fralda, Talco, Pão, Guarda-Chuva
2	Fralda, Talco
3	Cerveja, Fralda, Leite
4	Fralda, Cerveja, Detergente
5	Cerveja, Leite, Refrigerante

FONTE: PITONI, M. R. **Mineração de regras de associação nos canais de informação do direto**. Porto Alegre, 2002. 59 f. Monografia – Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul.

Nas figuras 2.3 e 2.4, simula-se a execução do algoritmo *Apriori* para lista de transações apresentada na tabela 2.3 Na primeira, o algoritmo gera os conjuntos de itens freqüentes com um suporte maior que o mínimo estabelecido de 40% (2/5). Já a segunda figura apresenta o algoritmo gerando as regras que possuem uma confiança mínima de 70%.

FIGURA 2.3 – *APRIORI* GERANDO CONJUNTOS DE ITENS FREQUENTES



Passo 1:

C1

Item ID	Itens	Suporte
I1	Cerveja	4/5(80%)
I2	Fralda	4/5(80%)
I3	Talco	2/5(40%)
I4	Pão	1/5(20%)
I5	Guarda-Chuva	1/5(20%)
I6	Leite	2/5(40%)
I7	Detergente	1/5(20%)
I8	Refrigerante	1/5(20%)

⇨

L1

Item ID	Itens	Suporte
I1	Cerveja	4/5(80%)
I2	Fralda	4/5(80%)
I3	Talco	2/5(40%)
I6	Leite	2/5(40%)

Passo 2:

C2

Item ID	Itens	Suporte
{I1, I2}	Cerveja, Fralda	3/5(60%)
{I1, I3}	Cerveja, Talco	1/5(20%)
{I1, I6}	Cerveja, Leite	2/5(40%)
{I2, I3}	Fralda, Talco	2/5(40%)
{I2, I6}	Fralda, Leite	1/5(20%)
{I3, I6}	Talco, Leite	0/5(0%)

⇨

Passo 3:

L2

ItemID	Itens	Suporte
{I1,I2}	Cerveja, Frald	3/5(60%)
{I1,I6}	Cerveja, Leite	2/5(40%)
{I2,I3}	Fralda, Talco	2/5(40%)

Passo 4 L2 não é vazio, repete passo 2

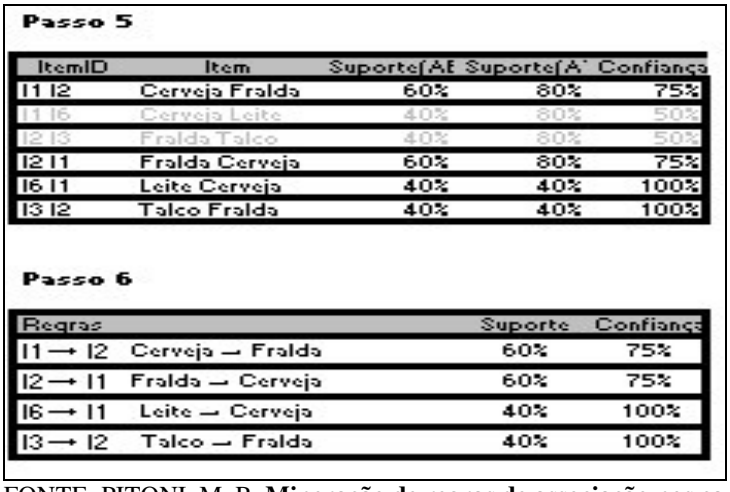
Passo 2:

C3

ItemID	Itens	Suporte
{I1,I2,I3}	Cerveja, Fralda, Talco	1/5(20%)
{I1,I2,I6}	Cerveja, Fralda, Leite	1/5(20%)
{I1,I3,I6}	Cerveja, Talco, Leite	0/5(0%)
{I2,I3,I6}	Fralda, Talco, Leite	0/5(0%)

FONTE: PITONI, M. R. **Mineração de regras de associação nos canais de informação do direto**. Porto Alegre, 2002. 59 f. Monografia – Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul.

FIGURA 2.4 – *APRIORI* GERANDO AS REGRAS



Passo 5

ItemID	Item	Suporte(AE	Suporte(A	Confiança
I1 I2	Cerveja Fralda	60%	80%	75%
I1 I6	Cerveja Leite	40%	80%	50%
I2 I3	Fralda Talco	40%	80%	50%
I2 I1	Fralda Cerveja	60%	80%	75%
I6 I1	Leite Cerveja	40%	40%	100%
I3 I2	Talco Fralda	40%	40%	100%

Passo 6

Regras		Suporte	Confiança
I1 → I2	Cerveja → Fralda	60%	75%
I2 → I1	Fralda → Cerveja	60%	75%
I6 → I1	Leite → Cerveja	40%	100%
I3 → I2	Talco → Fralda	40%	100%

FONTE: PITONI, M. R. **Mineração de regras de associação nos canais de informação do direto**. Porto Alegre, 2002. 59 f. Monografia – Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul.

O algoritmo *Apriori* pode ser implementado através de uma ferramenta de mineração de dados, que seja capaz de oferecer suporte as várias etapas do processo de mineração. Buscou-se para realização do presente trabalho uma ferramenta que suprisse as necessidades da base a ser estudada bem como sua acessibilidade em um ambiente acadêmico. A ferramenta apropriada para análise dos dados é o pacote Weka (*Waikato Enviroment for Knowledge Analysis*).

2.6.6 Ferramenta utilizada

O progresso da área de KDD e sua utilização nos mais variados domínios e pelas mais diversas organizações têm motivado o desenvolvimento de várias ferramentas comerciais além da elaboração de muitos protótipos de pesquisa. (REZENDE, 2003, p. 328).

O desenvolvimento de ferramentas comerciais de mineração de dados tem como objetivo fornecer aos tomadores de decisão das organizações, que geralmente não são especialistas em mineração de dados, ferramentas intuitivas e amigáveis, que ofereçam suporte às várias etapas do processo de mineração e disponibilizem apoio para diversas técnicas e tarefas.

Diante dessas informações a ferramenta selecionada para realização do estudo de caso é o pacote WEKA (*Waikato Enviroment for Knowledge Analysis*). Weka está implementado na linguagem Java, que tem como principal característica ser portável, desta forma pode rodar nas mais variadas plataformas e aproveitando os benefícios de uma linguagem orientada a objetos como modularidade, encapsulamento, reutilização de códigos dentre outros; além disso é um software de domínio público estando disponível em http://www.cs.waikato.ac.nz/ml/weka/. (WEKA, 2004)

A Weka é formada por um conjunto de pacotes, que são: *attribute selection, classifiers, clustering, association rules, filters e estimators.* O pacote *core*, possui

algumas classes (attribute, Instance e Instances) utilizadas por todas as demais classes. Objetos da classe *attribute* representam os atributos, com nome, tipo e, se for um atributo nominal, os possíveis valores. A classe Instance armazena os valores dos atributos e os objetos da classe Instances contêm um conjunto ordenado de instâncias, isto é, a base de dados.

A base de dados analisada no presente trabalho contém informações sobre o saneamento básico no Brasil e será detalhada na próxima seção.

2.7 SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO - SNIS

A informação, quando trabalhada de forma organizada, objetiva e direcionada, transforma-se em um instrumento fundamental para a eficácia dos empreendimentos públicos. No âmbito da prestação dos serviços públicos, a sistematização da informação prioriza, como objetivos principais, subsidiar a formulação de políticas e o planejamento das ações, orientar a aplicação de recursos e investimentos e aperfeiçoar a gestão elevando os níveis de eficiência e eficácia. (SNIS, 2004)

É dentro desse contexto que foi concebido pelo Governo Federal, em 1995, o Sistema Nacional de Informações sobre Saneamento – SNIS que apóia-se em um banco de dados administrado na esfera federal, onde as informações são coletadas junto aos prestadores de serviços de água e esgotos, de caráter operacional, gerencial, financeiro e, ainda, informações sobre a qualidade dos serviços.

Esses dados contribuem para a regulação e o controle da prestação dos serviços e para a elevação dos níveis de eficiência e eficácia na gestão das entidades prestadoras dos serviços, por meio do conhecimento de sua realidade, orientando investimentos, custos e tarifas, bem como incentivando a participação da sociedade no controle da qualidade, monitorando e avaliando os efeitos das políticas públicas. (SNIS, 2004)

2.7.1 Saneamento Básico No Brasil

Até o final da década de 60, o saneamento básico no Brasil era de competência exclusivamente municipal. Todos os investimentos eram promovidos a nível municipal, não havendo uma política unificada de provimento financeiro, nem políticas tarifárias para o setor, seja de âmbito nacional, regional ou estadual. Os recursos financeiros, normalmente consignados nos orçamentos públicos, geralmente eram irrisórios, em relação aos de outros setores. Sistemas de tarifação irreal mantinham um desequilíbrio acentuado entre a demanda crescente e a oferta insuficiente desses serviços públicos. (ABICALIL, et, al. p.47)

Devido a esses fatores em 1971 o Banco Nacional de Habilitação (BNH), criou o Plano Nacional de Saneamento (Planasa), que ficou responsável por todo planejamento de investimento do setor; e criou as Companhias Estaduais de água e Saneamento Básico (CESBs)[4], onde cerca de 3.200 municípios de 4.100 aderiram ao plano, causando a centralização dos serviços de saneamento. (MOTTA, 2004, p.6).

Porém no final da década de 80, o altamente centralizado sistema Planasa começou a apresentar um baixo desempenho, devido a um ambiente hiperflacionário e a inadimplência nos fundos para investimentos. As entidades representativas dos organismos e profissionais do setor, como a Associação Nacional dos Serviços Municipais de Saneamento (ASSEMAE), Associação das Empresas de Saneamento Básico Estaduais (AESBE), Associação Brasileira de Engenharia Sanitária e Ambiental (ABES) entre outras, passaram a debater entre si, e a discutir com a sociedade e o governo federal, a criação de um novo modelo de organização institucional para o saneamento urbano. O novo quadro institucional que se desenvolveu com a reforma constitucional de 1988 foi marcado pela

[4] "CAER/RR; CAERD/RO; CAESA/AP; COSAMA/AM; COSANPA/PA; DEAS/AC; SANEATINS/TO; AGESPI/PI; CAEMA/MA; CAERN/RN; CAGE/CE; CAGEPA/PB; CASAL/AL; COMPESA/PE; DESO/SE; EMBA/BA; CEDAE/ES; COPASA/MG; SABESP/SP; CASAN/SC; CORSAN/RS, SANEPAR/PR; CAESB/DF; SANEAGO/GO; SANESUL/MS". (ANA, 2004)

descentralização político-administrativa e fiscal, e a sua ênfase na descentralização tornou o esquema da Planasa obsoleto. A constituição então, declarou que os serviços públicos, incluindo água e saneamento, deveriam ser fornecidos pelas autoridades públicas, diretamente ou através de concessões[5], autorizando também os municípios a fazerem essas concessões. Assim, a descentralização, a flexibilização institucional e a desregulamentação, que implicam a abertura do setor a prestadores de serviços diversificados (empresas privadas, consórcio intermunicipais ou cooperativas, ao lado das CESBs e autarquias municipais), destacaram-se como diretrizes básicas reiteradas nas propostas para uma nova Política Nacional de Saneamento emanadas do governo federal a partir de 1990. (ABICALIL, et, al. p.131a)

Atualmente a prestação dos serviços pelos municípios ocorre, segundo três modelo concessão às Companhias Estaduais de Saneamento Básico (CESBs); operação direta pelos municípios que administram seus próprios serviços por meio dos departamento ou serviços de água e esgoto; e operações assistida pela Fundação Nacional de Saúde, onde os serviços prestados por autarquias municipais(serviços autônomos de água e esgoto) são administrado com assistência técnica daquele órgão federal. (ABICALIL, et al. p. 39b)

As CESBs tem menor presença na prestação de serviços de esgotamento sanitário. O serviço de esgotamento sanitário prestado pelas CESBs concentra-se nas capitais e nas maiores cidades de cada estado, o que explica o atendimento a 51% do total de habitantes servidos por redes coletoras de esgotos do País.(ANA, 2004)

Na indústria de saneamento básico ainda prevalece à organização definida pelo Planasa, onde os serviços estão concentrados nas 26 CESBs, conforme mostra a tabela 2. 4.

---

[5] "Transferência da execução do serviço, delegada a particulares, por ato administrativo bilateral". (TADINI, 1995, p.35)

TABELA 2.4 – DISTRIBUIÇÃO DOS PRESTADORES DE SERVIÇOS DE SANEAMENTO PARTICIPANTES DO DIAGNÓSTICO 2001, SEGUNDO CARACTERÍSTICA DO ATENDIMENTO

PRESTADOR DE SERVIÇO		POPULAÇÃO URBANA DOS MUNICÍPIOS ATENDIDOS		QUANTIDADE DOS MUNICÍPIOS ATENDIDOS	
Abrangência	Quant.	Água (GO6a) (milhões)	Esgoto (GO6b) (milhões)	Água (GO8)	Esgotos (GO9)
Regional	26	105,1	71,4	3.892	802
Microrregional	4	0,5	0,3	12	6
Local	230	23,0	21,1	230	127
Brasil	260	128,6	92,8	4.134	935

FONTE: SNIS 2001. Disponível em: < http://www.snis.gov.br/> Acesso em: 25 ago 2004.

NOTAS: GO6a: população urbana do(s) municípios(s) atendidos(s) pelo prestador de serviços com abastecimento de água. Em geral, é calculada a partir de projeções do Censo demográfico ou de dados e taxas de crescimento obtidos com a base nos últimos censos realizados pelo IBGE.
GO6b: população urbana do(s) município(s) atendido(s) pelo prestador de serviços com esgotamento sanitário. Em geral, é calculada a partir de projeções do Censo demográfico ou de dados e taxas de crescimento obtidos com base nos últimos censos realizados pelo IBGE.
GO8: quantidades de sedes municipais em que o prestador de serviços atua atendendo com o serviço o serviço de abastecimento de água.
GO9: quantidade de sedes municipais em que o prestador de serviço atua atendendo com o serviço de esgotamento sanitário.

As empresas municipais, a maior parte autarquias ou serviços municipais são responsáveis pela prestação de serviços no restante dos municípios brasileiros. Grande parte (64%) dos serviços municipais está concentrada na região Sudeste. Nas regiões Norte, Nordeste e Centro Oeste, representam apenas 26,5% do total. Em relação ao esgotamento sanitário, o nível de prestação de serviços pelas empresas municipais é elevado e, proporcionalmente, maior do que o relativo as CESBs. (ANA, 2004)

A participação do setor privado é ainda incipiente, pois se limita a cerca de 40 concessões municipais, plenas ou parciais, concentradas principalmente na região Sudeste e em cidades de porte médio. As maiores cidades com concessionários privados são Manaus (AM), Campo Grande (MS) e Niterói (RJ). A iniciativa privada participa também como acionista em empresas estaduais, como é o caso da

SANEPAR (PR) e da SANEATINS (TO). A participação do setor privado na prestação de serviços públicos tem os seguintes objetivos:

a) propiciar uma nova fonte de financiamento para ampliação da cobertura, complementar aos fundos atualmente utilizados;
b) aumentar a eficiência operacional e a capacidade financeira do setor;
c) agilizar a incorporação de novas tecnologias; e
d) elevar a qualidade dos serviços prestados.

Já é prática corrente no setor de prestação de serviços de saneamento a participação dos agentes privados em atividades de projetos, consultoria, construção e terceirização de serviços, tais como a operação de unidades de tratamento e bombeamento, a leitura de hidrômetros e entrega de contas, o corte e religação de conexões domiciliares e até atividades de manutenção. (ABICALIL, et, al. 1997 p. 38)

2.7.2 Indicadores de cobertura dos serviços

Embora quase 90% da população urbana do país já tenham acesso a água encanada, cerca de 15 milhões de pessoas ainda estão desprovidas deste serviço nas cidades brasileiras. Trata-se basicamente de populações de baixa renda que moram em assentamentos irregulares concentrados na periferia das grandes cidades, capitais e metrópoles, ou espalhadas em diversos municípios pobres de pequeno porte no interior. (ANA, 2004)

No esgotamento sanitário a situação é bastante crítica, apenas 56% dos domicílios urbanos têm acesso a redes públicas. Além das desigualdades de acesso em relação aos sistemas de água e esgotos a situação do atendimento dos serviços de saneamento no Brasil, apresenta grandes desigualdades regionais, que podem ser verificadas na Tabela 2.5, a seguir: (ANA, 2004)

TABELA 2.5 – CARACTERIZAÇÃO DA COBERTURA DOS SERVIÇOS DE SANEAMENTO NO BRASIL

Indicadores	Norte	Nordeste	Centro-oeste	Sudeste	Sul	Brasil
Abastecimento de Água						
Brasil – rede geral	48,01	66,39	73,19	88,33	80,06	77,82
Domicílios – rede geral	62,48	85,95	82,94	94,57	93,43	89,76
Domicílios rurais – rede geral	9,75	18,65	10,75	22,24	18,15	18,06
Esgotamento sanitário						
Brasil – rede de coleta	9,64	25,11	33,27	73,42	29,56	47,24
Brasil – rede + fossa séptica	35,62	37,95	40,79	82,33	63,78	62,2
Domicílios urbanos - rede de coleta	12,94	34,71	38,05	79,37	35,63	56,02
Domicílios urbanos - rede + fossa séptica	46,66	50,97	45,92	87,84	72,59	72,05
Domicílios rurais - rede de coleta	0,91	1,13	0,87	10,36	1,46	3,31
Domicílios rurais – rede + fossa sépticas	6,42	5,41	6,05	23,94	22,96	2,99

FONTE: ANA 2004. Disponível em: < http://www.ana.gov.br> Acesso em 25 ago 2004.
NOTA: Extraído da base de dados do IBGE

A partir da Tabela 2.5 observa-se que a cobertura dos serviços de saneamento é maior nas áreas urbanas das regiões Sul e Sudeste. Desta observação, depreende-se que os déficits de atendimento estão concentrados nas áreas rurais e nas regiões mais pobres do país. Sabe-se, também, que os segmentos populacionais de mais baixa renda localizados nas periferias dos grandes centros urbanos carecem dos benefícios do saneamento básico. (ANA, 2004)

Conforme estatística do Sistema Nacional de Informações sobre Saneamento (SNIS) no diagnóstico realizado em 2001 o percentual da população atendida com abastecimento de água e esgoto sanitário no Brasil é de 92,4% e 50,9%, sendo média de esgoto tratado de apenas 25,6% conforme mostra a tabela 2.6.

TABELA 2.6 – NÍVEIS DE ATENDIMENTO URBANO COM ÁGUA E ESGOTO DOS PRESTADORES DE SERVIÇOS PARTICIPANTES DO DIAGNÓSTICO 2001, SEGUNDO ABRANGÊNCIA

Abrangência	Índice de atendimento urbano (%)		
	Água (I23)	Coleta de esgoto (I24)	Tratamento de esgotos gerados
Regional	91,1	38,3	29,8
Microrregional	86,0	3,1	2,1
Local	97,8	77,4	17,0
Brasil	92,4	50,9	25,6

FONTE: SNIS, 2001 Disponível em: < http://www.snis.gov.br/> Acesso em: 25 ago 2004.
NOTAS: (I23): população atendida com abastecimento de água/população urbana dos municípios atendidos com abastecimento de água = (AO1/GO6a).
(I24): população atendida com esgotamento sanitário/população urbana dos municípios atendidos com abastecimento de água = EO1/GO6a).
(I46): volume de esgoto tratado/volume de água consumido – volume de água tratada exportado = EO6/A10-A19.

No entanto quando analisada a cobertura dos serviços por classes de renda observa-se um padrão bastante regressivo. A tabela 2.7 indica que a população com renda inferior a 2 salários mínimos (SM) apresenta índice de cobertura abaixo da média nacional. Já a população com mais de 10 salários mínimos, apresentam por sua vez uma cobertura 25% maior na água e mais de 40% maior no esgoto que as classes mais baixas, de até 2 salários mínimos.

TABELA 2.7 – COBERTURA DOS SERVIÇOS DE SANEAMENTO POR CLASSE DE RENDA - 2000

	Brasil	Até 2 SM	2 – 5 SM	5 –10 SM	> 10SM
Água	77,8	67,4	86,1	91,1	92,6
Esgoto	47,2	32,4	55,6	67,1	75,9

FONTE: IBGE, Censo Demográfico de 2000. Tabela extraída de MOTTA, R. S. In: Questões regulatórias do setor de saneamento no Brasil, p. 4. disponível em http://www.fazenda.gov.br/seal/NT_Saneamento%20IPEA_seroa.MF.pdf

2.7.3 Desempenho das empresas de saneamento

O desempenho das empresas de saneamento é peça fundamental para desenvolvimento dos serviços de saneamento básico. Dentro do contexto de prestação de serviço há dois aspectos que deve-se levar em consideração em relação as empresas: abrangência espacial (regional, microrregional e local) ; e natureza da gestão (pública e privada).

Segundo MOTTA (2004, p. 20), “a disputa pelo poder concedente coloca a questão de as vantagens e desvantagens dos serviços de saneamento serem geridos municipalmente ou com abrangência territorial de forma mais ampla”. Observando os indicadores da Tabela 2.8 podemos comparar o desempenho das empresas quanto a esta abrangência espacial.

TABELA 2.8 – INDICADORES DE EMPRESAS DE SANEAMENTO POR ABRANGÊNCIA DO SERVIÇO

Indicadores	Abrangência do serviço								
	Regional			Microrregional			Local		
	Média	Desvio	n	Média	Desvio	n	Média	Desvio	n
Gastos com energia sobre despesas de exploração (%)	13	4	26	25	10	4	21	14	173
Despesas com pessoal próprio e serviço de terceiros por despesas de exploração	67	10	26	60	13	4	63	17	174
Salário médio dos empregados próprios (R$/Ano)	33.648	10.348	26	13.565	4.161	4	12.576	6.871	223
Investimentos com recursos não onerosos pelo investimento total (%)	33	34	22	5	9	4	5	17	165
Proporção de impostos das despesas totais com os serviços (%)	6	4	25	22	36	4	2	5	165
Margem de lucro da arrecadação (%)	(-12)	78	26	28	8	4	19	19	169
Margem de lucro da receita (%)	(-23)	70	26	13	56	5	15	20	170
Tarifa média de água (R$/m)	1,1	0,2	25	0,9	0,4	4	0,7	0,5	160
Tarifa média de esgoto (R$/m)	1	0,2	22	0,8	0,3	4	0,6	0,3	99
Dias de faturamento comprometidos com contas a receber	248	239	26	136	142	4	126	133	197
Quantidade de água produzida e esgoto coletado por quantidade equivalente de pessoal total (1000m/ano/empregado)	103	40	22	68	49	4	65	42	188
Quantidade de água produzida e esgoto coletado por despesas como pessoal próprio e serviços de terceiros (1000m/ano/empregado)	0,003	0,001	22	0,005	0,003	4	0,006	0,005	204
Quantidade de água produzida e esgoto coletado por despesas de exploração (1000m/r$/ano)	0,002	0,001	22	0,003	0,001	4	0,003	0,002	170
Índice de atendimento de água (%)	86	16	26	90	12	4	94	14	217
Índice de coleta de esgoto (%)	32	20	21	6	7	3	71	38	92
Volume de água produzida por extensão de rede água (1000m/ano/km)	36	24	25	25	19	4	28	21	209
Volume de esgoto coletado por extensão da rede de esgoto (1000m/ano/km)	29	19	21	9	6	3	23	17	100
Índice de perdas na distribuição (%)	48	14	24	32	19	4	32	19	155

FONTE: Tabela extraída de MOTTA, R. S. In: Questões regulatórias do setor de saneamento no Brasil, p. 4. disponível em http://www.fazenda.gov.br/seal/NT_Saneamento%20IPEA_seroa.MF.pdf

Com base nas informações da tabela 2.8, levantou-se as seguintes observações:

a) as regionais apresentam gastos muito mais baixos de energia elétrica por despesa operacional (DEX), de quase 40% a menos. Isso poderia estar

indicando subsídios indiretos das distribuidoras estaduais de energia elétrica. Observa-se que a proporção na DEX dos gastos com pessoal próprio e serviços de terceiros, por outro lado, não difere muito por tipo de abrangência espacial. Já o salário médio do pessoal próprio é mais que o dobro nas regionais;

b) a participação dos recursos não-onerosos nas regionais no total dos investimentos chega a 30%, sendo seis vezes maior que nas micro e locais;
c) os impostos são mais gravosos nas micro, estão em torno de 22% enquanto nas regionais são de quase 6% e, nas locais, de apenas 2%;
d) as margens de lucro sobre a arrecadação (receita operacional arrecadada e DEX) e as margens sobre a receita operacional total (receitas e despesas totais) são negativas para as regionais e positivas para as micro e locais;
e) as tarifas de água e esgoto das regionais são mais elevadas que as das micro e locais. A tarifa de esgoto das regionais chega a ser quase o dobro e para a água, quase 70% em relação às tarifas das locais, embora em todos os casos se observe uma dispersão bem ampla em torno das médias. De qualquer forma, a tarifa média de esgoto das regionais chega a ser quase o dobro e para a água quase 70% em relação às tarifas das locais.
f) os índices de inadimplência das regionais são muito elevados e chegam a comprometer 250 dias com contas a receber; o dobro do observado nas micro e locais.
g) em termos de produtividade técnica poderíamos analisar o volume da produção de água e de esgoto coletado por quantidade equivalente de pessoal. Tal equivalência de pessoal, todavia, é medida como os gastos de terceiros divididos pelo salário médio do pessoal próprio mais o número de empregados próprios e, assim, esta equivalência está

fortemente enviesada pelo salário médio do pessoal próprio que, quanto maior for, menor será o número equivalente de pessoal, logo induzindo maior produtividade às empresas que pagam mais ao seu próprio pessoal. Dessa forma, induz a maior produtividade técnica das regionais;

h) já este mesmo volume de água produzida e esgoto coletado dividido por despesas de pessoal (incluindo próprio e de terceiros) e pelo total das despesas operacionais (DEX) é muito menor nas regionais, ou seja, a produção física por real gasto nas locais e micro é muito maior que nas regionais. Assim, parece plausível supor que as micro e locais parecem ser mais eficientes que as regionais;
i) o índice de abastecimento de água, o quanto da população municipal é atendido, é maior nas locais. O índice de coleta de esgoto nas locais é duas vezes maior que o observado nas regionais;
j) a produção de água e esgoto por extensão de rede, que oferece um indicador do efeito escala, é, conforme esperado, maior nas regionais, em torno de 25%, tanto na água como no esgoto. Surpreendentemente, contudo, esse efeito não é muito maior em relação às micro. Entretanto, o elevado desvio-padrão desse indicador para as locais mostra que as economias de escala neste tipo de empresa podem ser até maiores que nas regionais;
k) os índices de perdas de distribuição, que indicam o quanto de água produzida não é consumido, são de quase 50% nas regionais e em torno de 30% nas micro e locais.

Observou-se que as regionais apresentam um nível de lucratividade menor, com salários médios e níveis tarifários mais elevados, altas perdas de distribuição e de inadimplência e menor cobertura de serviço. Geram muito menos produção por real gasto, embora tenham uma contribuição maior de recursos não onerosos para investimentos.

Em relação ao desempenho dos prestadores microrregionais e locais segundo tipo de gestão (pública ou privada) verificou-se os seguintes aspectos na tabela 2.9: MOTTA (2004, p. 23)

TABELA 2.9 – INDICADORES DE EMPRESAS DE SANEAMENTO MICRORREGIONAIS E LOCAIS

Indicadores	**Natureza Jurídica**					
	Privada			**Pública**		
	Média	**Desvio**	**n**	**Média**	**Desvio**	**n**
gastos com energia sobre despesas de exploração (%)	22	13	14	21	14	163
Despesas com pessoal próprio e serviço de terceiros por despesas de exploração	48	7	14	64	17	164
Salário médio dos empregados próprios (R$/Ano)	16.897	5.759	17	12.245	6.795	210
Investimentos com recursos não onerosos pelo investimentos total (%)	0	0	14	6	18	155
proporção de impostos das despesas totais com os serviços (%)	18	20	14	1	3	155
Margem de lucro da arrecadação (%)	32	11	14	18	19	159
Margem de lucro da receita (%)	15	33	14	15	21	154
Tarifa média de água (R$/m)	0,7	0,3	14	0,7	0,5	150
Tarifa média de esgoto (R$/m)	0,5	0,3	10	0,6	0,3	91
Dias de faturamento comprometidos com contas a receber	78	85	17	130	136	184
Quantidade de água produzida e esgoto coletado por quantidade equivalente de pessoal total (1000m/ano/empregado)	97	41	16	63	41	176
Quantidade de água produzida e esgoto coletado por despesas como pessoal próprio e serviços de terceiros (1000m/ano/empregado)	0,006	0,003	17	0,006	0,005	191
quantidade de água produzida e esgoto coletado por despesas de exploração (1000m/r$/ano)	0,003	0,001	14	0,003	0,002	160
Índice de atendimento de água (%)	84	17	17	95	13	204
Índice de coleta de esgoto (%)	53	31	14	72	40	81
Volume de água produzida por extensão de rede água (1000m/ano/km)	30	17	17	28	21	196
Volume de esgoto coletado por extensão da rede de esgoto (1000m/ano/km)	33	30	14	21	13	89
Índice de perdas na distribuição (%)	26	20	15	33	19	144

FONTE: SNIS Diagnóstico 2001; Tabela extraída de MOTTA, R. S. In: Questões regulatórias do setor de saneamento no Brasil, p. 4. disponível em http://www.fazenda.gov.br/seal/NT_Saneamento%20IPEA_seroa.MF.pdf

a) a relação dos gastos de energia com DEX das públicas e privadas é muito próxima;
b) já a proporção na DEX dos gastos com pessoal próprio e serviços é quase 30% maior nas públicas. Todavia, o salário médio do pessoal próprio nas empresas privadas é 30% maior que nas públicas;
c) a participação dos recursos não-onerosos é muito baixa nas duas categorias;
d) enquanto os impostos chegam a 18% das despesas totais nas empresas privadas, estes não passam de 1% nas públicas;
e) enquanto a margem de lucro sobre a receita é quase a mesma nas públicas e privadas, a margem sobre a arrecadação é 80% maior nas privadas. Isso talvez seja explicado, em parte, pela inadimplência, quando os dias de faturamento comprometido das públicas (136 dias), mesmo com ampla dispersão, são quase duas vezes maiores que nas empresas privadas;
f) os níveis das tarifas de água e de esgoto são muito próximos;
g) em termos de eficiência observa-se que o volume de água produzida e esgoto coletado dividido por despesas de pessoal (incluindo próprio e de terceiros) é praticamente o mesmo nas duas categorias. Já este volume por total das despesas operacionais (DEX) é um pouco maior nas públicas, em torno de 10%, talvez refletindo a sua menor carga tributária;
h) o índice de abastecimento de água é 10% maior nas públicas e o de coleta de esgoto é quase 40% maior que o das privadas;
i) produção de água por extensão de rede, efeito escala, é similar em ambas, embora com maior dispersão nas públicas. Todavia, no caso da rede de esgoto, as empresas privadas apresentam, na média, um volume de esgoto coletado 60% maior por quilômetro que as públicas, o qual pode estar enviesado pela baixa cobertura da rede de esgoto;

j) os índices de perdas de distribuição são maiores nas públicas (33%) que nas privadas (26%).

Resumindo, as empresas micro e locais públicas, em relação as suas equivalentes empresas privadas, apresentam um desempenho financeiro menos favorável devido aos seus altos índices de perdas de distribuição e de inadimplência, mesmo pagando salários médios menores que as privadas. Em termos de eficiência, ambas parecem gerar um nível equivalente de produção por real gasto. As tarifas de água e esgoto são muito próximas nos dois tipos de empresa. Os índices de cobertura de água e esgoto são ainda mais baixos nas privadas e o atendimento de metas de expansão dessas empresas pode explicar parte do desempenho mais dinâmico das inversões privadas. (MOTTA, 2004, p. 25)

Há uma tendência no setor de saneamento básico de ampliar a participação da iniciativa privada na prestação dos serviços de abastecimento de água e esgotamento sanitário.

O grande desafio do setor é a viabilização da realização dos investimentos necessários para ampliação e modernização gerencial e operacional do setor, visando universalizar o atendimento à população, tanto em água quanto em esgotamento sanitário.

3 METODOLOGIA

3.1 TIPOS DE PESQUISAS

Esta pesquisa é de caráter exploratório que, de acordo com MARCONI e LAKATOS (1991 p. 34) “tem como objetivo aumentar o conhecimento em um determinado assunto, familiarizar o pesquisador com conceitos e esclarecê-los”. Para o desenvolvimento deste trabalho, foi utilizada também pesquisa bibliográfica e estudo de caso.

A Pesquisa bibliográfica deu base para a aquisição de conhecimento acerca dos temas envolvidos no projeto, como, por exemplo, Gerenciamento do Conhecimento, Descoberta de Conhecimento em Bancos de Dados e Mineração de dados. Envolveu, basicamente, consultas a livros de referência, teses e artigos científicos.

A pesquisa caracterizou-se como estudo de caso devido à seleção de um objeto de estudo restrito conforme cita SANTOS (2002 p. 31). No estudo de caso analisou-se a base de dados disponibilizada pelo SNIS, com objetivo de aprofundar o conhecimento em relação aos dados desta base.

3.2 PROCEDIMENTOS METODOLÓGICOS

Realizou-se, inicialmente, uma pesquisa bibliográfica dos assuntos discutidos nas seções anteriores para adquirir um embasamento teórico. A primeira etapa da pesquisa consistiu-se em realizar um levantamento dos recursos teóricos tais como: livros, artigos e internet, documentos que sirvam como fonte de informação e que abordem todas as etapas da mineração de dados. A pesquisa na internet realizou-se através de *sites* de busca tais como: google e altavista.

Para levantar a literatura pertinente, na busca de livros, artigos e trabalhos foram realizadas visitas em bibliotecas universitárias tais como:

a) UFPR – Universidade Federal do Paraná;
b) UNICENP – Centro Universitário Positivo;
c) CEFET – Centro Federal De Educação Tecnológica Do Paraná;
d) PUC – Pontifícia Universidade Católica Do Paraná.

Foi necessário conhecer o campo da Descoberta de Conhecimento em Bancos de Dados, e assuntos que abordam a Mineração de Dados desde processos, modelos, técnicas utilizadas, entre outros. Necessitou-se, também, conhecer os

problemas relacionados ao Saneamento Básico, tais como: controle da prestação dos serviços, investimentos, custos e tarifas. Os assuntos pesquisados foram:

a) descoberta do conhecimento em base de dados;
b) mineração dados;
c) aplicações da mineração de dados;
d) técnicas e ferramentas da mineração de dados;
e) regras de associação;
f) algoritmo *apriori*;
g) saneamento básico.

Após a coleta de informações necessárias para o embasamento teórico, foi realizada a análise na base de dados, ou seja, o estudo de caso. Dentre as etapas pré-definidas da técnica de DCBD, foram realizadas:

a) primeira etapa – seleção; os dados estavam distribuídos em quatro planilhas, armazenadas no Microsoft Excel, com as seguintes informações: prestadores de serviço de abrangência local de direito privado com administração pública; prestadores de serviço de abrangência regional; prestadores de serviço de abrangência local – empresa privada; prestadores de serviços de abrangência microrregional. Selecionaram-se apenas os dados relevantes para mineração;
b) segunda etapa - pré-processamento, nesta fase foram eliminados erros, lacunas e ruído. Tanto os dados ruidosos como os dados inconsistentes foram modificados.
c) terceira etapa - a próxima etapa foi realizada a transformação, onde foi necessário inserir intervalos nos atributos a fim de padronizar o formato da base;

d) quarta etapa – para efetuar a mineração utilizou-se o *software Weka*, a técnica de mineração de dados escolhida foi regras de associação. Portanto, o algoritmo apropriado para mineração foi o *apriori*. Para concretizar a mineração, os dados foram transportados para Weka no formato aceito pelo software. O software Weka foi descrito na seção 2.6.6

e) quinta etapa - interpretação dos resultados da mineração dos dados, transformando os resultados obtidos através de regras, em conhecimento para tomada de decisão.

4 ESTUDO DE CASO

A base na qual foi realizado o estudo de caso é SNIS que apóia-se em um banco de dados administrado na esfera federal e contêm informações coletadas junto aos prestadores de serviços de água e esgotos, de caráter operacional, gerencial e financeiro, inclusive dados de balanço e, ainda, informações sobre a qualidade dos serviços, atualizados anualmente desde 1995. Para realizar a pesquisa utilizou-se à base de dados que contén informações do diagnóstico realizado no ano de 2001.

4.1 Seleção

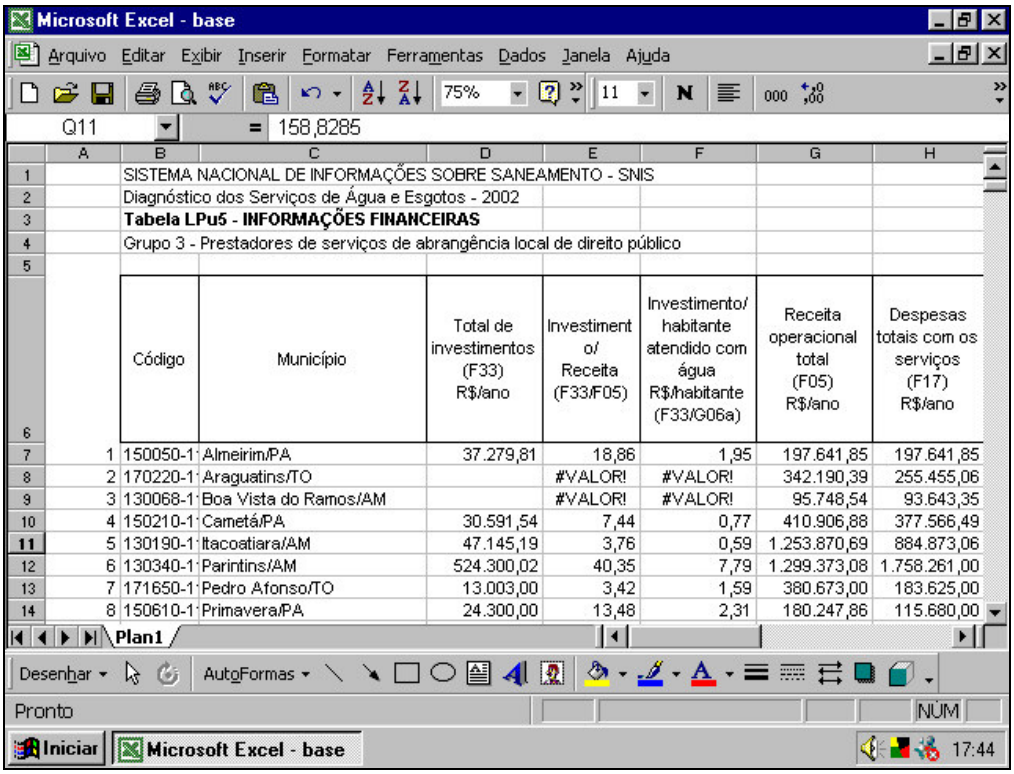
A seleção dos dados correspondeu à obtenção de dados que serviram de base para mineração. O conjunto de dados selecionado conteve apenas os dados relevantes ao processo de DCBD. Após análise das quatro bases selecionou-se os seguintes atributos:

a) municípios;
b) total de investimento r$/ano;

c) investimentos/receita;
d) investimento/habitante atendido com água r$/habitante;
e) receita operacional total r$/ano;
f) despesas totais com os serviços r$/ano;
g) receitas/despesas;
h) quantidade de empregados próprios empregado;
i) receita operacional total/quantidade de empregados próprios;
j) despesa com o serviço/m³faturado r$/m³;
k) tarifa média praticada r$/m3;
l) incidência da despesa pessoal + terceiros nas despesas totais de serviços;
m) despesa média anual/empregado r$/empregado;
n) receita/despesa/empregado (rendimento);
o) economias ativas por pessoal total/economia/empregado;
p) pessoal próprio/mil ligações de água/empregado/mil ligações;
q) pessoal próprio p/mil ligações /empregado /mil ligações;
r) índice de atendimento de água %;
s) índice de atendimento de esgoto %;
t) índice de tratamento de esgoto %;
u) margem despesa pessoal próprio %.

A etapa de seleção resulta num conjunto dos dados relevantes para a aplicação das técnicas de mineração de dados. Segue abaixo a figura 4.1 da base com o conjunto de dados alvo na planilha do Microsoft Excel:

FIGURA 4.1 – CONJUNTO DE DADOS



Microsoft Excel - base

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

Q11 = 158,8285

	A	B	C	D	E	F	G	H
1		SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO - SNIS						
2		Diagnóstico dos Serviços de Água e Esgotos - 2002						
3		**Tabela LPu5 - INFORMAÇÕES FINANCEIRAS**						
4		Grupo 3 - Prestadores de serviços de abrangência local de direito público						
5								
6		Código	Município	Total de investimentos (F33) R$/ano	Investimento/ Receita (F33/F05)	Investimento/ habitante atendido com água R$/habitante (F33/G06a)	Receita operacional total (F05) R$/ano	Despesas totais com os serviços (F17) R$/ano
7	1	150050-1	Almeirim/PA	37.279,81	18,86	1,95	197.641,85	197.641,85
8	2	170220-1	Araguatins/TO		#VALOR!	#VALOR!	342.190,39	255.455,06
9	3	130068-1	Boa Vista do Ramos/AM		#VALOR!	#VALOR!	95.748,54	93.643,35
10	4	150210-1	Cametá/PA	30.591,54	7,44	0,77	410.906,88	377.566,49
11	5	130190-1	Itacoatiara/AM	47.145,19	3,76	0,59	1.253.870,69	884.873,06
12	6	130340-1	Parintins/AM	524.300,02	40,35	7,79	1.299.373,08	1.758.261,00
13	7	171650-1	Pedro Afonso/TO	13.003,00	3,42	1,59	380.673,00	183.625,00
14	8	150610-1	Primavera/PA	24.300,00	13,48	2,31	180.247,86	115.680,00

Plan1

Desenhar AutoFormas

Pronto NÚM

Iniciar Microsoft Excel - base 17:44

FONTE: SNIS, 2001 Disponível em: < http://www.snis.gov.br/> Acesso em: 25 ago 2004.

4.2 Pré-processamento

Este item é dedicado a avaliar as estratégias de limpeza e pré-processamento utilizados na base de dados do SNIS. Primeiramente avaliou-se o estado em que se encontrava a base de dados no momento em que se iniciou o processo de tratamento:

a) retirou da base os seis primeiros campos com informações irrelevantes para mineração de dados tais como: nome do sistema que disponibilizou a base; a data da realização do diagnóstico; nome da tabela; o grupo que está sendo estudado, o código e nome dos municípios.

b) o algoritmo *Apriori* utilizado no estudo de caso trabalha com atributos nominais. Portanto, foi necessário transformar atributos numéricos em atributos nominais.
c) foi estudado, a possibilidade de agrupar os valores possíveis dos atributo em intervalos classificados como A,B,C,D,E,F,V;
d) tanto os dados inconsistentes como os campos em brancos foram preenchidos com o intervalo V;

A descrição dos agrupamentos adotados durante o pré-processamento pode ser visualizada a seguir na tabela 4.1.

TABELA 4.1 – NOME, INTERVALO E VALORES DE CADA ATRIBUTO

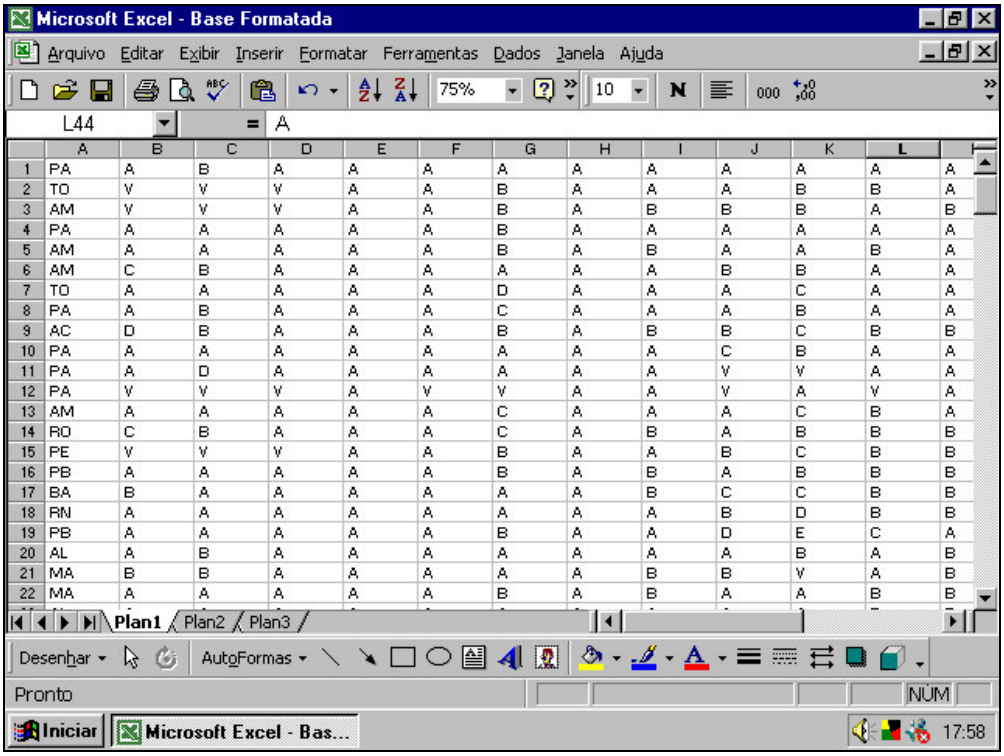
ATRIBUTOS	INTERVALOS e VALORES
Municípios	AC,AL,AM,BA,CE,ES,GO,MA,MG,MS,MT,PA,PB,PE,PI,PR,RJ, RN,RO,RS,SC,SE,SP,TO[6]
Total de investimentos	A (100000), B(100000-500000),C(500000-1000000), D(1000-10000000), E (10000000-20000000), F(20000000-40000000),V(VAZIO)
Investimento/Receita	A (10,19), B(10,19-51,51), C(51,51-106,88), D(106,88-1134,83), E (1134,83-900000000),V(VAZIO)
Investimentos/habitante atendido com água R$/Habitante	A (20,16),B(20,16-40,49), C(40,49-56,64), D(56,64-101,01), E (101,01-461,47),V(VAZIO)
Receita operacional total	A(10000000),B(10000000-50000000),C(50000000-100000000),D(100000000-150000000),E(150000000-200000000),V(VAZIO)
Despesas totais com os serviços	A(10000000),B(10000000-40000000),C(40000000-70000000),D(70000000-100000000),E(100000000-150000000),V(VAZIO)
Receita/Despesas	A(1000),B(1000-1483),C(1483-1983),D(1983-2153),E(2153-4489),V(VAZIO)
Quantidade de empregados próprios empregado	A(500),B(500-1000),C(1000-1500),D(1500-2000),E(2000-2500),V(VAZIO)
Receita operacional total /Quantidade de empregados próprios	A(30000),B(30000-80000),C(80000-130000),D(130000-200000),E(200000-800000),V(VAZIO)
Despesa com o serviço/m³ faturado R$ m³	A(0,30),B(0,30-0,60),C(0,60-0,90),D(0,90-1,50),E(1,50-28,00),V(VAZIO)
Tarifa média praticada R$ m³	A(0,40),B(0,40-0,80),C(0,80-1,20),D(1,20-1,60),E(1,60-155,00),V(VAZIO)
Incidência da despesa pessoal + terceiros nas despesas totais serviços	A(60),B(60-90),C(90-120),D(120-180),V(VAZIO)
Despesa média anual/empregado R$/empregado	A(10000),B(10000-20000),C(20000-30000),D(30000-40000),E(40000-50000),V(VAZIO)
Receita/Despesa/Empregado (rendimento)	A(5,00),B(5,00-10,00),C(10,00-20,00),D(20,00-30,00),E(30,00-90,00),V(VAZIO)
Economias ativas por pessoal total economia/empregado	A(200,00),B(200,00-400,00),C(400,00-600,00),D(600,00-800,00),E(800,00-1000,00),V(VAZIO)
Pessoal próprio p/mil ligações de água empregado/mil ligações	A(3,00),B(3,00-6,00),C(6,00-9,00),D(9,00-12,00),E(12,00-15,00),V(VAZIO)
Pessoal próprio p/mil ligações empregado/mil ligações	A(10),B(10-15),C(15-20),V(VAZIO)
Índice de atendimento de água %	A(3),B(3-6),C(6-9),D(9-12),V(VAZIO)
Índice de atendimento de esgoto %	A(10),B(10-40),C(40-70),D(70-10),V(VAZIO)
Índice tratamento de esgoto %	A,(10)B(10-40),C(40-70),D(70-100),E(100-130),V(VAZIO),
Margem despesa pessoal próprio %	A,(100)B,(100-250)C(250-500),D(500-750),V(VAZIO)

FONTE: O autor

[6] Para este atributo escolheu-se as siglas dos municípios como intervalo

Os intervalos foram resolvidos manualmente, padronizando assim o formato da base, conforme figura 4.2.

FIGURA 4.2 – ATRIBUTOS PADRONIZADOS



FONTE: O autor

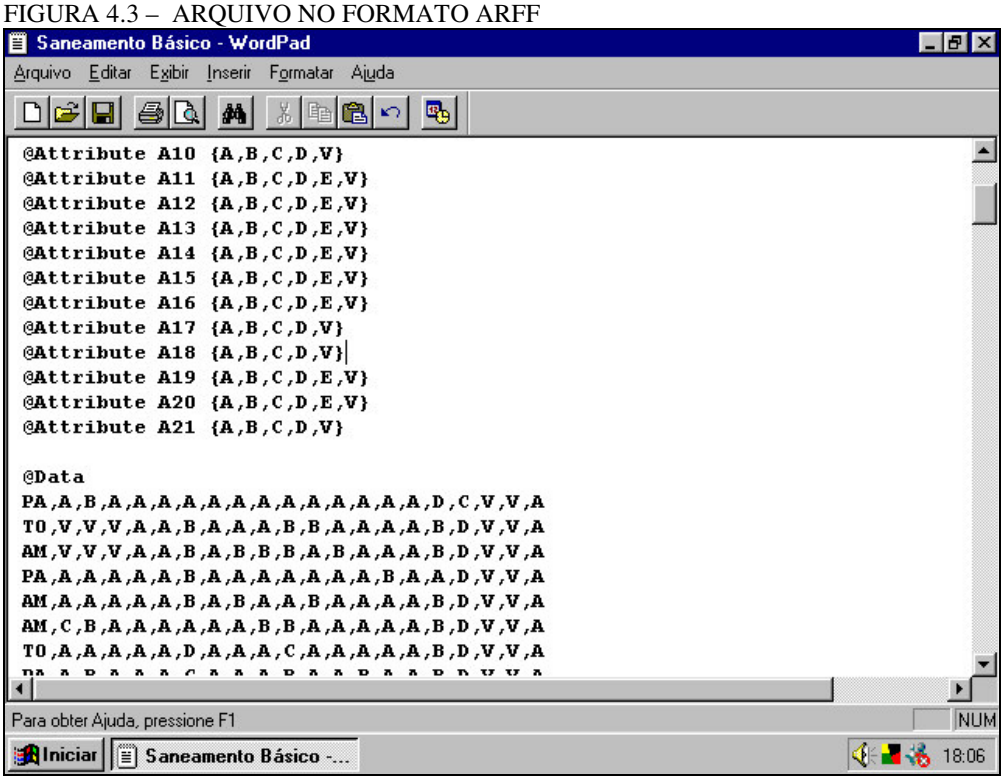
4.3 Transformação

Nesta fase podemos avaliar a estratégia do pré-processamento utilizada na base de dados do SNIS. A base estava em um arquivo Microsoft Excel que foi salvo como CSV[7], que separa as células da tabela com vírgula. A partir deste arquivo CSV foi introduzido cabeçalho adequado para transformá-la num arquivo do formato ARFF[8], que é o formato de arquivo aceito pelo *Weka*.

[7] Arquivo gerado pelo excel onde todas as células são separadas por virgula.
[8] Arquivo gerado pelo Weka [Weka].

O arquivo ARFF começa com o nome da relação identificado como @relation. A seguir vem um bloco definindo os atributos dos dados a minerar, identificados por uma linha começando por @attribute. Atributos nominais são seguidos pelo conjunto de valores que podem assumir, entre colchete, enquanto os numéricos são seguidos pela palavra chave numérica. Uma linha começando com @data indica que a partir da próxima linha estarão os dados, obedecendo à seqüência indicada pela definição dos atributos. A base de dados de SNIS gerou 21 atributos e 225 instâncias visualizados na a figura 4.3.

FIGURA 4.3 – ARQUIVO NO FORMATO ARFF



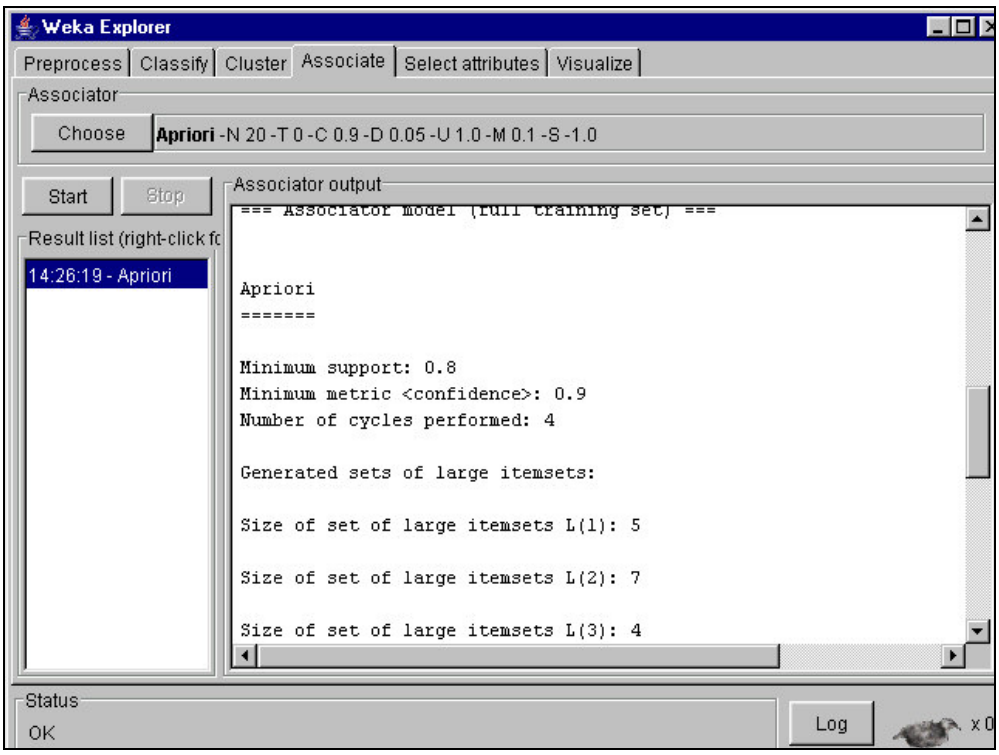
FONTE: O autor

4.4 Mineração

A mineração de dados é a etapa onde efetivamente são utilizados os métodos para a descoberta de conhecimento. Os métodos são definidos de acordo com o objetivo do estudo. Neste estudo de caso, o objetivo do processo focaliza a análise das informações financeiras das empresas privadas que prestam serviços de saneamento básico. Como entrada para esta etapa se tem a estrutura de dados construída pela etapa anterior, a saída consiste, por sua vez, em um conjunto de regras, que coincidem com os dados e critérios especificados pelo analista.

Foi sugerida a extração de regras de associação que explicassem o relacionamento entre os atributos conforme descritos a seguir. Segue abaixo a figura 4.4 que apresenta interface da ferramenta utilizada para mineração.

FIGURA 4.4 – INTERFACE DA FERRAMENTA *WEKA*



FONTE: O autor

4.4.1 Resultados Obtidos

A seguir é possível visualizar as regras geradas com o *Apriori* e a interpretação de duas regras:

Minimum support: 0.8
Minimum metric <confidence>: 0.9
Number of cycles performed: 4

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 7

Size of set of large itemsets L(3): 4

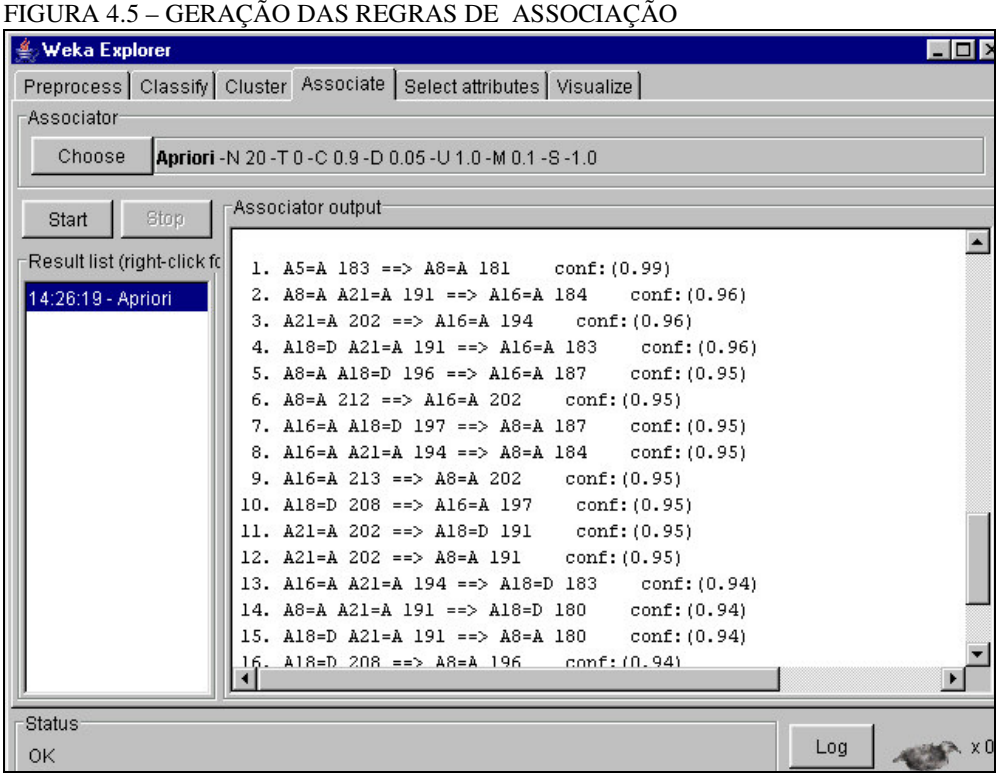
Best rules found:

1. A5=A 183 ==> A8=A 181 conf:(0.99)
2. A8=A A21=A 191 ==> A16=A 184 conf:(0.96)
3. A21=A 202 ==> A16=A 194 conf:(0.96)
4. A18=D A21=A 191 ==> A16=A 183 conf:(0.96)
5. A8=A A18=D 196 ==> A16=A 187 conf:(0.95)
6. A8=A 212 ==> A16=A 202 conf:(0.95)
7. A16=A A18=D 197 ==> A8=A 187 conf:(0.95)
8. A16=A A21=A 194 ==> A8=A 184 conf:(0.95)
9. A16=A 213 ==> A8=A 202 conf:(0.95)
10. A18=D 208 ==> A16=A 197 conf:(0.95)
11. A21=A 202 ==> A18=D 191 conf:(0.95)
12. A21=A 202 ==> A8=A 191 conf:(0.95)
13. A16=A A21=A 194 ==> A18=D 183 conf:(0.94)

14. A8=A A21=A 191 ==> A18=D 180 conf:(0.94)
15. A18=D A21=A 191 ==> A8=A 180 conf:(0.94)
16. A18=D 208 ==> A8=A 196 conf:(0.94)
17. A16=A A18=D 197 ==> A21=A 183 conf:(0.93)
18. A8=A A16=A 202 ==> A18=D 187 conf:(0.93)
19. A16=A 213 ==> A18=D 197 conf:(0.92)
20. A8=A 212 ==> A18=D 196 conf:(0.92)

No Weka a descrição das regras pode ser visualizada conforme a figura 4.5:

FIGURA 4.5 – GERAÇÃO DAS REGRAS DE ASSOCIAÇÃO



FONTE: O autor

4.4.2 Legenda

A5 – receita operacional total R$/Ano; Valor faturado anual decorrente das atividades-fim do prestador de serviços. Resultado da soma da Receita Operacional Direta[9] e da Receita Operacional Indireta[10]. R$/ano.
A8 – quantidade de empregados próprios empregados; Quantidade de empregados sejam funcionários do prestador de serviços, dirigentes ou outros, postos permanentemente – e com ônus – à disposição do prestador de serviços, ao final do ano.
A16 – pessoal próprio /mil ligações de água. Empregado/ mil ligações; Índice de Produtividade: Empregados Próprios por Mil Ligações de Água.
A18 – índice de atendimento de água %; população atendida com abastecimento de água/população urbana dos municípios atendidos com abastecimento de água.
A21 – margem despesa pessoal próprio %; despesas com o pessoal próprio/receita operacional direta (água + esgoto + água exportada).

4.5 Interpretação dos resultados obtidos

Para interpretação dos resultados foram selecionadas duas regras:

a) Regra 1: A5=A 183 ==> A8=A 181 conf:(0.99)
A8=A 181/225 sup: (0,80)

Tradução:

99% dos valores que ocorreram no intervalo A5 é acompanhada de A8, em 181 intervalos, ou seja, receita operacional total R$/Ano é

[9]Receita operacional direta total Valor faturado anual decorrente das atividades-fim do prestador de serviços, resultante exclusivamente da aplicação das tarifas. Resultado da soma da Receita Operacional Direta-Água Receita Operacional Direta-Esgoto e Receita Operacional Direta-Água Exportada.
[10] Receita operacional indireta Valor faturado anual decorrente da prestação de outros serviços vinculados aos serviços de água ou de esgotos, mas não contemplados na tarifação, como taxas de matrícula, ligações, religações, sanções, conservação e reparo de hidrômetros, acréscimos por impontualidade, entre outros.

acompanhada pela quantidade de empregados que sejam funcionários do prestador de serviços.

b) Regra 2: A8=A A21=A 191 ==> A16=A 184 conf:(0.96)
A16=A 184/225 sup(0,82)

Tradução:

96% dos valores que ocorreram simultaneamente em quantidade de empregados próprios que sejam funcionários dos prestadores de serviços e despesa com o pessoal próprio/receita operacional direta implica no índice de Produtividade: Empregados Próprios por Mil Ligações de Água.

Considerações:

A informação adicional contida nesta regra é a relação da quantidade de empregados próprios que sejam funcionários do prestador de serviços com as despesas com o pessoal próprio/receita operacional direta (água + esgoto + água exportada), em 191 das 225 transações processadas.

Utilizando a mineração chegamos a um certo conhecimento. Após algumas passagens pela base com o algoritmo *aprio*ri, verificou-se que a quantidade de empregados que sejam funcionários do prestador de serviços interfere no indicativo do Índice de Produtividade. O resultado poderá auxiliar a formulação de novos indicadores de desempenho para regulação dos serviços de saneamento.

5 CONSIDERAÇOES FINAIS

A mineração de dados e sua aplicação em grandes bancos de dados vêm sendo extensamente estudada, em virtude da dificuldade crescente de analisar grandes volumes de informações. Essa dificuldade nos mostra a necessidade de um processo para descobrir padrões desconhecidos em conjuntos de dados reais e transformar esses dados em conhecimento.

Ao realizar este estudo, percebeu-se a importância da mineração de dados na busca de padrões implícitos em bases de dados. Notou-se que é de extrema importância que o processo da descoberta do conhecimento passe por todas as suas etapas, pré-processamento, mineração e pós-processamento, formando assim uma boa base do conhecimento útil.

É importante ressaltar aqui alguns pontos que, durante a realização do trabalho foram possíveis identificar: a definição das etapas facilitou muito a condução do experimento, sendo determinante na obtenção do sucesso dos resultados; e o tratamento minucioso dos dados é muito importante para o sucesso da pesquisa.

As características levantadas com a pesquisa bibliográfica e as evidências colhidas com o estudo de caso fazem concluir que o foco do processo KDD está nas duas grandes áreas de atividade a preparação dos dados e a mineração de dados. Porém para o estudo foi dada ênfase à preparação de dados e padronização da base para concretizar a fase da mineração propriamente dita.

Com relação à preparação dos dados, verificou-se que na base de dados, existia um alto índice de presença de campos deixados em branco, isso acarretou um certo esforço para criar os intervalos para padronização, outra variável que dificultou a preparação foi os valores dos atributos, pois para cada atributo foi necessário criar um intervalo. Em suma, a limpeza da base foi à etapa mais demorada do processo, pois é etapa que requer muita atenção e cuidado para pequenos detalhes que podem

interferir no processo de mineração de dados. No entanto o processo de mineração de dados foi rápido. Porém, essa etapa requer entendimento na técnica utilizada, na escolha do algoritmo e na ferramenta usada, portanto para concretizar a mineração foi necessário um estudo sobre regras de associação, o algoritmo *apriori* e a ferramenta *Weka.*

O estudo de caso proposto descreve uma situação real da mineração de dados, aplicada a base de dados de domínio público o SNIS, onde, exercitou-se as situações adversas, tais como ausência de dados e ruídos, descritas nas cinco etapas padrão do processo KDD.

Com relação aos resultados obtidos, realizou-se o experimento em duas situações gerando 10 regras e 20 regras. Geras as regras buscou-se verificar um número maior de relações entre os atributos, um exemplo que pode ser citado é com relação ao indicativo receita operacional total R$/Ano, que conforme a regra é acompanhada pela quantidade de empregados que sejam funcionários do prestador de serviços, ou seja, a quantidade de empregados interfere na receita operacional. Portanto a geração de regras, a partir da aplicação da mineração de dados na base de dados do SNIS, foi de fundamental importância na identificação de relações entre os atributos.

O trabalho buscou descrever algumas técnicas importantes de mineração de dados, mostrando como cada uma pode contribuir para o processo de descoberta de padrões. Nota-se que as regras de associação utilizada para mineração expõem os resultados de maneira clara facilitando a compreensão até mesmo para não perito em mineração de dados.

A partir deste trabalho adquiriu-se maior habilidade em trabalhar com a ferramenta *Weka* que é muito utilizada no meio acadêmico. Essa ferramenta atendeu as necessidades do projeto mostrando completude em relação às técnicas de mineração de dados descritas na literatura.

Diante do referido estudo, deve-se ressaltar que o processo de mineração de dados vem se destacando nas áreas em que vem sendo empregada, devido a grande

necessidade existentes de analisar dados que estão dispostos de forma desordenada ou implícita dentro de uma base de dados e que se extraídos de forma correta, serão de grande utilidade.

Esta pesquisa foi de natureza exploratória visando familiarizar o assunto e aumentar o conhecimento referente ao processo KDD e a mineração de dados, realizando na prática o que se buscou na teoria.

Espera-se contribuir significativamente para pesquisa que está sendo desenvolvida na área de regulação dos serviços de saneamento no Brasil, já que este trabalho propõe o estudo inicial da aplicação de técnicas de descoberta de conhecimento através do estudo realizado na base do SNIS. Pretende-se também servir como referência, para os próximos trabalhos desenvolvidos na área de DCBD.

Quanto à continuidade do trabalho, deve-se enfatizar o estudo e aplicar outras técnicas na base estudada, possibilitando assim diversas visões e maneiras de extração de conhecimento e identificação de padrões, aumentando o suporte às decisões.

Sendo assim recomenda-se como trabalho futuros o estudo de outras técnicas e algoritmos para seram aplicados à base de dados do SNIS, buscando aprimorar os indicativos de desempenho das empresas prestadoras de serviços que podem ser retirados da base, a fim de auxiliar a tomada de decisão, seja de maneira gerencial ou operacional.

REFERÊNCIAS

ABICALIL, M.T. et al. Demanda, oferta e necessidade dos serviços de saneamento. **Série Modernização do Setor de Saneamento**, v. 4. Disponível em: http://www.snis.gov.br/pub_modernizacao.htm> Acesso em: 25 ago 2004.a

ABICALIL, M.T. et al. Diagnóstico do setor de saneamento: estudo econômico e financeiro. **Série Modernização do Setor de Saneamento,** v. 7. Disponível em: http://www.snis.gov.br/pub_modernizacao.htm Acesso 25 ago. 2004.b

AGRAWAL, R; IMIELINSKI, T.; SWAMI, A. **Mining Association Rules between Set of Itens in Large Databases**. In: ACM SIGMOD INT'L CONFERENCE ON MANAGEMENT OF DATA, 1993. **Proceedings.** Washington, p.207-216.

AGRAWAL, R. et al. Fast discovery of Association rules. In: **Advances in Knowledge Discovery and Data Mining.** Menlo Park: AAAI Press, 1996. 611 p. p.308-328.

AMARAL, F.C.N. Introdução. In: _____ **Data Mining**: técnicas e aplicações para o marketing direto. São Paulo: Berkeley Brasil, 2001.p. 10-18.

ANA - AGÊNCIA NACIONAL DE ÁGUAS. Disponível em:< http://www.ana.gov.br > Acesso em: 25 ago. 2004.

BERNARDES NETO, J. **Tecnologia da informação para o gerenciamento do conhecimento obtido das bases de dados de uma organização**. Florianópolis, 2001. 150 f. Dissertação (Mestrado em Engenharia da Produção) – Pós Graduação em engenharia da Produção, Universidade Federal de Santa Catarina. Disponível em: http://teses.eps.ufsc.br/defesa/pdf/7448.pdf

DAVENPORT, Thomas H.; PRUSAK, Laurence**. Ecologia da Informação.** Porque só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 2000**.**

DAVENPORT, T; PRUSAK, L. O que queremos dizer com conhecimento. In: _____ **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. Rio de Janeiro: Campus, 1998. p 1-28.

FAYYAD, Usama M.;PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An Overview. In:_____ **Advances in Knowledge Discovery and Data Mining.** Menlo Park: AAAI Press, 1996. 611 p. p.11-34.

FELIX, S. W. D. F. Introdução à gestão da informação. São Paulo: Alínea, 2003 p. 1-24.

FREITAS, O.G.; et al. Sistema de apoio à decisão usando a tecnologia data mining com estudo de caso da Universidade Estadual de Maringá. In: CONGRESSO BRASILEIRO DE COMPUTAÇÃO - CBComp 2001. Anais.

JAMIL, G. L. Sistemas de informações nos dias de hoje. In: _____**Repensando a TI na empresa moderna**. Rio janeiro: Axcel, 2001.p. 157-217.

MARCONI, M. A.; LAKATOS, E. M. Tipos de pesquisa. In: _____. **Fundamentos de metodologia científica.** 3. ed. São Paulo: Atlas, 1991.p. 30-39.

McGEE, J; PRUSAK, L. Informação e concorrência. In: _____**Gerenciamento estratégico da Informação:** aumente a competitividade de sua empresa utilizando a informação como uma ferramenta estratégica. Rio de Janeiro: Campus, 1994. p. 17-47.

MOTTA, R. S. Questões regulatórias do setor de saneamento no Brasil. disponível em : <http://www.fazenda.gov.br/seal/NT_Saneamento%20IPEA_seroa.MF.pdf> Acesso 30 ago. 2004.

NONAKA, I.; TAKEUCHI, H. **Criação do conhecimento na empresa** – como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro: Campus, 1997.

PITONI, M.R. **Mineração de regras de associação nos canais de informação do direto.** Porto Alegre, 2002. 59 f. Monografia. Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul. Disponível em:

http://www.inf.ufrgs.br/procpar/direto/trabalhos/Dissertacao_Pitoni.pdf Acesso em: 05 set. 2004.

REZENDE, O. S. **Sistemas Inteligentes**: fundamentos e aplicações. São Paulo, Manole, 2003.

SANTOS, A. R. Níveis e tipos de pesquisa. In: _____. **Metodologia científica a construção do conhecimento**. 5. ed. Rio de Janeiro: DP&A, 2002. p. 21-32.

SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO - SNIS. Disponível em: < http://www.snis.gov.br> Acesso em: 25 ago. 2004.

TADINI, V. Tendências recentes da regulação de serviços públicos no Brasil. In: ABICALIL, et al. **Proposta de Regulação da Prestação de Serviços de Saneamento.** Disponível em: < http://www.snis.gov.br/pub_modernizacao.htm> Acesso em: 25 ago. 2004.

Waikato Enviroment for Knowledge Analysis - WEKA. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/ Acesso em: 25 set. 2004.