UNIVERSIDADE FEDERAL DO PAMPA

MÁRCIO VERA DE ÁVILA

DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS DE GESTÃO PÚBLICA DO DEPARTAMENTO DE ÁGUA E ESGOTOS DE BAGÉ – DAEB

Bagé
2013

MÁRCIO VERA DE ÁVILA

DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS DE GESTÃO PÚBLICA DO DEPARTAMENTO DE ÁGUA E ESGOTOS DE BAGÉ – DAEB

Trabalho de conclusão de curso apresentado ao curso de Especialização em Sistemas Distribuídos com Ênfase em Banco de Dados da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista.

Orientador: Milton Roberto Heinen

Coorientador: Carlos Michel Betemps

Bagé
2013

MÁRCIO VERA DE ÁVILA

DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS DE GESTÃO PÚBLICA DO DEPARTAMENTO DE ÁGUA E ESGOTOS DE BAGÉ – DAEB

Trabalho de conclusão de curso apresentado ao curso de Especialização em Sistemas Distribuídos com Ênfase em Banco de Dados da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista.

Trabalho de Conclusão de Curso defendido e aprovado em: 08 de agosto de 2013.

Banca examinadora:

_____

Prof. Dr. Milton Roberto Heinen
Orientador
UNIPAMPA

_____

Prof. Dr. Sandro da Silva Camargo
UNIPAMPA

_____

Prof. Msc. Cristiano Cachapuz e Lima
UNIPAMPA

Dedico esse trabalho primeiramente a Deus pela inspiração, à minha esposa, à minha família, por todo o apoio e atenção. Ao professor Milton pela ajuda e orientação, e a todos os doutores, mestres e funcionários da UNIPAMPA que convivi durante o período do curso.

“A humildade é o primeiro degrau para a sabedoria.”

Tomás de Aquino

RESUMO

Através do Software de Gestão Municipal o Departamento de Água e Esgotos de Bagé detém uma base de dados com várias informações sobre todas as tarefas executadas em seus diversos setores. Esse grande volume de dados possui várias informações relevantes, porém a análise humana torna-se onerosa, e métodos tradicionais de recuperação de dados não são eficazes para obtenção de conhecimento ocultos em massas de dados. Assim a mineração de dados se apresenta como uma importante ferramenta para a descoberta de novos conhecimentos. Este conhecimento poderá servir de apoio para o desenvolvimento de importantes decisões para o desempenho de melhores práticas na administração da Autarquia. Este trabalho objetivamente irá analisar as questões financeiras das matrículas dos prédios municipais, sempre buscando a descoberta de novos conhecimentos.

Palavras-chave: Mineração de dados. Software de Gestão Municipal. Departamento de Água e Esgotos de Bagé.

ABSTRACT

The Management Software of Bage Municipal Department of Water and Sewerage holds a database with information of the various tasks performed in their various sectors. This large amount of data has various relevant information, but human analysis becomes costly and traditional methods of data retrieval are not effective to obtain knowledge hidden in masses of data. Thus data mining is an important tool for the discovery of new knowledge. This knowledge can serve as a support for the development of important decisions in the administration of the City. This work will analyze the financial issues of municipal buildings, always tring to discover new knowledge.

Keywords: Data Mining. Municipal Management Software. Bage Department of Water and Sewerage.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API - *Application Programming Interface* (Interface de Programação de Aplicativos)

BI - *Business Intelligence* (Inteligência de Negócios)

DAEB - *Departamento de Água e Esgotos de Bagé*

DB - *Data Bank* (Banco de Dados)

DCBD - *Descoberta de Conhecimento em Banco de Dados*

DNA - deoxyribonucleic acid (ácido desoxirribonucléico)

ETC - *Extract Transform Load* (Extração, Tratamento e Carga)

GUI - *Graphical User Interface* (Interface Gráfica do Usuário)

IPTU - *Imposto Predial Territorial Urbano*

ISSQN - *Imposto Sobre Serviço de Qualquer Natureza*

ITBI - *Imposto Sobre Transmissão de Bens Imóveis*

JRE - *Java Runtime Environment* (Ambiente de Tempo de Execução Java)

KDD - *Knowledge Discovery in Database* (Descoberta do Conhecimento em Banco de Dados).

LDO - *Lei de Diretrizes Orçamentárias*

OLAP - *On-line Analytical Processing* (Processamento Analítico Online)

PPA - *Plano Plurianual*

SIGM - *Sistema Informatizado de Gestão Municipal*

URL - *Uniform Resource Locator* (Localizador Padrão de Recursos)

WEKA - *Waikato Environment for Knowledge Analysis* (Waikato Ambiente para Análise de Conhecimento)

SUMÁRIO

1 INTRODUÇÃO

O avanço da tecnologia tem proporcionado ao homem a possibilidade de agregar o maior número de informações em sistemas informatizados. Um dos fatores que contribuíram para o armazenamento de grande massa de dados foi o barateamento do hardware, concomitantemente com a expansão das capacidades a níveis quase que inalcançáveis pelos pequenos usuários. Nesse sentido, as corporações passaram a armazenar uma grande quantidade de dados. Contudo, apesar dessas informações serem de grande valia, análises gerenciais que contribuam para a tomada de decisões são prejudicadas por não serem utilizadas ferramentas de mineração de dados que realizem a descoberta de novos conhecimentos através da análise destes dados. A mineração de dados surgiu para suprir essa necessidade, já que esta é uma das tecnologias mais promissoras para lidar com grandes quantidades de dados na atualidade. Isto é exemplificado pelo fato de que grandes corporações investem significativas quantias para fazer desta tecnologia uma forma de aumentar seus lucros.

Nesse contexto, a mineração de dados será aplicada ao banco de dados do Software de Gestão Municipal utilizado pelo Departamento de Água e Esgotos de Bagé, com o intuito de descobrir conhecimentos ocultos nessa grande massa de dados, que atualmente acumula mais de 200 *gigabytes* de informações Necessário esclarecer, que a Mineração de Dados será aplicada apenas aos dados financeiros do DAEB, não abragendo toda a base de dados.

Assim, a descoberta de conhecimento em bancos de dados surge como alternativa para auxiliar a descoberta automática de conhecimento através do processo completo de conversão de dados brutos em informações úteis (Tan, Steinbach & Kumar, 2009).

Diante disso, a principal etapa descoberta de conhecimento, onde são aplicados algoritmos inteligentes para extração de dados, é conhecida como mineração de dados.

Assim, espera-se que este trabalho possa trazer importantes informações, que sirvam para que o Departamento possa oferecer uma melhor prestação de serviços à comunidade bageense.

No auxílio desta pesquisa, será utilizada a ferramenta de mineração de dados WEKA (*Waikato Environment for Knowledge Analysis*), que através da aplicação de algoritmos, busca a descoberta de informações úteis.

Na abordagem deste relevante tema (mineração de dados), não se pretende esgotar o assunto, mas sim realizar uma pesquisa que traga de alguma forma benefícios à formação deste pesquisador, bem como possa ser descobertas informações que sejam de grande utilidade ao DAEB, para que ajudem o Departamento a alcançar os objetivos fundamentais.

1.1 Objetivos do Trabalho

1.1.1 Objetivo Geral

O objetivo deste trabalho é a aplicação da mineração de dados em banco de dados do Sistema Informatizado de Gestão Municipal (SIGM) do Departamento de Água e Esgotos de Bagé (DAEB), com a finalidade de extração de padrões não triviais.

O SIGM (e-cidade) é uma importante ferramenta para a administração pública do município, dentre outras finalidades, que atua principalmente na organização das atividades desenvolvidas pela administração pública, visto que ele contém funcionalidades que gerenciam rotinas corriqueiras do serviço público.

1.1.2 Objetivos Específicos

Quanto aos objetivos específicos, pode-se citar:

- Estudar e aplicar técnicas de mineração de dados aplicáveis à descoberta de conhecimento na base do SIGM (e-cidade).
- Sugerir a aplicação do conhecimento extraído, nas atividades desenvolvidas pelo DAEB.
- Demonstrar a importância da mineração de dados na administração pública do município.

1.2 Estrutura do Trabalho

O trabalho é dividido em mais 8 seções, detalhadas assim:

- Seção 2 (Mineração de Dados): apresenta definições, desafios, origens, tarefas e etapas da Mineração de Dados.
- Seção 3 (Dados): expõe sobre os tipos, qualidades e pré-processamentos dos dados.
- Seção 4 (WEKA): descreve como funciona e quais as principais características desta ferramenta.
- Seção 5 (SIGM): apresenta a estrutura e características do Sistema de Gestão Pública e-cidade.

- Seção 6 (Amostragem dos Dados Minerados): explica minuciosamente a extração dos dados do sistema e-cidade e quais os resultados desta extração.
- Seção 7 (Abordagem dos Dados): apresenta como funciona a tarefa de associação e explica a lógica dos algoritmos Apriori e Predictive Apriori.
- Seção 8 (Resultados): apresenta o resultados das análises realizadas nos dados e quais suas aplicações práticas.
- Seção 9 (Conclusão): apresenta as conclusões sobre o trabalho desenvolvido e aponta os trabalhos futuros a serem realizados.

2 MINERAÇÃO DE DADOS

No presente capítulo será abordado o processo utilizado para a descoberta de novos conhecimentos. Pretende-se, em uma abordagem sumária, conceituar, expor os desafios, a origem e as tarefas da Mineração de Dados.

2.1 O que é?

Análise de grande quantidade de dados procurando padrões e detectando relacionamentos entre informações, gerando novos subgrupos de dados.

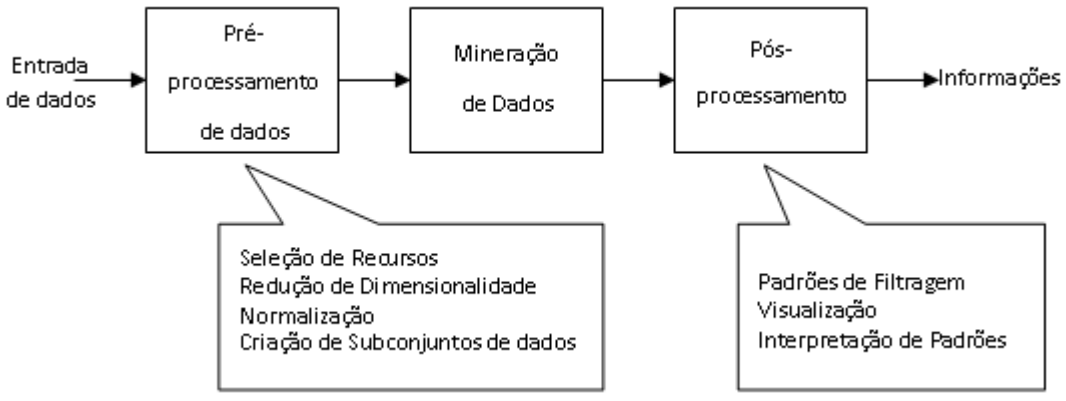
A mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados. Elas também fornecem capacidade de previsão do resultado de uma observação futura, como a previsão de que valores serão gastos por um cliente recém-chegado em uma loja de departamentos, por exemplo (Tan, Steinbach & Kumar, 2009).

A mineração de dados faz parte do processo de Descoberta de Conhecimento em Banco de Dados (KDD – *Knowlege Discovery in Databases*), que consiste em transformar dados brutos em informações úteis.

Segundo Goebel e Gruenwald (1999), o termo KDD é usado para representar o processo de tornar dados de baixo nível em conhecimento de alto nível, enquanto mineração de dados pode ser definida como a extração de padrões ou modelos de dados observados.

Na figura 01 a seguir pretende-se demonstrar como se dá este processo:

Figura 01 - Demonstra como se dá o processo de descoberta do conhecimento.



Fonte: (Tan, Steinbach & Kumar, 2009)

No pré-processamento os dados brutos de entrada são formatados nas análises subsequentes. Os passos envolvidos no pré-processamento de dados incluem a criação de novos subconjuntos de dados, oriundos de múltiplas fontes, a limpeza dos dados para remoção de ruídos, observações duplicadas, a seleção de registros e características que sejam relevantes à tarefa de mineração de dados. Por causa das muitas formas através das quais os dados podem ser coletados e armazenados, o pré-processamento de dados talvez seja o passo mais trabalhoso e demorado no processo geral de descoberta de conhecimento. (Tan, Steinbach & Kumar, 2009). Por este motivo, este assunto terá um capítulo específico, onde serão tratadas algumas questões relevantes ao pré-processamento.

Já no pós-processamento apenas os resultados válidos e úteis são incorporados ao sistema de apoio a decisões, com o fim de qualificar esta ferramenta.

2.2 Desafios, origens e tarefas

A mineração de dados surgiu devido a alguns desafios motivadores. Esses desafios dificultam a análise dos novos conjuntos de dados. Entre outros, pode-se citar: A escalabilidade; alta dimensionalidade; dados complexos e heterogêneos; propriedade e distribuição de dados; e Análises não tradicionais.

A escalabilidade: diante desta característica desejável a todo sistema, para manipular uma porção crescente de trabalho, de forma uniforme, resultando em conjunto de dados de grandes tamanhos, que chegam a ocupar *petabytes*, a mineração de dados permite analisar a manipulação destes dados, na busca do conhecimento. Com isso, grandes massas de dados, tornaram-se novos subconjuntos de dados, com padrões consistentes.

Alta dimensionalidade: neste grupo pode-se incluir os dados oriundos da bioinformática, que são responsáveis pela produção de conjuntos de dados de centenas ou milhares de atributos. Como exemplo cita-se um conjunto de dados que contenham medidas de temperaturas de diversos locais, se estas medidas forem feitas repetidamente por um período extenso, o número de dimensões aumenta na proporção do número de medidas realizadas, o que torna o conjunto de dados muito vasto para se analisar de forma simples. Para a obtenção de melhores resultados utiliza-se a mineração de dados.

Dados complexos e heterogêneos: os conjuntos de dados a serem analisados, possuem muitas vezes atributos do mesmo tipo, contínuos ou categorizados. Portanto as técnicas de mineração de dados devem evoluir para lidar com atributos heterogêneos. Com o

surgimento de dados mais complexos, como os dados de DNA com estrutura sequencial e tridimensional e dados sobre o clima que consistem em análises de temperatura, pressão, entre outros, sendo estes de diversos locais da superfície da Terra, a mineração de dados se torna ainda mais importante.

Portanto, na análise destes dados é importante que se leve em consideração os relacionamentos nos dados, como auto correlação temporal e espacial (Tan, Steinbach & Kumar, 2009).

Propriedade e distribuição de dados: muitas vezes os dados que deverão ser analisados estão em locais diferentes, ou não são de propriedade de uma organização apenas. Portanto, estão distribuídos geograficamente entre fontes pertencentes a múltiplas entidades. A aplicação de mineração de dados nestes casos se faz necessária com o desenvolvimento de técnicas distribuídas, como por exemplo, a ferramenta Grid Weka.

Nestas aplicações, a mineração de dados se dá com a utilização de algoritmos distribuídos, que encontram vários desafios, entre eles a tentativa de redução da quantidade de comunicação necessária para realizar a computação distribuída. Outro desafio, diz respeito a como consolidar de maneira eficaz os resultados da mineração de dados, já que são extraídos de múltiplas fontes. Por último, resta enfrentar a questão de segurança de dados.

Análises não tradicionais: Os conjuntos de dados analisados na mineração de dados são geralmente o resultado de um experimento projetado com cuidado e muitas vezes amostras oportunistas dos dados, em vez de amostras aleatórias. Portanto, as tarefas atuais de análise de dados muitas vezes requerem a geração e a avaliação de milhares de hipóteses, tendo como consequência a motivação de se automatizar o processo de geração e avaliação de hipóteses.

Por todos estes desafios, a mineração de dados surgiu devido à necessidade de lidar com estes diversos tipos de dados. Tendo se construído sobre a metodologia e algoritmos, que atraem a ideia como amostragem, estimativa e teste de hipóteses, oriundo de estatísticas e algoritmos de busca, técnicas de modelagem e teorias de aprendizagem da inteligência artificial, reconhecimento de padrões e aprendizagem de máquina.

Portanto, a mineração de dados surgiu para vencer estes desafios que através da evolução dos sistemas de armazenagem de dados foram propostos ao homem, visto que, desde o surgimento dos sistemas computacionais, um dos principais objetivos das organizações é o de armazenar dados. Ultimamente isto fica mais evidente tendo em vista a

queda dos custos para aquisição de hardware, tornando possível o armazenamento massivo de dados.

No que se refere às tarefas da mineração de dados, pode-se classificar em duas categorias (Tan, Steinbach & Kumar, 2009):

● Tarefas de Previsão: tem por objetivo prever o valor de um determinado atributo baseado em valores de outros atributos, sendo o atributo de predição chamado de *variável dependente* ou *alvo,* já os atributos utilizados para que se faça a previsão são chamados de *variáveis independentes ou explicativas.* Como exemplo, cita-se:

◦ Predizer o valor de uma ação dois meses adiante;

◦ Predizer o percentual de aumento de tráfego na rede se for aumentada a velocidade;

◦ Predizer o índice de alunos que irão concluir o curso de graduação em uma universidade.

● Tarefas Descritivas: tem por finalidade descrever os padrões e tendências revelados pelos dados. São muito utilizadas em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.

Destas duas categorias (Tarefas de Previsão e Tarefas Descritivas) podemos destacar quatro tarefas centrais da mineração de dados, que são: Modelagem de Previsão, Análise de Associação, Análise de Grupo e Detecção de Anomalias. Diante do exposto, passa-se a análise de cada uma destas tarefas.

Modelagem de previsão: visa construir um modelo para a variável alvo, como uma função das variáveis explicativas. Existem dois tipos de tarefas de modelagem de previsão: Classificação e Regressão. Classificação, a qual é usada para variáveis alvo discretas, e regressão, que é usada para variáveis alvo contínuas. Por exemplo, prever se um usuário *web* fará uma compra em uma livraria *online* é uma tarefa de classificação, porque a variável alvo é de valor binário. Por outro lado, prever o preço futuro de uma ação é uma tarefa de regressão, porque o preço é um atributo de valor contínuo. O objetivo de ambas as tarefas é aprender um modelo que minimize o erro entre os valores previsto e real da variável alvo. A modelagem de previsão pode ser usada para identificar clientes que responderão a uma campanha de vendas, prever perturbações no ecossistema da terra ou julgar se um paciente possui uma determinada doença baseado nos resultados de exames médicos (Tan, Steinbach & Kumar, 2009).

Análise de Associação: esta tarefa consiste em identificar quais atributos estão relacionados. É utilizada para descobrir padrões que descrevam características, altamente associadas dentro dos dados. É uma das tarefas mais conhecidas devido aos bons resultados obtidos. Um exemplo do uso da associação é a análise da "Cesta de Compras" *(market basket)*, onde se podem identificar quais produtos são em geral levados juntos pelos consumidores. Os padrões descobertos normalmente são representados na forma de regras de implicação ou subconjuntos de características.

O principal objetivo da análise de associação é extrair os padrões mais interessantes de forma eficiente, tendo em vista o tamanho exponencial do seu espaço de busca. Como exemplo, podemos citar a identificação de usuários de planos que respondem bem a oferta de novos serviços. Outro exemplo, já citado antes é a "Cesta de Compras".

Análise de grupo: visa identificar e aproximar registros similares. Um agrupamento é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa não tem a finalidade de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares. Conforme Tan, Steinbach & Kumar (2009), o agrupamento tem sido usado para juntar conjuntos de clientes relacionados, descobrir áreas do oceano que possuem um impacto significativo sobre o clima da terra e compactar dados, por exemplo.

Sua aplicação é das mais variadas possíveis: pesquisa de mercado, reconhecimento de padrões, processamento de imagens, análise de dados, segmentação de mercado, taxonomia de plantas e animais, pesquisas geográficas, classificação de documentos da Web, detecção de comportamentos atípicos (Camilo e Silva 2009).

Detecção de anomalias: É a tarefa de identificar observações cujas características sejam significativamente diferentes do resto dos dados. Tais observações são conhecidas como **anomalias** ou **fatores estranhos.** O objetivo de um algoritmo de detecção de anomalias é descobrir as anomalias verdadeiras e evitar que objetos normais sejam erroneamente rotulados como anômalos (Tan, Steinbach & Kumar, 2009).

2.3 Mineração de dados como parte da descoberta de conhecimento em banco de dados – DCBD

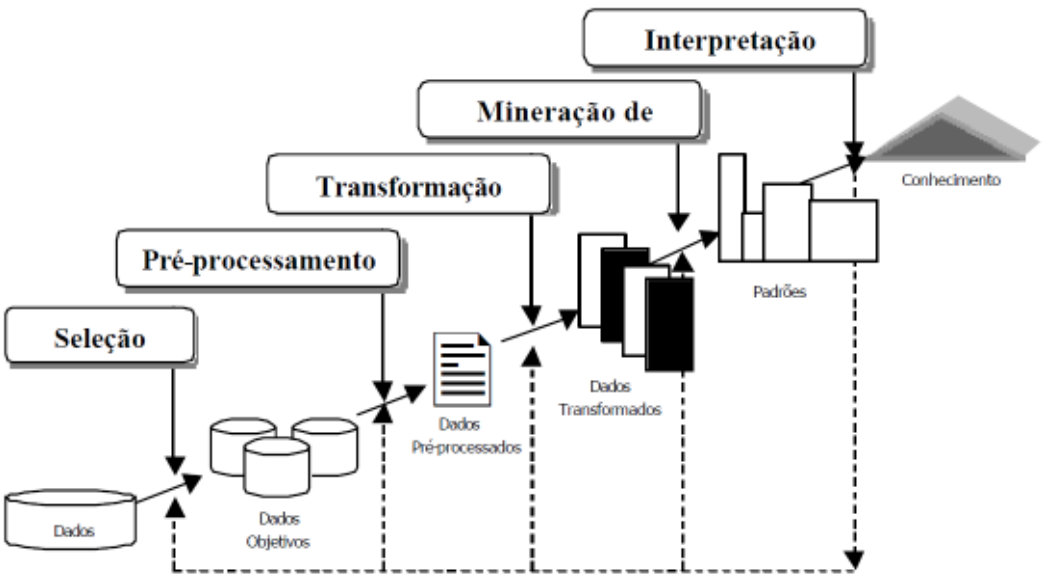
Freitas (1993) define que mineração de dados pode ser considerada como uma parte do processo de Descoberta do Conhecimento em Banco de Dados (KDD – *Knowledge Discovery in Databases*).

Como já referido, a quantidade de dados armazenados em um banco de dados que equivale a um volume potencial de informação sobre as atividades fins do DAEB, se estudadas separadamente podem não revelar tanto interesse, já que elas demonstram apenas a realidade de determinado Setor, ou ainda demonstram as atividades desenvolvidas pelo Setor para desempenhar sua função na Organização. Porém, a análise em conjunto destes dados, ou seja, o resultado da atuação de cada Setor do Departamento, registrando em um único banco de dados pode revelar detalhes interessantes, tornando possível a descoberta de novos padrões, disponibilizando informações úteis para que gere novos conhecimentos.

2.4 Etapas da mineração de dados segundo Fayyad

Esta abordagem define a iteratividade das etapas e a interatividade do usuário ao processo. A cada etapa o usuário analisa as informações geradas e procura incorporar sua experiência e tomar decisões para obter resultados cada vez melhores. O processo é composto de cinco etapas, como mostra a Figura 02 (Fayyad 1996).

Figura 02 - Etapas do processo de DCBD definidas por Fayyad



Fonte: (Fayyad 1996)

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

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

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

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

Mineração de dados (*Data Mining*): Etapa caracterizada pela busca de padrões nos dados. Nesta etapa, é escolhido o método e são definidos os algoritmos que realizarão a busca pelo conhecimento implícito e útil do banco de dados. É a fase mais importante do processo de DCBD onde dados são transformados em informação. Por isso, é importante que seja realizada quando os dados estiverem corretos e a tarefa seja adequada para alcançar o objetivo.

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

3 DADOS

Para uma mineração de dados satisfatória, é primordial que se conheça os tipos de dados a serem minerados. A boa técnica ensina que os dados são categorizados em dois tipos: quantitativos e qualitativos.

Cada categoria de dados possui um conjunto de dados com características, comuns ou especiais, que devem ser levadas em consideração para que se possa alcançar uma boa análise dos dados, tendo por consequência informações importantes.

Contudo, uma das primeiras atividades é obter uma visualização dos dados, para se ter uma visão geral, e assim, poder decidir quais as técnicas mais indicadas.

Nesse contexto os dados brutos podem-se tornar inapropriados para a análise, se fazendo necessário que sejam processados de forma que se adequem para oferecer melhores resultados. Por exemplo: o tamanho do imóvel, que nos dados a serem minerados no presente trabalho, é representado pela área do imóvel em metros quadrados. Na extração destes será necessária a aplicação de uma classificação em grupos com a seguinte nomenclatura: pequeno, médio e grande. Devendo-se aplicar esta técnica em outras informações que serão utilizadas na mineração dos dados abrangidos pelo presente trabalho.

3.1 Tipos de dados

Um **conjunto de dados** muitas vezes pode ser visto como uma coleção de **objetos de dados.** Outros nomes para um objeto de dados são *registros, ponteiros, vetores, padrões, eventos, casos, exemplos, observações ou entidades.* Por sua vez, objetos de dados são descritos por um número de atributos que capturam as características básicas de um objeto, como a massa de um objeto físico ou o tempo no qual um evento tenha ocorrido. Outros nomes para um atributo são *variável, característica, campo, recurso ou dimensão* (Tan, Steinbach & Kumar, 2009).

Assim, a pesquisa remete a análise do que é um atributo, já que este é o principal componente de um conjunto de dados.

Inicialmente, define-se que:

"Um atributo é uma propriedade ou característica de um objeto que pode variar seja de um objeto para outro ou de tempo para outro" (Tan, Steinbach & Kumar, 2009).

Diante disso, têm-se diferentes tipos de atributos, conforme se verifica na tabela abaixo:

Tabela 1 – Diferentes tipos de atributos

Tipo de Atributo		Descrição	Exemplos	Operações
Categorizados (Qualitativos)	Nominal	Os valores de um atributo nominal são apenas nomes diferentes; i.e., valores nominais fornecem apenas informação suficiente para distinguir um objeto de outro. (=,≠)	Códigos postais, números de ID de funcionário, cor dos olhos, sexo	Modo, entropia, correlação de contingência, teste x^2
	Ordinal	Os valores de um atributo ordinal fornecem informação suficiente para ordenar objetos. (>,<)	Dureza de minerais (boa, melhor, melhor de todas), notas, números de ruas	Medianas, porcentagens, testes de execução, testes de assinatura
Numéricos (Quantitativo)	Intervalar	Para atributos intervalares, as diferenças entre os valores são significativas, i.e., existe uma unidade de medida (+,-)	Datas de calendário, temperatura em Celsius ou Fahrenheit	Média, desvio padrão, correlação de Pearson, testes T e F
	Proporcional	Para variáveis proporcionais, tanto as diferenças quanto as proporções são significativas. (*,/)	Temperatura em Kelvin, quantidades monetárias, contadores, idades, mas, comprimento, corrente elétrica	Média geométrica, média harmônica, variação percentual

Fonte: Tan, Steinbach & Kumar (2009)

Como pode-se notar há uma divisão dos atributos em **qualitativos e quantitativos.**

Portanto, conforme se visualiza os atributos possuem propriedades de operações numéricas, que são usadas para descrevê-los:

- Distinção: = ou ≠
- Ordenação: <, ≤, > e ≥
- Adição: + e –
- Multiplicação: * e /

Com estas propriedades a tabela definiu quatro tipos de atributos: nominal, ordinal, intervalo e proporcional. Consequentemente cada tipo de atributo possui todas as operações que são válidas para o atributo nominal, que também são válidas para o atributo ordinal, e por consequência são válidas também para o atributo proporcional.

Segundo a análise da tabela proposta, temos que os atributos nominal e ordinal, são chamados coletivamente de atributos *categorizados ou qualitativos.*

Já os outros dois atributos, intervalar e proporcional são chamados coletivamente de atributos *quantitativos ou numéricos***.**

3.2 Qualidade dos Dados

Como na maioria das vezes os dados a serem analisados apresentam falhas, por serem incompletos ou duplicados, é necessário que na mineração de dados se utilize algoritmos que tolerem a baixa qualidade de dados.

Primeiramente, deve-se fazer a detecção e a correção dos dados, ou seja, limpeza dos dados. (Tan, Steinbach & Kumar, 2009).

Essas falhas podem ser produzidas por diversos fatores, seja devido a erro humano ou falha no processo de coleta dos dados. Valores faltando, são comumente encontrados em banco de dados, pois na maioria das vezes alguns dados não são informados, devido a sua pouca importância para o usuário. Porém, para a mineração de dados ele pode ser de grande valia. Contudo, é possível eliminar estes valores que faltam. Um exemplo que podemos referir, sendo inclusive bem interessante e já entrando na parte prática do trabalho, é que no Cadastro Geral do Município, muitos registros não dispõem o sexo informado, ficando o contribuinte sem essa informação no banco de dados, porém para a extração dos dados, haverá a tentativa de preencher este atributo de forma automática ao se analisar o nome da pessoa.

Cumpre salientar, que mesmo assim poderão ser extraídos dados com informações do sexo faltando, o que não trará grandes problemas para a pesquisa do presente trabalho.

Outro problema, que poderá ocorrer é a existência de dados duplicados, onde é preciso ter muito cuidado, pois as vezes os dados podem por exemplo possuir homônimos, que apesar de aparentarem ser da mesma pessoa, eles devem ser levados em consideração pois referem-se a pessoas diferentes.

3.3 Pré-processamento de dados

Como se constata, o presente trabalho tem por finalidade realizar mineração de dados no Sistema Informatizado de Gestão Municipal do Departamento de Água e Esgotos de Bagé. Assim, para que a tarefa seja realizada com intuito de se buscar novos conhecimentos, é evidente que não serão analisadas todas as tabelas que existem neste sistema, pois cumpre referir, que a base de dados hoje alcança aproximadamente 200 GB. Se

a mineração for aplicada nesta gama de dados, os resultados não serão os mais apropriados. É por isso que serão abordados apenas questões e dados mais relevantes, para que esta informação sirva para agregar e facilitar a tomada de melhores decisões gerenciais no Departamento, no que se refere à área financeira.

Nesse contexto, serão selecionados apenas atributos relevantes, sendo removidos ou não selecionados atributos irrelevantes para a pesquisa. Contudo, a seleção de atributos será tratada quando formos realizar a parte prática do trabalho.

Todavia, a escolha de dados irrelevantes, pode resultar na descoberta de padrões de baixa qualidade, sem dizer que o alto número de atributos irrelevantes ou redundantes aumenta o custo computacional do processo de DCBD (Descoberta de Conhecimento em Banco de Dados).

4 WEKA

WEKA é um produto da Universidade de Waikato da Nova Zelândia que começou a ser escrito em 1993 e foi implementado pela primeira vez em sua forma moderna em 1997. Ele utiliza a GNU *General Public License* (GPL). O software foi escrito na linguagem JAVA. Portanto, para que se possa utilizar o WEKA no computador é necessário ter um JRE (*Java Runtime Environment*) instalado no equipamento.

O objetivo do WEKA é agregar algoritmos provenientes de diversas abordagens. Esta ferramenta possui interface amigável, que agrega um conjunto de algoritmos de classificação, regras de associação, regressão, pré-processamento e clustering. O WEKA pode acessar dados oriundos de bancos de dados ou através da chamada de arquivos de dados próprios.

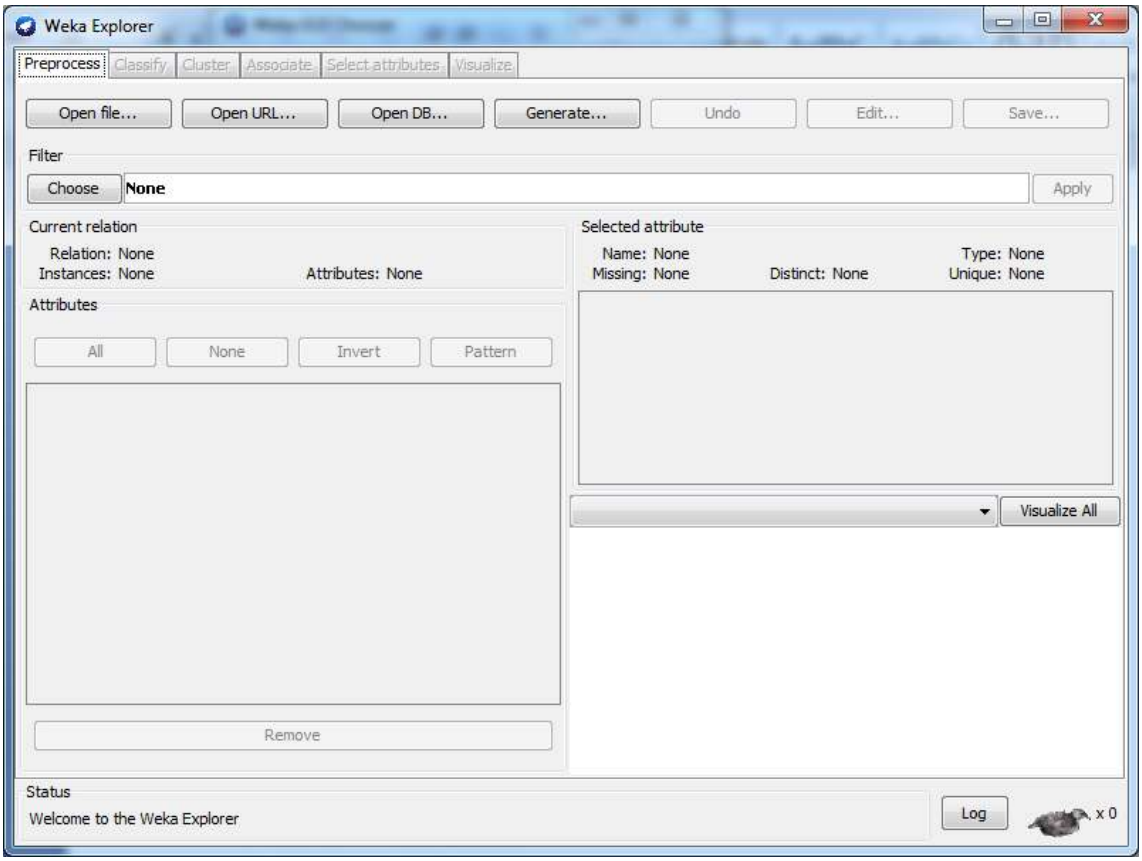
Esta ferramenta surgiu da necessidade sentida por um trabalho unificado que permitisse aos pesquisadores acesso fácil ao estado-da-arte em técnicas de inteligência artificial, mais conhecido como aprendizado de máquina. A inteligência artificial é um descobrimento interessante e de longo alcance em ciência da computação, no que diz respeito à invenção e aplicação de métodos de aprendizagem de máquina.

Assim, é possível através de um programa, no caso WEKA, analisar automaticamente uma grande massa de dados e decidir qual a informação é mais relevante, e assim se podem fazer previsões para que ajude na tomada de decisões, com bastante precisão.

Através desta ferramenta, um especialista pode utilizar o aprendizado de máquina e aplicá-lo no mundo real, obtendo conhecimento útil, que geralmente são muito grandes para serem analisadas manualmente.

A utilização desta ferramenta é muito simples e intuitiva. A começar por sua instalação que consiste em um arquivo de aproximadamente 50 megabytes (versão 3.6.9), após a instalação que é auxiliada por um assistente o software já está pronto para o uso. O arquivo padrão do software é .ARFF, com dados previamente inseridos, mas é possível também utilizar arquivos .CSV com delimitadores compatíveis com a ferramenta. Na figura 03 tem-se a demonstração da tela inicial do Weka.

Figura 03 - Tela inicial do WEKA



Fonte: WEKA

Seus algoritmos fornecem relatórios com dados analíticos e estatísticos do domínio minerado. Sua interface gráfica (GUI) possui recursos de fácil acesso. O fabricante disponibiliza ainda documentação online e também o código fonte.

Importante referir que o Weka possui boa portabilidade, já que é escrito em JAVA, podendo rodar em diferentes plataformas, sendo que no presente estudo irá se utilizar a plataforma Windows.

Para descrever sua interface gráfica, utilizaremos as figuras do trabalho de Silva (2004).

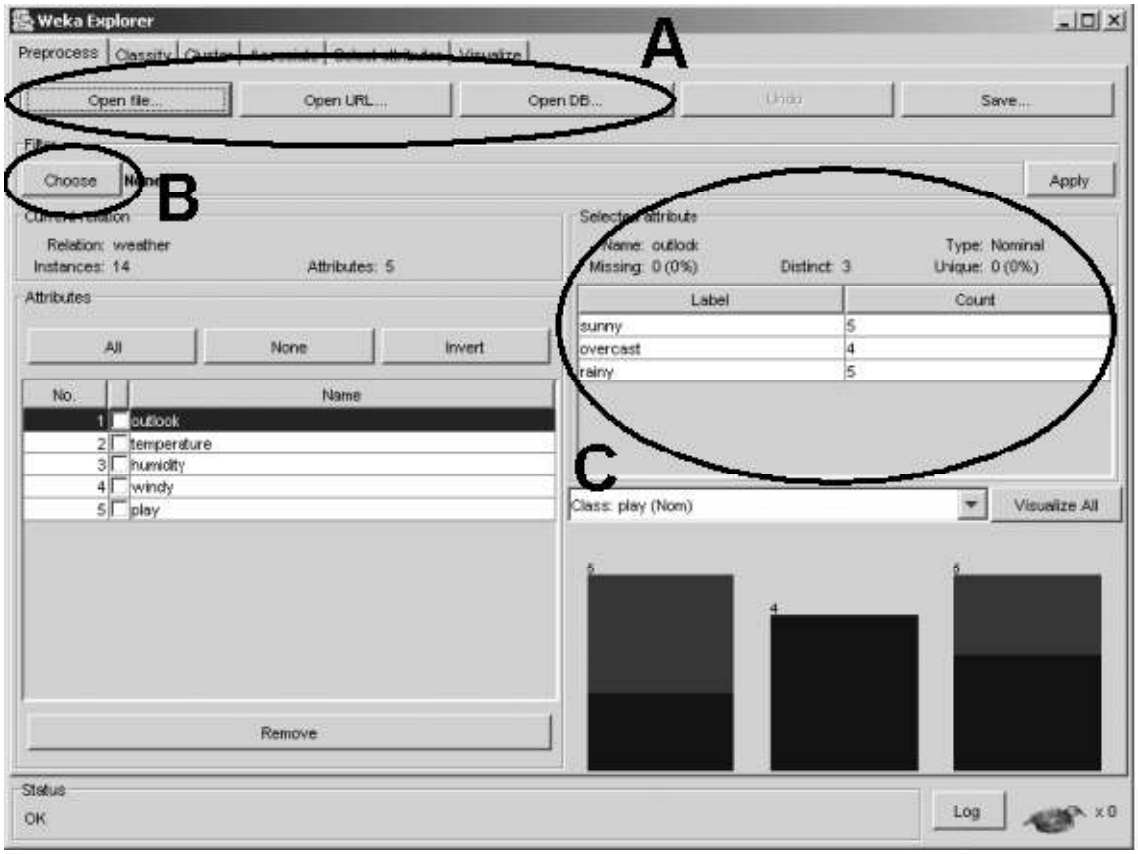
4.1 Interface e funcionalidades

A interface gráfica do Weka disponibiliza grande parte de suas funcionalidades.

Embora seja intuitiva, para uma abordagem inicial faz-se necessário reconhecer alguns elementos estratégicos da GUI Explorer, cujo guia do usuário encontra-se em http://aleron.dl.sourceforge.net/sourceforge/weka/ExplorerGuide.pdf

Esta primeira tela (Figura 04) apresenta elementos de pré-processamento:

Figura 04 - Interface do Weka (Preprocess)



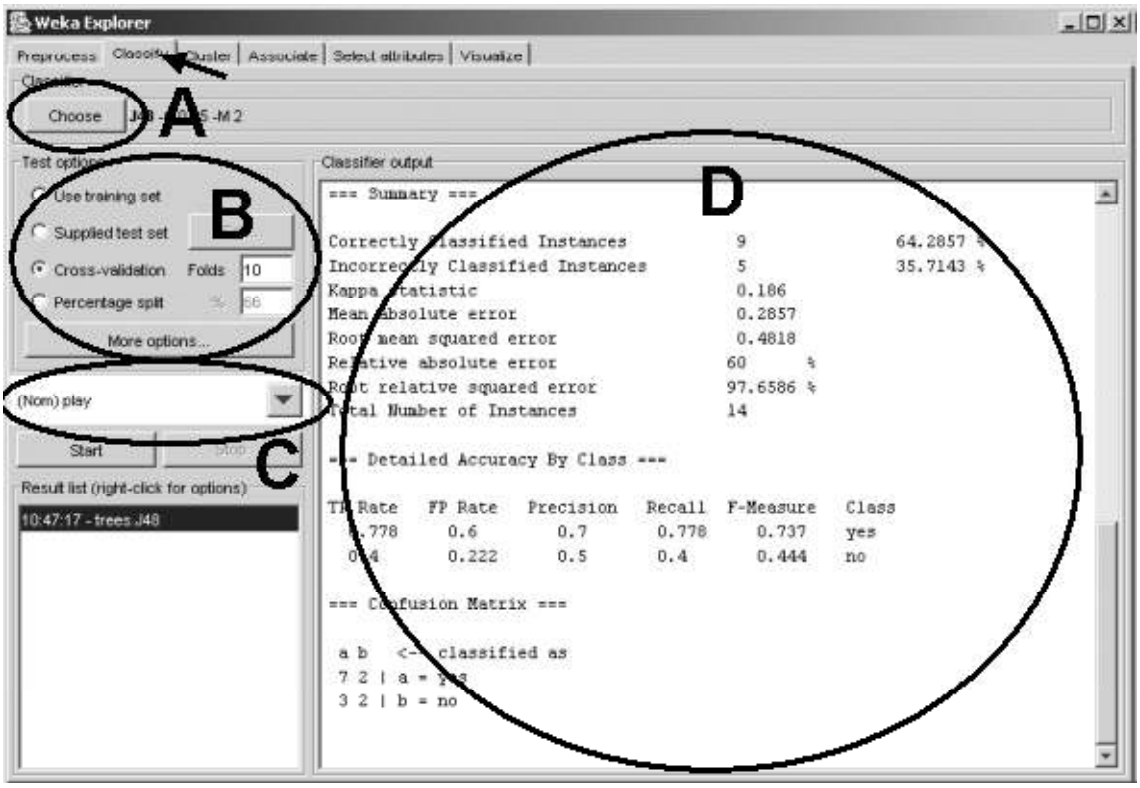
• (A) Open File, Open URL, Open DB: através destes botões é possível selecionar, respectivamente, bases de dados a partir de flat files locais (formato .arff), bases remotas (Web), e diferentes bancos de dados (via JDBC). Para acessar dados no MS Access, um roteiro de configuração está disponível em http://www.cs.waikato.ac.nz/~ml/weka/opening_windows_DBs.html . Uma breve descrição do formato .arff pode ser encontrada em http://www.cs.waikato.ac.nz/~ml/weka/arff.html ;

• (B) No botão filter é possível efetuar sucessivas filtragens de atributos e instâncias na base de dados previamente carregada (seleção, discretização, normalização, amostragem, dentre outros);

• (C) Navegando interativamente pelos atributos (quadro Attributes) é possível obter informações quantitativas e estatísticas sobre os mesmos (quadro Selected attribute);

Nesta interface, é possível desenvolver tarefas de classificação (Figura 05):

Figura 05 - Classificação no Weka



• (A) Seleção e parametrização do algoritmo a ser utilizado (Id3, C45, J48, BayesNet, Prism, Part etc);

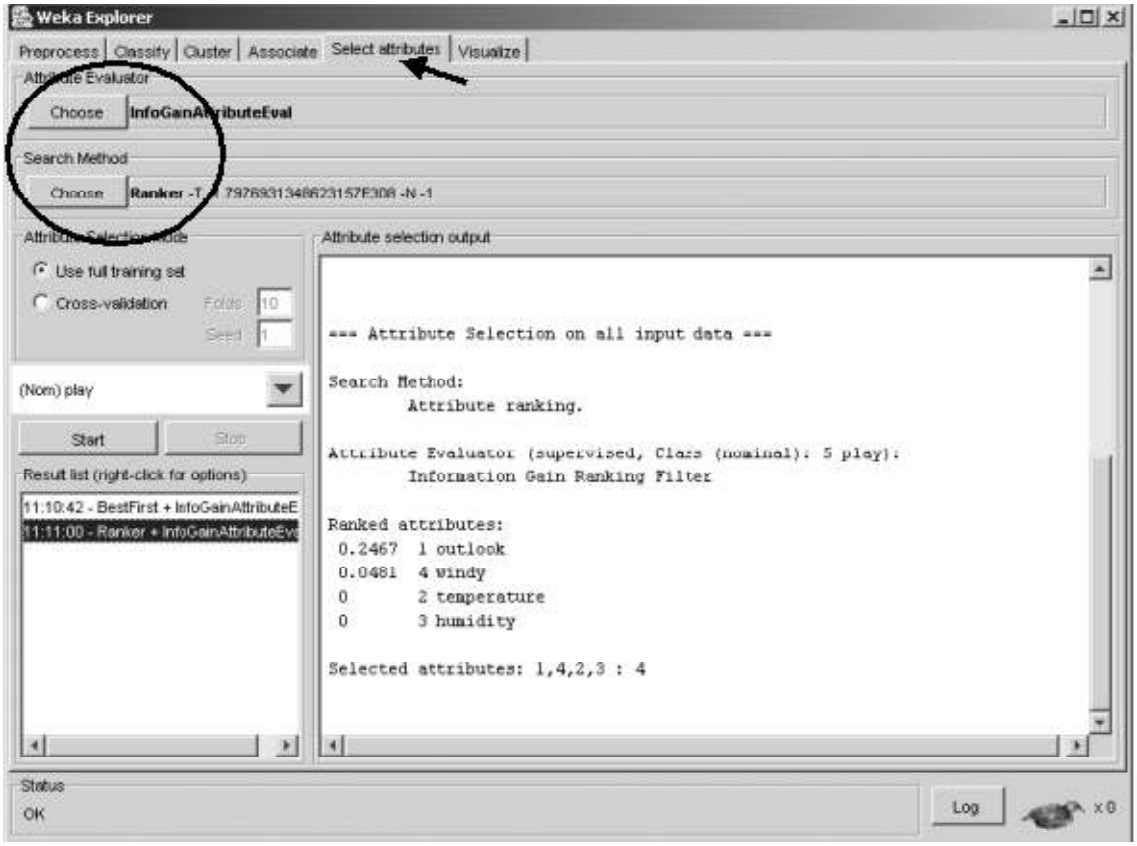
• (B) Permite selecionar a opção de teste e validação do modelo gerado (o próprio conjunto de dados do treinamento, um outro conjunto só para testes, crossvalidation, separar parte do conjunto de treinamento para teste);

• (C) Seleção do atributo classe para a tarefa de classificação;

• (D) Resumo da tarefa efetuada, com dados estatísticos, modelo, matriz de confusão etc.

As opções “Cluster”, “Associate” e “Select attributes” possuem interfaces semelhantes, fornecendo algumas opções peculiares a estas tarefas. No caso de tarefas de agrupamento (“Cluster”) a interface fornece a opção de ignorar atributos, pois é muito comum que neste tipo de tarefa um ou mais atributos gerem viés ou ruídos no processo de agrupamento. Já na seleção de atributos (“Select attributes”), é possível escolher o algoritmo avaliador de atributos e o método de busca para a tarefa (Figura 06). Faz-se necessário salientar que alguns avaliadores demandam métodos de busca específicos.

Figura 06 - Seleção de atributos



4.2 Instalação, configuração e documentação do Weka

A instalação do software é simples. Basta baixar o pacote de (Waikato 2013) e executar o instalador. Atividades de configuração podem ser encaradas como a própria parametrização dos algoritmos utilizados. O processo de escolha de algoritmos e a respectiva parametrização destes constituem um dos desafios na mineração de dados, pois dependem muito do conhecimento de cada algoritmo, da experiência do minerador e do domínio do especialista da área minerada (dados comerciais, científicos etc.). Na documentação abordada a seguir é possível encontrar informações que muito auxiliarão nesta tarefa.

Diferentes recursos de documentação podem ser encontrados no software e no site do projeto. Na instalação do Weka um pacote de documentação é disponibilizado, o qual contem informações da API (Figura 07). No pacote ainda está incluso um tutorial, que na realidade é o oitavo capítulo do livro escrito pelos líderes do projeto (Witten & Frank 2000).

No site do Weka diferentes recursos agregam informações e ajuda ao usuário (Figura 08):

- Página de trouble-shooting;
- Fórum de discussões (com arquivo das mensagens);
- Guia explicativo do formato ARFF adotado pelo Weka (e outros softwares);
- Introdução ao uso do Weka a partir da linha de comando (chamando diretamente componentes Java);
- Guia do usuário para a interface do Explorer;
- Descrição do pacote Bayes Net;
- Tutorial do Experimenter;
- Descrição de como carregar bases de dados do MS Access para o Weka.

São disponibilizadas ainda bases de dados para testes e aprendizagem, além de uma lista de projetos relevantes relacionados ao Weka.

Figura 07 - Documentação da API

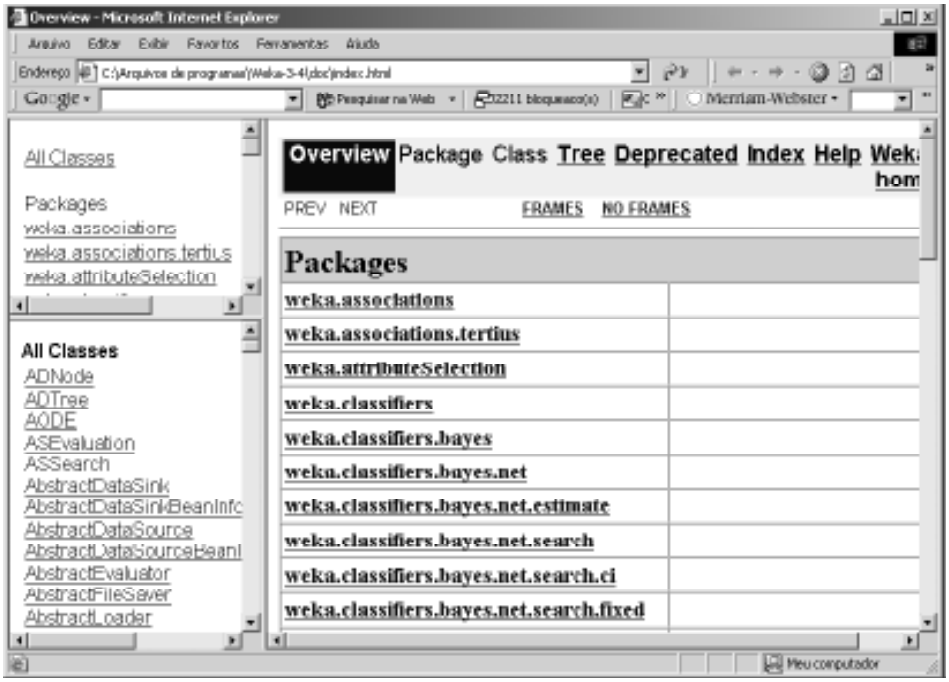
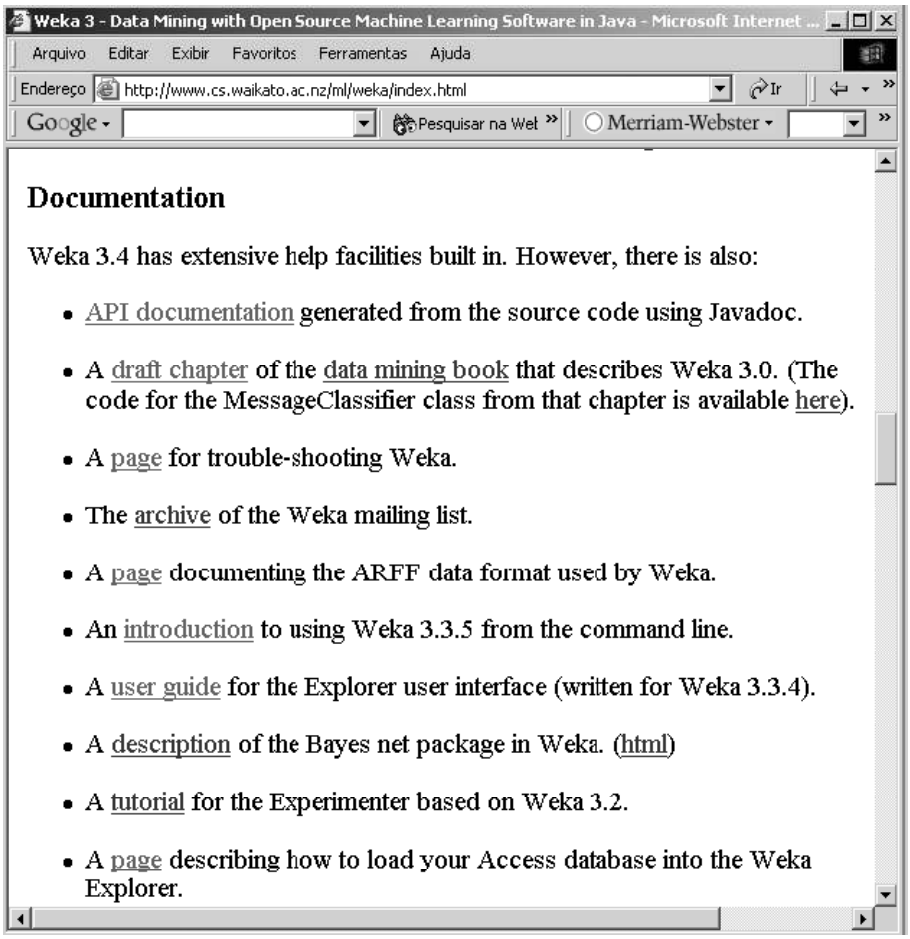


Figura 08 - Recursos e ajuda no site do Weka



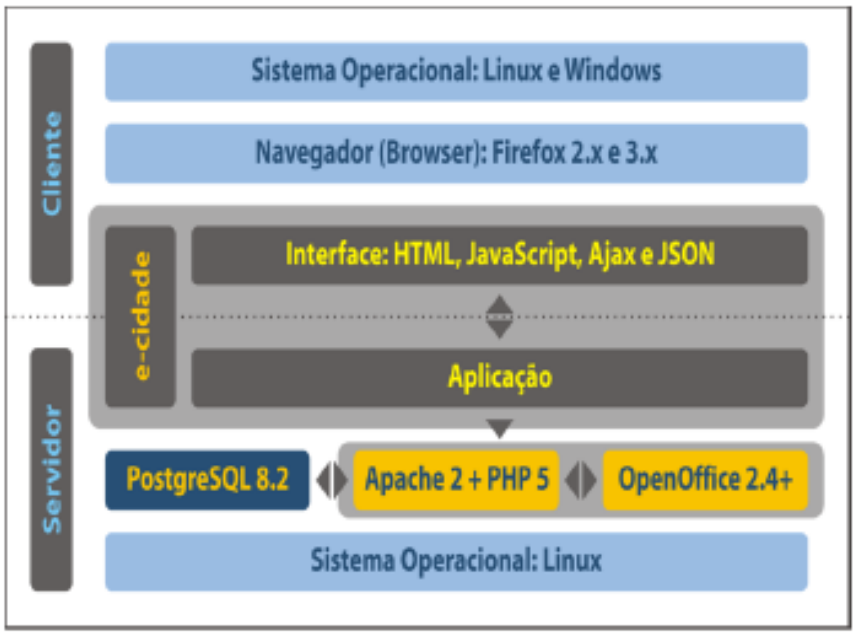
5 SISTEMA INFORMATIZADO DE GESTÃO MUNICIPAL – SIGM

O SIGM é um programa de propriedade da Prefeitura de Bagé –RS, que foi desenvolvido e segue recebendo atualização de versões pela empresa DBSeller Serviços de Informática Ltda., com sede em Porto Alegre – RS. Atualmente esta empresa publicou o presente software no Portal do Software Público Brasileiro, onde ele hoje é distribuído gratuitamente para quem quiser fazer uso desta importante ferramenta. Ele dispõe de uma estrutura modular e parametrizável, sendo o principal objetivo, organizar e otimizar a administração pública e seus mais diversos fatores, melhorando e facilitando o atendimento ao cidadão, buscando equilíbrio das contas públicas, a maximização da receita e o pleno cumprimento dos preceitos legais.

O sistema operacional em que roda a aplicação (e-cidade) é Linux, que tem licença GPL, distribuído gratuitamente, sem necessidade de licença de uso. A distribuição utilizada no servidor onde está instalado o software E-cidade é o UBUNTU Server 8.04 LTS, tendo como sistema de gerenciamento de banco de dados o Postgresql 8.2, o servidor web é o Apache 2 e o interpretador de linguagem é o PHP 5.

A figura a seguir demonstra como é a estrutura operacional para execução do programa:

Figura 09 - Estrutura operacional para execução do programa e-cidade.



Fonte: DBSELLER Serviços de Informática Ltda.

De acordo com a Figura 10, se verifica que o e-cidade além de ser uma importante ferramenta de apoio, é um grande armazenador de dados, pois este software contempla importantes funcionalidades para a gestão pública. Sua utilização se dá em praticamente todos os setores das instituições do município.

Através da gestão totalmente integrada, ele envolve:

● Gestão Financeira: PPA e LDO, Orçamento, Contabilidade, Empenhos, Tesouraria e Custos.

● Gestão Tributária: Arrecadação, Cadastro Técnico Municipal (IPTU), Cadastro Sócio Econômico (ISSQN), ITBI, Dívida Ativa e Jurídico, Projetos (Integra ao SISOBRANET), Fiscalização, Notificações e Cemitério.

● Gestão Patrimonial: Protocolo, Compras, Licitações, Patrimônio, Almoxarifados, Ouvidoria e Frotas/Veículos.

● Gestão de Recursos Humanos: Estágio Probatório, Recursos Humanos e Folha de Pagamento.

● Gestão de Educação: Biblioteca Pública, Bibliotecas das Escolas, Escolas, Secretaria, Merenda Escolar e Transporte Escolar.

● Gestão da Saúde: Vigilância Sanitária, Agendamento de Consultas e Exames, Atendimento Ambulatorial, Farmácias, Laboratório, Vacinas e PSF.

● BI – Business Intelligence: Indicadores, Análise Orçamentária, Análise Financeira, Análise Tributária, Análise Recursos Humanos, Análise Saúde e Análise Educação.

● Atendimento ao Cidadão: Segunda Via IPTU, Segunda Via ISSQN, Certidão Negativa, ISSQN Retenção na Fonte, Portal do Servidor, Nota Fiscal Eletrônica, Declaração Mensal de Serviços – DMS, Ouvidoria e Consulta de Processos.

Figura 10 - Áreas do e-cidade

Fonte: DBSELLER Serviços de Informática Ltda.

Além de todas estas áreas de atuação, existe ainda um módulo desenvolvido exclusivamente para saneamento, o Módulo Água. Este módulo permite que sejam cadastrados os imóveis da cidade, com o controle financeiro destes. Realiza a cobrança e a arrecadação dos débitos. Permite ainda, o controle da suspensão do fornecimento de água (corte de água).

Esta funcionalidade é de uso exclusivo do Departamento de Água e Esgotos de Bagé –DAEB que é Autarquia Municipal responsável pelo abastecimento de água e o recolhimento de esgotos nos prédios do município.

Por utilizar uma única base de dados, é possível o aproveitamento das informações, em tempo real, por todas as instituições do município (Prefeitura Municipal, Câmara Municipal, DAEB e outros). Portanto, esta característica faz com que sua base de dados esteja diariamente crescendo e agregando informações sobre o município, o que pode contribuir para que o processo de mineração de dados traga informações interessantes.

Uma das funcionalidades do e-cidade é a ferramenta de Business Intelligence - BI, que realiza diversos tipos de análises. Nesse sentido, se poderia questionar que os resultados do presente trabalho pudessem ser obtidos por esta ferramenta. Assim, para

esclarecer e afastar esta hipótese, é necessária uma comparação no intento de separar a utilização destas técnicas no âmbito do DAEB.

Por BI (inteligência de negócios) entende-se o processo de coleta, organização e análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. Ele permite que as empresas obtenham um conhecimento mais abrangente sobre os fatores que afetam os seus negócios. Porém, a ferramenta de BI do DAEB, utiliza apenas a OLAP (On-line Analytical Processing) no tratamento dos dados, trabalhando com operadores dimensionais, não utilizando a Mineração de Dados, que está mais relacionado com os processos de análise de inferência, do que com análise dimensional (OLAP), visto que a Mineração de Dados representa uma forma de busca de informação baseada em algoritmos que buscam o reconhecimento de padrões escondidos nos dados e não necessariamente revelados pelas abordagens analíticas, como a OLAP.

Portanto, o BI (DAEB) possui conceito ETC (Extract Transform Load) – Extração, Tratamento e Carga, fundamentais para o transformação do recurso de dados transacional em informacional, enquanto que a Mineração de Dados tem o objetivo de prover análises diferenciais.

Diante disso, acredita-se que através dos resultados obtidos no presente trabalho, aliado a ferramenta BI, será possível criar novas visões multidimensionais (CUBOS), com vista a alcançar melhores resultados.

Importante salientar, que a ferramenta de BI no DAEB está instalada a mais de dois anos e ainda enfrenta muita resistência por parte dos funcionários, diante desta realidade, a presente pesquisa visa impulsionar e fazer mais atrativa a utilização do BI por todos os colaboradores, através de resultados que auxiliarão na tomada de melhores decisões na seara administrativa.

Feita estas considerações, se encaminhará o presente estudo para que se realize a análise nestes dados com a utilização da Mineração de Dados.

6 AMOSTRAGEM DOS DADOS A SEREM MINERADOS NO SISTEMA E-CIDADE (SIGM)

Como a utilização do módulo água é de uso exclusivo do DAEB, a principal tarefa será realizada nos dados extraídos desta parte do banco de dados.

Importante salientar que apesar de ter módulo exclusivo (Módulo Água) o DAEB, utiliza praticamente todos os outros módulos, dentro do que lhe compete utilizar como ente público. Assim, pode-se citar a utilização dos módulos abrangidos pela Gestão Financeira, Gestão Tributária, Gestão Patrimonial, Gestão de Recursos Humanos, B. I. e Atendimento ao Cidadão.

Os dados foram extraídos do sistema com a ajuda de um profissional da empresa que presta suporte para o Departamento. Estes dados foram extraídos no formato .CSV (separado por vírgula), compatível com o WEKA. O arquivo gerado contém 52.540 registros (linhas, instâncias), com 12 campos (atributos) por instância.

Assim, pretende-se analisar os seguintes dados:

Dados relativos aos imóveis:

● Tamanho do imóvel: Possui três grupos, o Grande para imóvel com mais de 60 metros quadrados de área construída. Médio para imóveis com até 60 metros quadrados de área construída e Pequeno para imóveis com até 40 metros quadrados de área. Esta divisão se deu em virtude da legislação municipal assim dispor, com relação a cobrança de taxas de água.

● Água: Para imóveis com abastecimento de água ativo, será atribuído valor "SIM" e para imóveis com água inativa, será atribuído valor "NÃO".

● Esgoto: Para imóveis com rede coletora de esgotos instalada será atribuído valor "SIM" e caso contrário será atribuído valor "NÃO".

● Tipo: Este item tem a ver com a utilização do imóvel, se é "RESIDENCIAL" ou de uso "COMERCIAL".

● Hidrômetro: Se possui hidrômetro instalado será atribuído valor "SIM" caso contrário o valor será "NÃO".

● Corte: Irá listar quantas vezes o imóvel teve suspensão no abastecimento de água por falta de pagamento.

● Sexo: Neste conterá o sexo do proprietário do imóvel.

● Idade: Neste quesito haverá distribuição em três grupos: Idoso, para pessoas com mais de 60 anos de idade, Adulto para pessoas com menos de 60 anos de idade e Jovem para pessoas até 25 anos de idade.

● Parcelamento: Neste campo teremos informações referentes a quantidade de dívida que foi parcelada para o referido imóvel, tendo a seguinte divisão: “SEM” em caso de não haver dívida parcelada; “COM” para imóveis com parcelamento de divida.

● Dívida ativa: Neste campo teremos informações referentes a quantidade de dívida ativa (não parcelada) para o referido imóvel, tendo a seguinte divisão: “SEM” em caso de não haver dívida ativa; “COM” para inoveis que possuem divida ativa.

● Exercício: Será divido em dois grupos: “EM DIA” para quem não possui débitos atrasados no exercício; “ATRASADO” para quem tem débitos atrasados no exercício.

● Bairro: Constará o bairro onde está localizado o imóvel.

Para cada atributo registrou-se os seguintes resultados:

Tabela 2 – Resultados dos Dados

Atributo	Grupo	Quantidade
Tamanho	GRANDE	18.107
	MÉDIO	10.085
	PEQUENO	15.249
	TERRENO	9.099
Água	SIM	39.928
	NÃO	12.612
Esgoto	SIM	21.835
	NÃO	30.705
Tipo	RESIDENCIAL	40.846
	COMERCIAL	11.694
Hidrômetro	SIM	35.712
	NÃO	16.828
Corte	ZERO	43.308
	UM	4.342

	DOIS	2.012
	TRÊS	1.242
	QUATRO	757
	CINCO	411
	SEIS	185
	SETE	148
	OITO	73
	NOVE	18
	DEZ	19
	ONZE	11
	DOZE	6
	TREZE	3
	QUATORZE	2
	QUINZE	1
	DEZESSEIS	1
	VINTE E UM	1
Sexo	MASCULINO	33.546
	FEMININO	7.853
	INDEFINIDO	11.141
Idade	IDOSO	11.351
	ADULTO	7.214
	JOVEM	367
	INDEFINIDO	33.608
Parcelamento	COM	16.924
	SEM	35.615
Dívida Ativa	COM	20.988
	SEM	31.552
Bairro	Foram encontrados 117 bairros, onde se distribuem todos os 52.540 registros.	
Exercício	EM DIA	28.014
	ATRASADO	24.526

7 ABORDAGEM DOS DADOS

Este capítulo tratará sobre a tarefa que será utilizada para a mineração dos dados, bem como quais os algoritmos que auxiliarão na busca da descoberta do conhecimento.

7.1 Tarefa de associação

Regras de associação são expressões que indicam afinidade ou correlação entre dados. Esta regra, por exemplo, consiste em determinar quais itens tendem a serem comprados juntos em uma mesma transação.

"A tarefa de associação pode ser considerada uma tarefa bem definida, determinística e relativamente simples, que não envolve predição da mesma forma que a tarefa de classificação" (FREITAS, 2000, p. 65)

Esta tarefa tem em vista identificar associações entre registros de dados, que de alguma maneira estão ou devem estar relacionados. Ela busca elementos que implicam na presença de outros em uma mesma transação, ou seja, encontrar relacionamentos ou padrões frequentes entre conjunto de dados.

Estes padrões são encontrados nas transações (conjunto de dados) armazenadas.

Por exemplo: através do arquivo extraído do e-cidade, com uso desta tarefa (associação) tem-se a seguinte regra:

Imóveis com hidrômetro e que possuem os débitos do exercício em dia, com grau de certeza de 94% são utilizados como residência.

HIDRÔMETRO=SIM EXERCÍCIO=EM DIA 20393 ==> TIPO=RESIDENCIAL 19226 conf:(0.94)

Para a descoberta do conhecimento através da regra de associação, deve-se decompor o problema em duas partes:

● *Suporte:* encontrar todos os conjuntos de itens que possuem um suporte de transações acima de um limite mínimo. Assim, o suporte para um conjunto de itens é o número de vezes que ele ocorre. Chama-se conjunto de itens frequentes todos que tem suporte igual ou maior a um mínimo estabelecido.

Portanto, para que uma análise seja classificada como *"Boa"* ela precisa apresentar um número de suporte com certa razoabilidade de ocorrência.

● *Confiança:* quanto a confiança, as regras de associação geradas a partir de conjuntos frequentes, devem representar um grau mínimo a ser alcançado. Portanto, quanto

maior o nível de suporte e maior o grau de confiança, pode-se classificar a análise como uma "*Boa Análise"*.

A seguir, pode-se citar como boa a seguinte análise:

HIDRÔMETRO=SIM EXERCÍCIO=EM DIA 20393 ==> TIPO=RESIDENCIAL 19226 conf:(0.94)

Quanto ao suporte, num universo de 52.540 registros, a regra ocorre em 20.393 casos e possui confiabilidade de 94%. Portanto, a regra de associação, neste caso, serviu para evidenciar um bom padrão no que se infere a esta análise.

Já no exemplo a seguir, torna-se evidente que a análise apresenta um ótimo índice de confiabilidade (100%), porém o seu **suporte** não representa um índice considerável de ocorrência (apenas 113 num total de 52.540).

TIPO=RESIDENCIAL CORTE=SETE EXERCÍCIO=ATRASADO 113 ==> HIDRÔMETRO=SIM 113

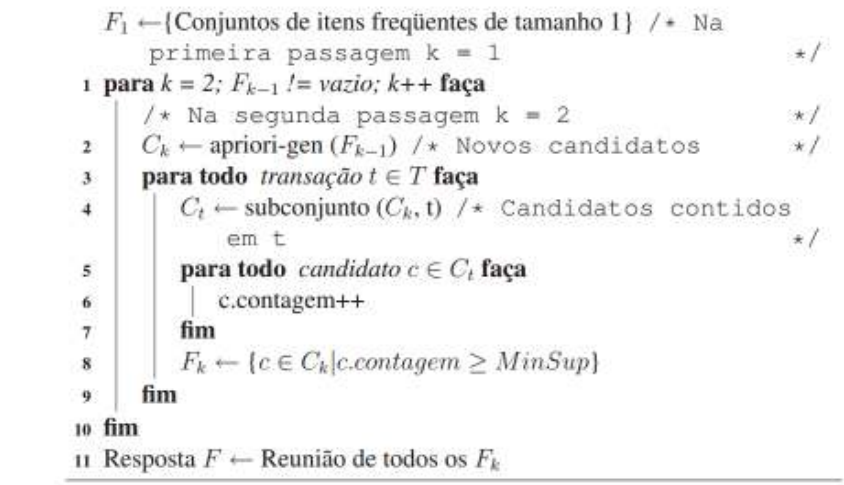
7.2 Algoritmo apriori

Este algoritmo emprega busca em profundidade e gera conjuntos de itens candidatos (padrões) de *k* elementos a partir de conjuntos de itens de *k-1* elementos, eliminando os padrões não frequentes.

Ele (algoritmo) rastreia toda a base de dados e os conjuntos de itens frequentes são obtidos a partir dos conjuntos de itens candidatos.

Na figura 11 se ilustra o núcleo do algoritmo:

Figura 11 – Núcleo do Algoritmo Apriori



$F_1 \leftarrow$ {Conjuntos de itens freqüentes de tamanho 1} /* Na primeira passagem k = 1 */
1 **para** $k = 2$; F_{k-1} != *vazio*; k++ **faça**
/* Na segunda passagem k = 2 */
2 $C_k \leftarrow$ apriori-gen (F_{k-1}) /* Novos candidatos */
3 **para todo** *transação* $t \in T$ **faça**
4 $C_t \leftarrow$ subconjunto (C_k, t) /* Candidatos contidos em t */
5 **para todo** *candidato* $c \in C_t$ **faça**
6 c.contagem++
7 **fim**
8 $F_k \leftarrow \{c \in C_k | c.contagem \geq MinSup\}$
9 **fim**
10 **fim**
11 Resposta $F \leftarrow$ Reunião de todos os F_k

1) F_k - conjunto de itens freqüentes de tamanho k (conjunto com k elementos) que atende o suporte mínimo estabelecido. Cada membro deste conjunto tem dois campos. O primeiro é conjunto de itens e o segundo é um contador para o suporte.

2) C_k - Conjunto de itens candidatos de tamanho k. Cada membro deste conjunto tem dois campos. O primeiro é conjunto de itens e o segundo é um contador para o suporte.

Fonte: Vasconcelos e Carvalho (2004)

Este algoritmo utiliza duas sub-rotinas:

● *Apriori-gen:* que é responsável por gerar o conjunto de itens candidatos. Considera todos os itens, independente deles atenderem o suporte mínimo especificado, eliminando os que não são frequentes.

● *Subconjuntos:* ele realiza a tarefa de associação, ou seja, enquanto os dados são separados, ele procura por relações entre os dados.

Assim, o algoritmo trabalha sobre a base de transações em busca de itens frequentes, ou seja, aqueles que possuem suporte maior ou igual ao suporte mínimo. Numa primeira passagem, o suporte do algoritmo para cada item individual (conjuntos de 1 item) é contado e todos que satisfazem o suporte mínimo são selecionados, construindo os conjuntos de 1 item frequentes (F_1).

Após, o algoritmo gera os conjuntos de 2 itens candidatos pela junção dos conjuntos de 1 item e seus suportes são determinados pela pesquisa no banco de dados, encontrando assim, os conjuntos de 2 itens frequentes.

O algoritmo prossegue até que o conjunto de *k* itens encontrados seja vazio.

Após, acontece o passo da poda, em que há a eliminação dos conjuntos candidatos *c* itens, baseada no suporte. Ou seja, os conjuntos que não alcançam o valor mínimo informado pelo suporte, são descartados.

7.3 Algoritmo predictive apriori

Foi definido para encontrar associações de grande valor, conveniência ou interesse entre os itens de dados. Tem por base o algoritmo *Apriori* (suporte e confiança) combinando em uma única chamada de *Predictive Accuracy* (Acurácia Preditiva) para encontrar as melhores regras de associação já ordenadas em sua exibição.

Através de uma distribuição binomial busca uma relação entre *Suporte e Confiança* de maneira a potencializar a geração das melhores regras.

8 RESULTADOS

Os resultados serão analisados por área de interesse e atuação do DAEB. Primeiramente serão destacados os resultados de interesse da área financeira, como perfil de contribuintes adimplentes e inadimplentes. Também se evidenciará possíveis estratégicas de recuperação de créditos.

Quanto aos micro medidores (hidrômetros), se buscará demonstrar a relação entre sua utilização e a adimplência dos débitos, no que concerne a conscientização dos contribuintes no cumprimento de suas obrigações, para o pleno exercício da cidadania.

8.1 Análise financeira quanto ao perfil dos contribuintes

Na análise que segue, infere-se que o suporte desta é bastante considerável (28.014 de 52.540) e o índice de confiança é bem elevado (75%), o que leva a considerar que esta é uma boa análise.

Análise 01:

EXERCÍCIO=EM DIA 28014 ==> CORTE=ZERO DÍVIDA ATIVA=SEM 21091 conf:(0.75) (Algoritmo utilizado: Apriori)

Neste resultado, pode-se constatar que todos os prédios que tem exercício em dia, em 75% desses casos nunca houve corte e também não possuem dívida ativa. Podendo ser criada uma classificação para este perfil de contribuinte, no intuito de que em caso de haver inadimplência, realizar uma cobrança de forma mais amigável. Já que pelo perfil, este contribuinte provavelmente deixou de pagar seu tributo por algum imprevisto. Assim, pela regra de cobrança atual, em caso de inadimplência o contribuinte (usuário) seria notificado através de carta que se não satisfazer o crédito será suspenso o fornecimento de água. Ao invés disso, poderia ser feita uma cobrança menos ofensiva, no intuito de manter a credibilidade junto a este perfil de usuário do serviço. Como por exemplo, o envio de e-mail lembrando a existência de débito vencido.

Já na próxima análise (analise 02) pode-se perceber que dos imóveis de utilização residencial que nunca sofreram suspensão do abastecimento de água por falta de pagamento (corte), 63% deles tem como proprietário pessoas do sexo masculino. Diante disso, o tipo de contribuinte que possui este perfil, em caso de inadimplência, também se poderia utilizar uma cobrança menos ofensiva.

Análise: 02

TIPO=RESIDENCIAL CORTE=ZERO 32161 ==> SEXO=MASCULINO 20414 conf:(0.63) (Algoritmo utilizado: Apriori)

Importante referir, que a análise 02 abrange 20.414 registros do sexo masculino (num total de 33.546). O que demonstra bastante credibilidade no resultado da pesquisa.

As próximas análises evidenciam imóveis que não possuem parcelamento de dívida, ou seja, sem renegociação dos débitos inadimplidos. Nesta realidade, seria importante a tomada de decisão no intuito de se buscar estratégias para que seja efetuado o parcelamento destes débitos, visando facilitar o adimplemento, bem como, proporcionar o aumento da arrecadação, para cada vez mais promover benefícios à população, através de investimentos. Portanto passa-se a análise.

8.2 Análise financeira quanto a novos parcelamentos

Análise 03:

PARCELAMENTO=SEM DÍVIDA ATIVA=COM 12624 ==> EXERCÍCIO=ATRASADO 9990 conf:(0.79) (Algoritmo utilizado: Apriori)

Nesta análise, percebe-se que existe uma considerável parcela de imóveis (12.624) que possuem dívida ativa, mas não efetuaram ainda o parcelamento destes débitos. Ainda fica evidente, que em 79% (9.990) destes casos, o imóvel apresenta pendência quanto aos débitos do exercício vigente (2013). Portanto, considerando-se que atualmente a média de valor da parcela nos parcelamentos existentes é de R$ 36,00 (trinta e seis reais), com a realização de mais 9.990 seria possível a arrecadação mensal de aproximadamente R$ 360.000,00 (trezentos e sessenta mil reais) que seriam investidos em serviços à população. No período de 12 meses, esta quantia se elevaria a quase R$ 4.300.000,00 (quatro milhões e trezentos mil reais).

Análise 04:

TAMANHO=TERRENO PARCELAMENTO=SEM DÍVIDA ATIVA=COM 6420 ==> EXERCÍCIO=ATRASADO 6058 acc:(0.45861) (Algoritmo utilizado: Predictive Apriori)

Na análise 04, é possível se perceber que mais de 50% dos casos citados na análise 03, trata-se de terrenos. Com isso, se a Direção do DAEB, entender que não há estrutura administrativa possível para atender toda a demanda gerada pela análise 03, pode focar a busca de parcelamentos nas matrículas abrangidas pela análise 04. Assim, a tarefa de parcelamento dos débitos elaborada pode ser fracionada, e o Departamento consiga, ainda que de forma parcial, buscar valores inadimplidos.

8.3 Análise financeira quanto ao bairro dos imóveis

Na presente análise, será feita uma abordagem quanto ao bairro dos imóveis. Dentre todos os bairros selecionou-se o bairro *Centro*, pois representa 14.036 registros. Portanto, devido a quantidade de registros este é o único bairro que alcança o número mínimo de suporte especificado no algoritmo (10% = 5.254 registros). Devido a esse fator, pode-se inferir que possivelmente poderá gerar boas análises.

Análise 05:

DÍVIDA ATIVA=SEM BAIRRO=CENTRO 10457 ==> EXERCÍCIO=EM DIA 8566 conf:(0.82) (Algoritmo utilizado: Apriori)

Na análise 05 evidencia-se que 10.457 (de 14.036) imóveis localizados no bairro Centro não possuem dívida ativa. Destes (10.457), 8.566 (82%) possuem débitos do exercício em dia.

Análise 06:

DÍVIDA ATIVA=SEM BAIRRO=CENTRO EXERCÍCIO=EM DIA 8566 ==> PARCELAMENTO=SEM 7243 conf:(0.85) (Algoritmo utilizado: Apriori)

Na análise 06, pode-se verificar ainda que grande parte dos imóveis elencados na análise 05 não possui parcelamento (85%).

Análise 07:

BAIRRO=CENTRO 14036 ==> PARCELAMENTO=SEM 10466 conf:(0.75) (Algoritmo utilizado: Apriori)

Já a análise 07, mostra que 75% dos imóveis do bairro Centro não possuem parcelamento.

Análise 08:

BAIRRO=CENTRO 14036 ==> DÍVIDA ATIVA=SEM 10457 conf:(0.75) (Algoritmo utilizado: Apriori)

A análise 08 demonstra que 75% dos imóveis no bairro centro não possuem dívida ativa.

Análise 09:

BAIRRO=CENTRO EXERCÍCIO=EM DIA 9789 ==> PARCELAMENTO=SEM DÍVIDA ATIVA=SEM 7243 conf:(0.74) (Algoritmo utilizado: Apriori)

Nesta análise (09), dos imóveis do bairro Centro com débitos do exercício em dia (9.789 registros), em 74% (7.243) deles não possui parcelamento e nem dívida ativa.

Análise 10:

PARCELAMENTO=SEM 35615 ==> BAIRRO=CENTRO 10466 conf:(0.29) (Algoritmo utilizado: Apriori)

Evidencia que de todos os imóveis que não possui parcelamento (35.615), 29% (10.466) são situados no bairro Centro.

Por todas as análises apresentadas (análise 05 a 10), se conclui que os imóveis situados no bairro Centro possuem bom índice de adimplência. Esta característica inclusive reflete no número de parcelamentos de dívida existentes nos imóveis deste bairro. Já que 29% de todos os imóveis que não possuem parcelamento estão situados neste bairro. Outro fator que induz a conclusão da existência de bons pagadores, é que 75% (análise 08) dos imóveis deste bairro não possuem dívida ativa.

8.4 Análise financeira quanto aos micro medidores (hidrômetros)

Uma das metas do DAEB é hidrometrar todos os imóveis da cidade, porém hoje isto já é realidade em mais de 68% dos imóveis. Portanto, a análise do comportamento financeiro das matrículas que possuem este equipamento é necessária para que seja destacado se o micro medidor é fator preponderante para que aumente a credibilidade do DAEB e assim faça com que os contribuintes realizem os pagamentos dos débitos.

Análise 11:

HIDRÔMETRO=SIM DÍVIDA ATIVA=SEM 23717 ==> TIPO=RESIDENCIAL 22400 conf:(0.94) (Algoritmo utilizado: Apriori)

Demonstra que dos imóveis que possuem hidrômetro e não possuem dívida ativa (23.717 imóveis), destes, 94% são utilizados para residência.

Análise 12:

HIDRÔMETRO=SIM DÍVIDA ATIVA=SEM EXERCÍCIO=EM DIA 17165 ==> TIPO=RESIDENCIAL 16179 conf:(0.94) (Algoritmo utilizado: Apriori)

Esta análise (12) expõe que os imóveis que possuem hidrômetro, sem dívida ativa e débitos do exercício em dia, representam a quantidade de 17.165 e destes, 94% são residenciais.

Análise 13:

HIDRÔMETRO=SIM EXERCÍCIO=EM DIA 20393 ==> DÍVIDA ATIVA=SEM 17165 conf:(0.84) (Algoritmo utilizado: Apriori)

Evidencia que imóveis com hidrômetro e que tem exercício em dia, destes, em 84% dos casos não possuem dívida ativa.

Análise 14:

HIDRÔMETRO=SIM EXERCÍCIO=EM DIA 20393 ==> TIPO=RESIDENCIAL PARCELAMENTO=SEM DÍVIDA ATIVA=SEM 12157 conf:(0.6) (Algoritmo utilizado: Apriori)

Nesta análise tem-se que dos imóveis que possuem hidrômetro e possuem exercício em dia (20.393 imóveis) em 60% dos casos são imóveis residenciais que não possuem parcelamento e dívida ativa.

Análise 15:

TIPO=RESIDENCIAL HIDRÔMETRO=SIM EXERCÍCIO=EM DIA 19226 ==> DÍVIDA ATIVA=SEM 16179 conf:(0.84) (Algoritmo utilizado: Apriori)

Imóveis do tipo residencial, com hidrômetro e exercício em dia (19.226 imóveis) em 84% não possuem dívida ativa.

Análise 16:

HIDRÔMETRO=NÃO PARCELAMENTO=SEM DÍVIDA ATIVA=COM 7747 ==> TIPO=COMERCIAL 6561 acc:(0.44858) (Algoritmo utilizado: Predictive Apriori)

Imóveis que não possuem hidrômetro, não possuem parcelamento e possuem dívida ativa (7.747 imóveis), em 6.561 imóveis são de utilização Comercial.

Análise 17:

HIDRÔMETRO=NÃO DÍVIDA ATIVA=COM EXERCÍCIO=ATRASADO 7918 ==> PARCELAMENTO=SEM 6902 acc:(0.46594) (Algoritmo utilizado: Predictive Apriori)

Os imóveis que não possuem hidrômetro possuem dívida ativa e exercício atrasado, representam 7.918 unidades e em 6.902 não possuem parcelamento.

Análise 18:

HIDRÔMETRO=NÃO DÍVIDA ATIVA=COM 8993 ==> EXERCÍCIO=ATRASADO 7918 conf:(0.88) (Algoritmo utilizado: Apriori)

Nesta análise (18) pode-se concluir que das matrículas que não possuem hidrômetro e possuem dívida ativa em 88% dos casos possui exercício atrasado.

As análises 11 a 18 evidenciam a importância da utilização dos micro medidores nos imóveis, pois além de ser conscientizador do uso responsável da água, ele é fator de credibilidade do Departamento junto aos contribuintes. Pois, nas análises em que os imóveis possuem hidrômetro (análise 11 a 15), há um grande índice de bons pagadores. Em contrapartida, os imóveis abrangidos pelas análises 16 a 18 (sem hidrômetros), evidenciam que imóveis sem micro medidores, tendem a possuir relevante índice de inadimplência.

Assim, a instalação de hidrômetros em todos os imóveis é importante devido a arrecadação dos tributos, bem como demonstra lisura na prestação do serviço. E ainda é importante instrumento de conscientização para o uso racional da água.

9 CONCLUSÃO

Através do presente trabalho, ficou evidente a importância da Mineração de Dados na descoberta de padrões não triviais. Com a utilização da ferramenta WEKA, através da técnica de análise de associação, possibilitou encontrar padrões consistentes de dados, revelando novos conhecimentos.

A utilização destes novos conhecimentos poderão ser fatores que auxiliem a Gestão Pública Municipal, a prestar um serviço de melhor qualidade para a população. Por isso, o presente trabalho serve como passo inicial para a implantação da mineração de dados no Departamento de Água e Esgotos de Bagé – DAEB, na busca de auxiliar os gestores nas tomadas de decisões.

Acredita-se que os resultados alcançados no presente trabalho sejam propagador desse novo objeto de estudo, e sua aplicação a nível municipal se estenda as outras Instituições que compõem o Poder Público no Município. Já que o Sistema e-cidade agrega hoje dados de todo o Município em uma única base, facilitando sobremaneira a exploração destes dados na busca de novos padrões e conhecimentos.

No que diz respeito ao presente trabalho, os resultados alcançados foram satisfatórios em nível de pesquisa, pois como foi exposto na introdução não tínhamos o propósito de esgotar o assunto aqui tratado, mas sim verificar a importância da aplicação da Mineração de Dados no sistema informatizado do município. E as descobertas realizadas apesar de na maioria das vezes apresentar resultados simples, serviu para demonstrar que o aprofundamento do estudo será muito valioso para o município. Pois conforme verificado no presente estudo, grandes corporações utilizam desta tecnologia para aumentar seus lucros e por consequência também melhoram seus serviços.

REFERÊNCIAS

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

BRAGA, Luis Paulo Vieira. **Introdução à Mineração de Dados**. 2ª edição revista e ampliada. Rio de Janeiro: E-Papers Serviços Editoriais, 2005.

CAMILO, C. O., SILVA, J.C. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Technical Report - RT-INF_001-09 - Relatório Técnico, Agosto 2009.

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

FREITAS, A. A. **Understanding the crucial differences between classification and discovery of association rules: a position paper**. ACM SIGKDD Explorations Newsletter, v.2 n.1, p. 65-69, junho de 2000.

FREITAS, Henrique. **A informação como ferramenta gerencial: um telessistema de informação em marketing para apoio à decisão**. Porto Alegre: Ortiz, 1993.

GOEBEL, M.; GRUENWALD, L. **A Survey of Data Mining and Knowledge Discovery Software Tools**. ACM SIGKDD Explorations, New York, v. 1, no. 1, p. 20-33, June. 1999.

SILVA, M. P. S. **Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka.** Mossoró, RN. 2004

SILVA, Paulo Ricardo. **Apresentação Software Público e-cidade.** Porto Alegre. 2010.

TAN, P., STEINBACH, M., KUMAR, V. **Introdução ao Data Mining.** Ciência Moderna, 2009.

VASCONCELOS L. M. R.; CARVALHO C. L. de. **Aplicação de Regras de Associação para Mineração de Dados na Web.** Relatório Técnico, 2004.

WAIKATO, U. O. WEKA. http://www.cs.waikato.ac.nz/ml/weka/, acessado em junho de 2013.

WIKIPEDIA. Mineração de Dados. http://pt.wikipedia.org/wiki/Minera%C3%A7%C3%A3o_de_dados, acessado em Maio de 2013.

WITTEN, I.; FRANK, E. **Data Mining – Practical Machine Learning Tools.** Morgan Kaufmann, 2000.