Matheus Emerick de Magalhães

Uma abordagem na adoção de BUSINESS INTELLIGENCE

Atena Editora
Ano 2022

Matheus Emerick de Magalhães

Uma abordagem na adoção de BUSINESS INTELLIGENCE

Atena Editora
Ano 2022

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

Editora chefe
Profª Drª Antonella Carvalho de Oliveira
Editora executiva
Natalia Oliveira
Assistente editorial
Flávia Roberta Barão
Bibliotecária
Janaina Ramos
Projeto gráfico
Bruno Oliveira
Camila Alves de Cremo
Luiza Alves Batista
Natália Sandrini de Azevedo
Imagens da capa
iStock
Edição de arte
Luiza Alves Batista

2022 *by Atena Editora*
Copyright © Atena Editora
Copyright do texto © 2022 Os autores
Copyright da edição © 2022 Atena Editora
Direitos para esta edição cedidos à Atena Editora pelos autores.
Open access publication by Atena Editora

Todo o conteúdo deste livro está licenciado sob uma Licença de Atribuição *Creative Commons*. Atribuição-Não-Comercial-NãoDerivativos 4.0 Internacional (CC BY-NC-ND 4.0).

O conteúdo do texto e seus dados em sua forma, correção e confiabilidade são de responsabilidade exclusiva do autor, inclusive não representam necessariamente a posição oficial da Atena Editora. Permitido o *download* da obra e o compartilhamento desde que sejam atribuídos créditos ao autor, mas sem a possibilidade de alterá-la de nenhuma forma ou utilizá-la para fins comerciais.

Todos os manuscritos foram previamente submetidos à avaliação cega pelos pares, membros do Conselho Editorial desta Editora, tendo sido aprovados para a publicação com base em critérios de neutralidade e imparcialidade acadêmica.

A Atena Editora é comprometida em garantir a integridade editorial em todas as etapas do processo de publicação, evitando plágio, dados ou resultados fraudulentos e impedindo que interesses financeiros comprometam os padrões éticos da publicação. Situações suspeitas de má conduta científica serão investigadas sob o mais alto padrão de rigor acadêmico e ético.

Conselho Editorial
Ciências Exatas e da Terra e Engenharias
Prof. Dr. Adélio Alcino Sampaio Castro Machado – Universidade do Porto
Profª Drª Alana Maria Cerqueira de Oliveira – Instituto Federal do Acre
Profª Drª Ana Grasielle Dionísio Corrêa – Universidade Presbiteriana Mackenzie
Profª Drª Ana Paula Florêncio Aires – Universidade de Trás-os-Montes e Alto Douro
Prof. Dr. Carlos Eduardo Sanches de Andrade – Universidade Federal de Goiás
Profª Drª Carmen Lúcia Voigt – Universidade Norte do Paraná

Prof. Dr. Cleiseano Emanuel da Silva Paniagua – Instituto Federal de Educação, Ciência e Tecnologia de Goiás
Prof. Dr. Douglas Gonçalves da Silva – Universidade Estadual do Sudoeste da Bahia
Prof. Dr. Eloi Rufato Junior – Universidade Tecnológica Federal do Paraná
Profª Drª Érica de Melo Azevedo – Instituto Federal do Rio de Janeiro
Prof. Dr. Fabrício Menezes Ramos – Instituto Federal do Pará
Profª Dra. Jéssica Verger Nardeli – Universidade Estadual Paulista Júlio de Mesquita Filho
Prof. Dr. Juliano Bitencourt Campos – Universidade do Extremo Sul Catarinense
Prof. Dr. Juliano Carlo Rufino de Freitas – Universidade Federal de Campina Grande
Profª Drª Luciana do Nascimento Mendes – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte
Prof. Dr. Marcelo Marques – Universidade Estadual de Maringá
Prof. Dr. Marco Aurélio Kistemann Junior – Universidade Federal de Juiz de Fora
Prof. Dr. Miguel Adriano Inácio – Instituto Nacional de Pesquisas Espaciais
Profª Drª Neiva Maria de Almeida – Universidade Federal da Paraíba
Profª Drª Natiéli Piovesan – Instituto Federal do Rio Grande do Norte
Profª Drª Priscila Tessmer Scaglioni – Universidade Federal de Pelotas
Prof. Dr. Sidney Gonçalo de Lima – Universidade Federal do Piauí
Prof. Dr. Takeshy Tachizawa – Faculdade de Campo Limpo Paulista

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

Uma abordagem na adoção de Business Intelligence

Diagramação: Natália Sandrini de Azevedo
Correção: Yaiddy Paola Martinez
Indexação: Amanda Kelly da Costa Veiga
Revisão: Daysianne Kessy Mendes Isidorio
Autor: Matheus Emerick de Magalhães

Dados Internacionais de Catalogação na Publicação (CIP)

M188 Magalhães, Matheus Emerick de
Uma abordagem na adoção de Business Intelligence / Matheus Emerick de Magalhães. – Ponta Grossa - PR: Atena, 2022.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-258-0726-3
DOI: https://doi.org/10.22533/at.ed.263222111

1. Inteligência competitiva (Administração). I. Magalhães, Matheus Emerick de. II. Título.
CDD 658.4038

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

DECLARAÇÃO DO AUTOR

O autor desta obra: 1. Atesta não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao conteúdo publicado; 2. Declara que participou ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certifica que o texto publicado está completamente isento de dados e/ou resultados fraudulentos; 4. Confirma a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhece ter informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autoriza a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

SUMÁRIO

RESUMO

Na sociedade atual em que a competitividade empresarial é cada vez mais frequente e desafiadora, a capacidade de transformar dados em informações e gerar conhecimento passou a ser o maior alvo de busca nas organizações pelo controle de seus negócios. Neste contexto, são empregados ferramentas e conceitos que organizam as informações de BI. Dentre estes conceitos e ferramentas, além do Data Warehouse (DW), destacam-se também Data Mart (DM), Data Mining e ferramentas OLAP, que constituem os pilares estratégicos dos Sistemas de Apoio à Decisão (SAD). São adotadas medidas de planejamento e gerenciamento que visam melhorar a segurança das informações obtidas e proteger os processos da organização, garantindo assim o melhor andamento dos negócios. O objetivo deste livro é mostrar as principais ferramentas e metodologias de BI, os desafios de sua utilização, e demonstrar sua utilização através da ferramenta de mineração de dados Weka.
PALAVRAS-CHAVES: Armazém de Dados, *OLAP e Data Mart*

ABSTRACT

In today's society where business competitiveness is increasingly frequent and challenging, the ability to transform data into information and generate knowledge has become the biggest target of search in organizations to control their business. In this context, tools and concepts are used to organize BI information. Among these concepts and tools, in addition to the Data Warehouse (DW), Data Mart (DM), Data Mining and OLAP tools also stand out, which constitute the strategic pillars of Decision Support Systems (DSS). Planning and management measures are adopted that aim to improve the security of the information obtained and protect the organization's processes, thus ensuring the best progress of the business. The objective of this book is presente the main BI tools and methodologies, the challenges of their use, and demonstrate their use through the Weka data mining tool.

KEYWORDS: *Data warehouse, OLAP e Data Mart*

INTRODUÇÃO

A competitividade está cada vez mais presente nas organizações e a informação tornou-se o bem de maior expressão no ambiente empresarial. A necessidade de relacionar informações para a realização de uma gestão empresarial eficiente é atualmente tão presente nas organizações, que se torna um importante diferencial na tarefa de tomada de decisões.

O atual interesse pelo *Business Intelligence* vem crescendo na medida em que sua utilização possibilita às organizações realizar uma série de análises, projeções e comparações, de modo a facilitar os processos pertinentes à tomada de decisões.

Contudo algumas premissas devem ser adotadas visando a utilização desta ferramenta, possibilitando uma análise efetiva dos dados para uma tomada de decisão eficaz, como um plano de implantação e comprometimento composto por um plano de ação, relacionando os 4 pilares necessários, pessoas, processos, tecnologia e informações.

Esse estudo tem como objetivo principal a conscientização da importância do *Business Intelligence* em ambientes corporativos, destacando a necessidade de se ter os processos internos da empresa bem definidos para a obtenção dos resultados esperados a fim de proteger a empresa dos equívocos gerados pela sua má utilização ou uma utilização ineficaz.

São empregados conceitos e ferramentas que organizam as informações de negócios inteligentes das empresas. Dentre estes conceitos e ferramentas, além de *Data Warehouse*, destacam-se também *Data Mart* (DM), *Business Intelligence* (BI) e ferramentas OLAP, que constituem os pilares estratégicos dos Sistemas de Apoio à Decisão.

Para isso o trabalho está organizado da seguinte forma: no Capítulo 2 são abordadas as aplicabilidades de utilização do *Business Intelligence* numa organização. O Capítulo 3 destaca os principais tipos de ferramentas no mercado que proporcionam o suporte necessário para a utilização do *Business Intelligence*. O Capítulo 4 demonstra a ferramenta de mineração de dados Weka e sua utilização em um estudo de caso real. Por fim têm-se as considerações finais e referências bibliográficas utilizadas na elaboração deste trabalho.

O BUSINESS INTELLIGENCE EM AMBIENTE ORGANIZACIONAL

Nesse capítulo são listados os conceitos, características e ideias sobre as aplicabilidades do BI, sua utilização no contexto atual nas organizações, as transformações decorrentes das mudanças de se trabalhar com informações, a necessidade de ter métodos para se tomar informações mais precisas e eficazes para a tomada de decisão, sendo explicado todo o ciclo que envolve esses conceitos.

CONCEITOS DE *BUSINESS INTELLIGENCE*

O termo SAD (Sistemas de Apoio a Decisão) está sendo utilizado cada vez menos, tanto em livros, revistas especializadas, quanto em sites na Internet. Em seu lugar tem sido cada vez mais frequente o uso do termo *Business Intelligence* (BI). BI é um conceito empregado a metodologias, ferramentas e tecnologias, que tem como objetivo fornecer informações estratégicas, que apoiam as organizações na tomada de decisão.

A literatura apresenta diversas definições para o termo *Business Intelligence* (BI). Segundo Primak (2008), "BI é um conceito que através de ferramentas específicas, ou seja, software, auxilia os gestores na tomada eficaz de decisões usando apenas os dados e informações que estão inseridos nos bancos de dados das empresas."

O conceito de BI é abstrato e gera margens a perguntas interessantes como, "Então é um software que precisa ser instalado na empresa?".

O BI não é necessariamente apenas um software a ser instalado nas máquinas da empresa, é uma mudança cultural na empresa, que deve obter o apoio integral da alta cúpula administrativa, a participação dos funcionários, principalmente na inserção de dados e no uso adequado de tecnologia de transformação de dados em informações.

Reformulando o conceito, entende-se sendo BI uma forma de ajudar os gestores na escolha de uma decisão para determinada situação, utilizando programas de computadores e uma metodologia que apresenta os dados e informações que foram inseridos pelos outros usuários numa forma mais adequada para a interpretação das informações.

O BI propõe ajudar as organizações em todos seus níveis a obter informações mais rápidas, melhores e eficazes, impulsionando os negócios a caminho do sucesso da organização, não sendo um campo exclusivo para grandes organizações, gerentes e analistas.

As organizações de ambos os portes e tamanhos, podem agora ter acesso ao BI, para o auxílio na tomada de decisão e gerenciamento de seus negócios, com o apoio de inúmeras ferramentas disponíveis no mercado a diversos segmentos e faixas de preço (HABERMANN, 2006).

DESAFIOS DE NEGÓCIO

Segundo Microsoft (2006), tradicionalmente, as empresas adquiriam e implementavam ferramentas de BI especializadas, conforme as necessidades. Geralmente, estas ferramentas eram departamentais, para usuários empresariais ou programadores que as selecionavam e que raramente tinham em mente uma estratégia de BI a nível organizacional, tratando apenas de uma necessidade particular ou de um setor específico da empresa.

Existem muitas razões para que a abordagem tradicional sobre BI não tenha cumprido totalmente a sua promessa de disponibilizar informações para toda uma organização. Algumas estão listadas a seguir (MICROSOFT, 2009):

- a necessidade de utilizar software de BI especializados, problemas de adaptação destas ferramentas, dedicação de recursos consideráveis em treinamento e tempo em aprendizado específico em cada software adotado;
- a utilização e o acesso a ferramentas de BI tem sido uma especialidade local, ou, departamental, sendo tipicamente de domínio exclusivo dos setores de decisões, departamento financeiro ou de analistas de negócio, deixando de lado uma das premissas para a obtenção de êxito do BI, a proposta da visão total do processo e integração das informações na organização.

Como resultado implementar e manter soluções de BI especializadas, que na maioria das vezes não produzem os resultados esperados ou tangíveis tem sido muitas vezes uma tarefa onerosa e com grande demanda do fator tempo de produção.

O *BUSINESS INTELLIGENCE* PARA PEQUENAS E MÉDIAS ORGANIZAÇÕES

Segundo Habermann (2006) ‘’não precisa ser uma grande organização para definir seu planejamento estratégico e iniciar um projeto de inteligência de negócios. O primeiro passo é avaliar se a informação a ser analisada já está disponível’’.

De acordo com Habermann (2006) a ênfase no planejamento estratégico poderia trazer benefícios significativos a pequenas e médias empresas, cujos processos de início geralmente são incompletos e informais, sendo em sua maioria decorrentes de empresas familiares e da falta de estruturação de abertura e implantação da empresa.

As empresas que conseguem superar as dificuldades e dar início a uma gestão voltada a informações, controlada e orientada para o mercado, podendo se adequar a realidade presente ou futura, serão capazes de oferecer com rapidez serviços e produtos precisos e a tempo esperado a seus clientes, proporcionando um diferencial de produtos e serviços ao mercado (HABERMANN, 2006).

Com a iniciativa de mudanças, as empresas poderão utilizar o BI para proporcionar

ganhos nos processos decisórios, e planejar futuras adequações, expansões e melhorias de seus produtos e serviços.

O desenvolvimento de uma estratégia é em essência o desenvolvimento de uma fórmula ampla que norteará o modo como uma empresa irá competir, quais serão suas metas a curto, médio e longo prazo, e quais serão as políticas necessárias para o cumprimento destas metas.

Além do planejamento estratégico, o cenário atual é dinâmico e exige que as pequenas e médias empresas tenham capacidade de resposta imediata, apoiada por um processo de tomada de decisões rápido, claro, objetivo e dinâmico. Para tanto, a informação precisa estar disponível para as pessoas certas, no formato esperado, no momento e local desejados, e no tempo apropriado.

Neste contexto, a informação representa um recurso de alto teor estratégico, que necessita ser aproveitado como gerador de diferenciais e vantagens competitivas.

No entanto, pressionadas por custos e pela falta de experiência, pequenas e médias empresas buscam soluções tecnológicas insuficientes, sem poder investir em esforços necessários para a iniciativa de uma estratégia voltada ao *Business Intelligence*. Projetos de *Business Intelligence* utilizam software e metodologias para análise de padrões e gestão da informação e alguns recursos mínimos muitas vezes fora do alcance de muitas pequenas e médias empresas (HABERMANN, 2006).

A grande maioria de pequenas e médias empresas ainda não tem uma infraestrutura ideal para a implantação de projetos de BI, uma vez que muitas informações estão armazenadas de forma desestruturada, em planilhas ou arquivos textos.

Também é comum a falta de procedimentos e processos bem definidos, que facilitam a gestão voltada aos negócios e viabilizam a programação de sistemas de informação para apoio operacional às rotinas de trabalho, deixando de lado a metodologia proposta do BI, que especifica que os processos devem estar claros e definidos.

Por isso, pensando em possuir uma ferramenta de informações gerenciais que possa dar suporte a decisões estratégicas, é necessário iniciar uma mudança na maneira de armazenar e trabalhar informações. É preciso utilizar sistemas de informação e bancos de dados estruturados, em que é possível obter respostas rápidas e mais confiáveis de uma determinada pesquisa, além de ter histórico das transações ocorridas.

O interessante desta mudança para as empresas é que ela não envolve grandes investimentos em infraestrutura tecnológica, uma vez que hoje já existem excelentes bancos de dados gratuitos e confiáveis que podem ser utilizados de maneira muito profissional.

Muitos sistemas gratuitos também são encontrados e atendem perfeitamente a processos internos de uma pequena e média empresa, como fluxos de caixa, contas a pagar e receber, estoque e etc. Há profissionais qualificados capazes de projetar sistemas

de informação para atender aos processos de trabalho das empresas.

Portanto, a importância do BI no planejamento estratégico começa a ser sentida a partir do momento em que a pequena e média empresa adota uma postura de trabalho mais voltada à gestão da informação.

Somente com a informação íntegra e confiável é possível criar estratégias que atendam melhor a seus clientes e colocar a empresa em um patamar de competitividade mais lucrativo.

SAD (SISTEMAS DE APOIO A DECISÕES)

O BI é parte da categoria de sistemas denominados SAD (Sistemas de Apoio à Decisão), que têm como função apoiar as tomadas de decisões nas organizações, sendo que sua utilização tem sido uma poderosa ferramenta no auxílio à busca, à extração e à armazenagem de informações.

Shim *et al*. (2002) afirmam que os SADs são soluções computacionais desenvolvidas para apoiar a tomada de decisões complexas durante a resolução de problemas. Ferramentas clássicas de SAD compreendem componentes para gerenciamento de sofisticados bancos de dados, poderosas funções de modelagem e poderosos, embora simples, projetos de interface com o usuário, que permitem trabalhar interativamente com questões, relatórios e funções gráficas.

Carlsson e Turban (2002) explicam que o termo SAD, propriamente dito, tem sido pouco empregado, tanto em revistas especializadas quanto em *Websites* de vendas, e no seu lugar tem sido cada vez mais frequente o uso de termos como *Business Intelligence* e OLAP[1]. Do mesmo modo, estes termos praticamente eliminaram o uso do termo EIS (*Executive Information Systems*).

Por outro lado, está crescendo o reconhecimento de que BI está se tornando um componente necessário na chamada segunda geração dos sistemas ERP, que claramente reconhece a necessidade de dar suporte não apenas ao processamento de transações operacionais, mas também ao processamento de análises.

Grigori *et al*. (2004) completam informando que com dados limpos e agregados sobre um determinado processo, armazenados em um *Data Warehouse*[2], é possível realizar análises utilizando-se tecnologias de BI e extrair conhecimento sobre as circunstâncias que levaram a determinado resultado no passado, tenha o resultado sido bom ou ruim.

Assim, é possível utilizar essas informações para explicar por que tais circunstâncias ocorreram e para predizer potenciais problemas nos processos em andamento.

1. OLAP é processo de análise analítico das informações em uma fonte de dados, em visão multidimensional.
2. *Data Warehouse* é um armazém ou repositório, em que os dados são armazenados.

Os *Data Warehouses*, OLAP e *Data Mining* surgiram no começo dos anos 90 como novas ferramentas para SAD, e formam a base dos sistemas de BI (SHIM *et al.*, 2002).

E-BUSINESS

E-business vem da palavra inglesa *Electronic Business*; é o termo que se utiliza para identificar os negócios efetuados por meios eletrônicos, geralmente aplicados na Internet.

Segundo Siegel (2000), atualmente os sistemas de *BI* tendem a estar integrados na Web sendo assim fazem parte de *e-business*. Pode-se definir *E-business* como negócios feitos através da Internet, desde contatos diretos com consumidores, fornecedores, como também análises de mercado, análises de investimentos, busca de informações, pesquisa de mercados, etc.

E-business tem como aplicação a criação de sistemas capazes de prover comunicação entre empresas e pessoas, agilizando os processos de compra e venda de produtos e serviços entre as mesmas, facilitando assim todo o processo de fabricação e venda, melhorando a disponibilidade de produtos de acordo com a demanda pelos mesmos, não havendo barreiras territoriais e promovendo um comércio internacional (SIEGEL, 2000).

"O negócio eletrônico não exige que uma empresa faça tudo on-line para os clientes, apenas que todos os funcionários utilizem as ferramentas da Internet para servir melhor ao cliente (SIEGEL, 2000)."

A demanda pelo *E-business* cresce com o aumento do número de *e-custumers* (compradores via Internet), cada vez mais exigentes na busca por ambientes diferenciados dos costumeiros locais de compras e serviços, sendo esse um desafio para o *e-business,* implantar uma empresa baseada na WEB e não simplesmente inserir sua empresa na Internet.

Grandes consequências são adquiridas com essa métrica de trabalho e conscientização, dentre elas vale ressaltar a busca pela fidelidade dos clientes, dando valor agregado à marca.

A Fig. 1 demonstra como o mundo ficou organizado antes e depois do ano 2000, tendo um diferencial na conduta dos negócios, sendo anteriormente baseado na oferta de mercado, e atualmente conduzido pelo cliente e suas necessidades reais.

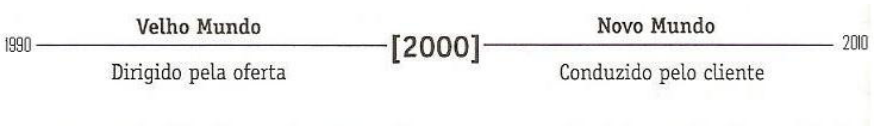

Figura 1 - Transição do velho mundo (SIEGEL, 2000).

CONCLUSÃO

Nesse capítulo foram abordados os principais conceitos e particularidades que envolvem o ambiente do BI no ambiente de negócios, mostrando as principais características, o contexto do BI em pequenas e médias organizações, o perfil dos atuais clientes, tendências e necessidades do mercado atual do ponto de vista de autores renomados e de grande expressão na literatura atual.

No próximo capítulo são abordadas as principais tecnologias do BI.

AS PRINCIPAIS TECNOLOGIAS DE BUSINESS INTELLIGENCE

A utilização e uso das ferramentas de BI crescem diariamente, acompanhando a evolução dos negócios, com escopos e tendências diferentes para cada necessidade particular das organizações.

Este capítulo apresenta as principais tecnologias de BI, começando na fase de preparação para o BI com a criação do *Data warehouse*, *Data Mart*, a utilização de tecnologias de OLAP (*On Line Analytical Processing* – Processamento *On-line* Analítico) e utilização da Mineração de Dados.

DATA WAREHOUSE

Trajetória histórica

Segundo Inmon (2002), os fundamentos de bancos de dados relacionais surgiram na empresa IBM, nas décadas de 1960 e 1970, através de pesquisas de funções de automação de escritório. Foi um período no qual empresas observaram que estava custoso empregar um número grande de pessoas para fazer trabalhos como armazenar e indexar os arquivos (dados). Por este motivo, houveram grandes investimentos em pesquisas de um meio mais barato para ter uma solução eficiente.

Diversas pesquisas foram conduzidas durante este período, no qual modelo hierárquico, de rede, relacional, e outros modelos foram descobertos, bem como muita tecnologia utilizada hoje em dia, como os discos de armazenamento, possibilitando um novo tipo de software de gerenciamento de banco de dados (DBMS) a data base *management system* (INMON, 2002).

Codd (1970) foi responsável pela primeira publicação de um artigo sobre base de dados relacionais, e o uso da álgebra relacional para que usuários não técnicos tivessem acesso à recuperação de grandes quantidades de informações, através de comandos (em inglês), iniciando assim uma nova fase em que o usuário começava a ter autonomia no acesso aos dados.

Conceitos

O primeiro passo é entender o conceito de *Data*, que são os dados presentes nas bases de dados geralmente em arquivos textos, e inseridos previamente pelos usuários do sistema. *Warehouse* significa uma espécie de armazém, ou lugar apropriado para armazenar determinado produto. Então pode-se entender por *Data Warehouse* como um armazém de dados, ou apenas um lugar em que os dados serão armazenados.

Muito é discutido sobre como começar um projeto de BI, o *Data Warehouse* (DW) hoje é considerado como um dos pilares do BI.

Segundo Inmon (2002), o Data warehouse é um conjunto de dados baseado em assuntos, integrado, não-volátil, e variável em relação ao tempo, de apoio às decisões "gerenciais".

Um DW pode ser definido como um conjunto de técnicas e de bancos de dados integrados, projetados para suportar as funções dos Sistemas de Apoio à Decisão (SAD), em que os determinados assuntos estão relacionados a uma mesma base de dados. Sua meta é fornecer subsídios e informações aos gerentes e diretores, para que assim possam analisar tendências de seus clientes e, com isso, melhorarem e unificarem os processos e agilizarem as tomadas de ações.

Segundo Kimball (2002) é necessário ter a convicção de que o *Data Warehouse* deve ser voltado às necessidades dos usuários das áreas de negócios e, portanto, construído e mostrado em uma perspectiva dimensional simplista.

Objetivos de um *Data Warehouse*

A definição de um DW consiste em armazenar os dados em vários graus de relacionamento e sumarização, de forma a facilitar e agilizar os processos de tomada de decisão por diferentes níveis gerenciais. Esses dados, oriundos de sistemas de informação de produção, deverão estar "mastigados", integrados e disponíveis, permitindo diversas formas de consultas, através dos mecanismos amistosos das ferramentas de usuários (BARBIERI, 2001).

Segundo Kimball (2002) os objetivos do DW podem ser adquiridos gradativamente percorrendo diversos setores da empresa e escutando a gerência. Inevitavelmente, as questões listadas a seguir tendem a aparecer.

- "Dispomos de um enorme volume de dados na empresa, mas não conseguimos acessá-los".
- "Temos que combinar os dados a partir de fontes diversas".
- "É preciso tornar mais fácil para os profissionais acessaram diretamente os dados".
- "Apenas mostre-me o que é realmente importante".
- "Fico louco quando duas pessoas apresentam exatamente os mesmos indicadores em uma reunião usando apenas números diferentes".
- "Queremos que as pessoas usem informações para dar um suporte à tomada de decisões baseando-se em fatos".

Transformando essas afirmativas em requisitos de DW tem-se:

- O DW deve fazer com que informações de uma empresa possam ser facilmente

acessadas. O conteúdo do DW deve ser compreensível. Os dados devem ser intuitivos e óbvios para o usuário da área de negócios e não apenas para o desenvolvedor. O conteúdo do DW precisa ser identificado de modo significativo.

- As ferramentas usadas para acessar o DW também devem ser simples, objetivando retornar para os usuários informações no menor tempo de espera possível.
- O DW deve apresentar as informações da empresa de modo consistente, sendo os dados confiáveis, filtrados e submetidos a um controle de qualidade e liberando apenas quando estiverem prontos para serem utilizados.
- O DW deve ser adaptável e flexível a alterações, pois mudanças são inevitáveis, podendo ser decorrentes do tempo de utilização do DW, devido a necessidades de usuários e particularidades departamentais. Para tal o DW deve ser planejado para lidar com mudanças.
- O DW deve ser o baluarte seguro que protege as informações, tendo-se em vista que o bem mais precioso das organizações está guardado nos DW, os dados.
- O DW deve funcionar como a base para uma melhor tomada de decisão, devendo conter os dados apropriados para dar suporte à tomada de decisões, pois essas decisões resultam do impacto e do valor comercial atribuído ao DW.

Para concluir cabe destacar que a implantação de um DW de sucesso demanda muito mais que apenas um conhecimento técnico em Tecnologia da informação. Os profissionais devem estar munidos de conhecimentos específicos em gerenciamento de negócio.

Data Mart

Na busca pela redução de custos e riscos na implantação de um Data Warehouse, são apresentados os *Data Mart*, que são pequenos Data Warehouse que fornecem informações a um menor grupo de pessoas, por ser de origem departamental.

Sendo que o tempo de investimento e implantação são menores se comparado aos Data Warehouse, e seu risco de implantação e desenvolvimento também é significamente reduzido.

De acordo com Inmon (2002) *Data Mart* é um tipo especial de armazém contendo dados específicos para uma área ou departamento da empresa. É um subconjunto dos dados empresariais que contém dados úteis apenas para uma unidade de negócio específica ou departamento, criado para dar suporte ao processo de tomada de decisão.

Segundo especialistas no assunto a única diferença entre DW e *Data Mart* é o

escopo a ser desenvolvido. Um *Data Mart* trata das questões departamentais ou locais (de um departamento específico), enquanto um DW envolve as necessidades de toda a companhia de forma que o suporte à decisão atue nos diversos níveis da organização (INMON, 2002).

Existe um interessante ponto de vista do autor Kimball (2002) considerado um dos mais influentes gurus do BI, que discorda dessa definição e argumenta que os Data Marts não devem ser departamentais, mas sim orientados aos dados ou a fontes de dados. Ele defende a ideia de que os *Data Marts* devem ser orientados a assuntos e não departamentais. Ele exemplifica o caso de uma instituição bancária que dispõe de uma fonte de dados de contas correntes e poupança.

Nesse caso deveria ser criado um *Data Mart* de Contas, que não será um *Data Mart* proprietário da área financeira, e nem da área de *marketing*, mas sim um repositório de dados que terá como público todos os usuários dos departamentos que lidam com aquele assunto.

Kimball (2002) defende ainda que na maioria das vezes os *Data Mart* devem ser construídos isoladamente, para se unirem na criação de um único DW.

Componentes de um *Data Warehouse*

Cada componente do DW possui uma função específica, e para compreendê-los deve-se aprender a controlar cada um de modo eficaz para obter êxito na sua utilização.

Segundo Kimball (2002), os principais componentes a serem considerados são:

- sistemas operacionais de origem: capturam pequenas quantidades de dados da empresa, sem nenhum ou pouco controle sobre o conteúdo inserido, sendo externos ao DW. As prioridades desses sistemas são desempenho e disponibilidade de processamento, sendo suas consultas limitadas, feitas de um registro por vez, diferentemente de como os DW costumam ser;
- *data staging área:* é a área de armazenamento do DW, geralmente denominada de ETL (*Extract-Transformation-Load*) extração, transformação e carga. Transformando os dados brutos presentes nos sistemas operacionais em formatos de DW, prontos para serem utilizados por profissionais qualificados, preservando suas informações dos usuários finais ou denominados clientes. A extração é o primeiro passo para enviar os dados para a *staging área*, ocorrendo transformações em potencial como filtragem dos dados, principalmente para correção de erros e discrepância, decorrentes principalmente de fontes externas aos sistemas como digitalizações incorretas, dentre outras. Sendo que a última parte consiste em carregar os dados para os *Data Marts* de forma que os dados possam ser publicados e apresentados para os usuários;

- apresentação dos dados: é o local em que os dados ficam organizados e armazenados, tornando disponíveis para serem consultados diretamente pelos usuários, através de relatórios e outras aplicações, usa-se o termo *"getting the data out"* para a obtenção de dados;
- ferramentas de acesso a dados: são as ferramentas específicas de acesso aos dados, sendo de uso destinado aos profissionais que necessitam da informação em tempo hábil.

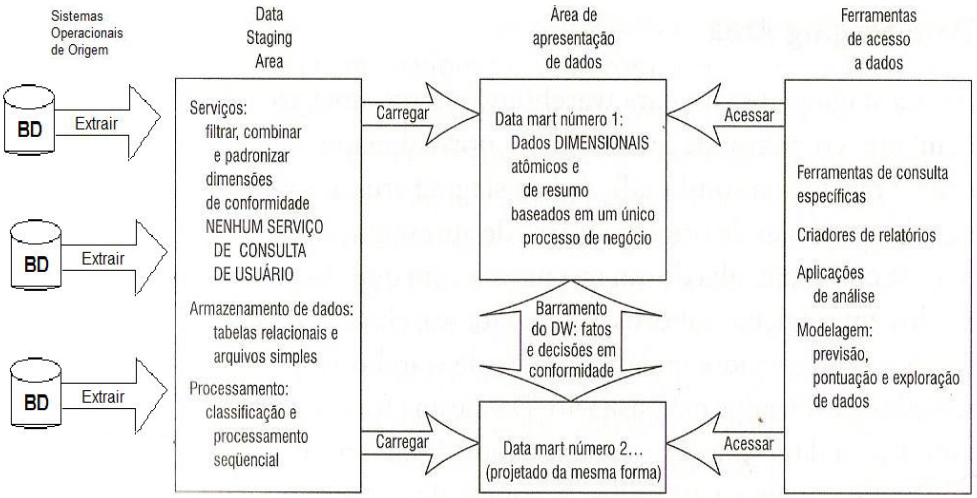

Figura 2- *The Data Warehouse Toolkit* (KIMBALL, 2002).

Adversidades Encontradas

Segundo Kimball (2002), o maior problema do *Data Warehouse* é a sua grande complexidade. Sua criação requer pessoas altamente especializadas e comprometidas com o processo, uma metodologia consistente, computadores eficientes, banco de dados, ferramentas de *front-end* (sistemas transacionais para captura dos dados), ferramentas para extração e limpeza dos dados, e treinamento dos usuários. É um processo complicado e demorado, que requer altos investimentos e que se não for corretamente planejado e executado, pode trazer prejuízos enormes dentro da organização.

Deve-se observar atentamente para a necessidade ou não da criação de um DW, efetuar estudos sobre viabilidade e a real necessidade para sua criação.

Segundo Inmon (2002) mensurar o custo de desenvolvimento de um DW antes de sua criação é uma tarefa árdua, felizmente o DW pode ser construído incrementalmente, ou seja, por fases, podendo ser mais rápido e de menor custo financeiro se incrementado de acordo com a necessidade do negócio.

OLAP

Introdução

A aplicação OLAP proporciona análise e consolidação de dados, pois é o processamento analítico *online* dos dados. Tem capacidade de visualizações das informações a partir de perspectivas diferentes.

A visualização é realizada em dados agregados, e não em dados operacionais pois a aplicação OLAP tem por finalidade apoiar os usuários finais a tomar decisões estratégicas. Os dados são apresentados em termos de medidas e dimensão (ORACLE, 2005). Neste tópico está sendo demonstrada a tecnologia OLAP, o entendimento de sua origem, seus conceitos e características, suas funções e sua ligação com DW e DM. Além disto, mostrará como o OLAP auxilia o gestor na tomada de decisões.

Origem

A análise Multidimensional para OLAP (*Online Analytical Processing* – Processamento Analítico Online) não é nova. De fato, ela inicia com a publicação do livro *A Programming Language* (IVERSON, 1966). A IBM desenvolveu a primeira linguagem multidimensional, chamada APL, sendo amplamente utilizada nos negócios nas décadas de 80 e 90. Na década de 90 devido a avanços tecnológicos e a utilização de conceitos da APL foi criado uma nova classe de ferramentas chamada de OLAP, que possuía ferramentas desenvolvidas por grandes empresas como IBM, Microsoft, Microstrategy, Oracle, entre outras.

O termo OLAP foi citado pela primeira vez por Codd (1970), instituindo regras que as aplicações deveriam atender. A visão conceitual multidimensional dos negócios de uma empresa foi umas das regras citadas, sendo fundamental no desenvolvimento destas aplicações. A visão multidimensional consiste de consultas que fornecem dados a respeito de medidas de desempenho, decompostas por uma ou mais dimensões dessas medidas. Podendo também ser filtradas pela dimensão e/ou pelo valor da medida. As visões multidimensionais fornecem as técnicas básicas para cálculo e análise requeridos pelas aplicações de BI.

Multidimensionalidade

De acordo com Codd (1970), para compreender OLAP e seu funcionamento são necessários conhecimentos dos conceitos sobre visão multidimensional, que é a forma que o OLAP funciona, como uma espécie de cubo.

Segundo Microsoft (2002), um cubo agrega os fatos em níveis de dimensão e fatos, o BI usa a palavra “cubo”, porque descreve melhor os dados resultantes da relação dos dados.

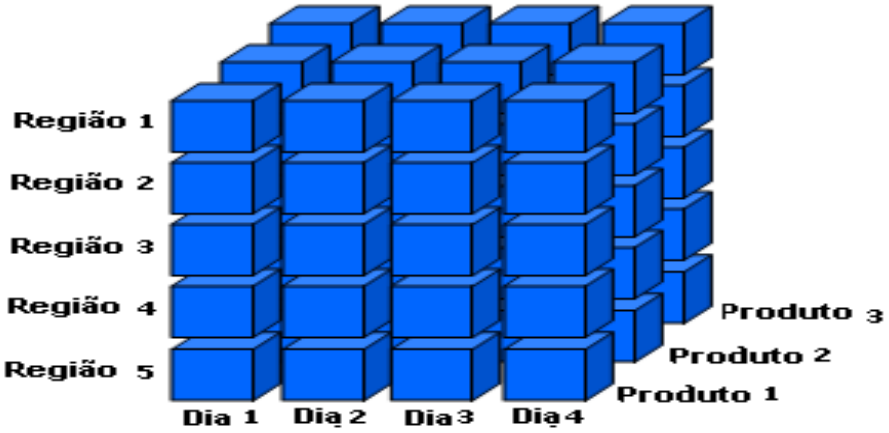

Figura 3– O Que São Cubos (MICROSOFT, 2002).

Segundo Codd (1970), alguns dos fatores que compõem a tecnologia OLAP são:

- Cubo é uma estrutura que armazena os dados de negócio em formato multidimensional, facilitando sua análise.
- Dimensão são os parâmetros aplicados no cubo, como exemplo, tempo, hora mês.
- Hierarquia é a divisão em níveis, ou seja, e o agrupamento dos dados em grupos, sendo esses grupos divididos em níveis hierárquicos de acordo com sua característica. Com o exemplo em um ambiente de clientes o cliente X não possui casa própria e o cliente Y possui.
- Membro é um subconjunto de uma dimensão. Cada nível hierárquico tem membros apropriados para aquele nível.
- Medida é uma dimensão utilizada para realizar comparações. Ela inclui membros tais como: custos, lucros ou taxas.

Segundo Microsoft (2009) um cubo OLAP deve prover as seguintes operações:

- *drill down*: significa descer um nível hierárquico em uma dimensão. Ex: dimensão tempo, ano para o trimestre e trimestre para o mês;
- *drill up/ roll up*: significa subir um nível hierárquico em uma dimensão. Ex: dimensão produto, subir o nível do produto para categoria do produto;
- *drill across*: significa analisar um nível intermediário dentro de uma mesma dimensão. Ex: dimensão produto, venda dos produtos num determinado ano, venda de um produto X no ano Y;

- *drill throught*: significa alternar a análise de uma dimensão para outra. Ex: produto para região ou mesmo de uma agregação (todas as vendas de agosto de 2006) para os detalhes (tabela com cada venda no período citado);
- *drill back/ Write back*: é bastante utilizado em previsões e consiste na ação de alterar os valores existentes em um cubo OLAP. Pode ser usado, por exemplo, para medir o impacto na empresa do aumento em 10% do orçamento para o ano seguinte;
- *slice*: significa analisar determinada fatia do cubo OLAP. Ex: analisar determinado produto em uma determinada região;
- *dice*: significa alterar a visão de um cubo OLAP, alterando a análise de vendas dos produtos por região para vendas por faixa etária de cada mês.

Solução OLAP

Segundo Microstrategy (2009) a tecnologia OLAP fornece uma forma de análise mais simples dos dados, permitindo fazer o "*slice e dice*", ou seja, é possível executar tarefas de análise de subconjuntos dos dados ou cubos inter-relacionados com pouco esforço empregado.

Os usuários podem analisar os dados usando recursos OLAP para ter acesso a uma série de visões dos dados de acordo com as necessidades presentes. A análise OLAP oferece aos usuários acesso aos dados contidos nos DW ou *Data Mart* em lugar de funções complexas e avançadas preferidas pelos analistas e usuários mais exigentes.

Microsoft (2009) destaca que os sistemas OLAP baseiam-se não apenas em informações provenientes dos sistemas operacionais, mas também dos diversos software utilizados pelas organizações (como planilhas Excel, arquivos textos, XML, dentre outros) e unificadas com as informações presentes nos DW e *Data Mart.*

A Fig. 4 mostra as fases dos dados até o uso de ferramentas de OLAP.

Figura 4 - Elevando a gestão dos negócios a um novo patamar (MICROSOFT, 2009).

Segundo Inmon (2002) a modelagem de dados é muito importante para as respostas esperadas em consultas complexas, devendo considerar flexibilidade e requisitos propostos pelos usuários.

São adotadas topologias para uso em determinado contexto sendo considerado as vantagens, desvantagens e aplicabilidade oferecidas por cada tipo. Dentre as quais vale destacar: Esquema Estrela ou *Star Schema.*

Segundo Microsoft (2002) são ligeiramente normalizados e compostos por dois tipos básicos de tabelas, de fatos e dimensão.

A tabela de fatos é a tabela central normalizada, que representa as transações contendo os valores que estão sendo analisados e as chaves estrangeiras das tabelas de dimensão. Por exemplo, considerar que será adotado o esquema estrela. Quando cria um cubo a partir desse esquema, deve-se tomar a carga, a quantidade, o desconto, e outros fatos e adicioná-los por cidade, por ano, por cidade e ano, e por todas as outras possíveis combinações de dimensão e nível hierárquico. Essas informações produzem os seguintes tipos de dados estrutura, como exemplificado na Fig. 5.

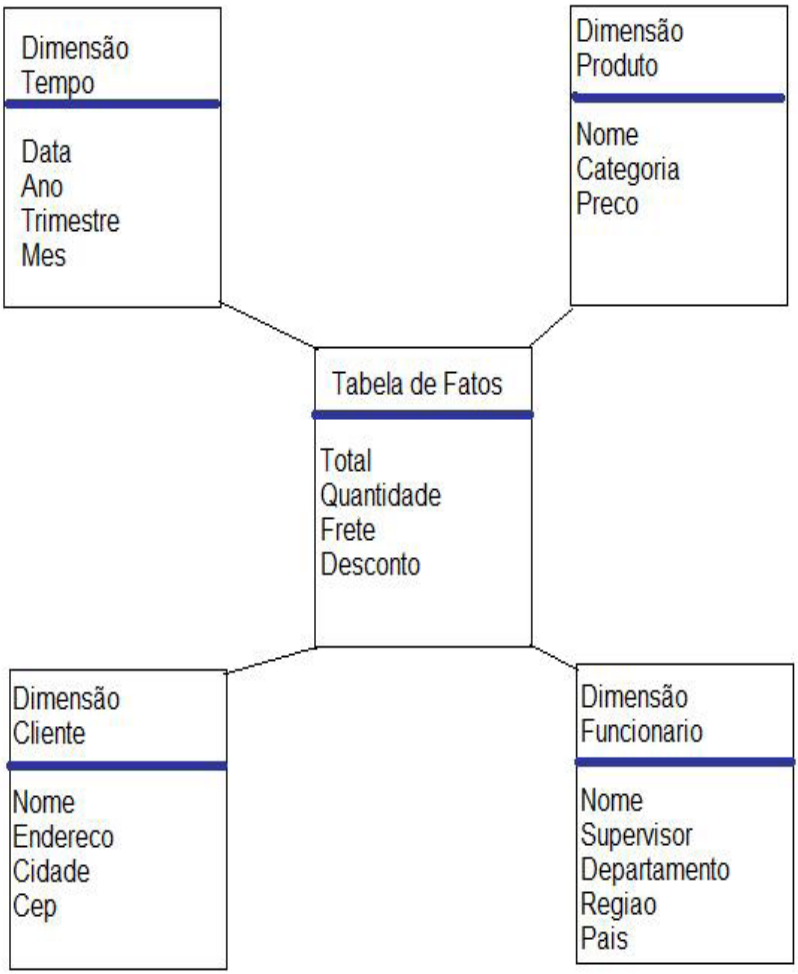

Figura 5 – O Que São Cubos (MICROSOFT, 2002).

De acordo com Oracle (2005), o esquema *Snow Flake* é uma variação do modelo estrela, com visões ajustadas para o enfoque desejado em níveis de hierarquia, permitindo que a visualização permita níveis de visão das informações, no qual as tabelas de dimensão também são normalizadas. A Fig. 6 demonstra um exemplo de esquema flocos de neve:

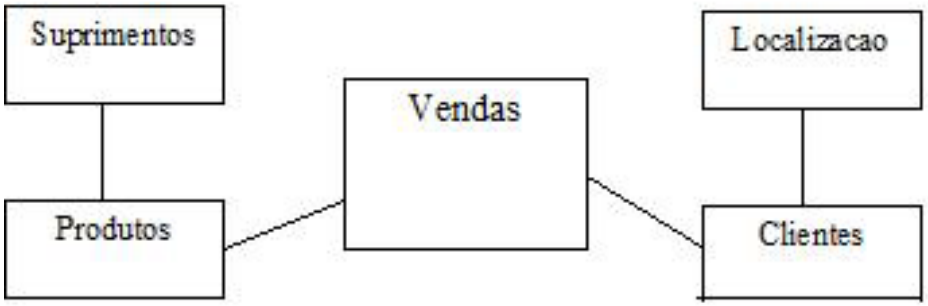

Figura 6 – Guia de Desenvolvimento OLAP (ORACLE, 2005).

Ferramentas e suas Características

Existem muitas ferramentas de OLAP e mudanças têm ocorrido em um ritmo acelerado. Na maioria das ferramentas observa-se a existência de dois grupos de componentes (ORACLE, 2005):

- o grupo de componente do administrador, que é utilizado para gerenciar e gerar os cubos de dados a serem acessados (cubos OLAP);
- o grupo de componente do usuário final que possui acesso aos dados para extraí-los de suas bases de dados, com os quais geram relatórios capazes de responder às suas questões. As ferramentas surgiram juntamente com os sistemas de apoio à decisão (SAD) para fazerem a extração e análise dos dados contidos nos ***Data Warehouse*** e Data ***Mart***.

As principais características destas ferramentas OLAP são: consultas *ad-hoc*: geradas pelos usuários finais de acordo com as suas necessidades de cruzar informações de uma forma não vista e que o leve à descoberta do que procura. Segundo Inmom (2002) "são consultas com acesso casual único e tratamento de dados segundo parâmetros nunca antes utilizados de forma iterativa e heurística"; e as ferramentas de "Slice and Dice" e "Drill down/up", que analisam as bases por determinada perspectiva, de acordo com a necessidade de extração.

Relação do *Data Warehouse* e OLAP

O *Data Warehouse* como visto previamente é utilizado como uma espécie de armazém de informações e o OLAP para recuperar essas informações. As duas tecnologias são complementares, desta forma, para explorar o DW completamente é necessário o OLAP que irá extrair e proporcionar o acesso total às informações nele contidas.

DATA MINING

Introdução

A mineração de dados ou *Data Mining* é parte do ciclo de geração de descoberta em bases de dados KDD (*knowlegde Discovery in databases*), com o objetivo de identificar padrões e modelos aplicáveis em determinada base de dados, sendo estes válidos, novos, úteis e compreensíveis. A mineração de dados assim como outros processos de BI teve maior criação e expansão devido à grande quantidade de dados, informações sem uso e a busca de conhecimento presentes em bases de dados (HAM, 2001).

Segundo Fayyad (1996) "KDD é o processo não trivial de identificação de padrões em dados que sejam válidos, novos, potencialmente úteis e compreensíveis."

Se comparar a mineração de dados com outras formas de mineração, como a de ouro, é possível compreender que se trata de um processo trabalhoso, embora o resultado seja de grande valor.

Outras denominações são utilizadas no processo de busca de conhecimento em bases de dados como:

- *Knowledge mining from databases* (mineração de conhecimento em bases de dados);
- *Knowledge extraction* (extração de conhecimento);
- *Data/pattern analysis* (análise de dados padrão);
- KDD (embora seja apenas uma parte desse processo);

Data Mining, ou Mineração de Dados, pode ser entendida como o processo de extração de informações, de um armazém de dados e seu uso deve ser empregado na tomada de decisões. É uma metodologia aplicada em diversas áreas que usam a gestão do conhecimento, como na geografia espacial, biologia molecular, meios de multimídia, análise de produção, estatística, textos, dentre outros. *Data Mining* define o processo de captura para análise de grandes conjuntos de dados visando extrair um significado (informações úteis), sendo usado tanto para descrever características do passado ou presente, chamado de descrição, ou predizer tendências de futuro, chamada de predição (HAM, 2001).

Fases do KDD

As fases do KDD, variam de acordo com o autor, embora sempre objetive o mesmo resultado. Nesse caso foi adotada a Fig. 7, que destaca com clareza as fases que compõem o ciclo de descobrimento de conhecimento em base de dados:

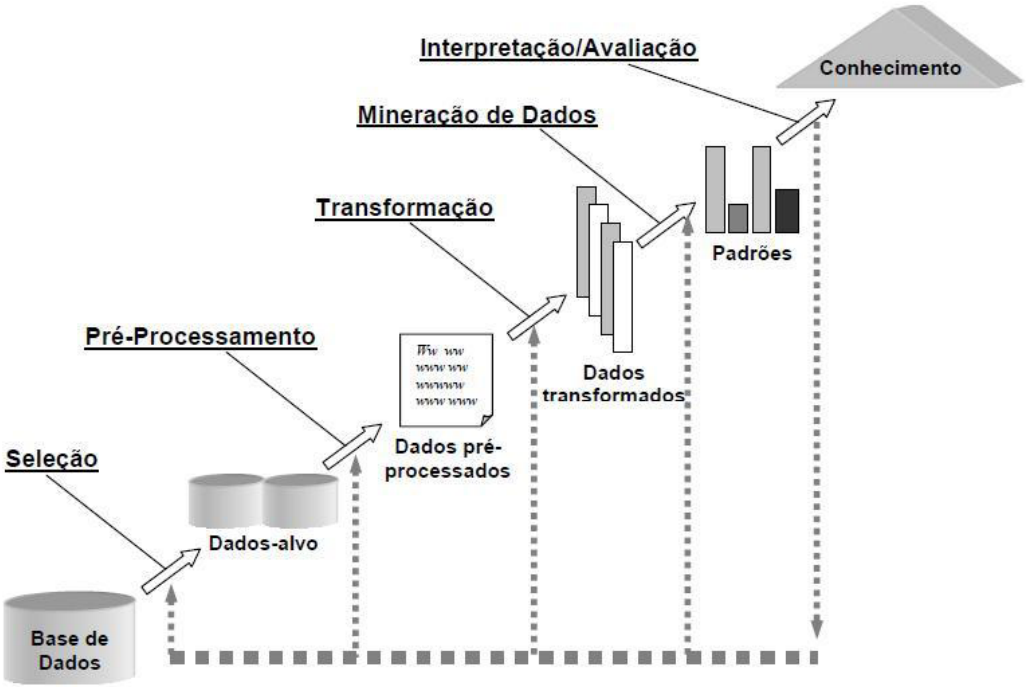

Figure 7 - Fases que compõem o ciclo de descobrimento de conhecimento em base de dados.

Segundo Fayyad (1996), o processo do KDD é iterativo e interativo, composto por várias etapas interligadas, permitindo retornos a fases anteriores para melhor utilização de fases seguintes, como visto no exemplo da Figura 7 que utiliza as setas de transição. As etapas do processo podem ser descritas como:

- entendimento do negócio: sendo o primeiro passo o entendimento e compreensão do negócio, e da real necessidade de se utilizar KDD;
- base de dados: essas bases podem ser uma ou mais bases de dados, data *Warehouse, Data Mart* ou outros tipos de repositórios;
- seleção: consiste no conjunto de dados a serem utilizados, sendo selecionados dados que são relevantes a serem analisados sobre determinada ótica e obtido através da base de dados;
- dados alvos: dados relevantes ao processo. Por exemplo, a necessidade de identificar quais produtos o cliente X compra no supermercado, são encontrados dados alvos como nome dos produtos, tipo do produto, fabricante, e são removidos os dados como CPF do cliente peso do produto, etc;
- pré-processamento: limpeza dos dados, suavizando ruídos, tratando os dados

ausentes (campos vazios), identificando valor com discrepância e verificando a sua consistência. Essas atividades constituem em um conjunto de atividades realizadas sobre os dados extraídos dos dados alvos, de modo a corrigir o uso incorreto ou ineficaz dos códigos e caracteres especiais, resolverem problemas de conflito de domínios, tratarem dados perdidos, corrigir os valores duplicados ou discrepantes. Independente do problema a ser solucionado pela etapa de pré-processamento, a finalidade é deixar os elementos de dados dentro dos padrões adotados, não duplicados, corretos, consistentes e retratando a realidade;

- dados pré-processados: dados sem valores equivocados e/ou discrepantes, isenção de valores ausentes, como visto na Fig.8.

Produto	Valor	Produto	Valor
Fogão	100,00	Fogão	100,00
FOGAO	90.000,00	Fogão	900,00
GELADEIRA		Geladeira	1500,00

Figura 8 - Antes do pré-processamento x Depois do pré-processamento.

- transformação: uma vez que os dados tenham sido extraídos dos sistemas de origem, um conjunto de transformações deve ser processado sobre esses dados. A transformação dos dados pode ser simples ou complexa, dependendo da natureza dos sistemas fontes. Em algumas situações, múltiplos estágios de transformações são necessários (PEREIRA, 2004). Dados são transformados ou consolidados em formas apropriadas para a mineração de dados, visando minimizar o desempenho da sumarização ou funções e agregações para as instâncias (HAM, 2001);
- dados transformados: dados prontos para serem minerados por algoritmos de acordo com sua aplicação e relevância no contexto a ser explorado;
- mineração de dados: o objetivo principal desse passo a aplicação de técnicas (algoritmos) de mineração nos dados pré-processados e transformados, o que envolve ajuste de modelos e/ou determinação de características nos dados. Em outras palavras, exige o uso de métodos inteligentes para a extração de padrões ou conhecimento a partir dos dados. É importante destacar que cada técnica de *Data Mining* utilizada para conduzir as operações de Mineração de

Dados adapta-se melhor a alguns problemas do que a outros, o que impossibilita a existência de um método de *Data Mining* universalmente melhor. Para cada problema tem-se uma técnica particular. Portanto, o sucesso de uma tarefa de *Data Mining* está diretamente ligado à experiência e à intuição do analista;

- padrões: após a utilização da mineração de dados são encontrados padrões válidos e úteis a serem adotados;
- interpretação e avaliação: de nada adianta possuir as informações sem as pessoas para analisar, ou seja, a fase de interpretação e avaliação depende de um analista humano e seus conhecimentos. Por exemplo, um gráfico em um documento é apenas um conjunto de dados, mas o analista possui capacidade de extrair informações e tirar conclusões sobre o assunto em questão;
- conhecimento: utiliza técnicas de visualização e representação do conhecimento para apresentação e demonstrar os conhecimentos adquiridos pelos usuários, e introduzi-los ao nicho estudado.

Aplicação

Segundo Pereira (2004) as aplicações da mineração de dados podem ser vistas do ponto de vista acadêmico e comercial. Dentre elas pode-se citar:

- mineração em Data Warehouse: repositórios com dados de boa qualidade, integrados, estratégicos, históricos, disponibilidade de metadados, infraestrutura de processamento;
- mineração em bancos de dados espaciais: aplicável sobre elementos geográficos, imagens de sensoriamento remoto, imagens médicas, *layout* de chips VLSI, etc. No caso de dados geográficos aplicações relevantes contemplam estudos ambientais, vigilância territorial, detecção de desmatamentos, sistema de água e esgoto, etc.;
- mineração de dados multimídia: extração de padrões relevantes a partir de animações, áudio, vídeo, imagens e textos, busca por similaridades, análise multidimensional, classificatória, preditiva, dentre outros;
- mineração de dados (séries) temporais: mercado de ações, processos de produção, experimentos científicos, tratamentos médicos, análises de tendências, de históricos e de similaridades. Envolve ainda a avaliação de eventos cíclicos, sazonais e aleatórios;
- mineração de textos: muitas informações estão disponíveis em documentos

(artigos de jornais ou científicos, livros, *e-mails*, páginas Web, etc.). Dentre as abordagens destaca-se a recuperação de informações e a aplicação de técnicas de mineração em informação semiestruturada;

- mineração na Web: mineração de conteúdo, uso e estrutura. Configura-se num repositório imenso, distribuído e global que contém uma ampla e rica coleção de hiperlinks. Seu tamanho, complexidade e dinamismo oferecem grandes desafios científicos. Os mecanismos de busca como Google, por exemplo, apresentam abordagens interessantes utilizando a mineração de dados.

Regras de Extração de Conhecimento em Armazém de Dados

Segundo Fayyad (1996) para encontrar respostas ou extrair conhecimento interessante nas bases de dados, existem diversos métodos para o uso do *Data Mining* disponíveis no mercado e na literatura atual. Mas, para que a descoberta de conhecimentos seja relevante, é importante estabelecer metas bem definidas e coesas. Essas metas são alcançadas por meio dos seguintes métodos:

- classificação: associa ou classifica uma instância a uma ou várias classes categóricas pré-definidas com o objetivo de predizer a classe de um item. A ideia é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada. Como exemplo a classificação pode construir um modelo para categorizar os empréstimos feitos pelo banco X e categorizados como seguro ou não, de acordo com os dados presentes, efetuando uma tarefa de predição;

- clusterização ou agrupamento: associa uma instância a um ou vários clusters em que as classes são determinadas pelos dados, diferentemente da classificação em que as classes são pré-definidas. Os clusters são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos. A análise de cluster (ou agrupamento) é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles;

- associação: determina regras de associação e correlação para os Items Sets (campos) frequentes no armazém de dados, com o objetivo de subsidiar a tomada de decisões, como o marketing cruzado e análise de perdas. Um tipo exemplo de regras de associação é análise de cesta de compras em um mercado, em que o objetivo é saber que produtos são frequentemente comprados em associação a outros pelos clientes. Acompanhe o exemplo: se cliente X compra arroz então compra feijão; se cliente Y compra cerveja então não compra fralda;

- sumarização: determina uma descrição compacta para um dado subconjunto. As medidas de posição e variabilidade são exemplos simples de sumarização. Funções mais sofisticadas envolvem técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são frequentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas – como mínimo, máximo, média, moda, mediana e desvio padrão amostral –, no caso de variáveis quantitativas, e, no caso de variáveis categóricas, por meio da distribuição de frequência dos valores. Técnicas de sumarização mais sofisticadas são chamadas de visualização, que são de extrema importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Exemplos de técnicas de visualização de dados incluem diagramas baseados em proporções, diagramas de dispersão, histogramas, entre outros;
- análise de desvio: encontrar alterações nos dados;
- regressão: prever um valor numérico contínuo para um determinado atributo, muito utilizado com valores que tendem a permanecerem inalterados.

Conforme definição efetuada por Fayyad (1996), é importante ressaltar que a maioria desses métodos é baseada em técnicas das áreas de aprendizado de máquina, reconhecimento de padrões e estatística aplicada. Essas técnicas vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos.

Relação entre Data Mining e OLAP

O OLAP e *Data Mining* são partes integrantes de qualquer processo de BI. Ainda, nos dias de hoje, a maioria dos sistemas de OLAP tem o foco no provimento de acesso à informações em níveis multidimensionais (proposta principal do OLAP), enquanto os sistemas de *Data Mining* lidam com a análise de influência para os dados de uma única dimensão. “Quando os usuários possuem ferramentas de OLAP e não de mineração de dados, eles gastam boa parte de seu tempo fazendo as tarefas pertinentes a um DM, como classificações e predições das informações recebidas.” De modo que a tarefa de OLAP deve ser utilizada em conjunto com a mineração de dados, proporcionando a relação entre as 2 tecnologias na busca de informações.

CONCLUSÃO

Nesse capítulo foram abordados os conceitos e características das principais ferramentas de BI, Data Warehouse, Data Mart, OLAP e mineração de dados, sendo demonstradas as relações entre as tecnologias e sua exemplificação no contexto real de uso. O próximo capítulo apresenta a utilização da mineração de dados com a ferramenta Weka, sendo demonstrado a sua utilização e benefícios em um contexto empresarial.

UTILIZAÇÃO DE DATA MINING COM A FERRAMENTA WEKA

Nesse capítulo, para demonstrar a utilização do Data Mining é utilizado o pacote Weka (*Waikato Environment for Knowledge Analysis*) na versão 3.6.0, formado por um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados.

O Weka está implementado na linguagem Java, que tem como principal característica ser portável e adaptável a diversas tecnologias. Desta forma pode ser executável nas mais variadas plataformas e, além disso, é um software de domínio público estando disponível em http://www.cs.waikato.ac.nz/ml/weka/.

O Weka suporta várias tarefas de mineração de dados, mais especificamente, pré-processamento, agrupamento, classificação, regressão, visualização, seleção de características.

TAREFAS DE MINERAÇÃO DE DADOS NO WEKA

Para iniciar a utilização, o primeiro passo e instalar corretamente o programa no computador ou servidor, na opção instalação completa (*Full*).

Com a ferramenta devidamente instalada e iniciada, será acessada a opção ‘’*Explorer”* no console principal, como demonstrado na Fig. 9, dando início à possibilidade de executar as tarefas de mineração de dados e efetuar as análises pertinentes.

Figura 9 - Tela inicial Weka.

A primeira atividade de mineração de dados utilizada é o pré-processamento do

conjunto de dados alvos a serem estudados, ou *dataset.*

Um *dataset* que é uma coleção de instâncias em uma base de dados, por exemplo, TRANSPORTE, contendo um conjunto de atributos, como exemplo, CARRO, AVIÃO, NAVIO. Os atributos podem ainda ser dos tipos: números (números reais ou inteiros), literal (valores literais) e nominal (lista predefinida de valores).

Para fins de demonstração neste trabalho foi utilizado o *dataset* denominado viajar. ARFF.

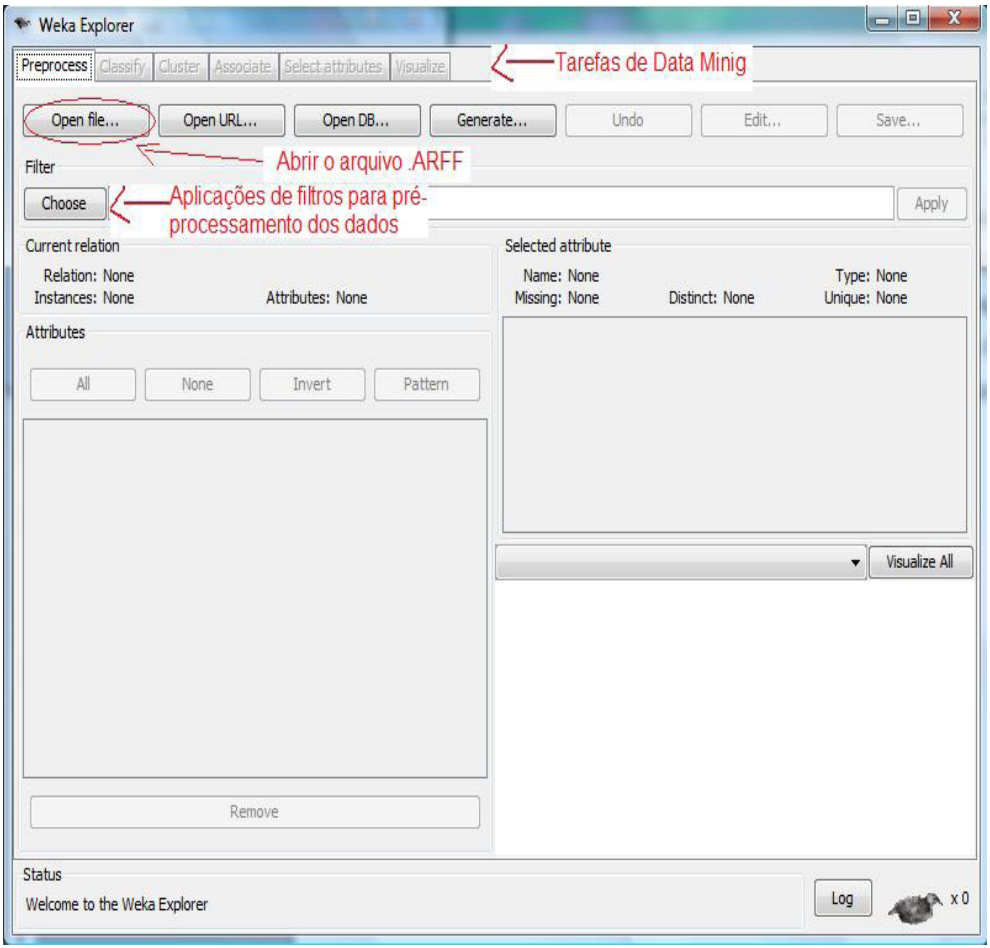

Figura 10 - Pré-processamento.

ENTENDIMENTO DO NEGÓCIO

Os dados alvos utilizados neste trabalho são de uma empresa no ramo de locação de veículos, situada na cidade de Juiz de Fora. A empresa trabalha a mais de 10 anos no

ramo de transporte, com uma cartela diversificada de clientes, atendendo a empresas, clientes particulares e na terceirização de serviços para outras empresas.

O *dataset* armazena as informações referentes a 51 clientes que fizeram orçamento na empresa, a quantidade de dias da viagem, se a data da viagem é em dia de semana ou final de semana, se o orçamento foi agendado pelo menos com três dias de antecedência da possível viagem e, como classe, é utilizada a informação se o cliente viajou ou não.

São seis atributos (dia, feriado, nome, agendado, veiculo, tipo e viagem), sendo a classe definida como ‘’viagem’’, com as seguintes faixas de valores:

- atributo dia {segunda, terça, quarta, quinta, sexta, sábado, domingo};
- atributo feriado {sim, não};
- atributo agendado {sim, não};
- atributo veículo {van, micro-ônibus, automóvel};
- atributo tipo {particular, empresa, terceirizado};
- atributo viagem {sim, não}.

ESTRUTURA DO ARQUIVO DO DATASET

A seguir são demonstradas as partes de um arquivo em formato .ARFF necessários para a utilização da ferramenta:

- @relation nome_do_dataset

A sintaxe @relation define o nome do dataset para o WEKA, como exemplo:

- @relation viajar
- @attribute nome_do_atributo {valores possíveis}

A sintaxe @attribute define os atributos do dataset e seus possíveis valores, como exemplo:

- @attribute pais {Brasil, Argentina, Alemanha, França}
- @attribute fileira {01, 02, 03}
- @data + enter + atributo1, atributo2, atributo3, numero_de_atributos_possiveis

A sintaxe @data define os dados presentes no *dataset*, sendo mantida a ordem de precedência do atributo, como exemplo:

- @data
- Alemanha, 03

- Franca, 02
- Brasil, 01
- Argentina, 02.

Utilizando-se o dataset “viajar.ARFF” é obtida a seguinte estrutura, na Fig. 11:

```
@relation viajar

@attribute dia {segunda, terca, quarta, quinta, sexta, sabado, domingo}
@attribute feriado {sim, nao}
@attribute agendado {sim, nao}
@attribute veiculo {van, microonibus, automovel}
@attribute tipo {particular, empresa, terceirizado}
@attribute viagem {sim, nao}

@data
quarta, nao, nao, van, particular, nao
quinta, nao, nao, van, particular, sim
sexta, nao, sim, microonibus, particular, sim
sexta, nao, sim, microonibus, particular, nao
sabado, sim, sim, microonibus, particular, sim
sabado, nao, sim, microonibus, particular, nao
sabado, nao, sim, van, particular, nao
sabado, sim, sim, microonibus, particular, sim
sexta, sim, sim, microonibus, particular, sim
sexta, sim, sim, microonibus, particular, sim
sexta, nao, sim, van, particular, nao
quinta, nao, sim, microonibus, particular, sim
quinta, nao, sim, microonibus, particular, sim
quinta, nao, sim, microonibus, particular, nao
sexta, nao, sim, van, particular, nao
sabado, nao, sim, automovel, terceirizado, nao
terca, nao, nao, microonibus, empresa, nao
sabado, nao, sim, microonibus, particular, nao
sabado, sim, sim, van, particular, sim
sexta, nao, sim, microonibus, particular, nao
quinta, sim, nao, automovel, empresa, sim
domingo, nao,  sim, van, particular, nao
quinta, sim, sim, automovel, empresa, sim
domingo, nao, sim, microonibus, particular, nao
sabado, nao,  sim, microonibus, particular, sim
sabado, nao, sim, microonibus, particular, sim
sabado, nao, sim, van, particular, nao
sexta, nao, sim, automovel, terceirizado, nao
quinta, nao, nao, van, empresa, nao
sexta, nao, sim, microonibus, particular, sim
segunda, nao, nao, microonibus, empresa, nao
quarta, sim, sim, microonibus, particular, sim
```

Figura 11 - Instâncias do Dataset Viajar.ARFF

PRÉ-PROCESSAMENTO

O painel Pré-Processamento da ferramenta Weka é utilizado para importar o *dataset*,

nos formatos de arquivo de extensão ARFF, CSV e outros arquivos de bancos de dados. Para o pré-processamento dos dados pode-se usar a filtragem através de algoritmos ou de uso manual de acordo com a necessidade ou objetivo da mineração de dados.

Esses filtros podem ser utilizados para transformar os dados (por exemplo, transformar atributos numéricos contínuos em valores discretos), e tornar possível a remoção de instâncias e atributos de acordo com critérios específicos. A Fig. 12 mostra a tela de pré-processamento, após selecionada a opção *Open File*(abrir arquivo) e atribuído o *dataset* viajar.ARFF.

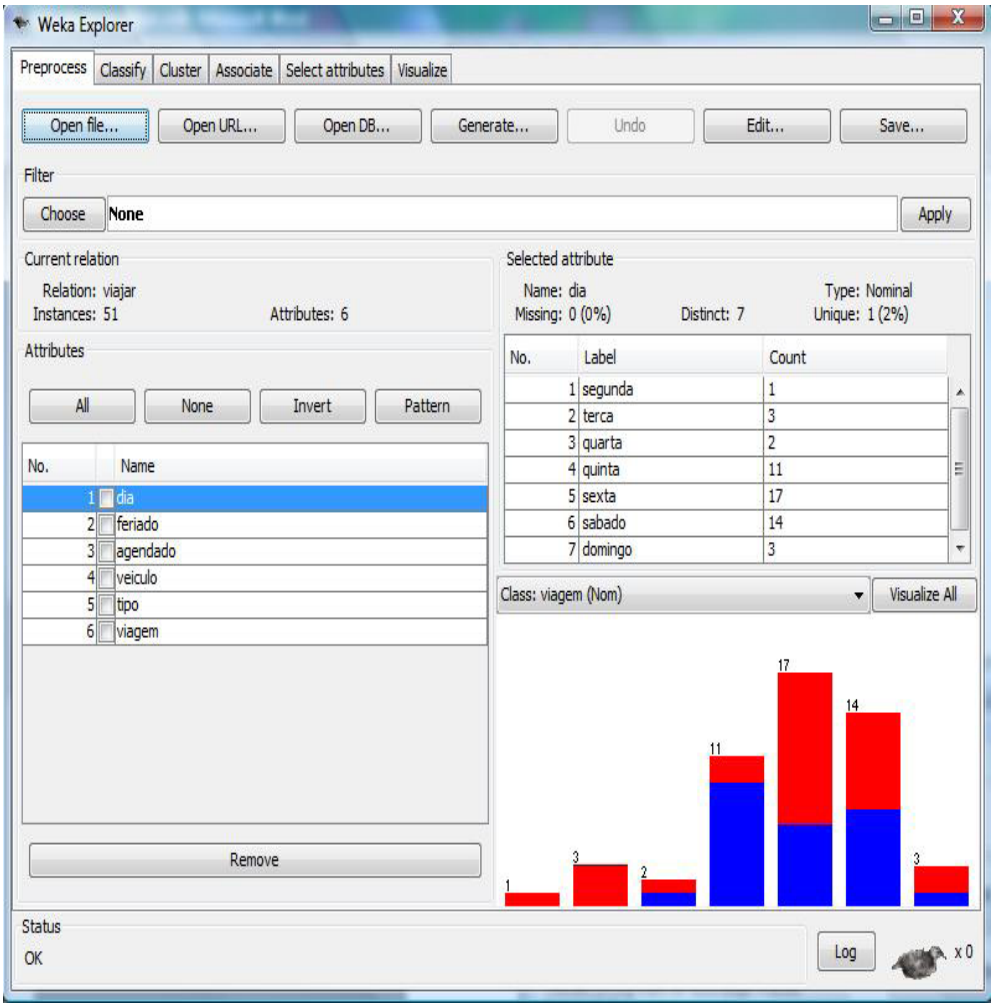

Figura 12 - Pré-processamento na ferramenta Weka.

São mostradas as instâncias no lado esquerdo da Fig. 11. Na parte superior direita

os atributos das instânias selecionadas e suas faixas de valores, e na parte inferior direita a representação gráfica das instâncias de acordo com a classe. Neste caso a classe é viagem.

Clicando na opção *Visualize All*, obtêm-se a representação gráfica das instâncias e sua aderência à classe, em que a cor azul corresponde ao valor SIM e a cor vermelha ao valor NÃO, conforme visto na Fig. 13.

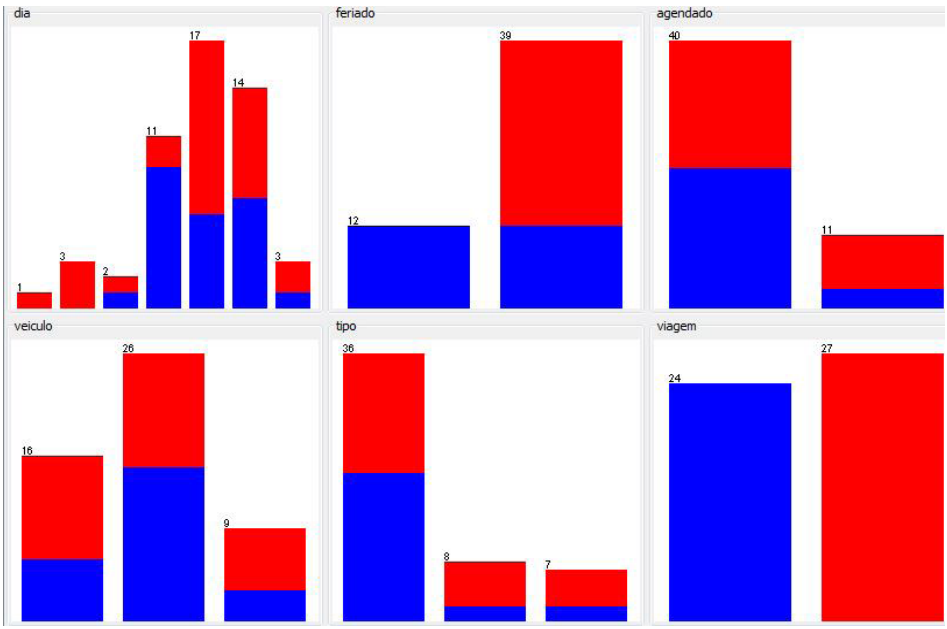

Figura 13 - Representação gráfica das instâncias.

CLASSIFICAÇÃO

O painel de classificação da ferramenta Weka permite ao usuário aplicar algoritmos de classificação para os dados resultantes do pré-processamento, dessa forma busca categorizar as instâncias em uma ou mais classe, com o objetivo de predizer futuros itens.

Para uma classificação mais elaborada, utiliza-se a validação Cross-Validation, onde um número n de dobras é especificado pelo usuário da ferramenta Weka, e o conjunto de dados é aleatoriamente reordenados e, em seguida, dividida em n dobras de igual tamanho.

Em cada iteração, uma dobra é utilizada para os testes e as outras n-1 dobras são usadas para o treinamento da classe na classificação. Os resultados do teste são coletados e calculados sobre todas as dobras. Isto dá a validação cruzada uma maior estimativa da precisão (SEEWALD, 2008).

Escolhemos o algoritmo J48, que é uma implementação do algoritmo C4.8, que gera a árvore de decisão e é considerado o mais popular algoritmo da Weka (SEEWALD, 2008). O J48 constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, sendo que esse modelo é utilizado para classificar as instâncias do conjunto de teste.

Durante o processo de utilização do algoritmo J48 é interessante conhecer alguns parâmetros que podem ser modificados para proporcionar melhores resultados, como por exemplo, o uso de podas na árvore, o número mínimo de instâncias por folha e a construção de árvore binária. Na Fig. 14, pode ser observado o resultado da utilização do algoritmo de classificação J48 no dataset viajar.ARFF.

```
                dia
                feriado
                agendado
                veiculo
                tipo
                viagem
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree (J48 poda de arvore)
feriado = sim: sim (12.0)
feriado = nao: nao (39.0/12.0)
Number of Leaves  :  2
Size of the tree :  3  (tamanho da arvore)
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          39         76.4706 % (percentual das instâncias classificadas corretamente)
Incorrectly Classified Instances        12         23.5294 % (percentual das instâncias classificadas incorretamente)
Kappa statistic :                 0.5143     (A estatistica Kappa mede a concordância do previsão com a classe de verdade - 1,0 significa total concordância.)
Mean absolute error                0.3284
Root mean squared error            0.408
Relative absolute error           65.6887 %
Root relative squared error       81.4552 %
Total Number of Instances         51
=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.5       0          1        0.5       0.667      0.659    sim
                   1         0.5       0.692     1         0.818      0.659    nao
Weighted Avg.    0.765     0.265      0.837     0.765     0.747      0.659
=== Confusion Matrix (matriz de confusão) ===
  a  b   <-- classified as
 12 12 |  a = sim
  0  27 |  b = nao
```

Figura 14 - Algoritmo J48 no Dataset Viajar.ARFF (continuação).

A matriz de confusão é normalmente denominada tabela de contingência. Neste caso, temos duas classes e, portanto, uma matriz de confusão 2x2. A matriz pode ser arbitrariamente grande dependendo no número de atributos da instância classe. O número de casos incorretamente classificados são:

- Classe “A” recebe erroneamente classificada como “b”, exatamente 27 vezes;
- Classe “B” recebe erroneamente classificada como “a” em 0 vezes, ou seja, não ocorre).

CLUSTERIZAÇÃO

O painel de clusterização (*Cluster*), permite o agrupamento das instâncias de acordo com a necessidade a ser utilizada, nesse caso foi utilizado o algoritmo K-means

ou K-media. O K-means é uma técnica que usa o algoritmo de agrupamento de dados por K-médias (*K-means clustering*). O objetivo deste algoritmo é encontrar a melhor divisão de informações em K grupos, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somado por todos demais grupos seja minimizado.

Em outras palavras, o algoritmo atribui aleatoriamente os P pontos (instâncias) a K grupos (*clusters*) e calcula as médias dos vetores de cada grupo. Em seguida, cada ponto é deslocado para o grupo correspondente ao valor médio do qual ele está mais próximo. Com este novo rearranjo dos pontos em K grupos, novos valores médios são calculados. O processo de re-alocação de pontos a novos grupos cujos valores médios são os mais próximos deles continua até que se chegue a uma situação em que todos os pontos já estejam nos grupos dos seus valores médios mais próximos. Na Fig. 15, pode ser observado a utilização do K-means no dataset viagem.ARFF :

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     viajar
Instances:    51
Attributes:   6
              dia
              feriado
              agendado
              veiculo
              tipo
              viagem
Test mode:    evaluate on training data
=== Model and evaluation on training set ===
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 102.0
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute         Full Data          0          1
                            (51)       (42)        (9)
=================================================
dia                       sexta      sexta       terca
feriado                    nao        nao         nao
agendado                   sim        sim         nao
veiculo         microonibus microonibus microonibus
tipo               particular  particular     empresa
viagem                     nao        sim         nao

Clustered Instances

0      42 ( 82%)
1       9 ( 18%)
```

Figura 15 - Algoritmo K-means no Dataset Viajar.ARFF.

Foram identificados 2 clusters e um conjunto de todas as instâncias denominado *full data*. O primeiro grupo é denominado como cluster zero, é formado pela associação dos atributos e valores: dia é sexta, feriado é não, agendado é sim, veículo é micro-ônibus,

tipo é particular e a classe viagem é sim. O segundo grupo é o denominado cluster um, é formado pela associação dos atributos e valores: dia é terça, feriado é não, agendado é não, veículo é micro-ônibus, tipo é empresa e a classe viagem é não.

O resultado mostra o agrupamento das instâncias em que 82% em que estão presentes no cluster zero e 18% estão inseridas no cluster um.

A Fig. 16 apresenta graficamente a relação entre as instâncias. Nesse caso, é demonstrado utilizando como matriz X o atributo TIPO, que é correspondente ao tipo do cliente e matriz Y o atributo VEICULO, que corresponde ao veículo objeto da viagem:

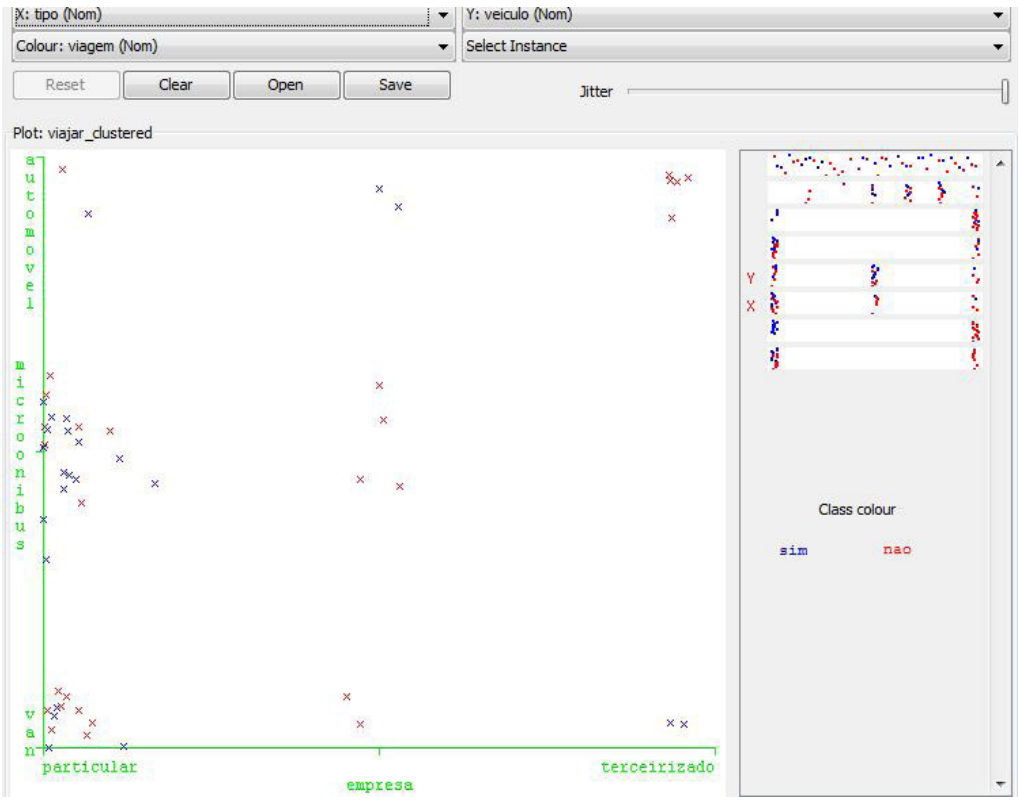

Figura 16 - Representação gráfica da clusterização.

Sendo que os valores referentes ao cluster com viajar igual a SIM, são representados com a cor azul e ao contrário com a cor vermelha.

ASSOCIAÇÃO

O painel de associação na ferramenta Weka contém esquemas de aprendizagem de

regras de associação através de algoritmos. Os algoritmos de aprendizagem são escolhidos na parte superior do Weka, assim como a classificação e a clusterização.

O objetivo da associação é encontrar todas as regras relevantes ao contexto empregado entre os itens utilizados, as regras encontradas são do tipo X (antecedente) □ Y (consequente), ou seja, Se X acontece então Y também deverá acontecer, construindo uma relação entre dois ou mais itens. Para tratar desta questão, Agrawal *et.all* (1993) propuseram um modelo matemático, em que as regras de associação geradas devem atender a um suporte e confiança mínimos, sendo especificados pelos profissionais responsáveis de acordo com o nível de confiabilidade e segurança exigidas para determinada situação, por exemplo, para situação de risco elevado os valores de suporte e confiança tendem a serem altos, pelos riscos envolvidos, em casos com risco baixo os valores de suporte e confiança tendem a serem menores.

A medida de suporte corresponde à frequência com que ocorrem os padrões em toda a base de dados. O suporte mínimo (minsup) é a fração das transações que satisfaz a união dos itens do consequente com os do antecedente, de forma que estejam presentes em pelo menos s% das transações no banco de dados.

A confiança mínima (minconf) garante que ao menos c% das transações que satisfaçam o antecedente das regras também satisfaçam o consequente das regras.

Na ferramenta Weka o valor suporte e confiança e definido na Fig. 17:

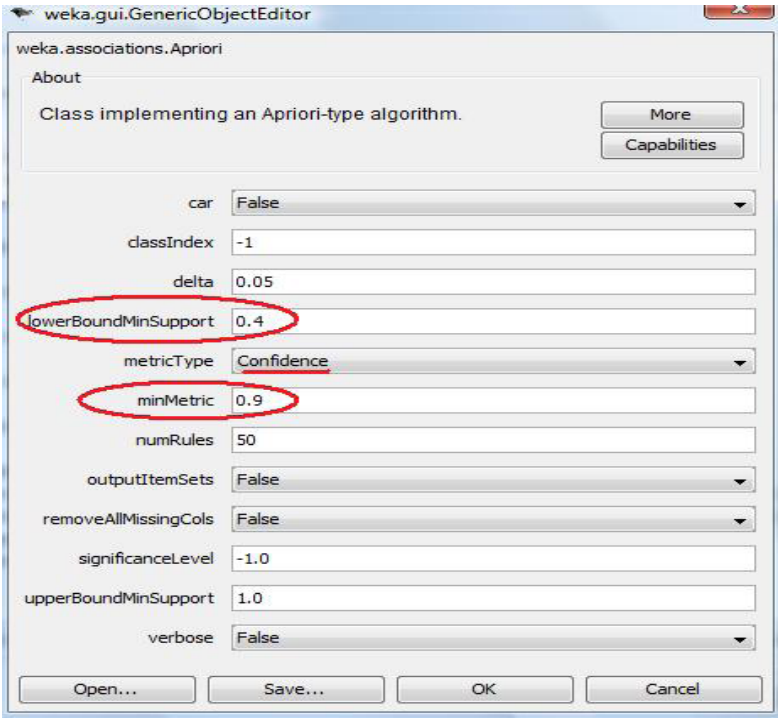

Figura 17 - Regras de suporte e confiança.

Na Fig. 17, o valor lowerBoundMinSupport determina o valor de suporte mínimo, minMetric atribui o valor da confiança, numRules gera o número máximo de regras de associação.

Como aplicação do algoritmo Apriori seguindo os valores de suporte e confiança estabelecidos na Fig. 17, foram gerados os resultados dados demonstrados na Fig. 18:

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.4 -S -1.0 -c -1
Relation:     viajar
Instances:    51
Attributes:   6
              dia
              feriado
              agendado
              veiculo
              tipo
              viagem
=== Associator model (full training set) ===

Ilustração 18 - Algoritmo Apriori no Dataset Viajar.ARFF

Apriori
=======

Minimum support: 0.4 (20 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 2

Best rules found:

 1. viagem=nao 27 ==> feriado=nao 27    conf:(1)
 2. veiculo=microonibus tipo=particular 22 ==> agendado=sim 22    conf:(1)
 3. agendado=sim veiculo=microonibus 22 ==> tipo=particular 22    conf:(1)
 4. feriado=nao tipo=particular 27 ==> agendado=sim 25    conf:(0.93)
 5. tipo=particular 36 ==> agendado=sim 33    conf:(0.92)
```

Figura 18 - Algoritmo Apriori no Dataset Viajar.ARFF (continuação).

Nesse caso foram obtidas as seguintes regras do dataset:

- se feriado é igual a não, então viagem é igual a não em 100% dos casos ou confiança igual a 1, valor válido em 27 instâncias;
- se veículo é igual a micro-ônibus e tipo igual à particular, então agendado é igual a sim em 100% dos casos ou confiança igual a 1, valor válido em 22 instâncias;
- se agendado é igual a sim e veiculo é igual a micro-ônibus, então tipo é igual à particular em 100% dos casos ou confiança igual a 1, valor válido em 22 instâncias;
- se feriado é igual à não e tipo é igual à particular, então agendado é igual à particular em 93% dos casos ou confiança igual a 0,93, valor válido em 27 instâncias;

- se tipo é igual à particular, então agendado é igual a sim em 92% dos casos ou confiança igual a 0,92, valor válido em 36 instâncias.

VISUALIZAÇÃO

A opção *visualize* na ferramenta Weka, permite a visualização gráfica em 2D, das instâncias no *dataset,* com objetivo de facilitar o entendimento dos dados e as interações entre as instâncias.

A Fig. 19 demonstra um exemplo, da utilização da opção *visualize,* no dataset viajar. ARFF.

Os pontos são para demonstrar a relação com a classe viajar, em que os valores em azul são correspondentes a SIM e vermelho NÃO.

Na parte inferior da Fig. 22, são apresentadas as variáveis de ajuste de tamanho, visualização e a escolha da classe.

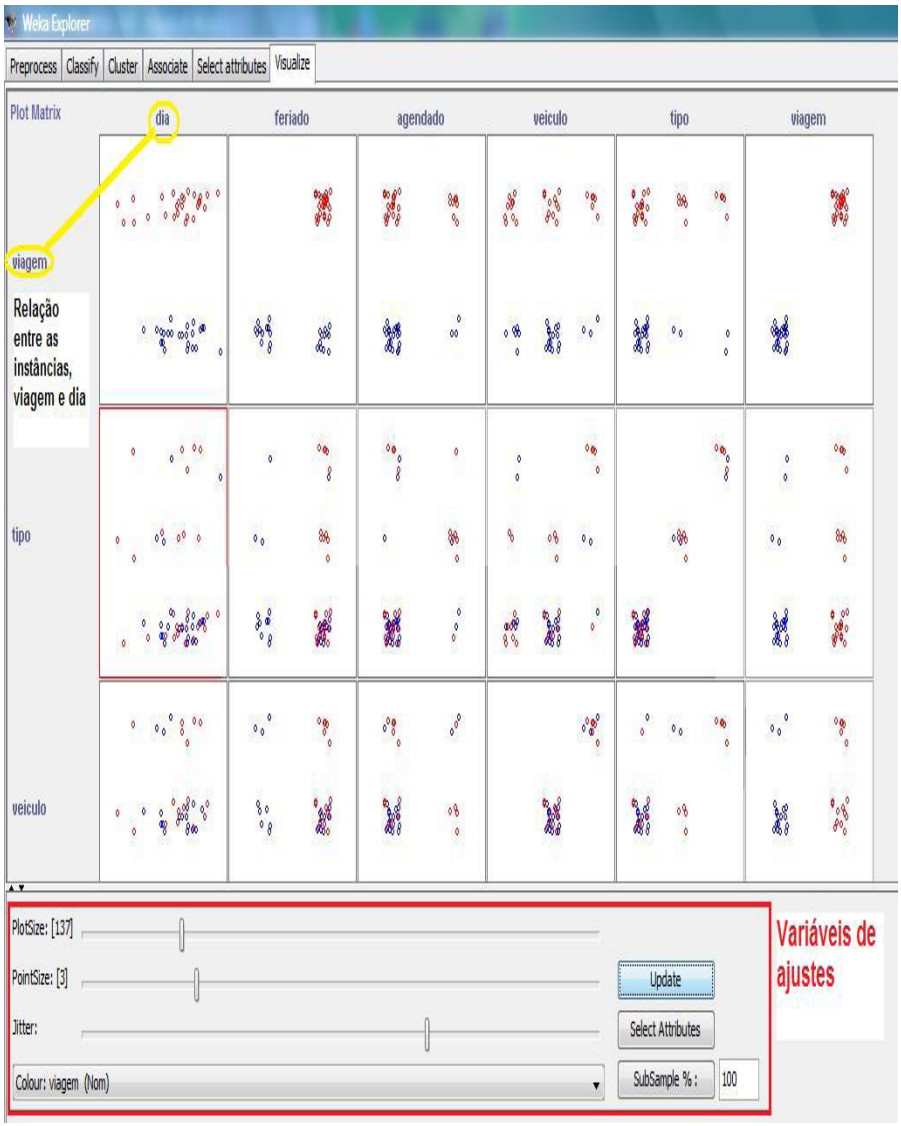

Figura 19 - visualização das Instâncias

INTERPRETAÇÕES DOS RESULTADOS ENCONTRADOS

A execução dos algoritmos de mineração de dados utilizados deste trabalho, geraram as seguintes interpretações de classificação, clusterização e associação.

Classificação: de acordo com resultados obtidos na classificação através da utilização do algoritmo J48, pode-se efetuar a predição em futuras instâncias da base viajar .ARFF, que determina se o valor referente a feriado for igual a sim, então o resultado da classe será sim, em 100% dos casos.

Esse dado é de grande importância para o negócio da empresa de transporte, pois a empresa pode planejar maneiras de aumentar o lucro, escalonar maior número de funcionários e planejar com maior antecedência essas datas, devido principalmente ao elevado grau da precisão da informação em dizer que em todos os casos estudos as viagens foram realizadas, ou seja, a empresa prestou o serviço.

Clusterização: os resultados obtidos na tarefa de clusterização através da utilização do algoritmo k-means, foi que no *cluster* zero, compostos pelos atributos dia igual à sexta, feriado igual à não, agendado igual à sim, veículo igual à microônibus e tipo do cliente igual à particular, obteve como conclusão a classe viagem igual à sim. Já no *cluster* um, compostos pelos atributos dia igual à terça, feriado igual à não, agendado igual à não, veículo igual à micro-ônibus e tipo do cliente igual à empresa, obteve como conclusão a classe viagem igual à sim.

Com base nos grupos encontrados é possível identificar os agrupamentos das instâncias, e para a empresa permite identificar em que grupos estão seus clientes.

Associação: os resultados obtidos na tarefa de associação através da utilização do algoritmo apriori, como valor de suporte mínimo de 40% e confiança mínimo de 90%, com o objetivo de identificar regras de associação entre as instâncias, sendo úteis para a necessidade da empresa.

Foram identificadas cinco regras que atendem aos valores estabelecidos, e para a empresa a de maior importância é a regra se feriado é igual a não, então viagem é igual a não em 100% dos casos ou confiança igual a 1, em mais de 50% do número total de instâncias, pois ela determina uma possível deficiência da empresa em prestar serviços de viagem nos dias de semana, permitindo a empresa rever e analisar seus procedimentos, processos e preços aplicados na tentativa de melhorar este quadro.

Necessidade esta que não havia sido percebida pela utilização de métodos empíricos da empresa, mostrando o real valor e eficiência de uma das ferramentas de BI, o *Data Warehouse*.

É importante ressaltar que os dados são apenas informações na tela, e necessários pessoas para interpretar, demonstrando a necessidade dos profissionais de sistemas de informação.

CONCLUSÃO

Nesse capítulo foram abordados os conceitos e características da utilização do *Data Mining* na ferramenta Weka, utilizando os algoritmos J48, K-means e Apriori, sendo demonstrado os resultados obtidos no contexto de uma empresa do ramo de transporte, e proporcionado uma visão realista de como funciona uma das ferramentas de BI.

CONSIDERAÇÕES FINAIS

O *Business Intelligence* representa um importante passo na computação relacionada aos negócios. Aliada a diversas fontes de dados gerados pelos sistemas ERPs e CRMs, representa um aumento considerável na confiabilidade dos negócios e na tarefa de tomada de decisões.

O objetivo deste trabalho foi mostrar a real necessidade da utilização do BI e suas aplicações nos mais diferentes ambientes organizacionais, demonstrando resultados obtidos na utilização da mineração de dados em uma empresa real.

As conclusões da presente pesquisa não esgotam o tema abordado, procuram apenas responder algumas questões de propostas. Sugere-se dessa forma, que outros estudos sejam efetuados, a fim de complementar os resultados obtidos neste trabalho.

Como propósito de trabalhos futuros, pode-se efetuar uma pesquisa que visa identificar se a utilização de técnicas, ferramentas e metodologias de *Business Intelligence*, no auxílio à gestão organizacional de pequenas empresas do estado de Minas Gerais e um diferencial no processo de manter a empresa em pleno funcionamento no mercado atual, visto que cria um diferencial nessas empresas.

REFERÊNCIAS

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Mining association rules between sets of items in large databases, p. 207-216, 1993. Disponível em:. <http://www.almaden.ibm.com/u/ragrawal/pubs.html.> Acessado em: nov. de 2007.

BARBIERI, Carlos. ***Business Intelligence*: modelagem e tecnologia**. Rio de Janeiro: Axcel Books, 2001.

CARLSSON, C.; TURBAN, E. DSS: directions for the next decade. **Decision Support Systems**, v. 33, n. 2, p. 105-110, 2002.

CODD,E.F. **A relation model of data large shared data banks**, C. ACM, 1970.

FAYYAD,U.M. **Knowledge Discovery and Data Mining**. Oregon 1996.

GRIGORI, D., CASATI, F.; CASTELLANOS, M.; DAYAL, U.; SAYAL, M.; SHAN, M. C. **Business Process Intelligence. Computers in Industry**, v. 53, n. 3, p. 324-343,2004.

HABERMANN, R. ***Business Intelligence* para pequenas empresas**, 2006. Disponível em: <http://webinsider.uol.com.br/index.php/2006/07/08/business-intelligence-parapequenas-empresas/>. Acessado em: nov. de 2007.

HAN, J., KAMBER. M. **Data mining: concepts and techniques**. 2001.

IVERSON, K. **A programming language**, Estados Unidos, 1966.Disponivel em: <http://hopl.murdock.edu.au/showlanguage.prx?exp=421&language=iverson`s%20language> Acessado em : abr. de 2009.

INMON, W. H. **Building the Data Warehouse**, Willey Conputer Publishing. Canada, 2002.

KIMBALL, R. **The Data Warehouse Toolkit** , Willey Conputer Publishing. Canada, 2002.

LAUDON, K.C.; LAUDON, J.P. **Gerenciamento de Sistemas de Informação**, 3º edição, Rio de Janeiro, 2002.

LOPES, C., Business **Inteligence e Gerenciamento de Projeto Modulo 2**, 2005.Disponível em: <http://www.datawarehouse.com.br/portal/modules.php?name=News&file=article&sid=2> Acessado em: mai. de 2009.

MICROSOFT. **Optimização da infra-estrutura de plataforma de aplicações**, 2009. Disponível em:<http://www.microsoft.com/portugal/business/peopleready/appplat/capability_busintel.mspx > Acessado em : set. de 2007.

_____. **Elevando a gestão dos negócios a um novo patamar**, 2006**.** Disponível em:<http://technet.microsoft.com/pt-br/library/cc668463.aspx> Acessado em : jul. de 2009.

MICROSTRATEGY. **Business Intelligence Solutions**, 2009**.** Disponível em: <http://www.microstrategy.com.br/Solutions/5Styles/olap_analysis.asp> Acessado em : jun. de 2009.

ORACLE CORPORATION. **OLAP Application Developer's Guide**, 2005. Disponível em: em: <http://www.filibeto.org/sun/lib/nonsun/oracle/10.2.0.1.0/B19306_01/olap.102/b14349/wiz.htm> Acessado em : jun. de 2009.

PEREIRA, M., **Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka**, 2004. Disponível em: <http://www.dpi.inpe.br/~mpss/artigos/MineracaoDeDados2004.pdf> Acessado em : mai. de 2009.

PEROTTONI, R., OLIVEIRA M., LUCIANO E. M. e FREITAS H., **Sistemas de informações: um estudo comparativo das características tradicionais às atuais**. Rio Grande do Sul, 2001.

PRIMAK, F. **Analisando o Conceito de B.I**, 2008. Disponível em: <http://www.weblivre.net/artigo/business-intelligence/analisando-o-conceito-de-bi/> Acessado em: mar. de 2009.

SHIM, J. P.; WARKENTIN, M.; COURTNEY, J.; POWER, D. J.; SHARDA, R.; CARLSSON, C. **Past, present, and future,** p. 111-126, 2002.

SIEGEL, D. **Futurize sua empresa**, p.17-62, São Paulo, Futura, 2000.

SEEWALD, A. **Weka Manual for version 3-6-0**, 2008.Disponível em: <http://ufpr.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3.6.0.pdf> Acessado em: nov. de 2009.

SOBRE O AUTOR

MATHEUS EMERICK DE MAGALHÃES – Mestre e Doutorando em Engenharia da Computação pela Universidade Federal do Rio de Janeiro e Pós-Graduado em Business Intelligence e Gerenciamento de Projetos. Há mais de 10 anos trabalha na área de Business Intelligence e Banco de Dados. Tem experiência em sistemas de apoio a tomada de decisão, data warehouse, business intelligence, tratamento de dados não-estruturados, BigData e gerenciamento de projetos. Atualmente é Capitão de Corveta da Marinha do Brasil, sendo condecorado com a premiação de primeiro colocado em todo o Brasil no curso de formação de oficiais de notório saber.

Uma abordagem na adoção de

BUSINESS INTELLIGENCE

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2022

Uma abordagem na adoção de

BUSINESS INTELLIGENCE

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2022