UNIVERSIDADE DO VALE DO RIO DOS SINOS

UNISINOS

CENTRO DE CIÊNCIAS ECONÔMICAS

DATA MINING

GILMARA MACHADO DE OLIVEIRA
LUCIMARA MENGUE PEREIRA

quarta-feira, 20 de Junho de 2007

DISCIPLINA – GESTÃO DA INFORMAÇÃO

GRAU
B

SUMÁRIO

INTRODUÇÃO

A tecnologia estudada neste trabalho é o *Data Mining*, visando mostrar sua origem conceitos, objetivos, vantagens, desvantagens, aplicações em empresas, etapas do processo, entre outros, para melhor entendimento do seu processo. *Data Mining* ou Mineração de Dados é uma nova metodologia para melhorar a qualidade e eficiência das decisões, quer sejam elas científicas ou de negócios.

Atualmente, as organizações têm se mostrado extremamente eficientes em capturar, organizar e armazenar grandes quantidades de dados, obtidos de suas operações diárias ou pesquisas científicas, porém, ainda não usam adequadamente essa gigantesca montanha de dados para transformá-la em conhecimentos que possam ser utilizados em suas próprias atividades, sejam elas comerciais ou científicas.

A rápida taxa de inovação nas tecnologias de informática está exigindo que, cada vez mais, os profissionais estejam preparados e atualizados para conhecer e enfrentar os desafios da Tecnologia da Informação.

O conceito de *Data Mining* está se tornando cada vez mais popular como uma ferramenta de gerenciamento de informação, que deve revelar estruturas de conhecimento, que possam guiar decisões em condições de certeza limitada. Recentemente, tem havido um interesse crescente em desenvolver novas técnicas analíticas, especialmente projetadas para tratar questões relativas à *Data Mining*. No entanto, *Data Mining* ainda está baseado em princípios conceituais de Análise de Dados exploratórios e de modelagem.

Data Mining é parte de um processo maior de conhecimento denominado *Knowledge Discovery in Database* (KDD). KDD consiste, fundamentalmente, na estruturação do banco de dados; na seleção, preparação e pré-processamento dos dados; na transformação, adequação e redução da dimensionalidade dos dados; no processo de *Data Mining*; e nas análises, assimilações, interpretações e uso do conhecimento extraído do banco de dados, através do processo de *Data Mining*.

1. ORIGEM

Na década de 1960, o tema tecnológico que rondava as organizações era o “processamento de dados”. Nessaépoca, a maioria das empresas direcionava os recursos para o processamento centralizado de dados em *mainframes* (grandes computadores) e para os sistemas de controles operacionais, tais como faturamento, estoque, folha de pagamento, finanças e contabilidade. Tais sistemas eram processados de forma mecanizada e em *batch* (processamento em grupos ou lotes). O processamento de dados (PD) era utilizado nas empresas para a substituição de mão-de-obra e redução de custos. As funções de informática praticamente não existiam e os poucos recursos eram totalmente centralizados na área de PD. Aos poucos, porém, as empresas foram se sensibilizando para a importância da informação na gestão de negócios.

Contagiadas pela “informática”, que passa a substituir o tradicional “processamento de dados”, as empresas superam resistências e incorporam essa nova ferramenta empresarial. Com a “informática”, as empresas integraram os seus sistemas, mesmo com algumas redundâncias.

2. CONCEITOS

Data Mining ou Mineração de Dados consiste em um processo analítico projetado para explorar grandes quantidades de dados, filtrando informações gerenciáveis e de interesse do gestor.

O *Data Mining* esta se tornando uma importante ferramenta de auxilio à tomada de decisões. Segundo Groth (1998), "*Data Mining* é uma área de pesquisa da inteligência artificial que busca encontrar padrões em bases de dados". Para Kremer (1999), "*Data Mining* é uma tecnologia usada para revelar informação estratégica escondida em grandes massas de dados".

Segundo Fernandes (2003), "Uma empresa que emprega a técnica de *Data Mining* está muito à frente das outras, pois é capaz de criar parâmetros para entender o comportamento do consumidor; identificar afinidades entre as escolhas de produtos e serviços; prever hábitos de compras; analisar comportamentos habituais para se detectar fraudes".

Ainda segundo Groth (1998), "Vários analistas atualmente utilizam o equivalente a uma lanterna para localizar informações importantes em seus bancos de dados. Uma vez que existam ferramentas para pesquisar, acessar e manipular dados, o usuário é levado a apontar a lanterna para onde ele acha que existe informação útil, ou seja, *Data Mining* automatiza o processo de descoberta, varredura, de tendências e padrões úteis escondidas em grandes bases de dados".

Conforme análise nos dados apresentados pelos autores pode se perceber a importância desta ferramenta para a gestão de marketing, e gerenciamento das mais diversas áreas. Podendo ser utilizado para detectar as mais diversas variáveis desejadas para obtenção de dados e captação de clientes potenciais para cada produto da sua empresa, ou seja, permite focar o produto certo ao cliente certo.

3. OBJETIVOS DO *DATA MINING*

- Apresentar e explorar as principais funcionalidades, técnicas e algoritmos utilizados em *Data Mining*;

- Mostrar como a tecnologia de *Data Mining* pode ajudar a extrair informações valiosas de grandes bases de dados;

- Introduzir os principais conceitos e tecnologias necessárias para melhorar a tomada de decisões nas empresas com base em seus acervos de dados;

- Apresentar os conceitos, técnicas, ferramentas e aplicações de *Data Mining*;

- Mostrar como utilizar a tecnologia de *Data Mining* no contexto de *Business Intelligence* (BI);

- Capacitar os participantes para atuarem de forma ativa em um projeto de *Data Mining*;

- Fornecer uma visão gerencial das tecnologias de informática em um contexto de tomada de decisões e incorporação de informações em seus negócios;

- Apresentar exemplos de aplicações de *Data Mining* para Market Basket Analysis, Segmentação de Mercado, Modelagem de *Churn/Attrition, Credit Scoring*, Detecção de Fraude, *Webmining*, etc.

4. PÚBLICO ALVO

- Gerentes, Analistas de Negócio e profissionais de Tecnologia de Informação preocupados em obter vantagens competitivas a partir da utilização e exploração de informações sobre clientes e mercados;

- Profissionais que trabalhem ou pretendam trabalhar com Análise de Dados;

- Profissionais de Empresas e Pesquisadores interessados em melhor explorar um acervo de dados para potencializar sua atuação;

- Profissionais de informática familiarizados com sistemas de informação voltados para Análise de Dados e/ou Tomada de Decisões;

- Profissionais de Informática voltados para o processo de Descoberta de Conhecimento em Bancos de Dados (KDD);

- Usuários de informática que necessitem entender melhor os processos da construção e exploração de dados para busca de conhecimento e tomada de decisões;

- Profissionais de marketing;

- Profissionais que ocupem ou pretendam ocupar cargos de nível gerencial.

5. ÁREAS DE APLICAÇÃO DE *DATA MINING*

Técnicas de *Data Mining* têm sido aplicadas com sucesso para a solução de problemas em diversas áreas, como as descritas a seguir:

Vendas

- Retenção de clientes: identificar clientes que podem "migrar" para o competidor e tentar retê-los;

- Detectar associações entre produtos;

- Identificar padrões de comportamento de consumidores;

- Encontrar características dos consumidores de acordo com a região demográfica;

- Prever quais consumidores serão atingidos nas campanhas de marketing (nestes casos, basta enviar mala direta anunciando o produto apenas para aqueles prováveis compradores de mala direta direcionada).

Finanças

- Detectar padrões de fraudes no uso dos cartões de crédito;
- Identificar os consumidores que estão tendendo a mudar a companhia do cartão de crédito;
- Identificar regras de estocagem a partir dos dados do mercado;
- Encontrar correlações escondidas nas bases de dados.

Seguros e Planos de Saúde

- Determinar quais procedimentos médicos são requisitados ao mesmo tempo;
- . Prever quais consumidores comprarão novas apólices;
- Identificar comportamentos fraudulentos.

Transporte

- Determinar a distribuição dos horários entre os vários caminhos;
- Analisar padrões de sobrecarga.

Medicina

- Caracterizar o comportamento dos pacientes para prever novas consultas;
- Identificar terapias de sucessos para diferentes doenças;
- Prever quais pacientes têm maior probabilidade de contrair uma certa doença, em função de dados históricos de pacientes e doenças.

Telecomunicação

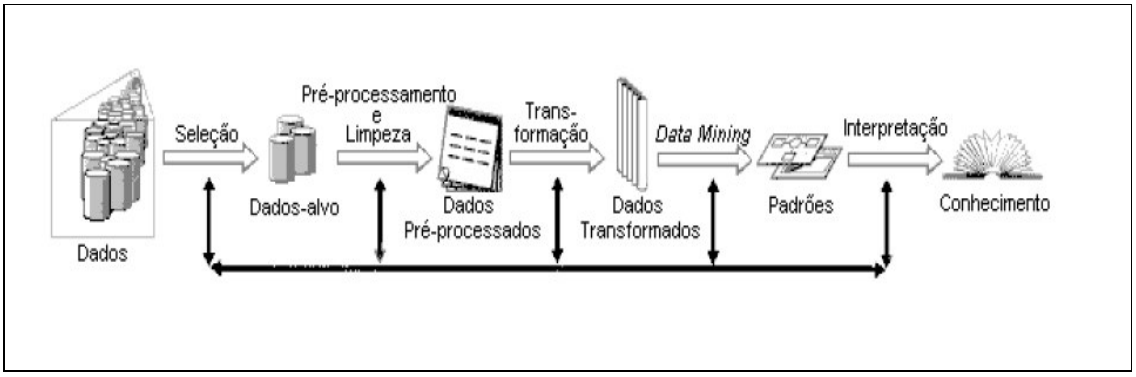
- Identificar fraudes em ligações telefônicas (particularmente em celulares), dentre um enorme número de ligações efetuadas pelos clientes.

Mercado Financeiro

- Prever quais ações irão subir ou descer na bolsa de valores, em função de dados históricos com preços de ações e valores de índices financeiros.

6. ETAPAS DO DATA MINING

Processo *Knowledge Discovery Database* (KDD)



Fonte: FAYYAD, 1996.

Para facilitar a recuperação e o uso desses dados, uma das alternativas possíveis é a aplicação do processo de *Knowledge Discovery in Database* (KDD). Segundo Fayyad (1996), esse processo é composto das seguintes etapas:

6.1. Seleção de Dados: prevê a coleta e seleção dos dados;

6.2. Limpeza: prevê a análise dos dados coletados verificando a existência de ruídos, tratamento de valores ausentes etc.;

6.3. Transformação ou Enriquecimento dos Dados: trata-se da incorporação de novos dados criados a partir dos já existentes;

6.4. *Data Mining* ou Mineração de Dados: consiste em aplicar um algoritmo que efetivamente procura por padrões/relações e regularidades em um determinado conjunto de dados;

6.4.1. Tarefas de Mineração: tarefa, no contexto da mineração de dados, é um tipo de problema de descoberta de conhecimento a ser solucionado. A escolha da tarefa mais adequada depende da natureza da aplicação que se pretende desenvolver. Dentre as principais tarefas utilizadas em mineração de dados, destacam-se: Agrupamento, Classificação, Sumarização e Regras de associação.

6.4.2. Classificação: seu objetivo é classificar casos em diferentes classes, baseado em propriedades (atributos) comuns entre um conjunto de objetos em uma base de dados. O modelo de classificação construído é utilizado para predizer classes de novos casos que serão incluídos em um banco de dados. Essa tarefa possui uma abordagem de aprendizado dita supervisionada.

6.4.3. Agrupamento, também denominado *Clustering*: é definido por Fayyad como uma tarefa comum de descrição onde se busca identificar um conjunto finito de categorias para descrever os dados. O agrupamento identifica a classe de cada objeto de maneira que os objetos dentro de uma mesma classe apresentem alta similaridade entre si, e ao mesmo tempo, baixa similaridade em relação aos objetos das outras classes, segundo Han e Kamber. Diferentemente da classificação, onde os dados de treinamento estão devidamente classificados e as etiquetas das classes são conhecidas, a análise de agrupamento trabalha sobre dados onde às etiquetas das classes não estão definidas. Essa tarefa possui uma abordagem de aprendizado dita não-supervisionada.

6.4.4. Regras de associação**:** são tipicamente importantes ao tentar encontrar associações relevantes entre itens transacionais durante um período de tempo. Um exemplo da saída de tal mineração é a sentença que diz que X% das transações que negociam oproduto1 e produto2 também lidam com produto3, onde o número X% é o fator de confiabilidade da regra. O conhecimento que é extraído através dessa tarefa pode ser aplicado com o objetivo de aumentar potencialmente o volume de vendas: criação de uma promoção (por exemplo: oferecer o produto y a quem comprar x); colocar os produtos x e y lado a lado na loja para promover a compra por impulso; etc.

6.4.5. Sumarização: **é** utilizada para encontrar uma descrição compacta de uma classe, encontrando as qualidades interessantes da mesma. Existem duas formas de obtenção dessas descrições das classes: a caracterização e a discriminação. A caracterização descreve o que os registros de uma classe possuem em comum entre eles, não levando em consideração as demais classes existentes. Já a discriminação descreve como duas ou mais classes diferem entre si.

7. Interpretação e Avaliação**:** verifica a qualidade do conhecimento (padrões) descoberto, procurando identificar se o mesmo auxilia a resolver o problema original que motivou a realização do processo KDD.

8. VANTAGENS DO *DATA MINING*

O uso de *Data Mining* para construção de um modelo traz as seguintes vantagens:

- Modelos são de fácil compreensão: pessoas sem conhecimento estatístico (por exemplo, analistas financeiros ou pessoas que trabalham com *data base marketing*) podem interpretar o modelo e compará-lo com suas próprias idéias. O usuário ganha mais conhecimento sobre o comportamento do cliente e pode usar esta informação para otimizar os processos dos negócios.

- Grandes bases de dados podem ser analisadas: grandes conjuntos de dados, de até vários gigabytes de informação podem ser analisados com *Data Mining*.

- *Data Mining* descobre informações não esperadas: como muitos modelos diferentes são validados, alguns resultados inesperados podem surgir. Em diversos estudos, descobriu-se que combinações de fatores particulares apresentaram resultados inesperados.

- Variáveis não necessitam de recodificação: *Data Mining* lida tanto com variáveis numéricas (quantitativas) quanto categóricas (qualitativas). Estas variáveis aparecem no modelo exatamente da mesma forma em que aparecem na base de dados.

- Modelos são precisos: os modelos obtidos por *Data Mining* são validados por técnicas de estatística. Desta forma, as predições feitas por modelos são precisas.

- Diversos segmentos da economia ganham novo fôlego no contato com o cliente e buscam ampliar oferecimento de serviços.

- Melhoria do relacionamento com clientes proporcionando aumento da rentabilidade.

- Agilidade;
- Confiança;
- Prevenção;
- Comparação.

9. DESVANTAGENS DO *DATA MINING*

- Alto custo das soluções: dificuldade da disseminação das ferramentas entre as corporações;

- Necessidade de grandes volumes de dados armazenados em poderosos servidores;

- Complexidade das Ferramentas de DM para pessoas que não fossem altamente especializadas. Isso significa que o DM ainda tem que ser feito no contexto da área de sistemas, a quem os usuários têm que submeter as suas solicitações, enquanto um expert processa os dados, para então receberem e examinarem a saída sumarizada. Se os resultados não satisfizerem, todo processo tem que ser recomeçado. Felizmente, há técnicas de DM mais compreensíveis, como, por exemplo, as árvores de decisão, que permitem que qualquer um, com conhecimentos básicos de computadores, possa utilizar e compreender o processo.

- O desafio da preparação dos dados para a mineração: A preparação dos dados para se realizar a mineração envolve muitas e trabalhosas tarefas num projeto de DM, sendo considerada como 80% do trabalho.

- As dificuldades de se realizar a análise custo/benefício do projeto de DM: estimar a taxa de retorno do investimento de um projeto de DM é complicado devido ao fato que, como o objetivo da tecnologia é descobrir tendências (em dados) que não seriam visíveis de outra maneira, torna-se virtualmente impossível estimar tal taxa a partir de algo que é desconhecido. Visto que normalmente um projeto de DM é razoavelmente caro, pode ser um tanto arriscado se decidir por um projeto desse tipo.

- Viabilidade dos fornecedores de ferramentas de DM: a viabilidade de mercado da maioria das ferramentas é uma preocupação das empresas que procuram uma ferramenta confiável, para hoje e para o futuro. Assim como qualquer nova tecnologia, a escolha do fornecedor é tão importante quanto à escolha da ferramenta.

10. CASOS REAIS UTILIZANDO *DATA MINING*

10.1.Wal-Mart

Embora recente, a história da Mineração de Dados já tem casos bem conhecidos. O mais divulgado é o da cadeia americana *Wal-Mart*, que identificou um hábito curioso dos consumidores. Ao procurar eventuais relações entre o volume de vendas e os dias da semana, o software apontou que, às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas. Crianças bebendo cerveja? Não. Uma investigação mais detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana.

10.2.Vestibular PUC-RJ

Utilizando as técnicas da mineração de dados, um programa de obtenção de conhecimento depois de examinar milhares de alunos forneceu a seguinte regra: se o candidato é do sexo feminino, trabalha e teve aprovação com boas notas no vestibular, então não efetivava a matrícula. Estranho, mas uma reflexão justifica a regra oferecida pelo programa: de acordo com os costumes do Rio de Janeiro, uma mulher em idade de vestibular, se trabalha é porque precisa, e neste caso deve ter feito inscrição para ingressar na universidade pública gratuita. Se teve boas notas provavelmente foi aprovada na universidade pública onde efetivará matrícula. Claro que há exceções: pessoas que moram em frente à PUC, pessoas mais velhas, de alto poder aquisitivo e que voltaram a estudar por outras razões que ter uma profissão, etc. Mas a grande maioria obedece à regra anunciada.

11. SOFTWARES

- *Enterprise Miner:* Ferramenta de data mining do SAS.

- *IlliMine:* Projeto de mineração de dados escrito em C++.

- *InfoCodex:* Aplicação de mineração de dados com uma base de dados linguística..

- KDB2000: Uma ferramenta livre em C++ que integra acesso à bases de dados, pre-processamento, técnicas de transformação e um vasto escopo de algoritmos de mineração de dados.

- *KXEN:* Ferramenta de mineração de dados comercial, utiliza conceitos do Profesor Vladimir Vapnik como Minimização de Risco Estruturada (*Structured Risk Minimization* ou SRM) e outros.

- *KNIME:* Plataforma de mineração de dados aberta que implementa o paradigma de *pipelining* de dados.

- *LingPipe:* API em Java para mineração em textos distribuída com código-fonte.

- MDR: Ferramenta livre em Java para detecção de interações entre atributos utilizando o método da *multifactor dimensionality reduction* (MDR).

- **Orange:** *Tookit* livre em Python para mineração de dados e aprendizado de máquina.

- Pimiento: Um ambiente para mineração em textos baseado em Java.

- Tanagra: Software livre de mineração de dados e estatística.

- WEKA: Software livre em java para mineração de dados. O *software* WEKA é formado por um conjunto de algoritmos de diversas técnicas para resolver problemas concretos de *Data Mining*. O WEKA está implementado em linguagem Java e foi desenvolvido no meio acadêmico da Universidade de Waikato, na Nova Zelândia, em 1999. È um software de domínio público estando disponível em: http://www.cs.waikato.ac.nz/ml/weka.

CONCLUSÃO

A construção de um banco de dados de cunho corporativo que integre dados operacionais com dados sobre clientes, fornecedores, e informações de mercado têm resultado em uma "explosão" de informações. E o panorama competitivo atual requer dos empresários brasileiros, investimento em tempo e em sofisticadas análises dentro de uma visão interativa dos dados.

Entretanto, existe ainda uma lacuna crescente entre a maior capacidade de armazenamento dos dados, sistemas de restauração e a habilidade efetiva dos executivos e empresários brasileiros em analisar e agir com as informações que contenham em suas bases de dados.

Um novo salto tecnológico se faz necessário quando se quer estruturar e priorizar informações críticas de marketing para a resolução de problemas muito específicos encontrados pelos analistas de mercado e pelos responsáveis pelos processos de tomada de decisão dentro das empresas. As ferramentas *Data Mining* podem promover este salto.

BIBLIOGRAFIA

http://pt.wikipedia.org/wiki/Minera%C3%A7%C3%A3o_de_dados#Wal-Mart (19/05/07)

http://pt.wikipedia.org/wiki/Data_Mining (19/05/07)

http://conged.deinfo.uepg.br/artigo3.pdf (20/05/07)

http://www.cpgmne.ufpr.br/dissertacoes/D070_Eliane_Prezepiorski_Lemos15072003.pdf (21/05/07)

http://www.dwbrasil.com.br/html/dmining.html (22/05/07)