EXTRAÇÃO DE CONHECIMENTO EM BASES DE DADOS

KNOWLEDGE DISCOVERY IN DATABASE

Luiz Sérgio dos SANTOS*

RESUMO: Dentro dos Banco de Dados das empresas residem informações extremamente valiosas. Informações de operações e comportamento dos clientes, transações contábeis, dados de faturamento, dados financeiros, contas a receber, contas a pagar, histórico dos principais produtos, enfim, informações que geram um conhecimento interno muito importante e de grande potencial. Quando analisadas corretamente, elas podem ser de grande valor e, às vezes, permanecem escondidas porque as relações e certas características entre elas não são óbvias nem mesmo para os grandes especialistas das empresas. As ferramentas de consulta e visualização no processo KDD (Knowledge Discovery in Databases) tentam descobrir o valor escondido nos dados, proporcionando às empresas vantagens competitivas e benefícios organizacionais. Benefícios estes de que as ferramentas e técnicas convencionais de exploração de Bancos de Dados não conseguem perceber os detalhes em sua totalidade.

UNITERMOS:Extração de Conhecimento em Bases de Dados (KDD); Mineração de Dados (DM).

ABSTRACT: Inside of the database of the companies extremely valuable informations reside. Informations of operations and the customers' behavior, all the accounting transactions, data of revenue, bills to receive,

* Formado em Administração pela EAESP/FGV e Mestre em Gestão de Sistema de Informação pela PUCCAMP, Campinas, SP - Brasil.

bills to pay, historical of the principal products, finally, informations of a knowledge in potential. When correctly analyzed, these informations can be of great value, and sometimis stay hidden because the relationships and tendencies are not obvious not even for the great specialists of the companies. The tools of consultation and visualization in the KDD process (Knowledge Discovery in Databases) will try to discover the hidden value in the data, providing the companies competitive advantages and organizational benefits which the tools and conventional techniques of database don't get to notice the details in your totality.

UNITERMS: Knowledge Discovery in Databases (KDD); Data Mining (DM).

INTRODUÇÃO

Uma grande quantidade de dados está sendo colecionada em Banco de Dados (BD ou DB), acumulando uma grande quantidade de conhecimentos. Há uma necessidade crescente e urgente de novas teorias computacionais e de ferramentas para ajudar o homem a extrair informações úteis (conhecimentos). Estas informações estão levando ao surgimento do campo da Descoberta de Conhecimento em Banco de Dados (KDD).

A noção de encontrar padrões novos e úteis em dados recebeu uma série de nomes: Extração de Conhecimentos, Descoberta de Informação, Arqueologia de Dados, Processo Padrão de Dados e Mineração de Dados.

Mineração de Dados é um passo no processo de KDD que consiste em aplicar análise de dados e algorítmos de descoberta que produzem uma enumeração particular de padrões, ou modelos, em cima de dados.

Devemos saber que a palavra *dados* isoladamente não representa muito, o que importa realmente é a informação. Então é necessário fazer uma distinção entre dados e informação. Dados

são os componentes básicos a partir dos quais a informação é criada. Informação são dados que estão sendo analisados em uma determinada situação. A partir da informação vem o conhecimento, que permite tomar decisões adequadas, trazendo vantagens competitivas.

Há dois tipos de ferramentas de Mineração de Dados (DM) que permitem a consulta e visualização no processo KDD: as ferramentas de *descoberta* e *verificação*.

A maioria das ferramentas de descoberta surgiu de algoritmos de Inteligência Artificial (AI) para simular nos sistemas de computadores a atividade do cérebro humano. Permitem descobrir padrões e tendências em dados que podem, de acordo com tendências de mercado, ajudar as companhias a descobrir conhecimentos novos usados para obter-se vantagens competitivas.

As ferramentas de verificação podem ajudar a tomar decisões de negócio corretas e seguras, uma vez que ajudam a validar as descobertas na Mineração de Dados, ao passo que as ferramentas de descoberta podem encontrar novas situações interessantes, executar uma série de testes para procurar as diferenças, mas não podem explicar se as descobertas são realmente válidas.

Data Mining

Assumindo em português a expressão *mineração de dados* este conceito significa a manipulação de dados do negócio de forma a descobrir fatos que até então eram desconhecidos pela empresa. Dessa maneira, a mineração vem enriquecer o manancial de informações estratégicas já fornecidas pelo Data Warehouse (DW) ao responsável por tomadas de decisões estratégicas. A mineração de dados utiliza

técnicas estatísticas a fim de realizar inferências sobre os dados, por exemplo, do cliente. Essas inferências são então validadas de forma a serem utilizadas para descoberta de novas inferências. O processo Data Mining é capaz de localizar tendências, modelar agrupamentos e diferenças entre os dados, a fim de fornecer informações ao executivo que o previnam de acontecimentos não previstos.

Uma questão que pode surgir aqui é: por que só recentemente se utiliza o termo mineração quando as técnicas estatíticas já estão disponíveis há décadas? A resposta é que hoje as ferramentas de *mining* vêm acompanhadas de recursos gráficos sofisticados que permitem análises de cálculos intensos, incluindo "features" para marketing, planejamento estratégico e finanças. Os dados utilizados para mineração são geralmente provenientes de um Data Warehouse (DW).

O motivo para tal é que no DW os dados já se encontram orientados por assunto, já estão consolidados (pois são provenientes de múltiplos sistemas legados) e já se encontram com as limpezas necessárias afetuadas. Entretanto, não existe impedimento para realização de mineração em dados provenientes de sistemas operacionais do mundo OLTP (On-line Transaction Processing).

Um exemplo clássico de *data mining* foi desenvolvido pela Wal-Mart. A empresa descobriu que o perfil do consumidor de cervejas era semelhante ao de fraldas. Eram homens casados, entre 25 e 30 anos, que compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa. Com base na verificação destas hipóteses, a Wal-Mart optou por uma otimização das atividades junto às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas. Resultado: o consumo cresceu 30% às sextas-feiras com a redefinição de lay-out baseada na conexão de hipóteses desenvolvidas pelo *data mining*.

Técnicas de Mineração

Entre as técnicas de mineração existentes destacam-se: Redes Neurais, Árvores de Decisão, Cluster (agrupamento), Associação e outras.

Redes Neurais é uma técnica que tenta imitar a forma como o cérebro humano aprende a reconhecer objetos, situações, sons, imagens, etc. Uma rede neural é formada pela interconexão dos chamados neurônios artificiais que simulam o comportamento dos neurônios humanos. A rede é treinada com dados (padrões) que a ensinam a reconhecer ou prever novos dados por ela ainda não conhecidos. A rede possui algumas camadas de neurônios que recebem os dados desde a camada de entrada, propagando essa informação até a camada de saída. Cada neurônio tem conexões que possuem pesos associados.

A informação sobre os dados chega até o neurônio através dessas conexões, sofre cálculo no interior do neurônio e vai sendo transformada entre a rede até chegar a camada de saída. O valor da camada de saída é então comparado com o valor que seria o correto para uma dada informação de entrada na rede. Mediante a diferença entre o valor de saída e o valor correto, os pesos são modificados de forma que a rede possa aprender o comportamento dos dados. Dessa forma, a rede neural estará apta a descobrir e prever relacionamentos entre os dados.

As árvores de decisão são muito utilizadas em problemas onde há escolhas a serem tomadas. Por exemplo, numa empresa de cartão de crédito deseja saber de antemão a probabilidade que tem um determinado cliente de pagar as suas contas em dia. Nesse caso, a variável alvo ou dependente é: pagamento do cliente (em dia ou não pago). Outras variáveis chamadas de independentes, que influenciam no comportamento da variável dependente, são analisadas de modo a prever o comportamento de novos clientes.

Cluster é o processo de particionar o Database de forma que os registros que têm características semelhantes sejam colocados no mesmo grupo. Uma outra medida importante é que os grupos sejam separados por características diferentes na mesma proporção. A técnica de Associação usa o Database com transações de um determinado negócio e descobre quais tipos de transações implicam nas outras. Tomando como exemplo um supermercado, de todos os clientes que compraram queijo, 55% deles também compraram algum tipo de pão e 35% compraram leite.

Exemplo de Aplicação

Na indústria de telecomunicações a perda de um percentual de clientes, como também a aquisição de novos outros, tem um alto custo. Dessa forma, reduzir o que é conhecido como "churn defection" (perda do cliente para outra companhia de telecom) é uma excelente oportunidade de negócios. A idéia então é atacar os clientes que têm maior probabilidade de "churning" e oferecer facilidades (utilidades) para retê-los na companhia.

Há várias razões que fazem com que um consumidor procure outra companhia. Muitas vezes é por causa de um serviço de má qualidade, outras vezes é por causa de um serviço mais barato e algumas vezes porque o serviço não acompanha as mudanças na atividade do cliente.

O problema de "churning" implica que a empresa procure por um melhor entendimento sobre os seus clientes a fim de ter competitividade com as outras empresas do ramo. Isso pode ser obtido através do discernimento do critério-chave que leva à saída do consumidor da companhia e da criação de estratégias para manter a lealdade dos clientes. Assim, é necessário identificar características de consumidores cujos relacionamentos são mais prováveis de serem de curta duração. Essas descobertas podem

servir de instrumento para medição do impacto de competição, baseado em promoções realizadas por competidores, no tipo de oferta, no tipo de cliente e nos padrões relacionados geograficamente.

Reter um cliente custa de 5 a 10 vezes menos que adquir um novo. Dessa forma, o ideal é manter os melhores clientes. Para a endereçar o problema de "churning" é necessário identificar clientes que têm perfil valioso, oferecer novos serviços a clientes baseados nos seus perfis, identificar clientes que possuem risco de "churn" e em alguns casos encorajá-los a migrar para o competidor e fornecer informação para vendas e campanhas de marketing. A fim de atingir esses objetivos é preciso analisar clientes e seus padrões de utilização, predizer quais são os clientes que possuem alto custo de "churn", atribuir valores aos clientes e fornecer ajuda para promover a retenção ou migração de um cliente.

A melhor informação que uma empresa tem sobre os seus consumidores e suas formas de utilização é o tipo de decisão que eles podem tomar. Todas as soluções de gerenciamento de "churning" incluem análise de frequência dos fatores mais comuns que levam ao "churn". As variáveis que implicam a ocorrência de "churn" não são estáticas. O relativo impacto destas variáveis varia de região para região ou de um país a outro, dependendo dos fatores de comportamento de uso, condições econômicas, competição, marketing, legislação e tecnologia. Assim, os modelos devem ser atualizados dinamicamente.

Hoje o serviço pode ser considerado excelente devido a uma série de fatores, entre os quais destacam-se o uso da tecnologia sem fio e a competitividade acirrada dos serviços oferecidos pelas telefônicas, o que torna imprescindível a rápida adaptação a novas oportunidades de mercado. Dessa forma, os serviços de telecom devem combinar performance com flexibilidade e modularidade. A mineração de dados vem de forma a fornecer os subsídios necessários para competição de

igual para igual no mercado de telecom. DM é a tecnologia chave para empresas que querem melhorar a qualidade de tomada de decisão e ganhar competitividade explorando o dado disponível em suas bases de dados.

A mineração provê gerenciamento de "churn" modelando e prevendo capacidades através da procura em grandes volumes de dados de informação, procurando por padrões escondidos e tendências. O software de mineração automaticamente examina as razões que levaram certos tipos de consumidores a sairem da empresa e também os motivos que fizeram os outros consumidores permanecerem na empresa. A ferramenta de mineração usa até mesmo centenas de variáveis tais como demografia, sazonalidade, tipo do serviço, padrões de utilização, número de telefones, a presença de competidores eventuais, etc. Após a mineração é possível descobrir quais são os consumidores que se encaixam num determinado critério, identificando que são altamente prováveis de saírem da empresa num determinado mês. Assim, a empresa toma a ação apropriada a reter esses clientes, se for o caso.

CONCLUSÃO

A descoberta de conhecimento em bases de dados – KDD – pode significar um grande diferencial competitivo para as empresas. O uso de ferramentas no processo KDD pode gerar valores importantes em áreas como:

Previsão - Os modelos gerados pela mineração podem, a partir da série histórica dos dados, prever o comportamento futuro de novos dados

Detecção de Fraudes - Modelos podem identificar, através das diversas técnicas de mineração, algum comportamento fora do padrão nos dados analisados.

Retenção e Aquisição de Novos Consumidores - Baseado

no comportamento passado dos clientes de uma determinada empresa, as campanhas para retenção e aquisição de novos clientes podem ser mais efetivas.

Cross-Selling - De posse do entendimento sobre o comportamento dos clientes é possível oferecer a cada tipo de segmento descoberto um produto adequado.

Segmentação de Mercado - Técnicas de "clustering" podem ser utilizadas para agrupar tipos de clientes. Com essa informação é possível aumentar o retorno sobre o investimento e obter resultados máximos.

REFERÊNCIAS BIBLIOGRÁFICAS

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH P. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, Fall,1996.

FAYYAD, U. M.; PIATETSKY-SHAPIRO G.; SMYTH P. **From Data Mining to Knowledge Discovery:** An Overview. In: ____. **Advances in Knowledge Discovery and Data Mining.** AAAI Press,1996.

EDELSTEIN, H. **Technology how to: Mining data warehouses.** CMP Publications: Jan. 98, 1996.

SRIKANT R.; VU Q.; AGRAWAL R. **Mining Association Rules winth Item Constraints. 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining.** California: Newport Beach, August 1997.