Introdução

As duas últimas décadas acompanharam um aumento dramático na quantia de informações ou dados que são armazenadas em formato eletrônico. Esta acumulação de dados aconteceu a uma taxa explosiva. Foi calculado que a quantia de informação no mundo dobra a cada 20 meses e o tamanho e número de bancos de dados estão aumentando ainda mais rapidamente. O valor destes dados armazenados está tipicamente ligado à capacidade de extrair informações de mais alto nível que se encontra subjacente a estes dados, ou seja, informação útil que sirva para dar suporte a decisões, e para exploração e melhor entendimento do fenômeno gerador dos dados. Podem existir padrões ou tendências úteis interessantes que, se descobertos, podem ser utilizados, por exemplo, para otimizar um processo de negócio em uma empresa, ajudar no entendimento dos resutados de um experimento científico, ajudar médicos a entender efeitos de um tratamento entre outros.

Dentro deste contexto, *data mining* (mineração dos dados - extração de informações implícitas, padrões ocultos em bases de dados) tem ganho muita atenção de diversas áreas de interesse. Elas o consideram como um campo crítico para seus negócios. O uso de informações valiosas obtidas por mineração dos dados é necessário para manter a competitividade no ambiente comercial atual. Com o advento do *"data warehousing"* que faz a armazenagem de grandes quantidades de dados em um local comum e do contínuo avanço no aumento do poder de processamento dos computadores, os empresários procuram por tecnologias e ferramentas para extrair informações úteis dos dados.

1

O que é Data Mining?

É o processo de descobrir informações relevantes, como padrões, associações, mudanças, anomalias e estruturas, em grandes quantidades de dados armazenados em banco de dados, depósitos de dados ou outros repositórios de informação. Devido a disponibilidade de enormes quantias de dados em formas eletrônicas, e à necessidade iminente de extrair delas informações e conhecimentos úteis a diversas aplicações, por exemplo na análise de mercado, administração empresarial, apoio à decisão, etc, *data mining* foi popularmente tratado como sinônimo de descoberta de conhecimento em bases de dados, apesar de, na visão de alguns pesquisadores, *data mining* será considerado como um passo essencial da descoberta de conhecimento. Em geral, um processo de descoberta de conhecimento consiste em uma iteração das seguintes etapas:

- **Preparação:** é o passo onde os dados são preparados para serem apresentados às técnicas de *data mining*. Os dados são selecionados (quais os dados que são importantes), purificados (retirar inconsistências e incompletude dos dados) e pré-processados (reapresentá-los de uma maneira adequada para o *data mining)*. Este passo é realizado sob a supervisão e conhecimento de um especialista, pois o mesmo é capaz de definir quais dados são importantes, assim como o que fazer com os dados antes de utilizá-los no *data mining*.

- **Data Mining:** é onde os dados preparados são processados, ou seja, é onde se faz a mineração dos dados propriamente dita. O principal objetivo desse passo é transformar os dados de uma maneira que permita a identificação mais fácil de informações importantes.

- **Análise de Dados:** o resultado do *data mining* é avaliado, visando determinar se algum conhecimento adicional foi descoberto, assim como definir a importância dos fatos gerados. Para esse passo, várias maneiras de análise podem ser utilizadas, por exemplo: o resultado do *data mining* pode ser expresso em um gráfico, em que análise dos dados passa a ser uma análise do comportamento do gráfico.

Data mining é uma das ferramentas mais utilizadas para extração de conhecimento através de bancos de

dados (Knowledge Discovery in Databases - KDD), tanto no meio comercial quanto no meio científico.

Extração de Conhecimento em Base de Dados

A extração de conhecimento em bases de dados consiste na seleção e processamento de dados com a finalidade de identificar novos padrões, dar maior precisão em padrões conhecidos e modelar o mundo real. *Data mining,* em português, *mineração de dados* se refere ao exame de grandes quantidades de dados, procurando encontrar relações entre dados não explícitas que possam ser usadas em modelos do mundo com capacidade preditiva e explanatória. Espera-se que o conhecimento extraído seja utilizado. Neste caso, seu uso dará frutos que poderão ou não interferir com novos dados a serem obtidos, como foi visto em Barreto [3].

O objetivo deste capítulo é apresentar alguns conceitos e definições do contexto de *data mining* e principalmente dar uma introdução ao processo completo de descoberta do conhecimento incluindo esquemas gráficos e a descrição de algumas de suas tarefas básicas. No final do capítulo será apresentada a descrição de algumas aplicações de sucesso que mostram como as técnicas de *data mining* atingem todas as áreas do conhecimento.

Um ciclo completo está representado na Figura 4.1 (Inspirada em Fayyad [11]).

O ponto de partida do ciclo consiste em tomar todos os dados referentes a um assunto que seja possível obter, o que está representado na figura por dados brutos.

O seguinte passo é consolidar estes dados procurando dar uma estrutura conveniente para serem explorados e armazenados. Esta fase, de grande importância é conhecida por *data warehouse,* ou armazém de dados.

Neste momento é conveniente que se tenha alguma hipótese sobre o possível modelo que se vai obter, para que um pré-processamento coloque os dados de modo conveniente à obtenção deste modelo *(data minig),* que deve ser interpretado para extrair o conhecimento desejado.

Definições do Termo

Descobrir padrões úteis em dados é conhecido em diversas comunidades por nomes diferentes como: extração de conhecimento, descoberta de informação, colheita de informação, arqueologia de dados e processamento de padrão de dados inclusive *data mining.* O termo *data mining* é muito usado por estatísticos, pesquisadores de banco de dados e comunidades de negócio.

O termo KDD (Knowledge Discovery in Databases) refere-se ao processo global de descobrimento de conhecimento útil em bases de dados. *Data mining* é um passo particular neste processo-aplicação de algoritmos específicos para extrair padrões (modelos) de dados. Os passos adicionais no processo KDD, como preparação de dados, seleção de dados, limpeza de dados, incorporação de conhecimento anterior apropriado e interpretação formal dos resultados de mineração assegura

aquele conhecimento útil que é derivado dos dados. A aplicação cega de métodos de *data mining* pode ser uma atividade perigosa que conduz a descoberta de padrões sem sentido.

O KDD evoluiu e continua evoluindo da interseção de pesquisas em campos como bancos de dados, aprendizado de máquinas, reconhecimento de padrões, estatísticas, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados, descoberta científica, recuperação de informação e computação de alto-desempenho. Sistemas de software KDD incorporam teorias, algoritmos e métodos de todos estes campos.

Data Warehouse é um armazém centralizado de dados. *Data Warehousing* referese à organização dos dados para os tornar disponíveis para análise *on une.* Uma das ferramentas que vem apresentando vantagens em relação à SQL (uma linguagem de definição e manipulação de dados) é a OLAP (On line Analytical Processing). Existem similaridades e diferenças entre OLAP e *data mining.*

44

O termo *data mining* teve conotações negativas em estatísticas desde a década de 1960, quando o computador baseado em técnicas de análise de dados foi introduzido primeiro. A preocupação surgiu em cima do fato de que pesquisas minuciosas em qualquer conjunto de dados, podem identificar padrões que parecem ser estatisticamente significantes mas de fato não o são. *Data mining* produz resultados eficazes desde que usado corretamente.

Data mining e Reconhecimento de Padrões:

Alguns textos da área, como por exemplo Kennedy [9], utilizam os termos *"data mining"* e *"pattern recognition"* com o mesmo significado, pois ambos se concentram na extração de informações ou relacionamentos dos dados. O termo *"data mining"* é originário principalmente das aplicações em bases de dados comerciais, enquanto *"pattern recognition"* foi derivado dos campos de engenharia tais como controle de processos e inspeção de qualidade. Os dois termos tratam essencialmente das mesmas idéias, mas representam a nomenclatura desenvolvida em diferentes áreas visto em Kennedy [9].

Principais Tarefas

Em geral, as tarefas do *data mining* podem ser classificadas em duas categorias:

descriptive data mining e *predictive data mining.* O primeiro descreve o conjunto de dados de uma maneira concisa e resumida e apresenta propriedades gerais interessantes dos dados; o segundo constrói um ou um conjunto de modelos, realiza inferências sobre o conjunto de dados disponíveis e tenta predizer o comportamento de novos conjuntos de dados.

Um sistema de *data mining* pode realizar pelo menos uma das seguintes tarefas:

> **1. Descrição de classes** - provê um resumo conciso e sucinto de uma coleção de dados e a distingue de outras. O resumo de uma coleção de dados é chamado de caracterização de classe; enquanto a comparação entre duas ou mais coleções de dados é chamada comparação

ou discriminação de classe. A descrição de classe não só deveria cobrir suas propriedades de resumo tal como a contagem, somas, e cálculos de médias, mas também suas propriedades sobre a dispersão dos dados, tais como a variância, desvio padrão, quartis, etc.

Por exemplo, a descrição de classe pode ser usada para comparar as vendas européias e asiáticas de uma companhia, identificar os fatores importantes que discriminam as duas classes e apresentar um resumo conciso.

2. Associação - é a descoberta de relações de associação ou correlações entre um conjunto de itens. Eles são expressados frequentemente na forma de regras que mostram as condições atributo-valor que acontecem frequentemente juntas em um determinado conjunto de dados. Uma regra de associação da forma X —+ Y é interpretada como "tuplas (conjunto de valores de atributos) de base de dados que satisfazem X são prováveis que satisfaçam *Y*." Análise de associação é extensamente usada em "transaction data analysis for directed marketing", design de catálogo e outros processos de decisões comerciais.

Significativo esforço de pesquisa foi desenvolvido em análise de associações com a proposição de algoritmos eficientes, incluindo "level-wise", mineração em múltiplos níveis, associações multidimensionais, mineração de associações numéricas, categóricas e de intervalos de dados, mineração baseada em restrições além de mineração de correlações como Elmasri & Navathe.

3. Classificação - analisa um conjunto de dados de treinamento (i.e., um conjunto de objetos cuja classificação **já** é conhecida) e constrói um modelo para cada classe baseado nas características dos dados. Uma árvore de decisão ou um conjunto de regras de classificação é gerado por tal processo de classificação, que pode ser usado para entender melhor cada classe no banco de dados e para classificação de futuros dados. Por exemplo, alguém pode classificar doenças e ajudar a prever tipos de doenças baseados nos sintomas dos pacientes.

Houveram muitos métodos de classificação desenvolvidos nos campos de aprendizagem de máquina, estatística, banco de dados, redes neurais, conjuntos rough sets", e outros. A classificação foi usada em segmentação de clientes, modelagem de negócios e análise de crédito.

4. Previsão - esta função de mineração prediz os possíveis valores de alguns dados perdidos ou a distribuição de valores de certos atributos em um conjunto de objetos. Ela envolve a descoberta de um conjunto de atributos relevantes para o atributo de interesse (e.g., por algumas análise estatística) e prediz a distribuição do valor baseada no valor do conjunto de dados semelhantes ao(s) objeto(s) selecionado(s). Por exemplo, o salário potencial de um empregado pode ser predito baseado na distribuição do salário de empregados semelhantes na companhia. Usualmente, análise de regressão, modelo linear generalizado, análise de correlação e árvores de decisão são ferramentas úteis em predição de qualidade. Também são usados algoritmos genéticos e redes neurais com bastante sucesso.

5. Agrupamento - análise de "clusters" ou de agrupamento consiste em identificar possíveis agrupamentos nos dados, onde um agrupamento é uma coleção de objetos que são "semelhantes" um ao outro. Diferentes medidas de similaridade, baseadas em funções de distância podem ser especificadas para diferentes contextos de aplicação. Um bom método de "cluster" assegura que a similaridade inter-cluster é baixa e a similaridade intra-cluster é alta. Por exemplo, pode-se agrupar as casas de uma área de acordo com sua categoria, área construída e localização geográfica.

Data mining têm enfocado suas pesquisas em métodos de "clustering" de alta qualidade para grandes bases de dados e armazém de dados *(data warehouse)*.

6. Análise de série temporal - analisa um grande conjunto de dados de séries temporais para encontrar certas regularidades e características interessantes,

incluindo a pesquisa de sequências ou subsequências semelhantes e descobrindo assim

padrões sequenciais, periodicidades, tendências e divergências. Por exemplo , pode-se predizer a tendência dos valores acionários para uma companhia baseando-se em sua história acionária, situação empresarial, desempenho dos competidores e mercado atual.

Há outras tarefas do *data mining,* como análise de "outlier". A Identificação de novas tarefas para fazer melhor uso melhor uso dos dados coletados é um tópico de pesquisa interessante.

Principais Tecnologias usadas em KDD

• Organização de dados *(data warehousing).*

• Banco de dados distribuídos são úteis pois frequentemente vê-se obrigado a trabalhar com grandes volumes de dados que se encontram distribuídos em diferentes plataformas.

• IA e sistemas especialistas.

• Redes neurais e seus paradigmas de aprendizado supervisionado e não supervisionado. Principalmente neste segundo caso estas redes se mostram úteis por suas características de identificação de agrupamentos de dados semelhantes, fato dificilmente detectável sem seu auxílio.

• Interfaces amigáveis incluindo realidade virtual.

Sistemas de Informação e KDD

A habilidade peculiar do ser humano de juntar e armazenar mais dados do que pode analisar e entender demanda o uso de técnicas de aquisição, organização, armazenagem e recuperação de informações. Figurando como uma destas técnicas está a utilização de Sistemas de Informações (SI). SI's são constituídos de um conjunto de dados com atributos relevantes e disponíveis. Os dados de um SI são provindos de uma fonte de dados e armazenados em uma memória não volátil (permanente). Utilizam regras para combinar estes dados em informações sumarizadas e visões sobre os dados manipulados como visto em Barreto [3].

A extração de conhecimentos em bases de dados (KDD) consiste na seleção e processamento de dados com a finanlidade de identificar novos padrões, dar maior precisão em padrões conhecidos e modelar fenômenos do mundo real, utilizando-se de, entre outras técnicas, mineração de dados *(data mining).*

SI's e KDD são tópicos intimamente ligados, visto que sistemas de informação permitem o armazenamento, recuperação e organização de grandes volumes de dados e as técnicas de extração de conhecimento em bases de dados obtêm melhores resultados quando aplicados a massivos repositório de dados.

Exemplos Aplicação

Nesta seção serão mencionados alguns exemplos de sucesso onde foram aplicadas técnicas de *data mining.*

Wal- Mart

Embora recente, a história do *data mining* já tem casos bem conhecidos. O mais divulgado é o da cadeia americana Wal-Mart, que identificou um hábito curioso dos consumidores. Há cinco anos, ao procurar eventuais relações entre o volume de vendas e os dias da semana, o software de *data mining* 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.

Bank of America

Há quem consiga detectar fraudes, cortar gastos ou aumentar a receita da empresa. O Bank of America usou essas técnicas para selecionar entre seus 36 milhões de clientes aqueles com menor risco de dar calote num empréstimo. A partir desses relatórios, enviou cartas oferecendo linhas de crédito para os correntistas cujos filhos tivessem entre 18 e 21 anos e, portanto, precisassem de dinheiro para ajudar os filhos a comprar o próprio carro, uma casa ou arcar com os gastos da faculdade. Resultado : em três anos, o banco lucrou 30 milhões de dólares.

Telecomunicações

Atualmente, em telecomunicações, existe uma explosão nos crimes contra a telefonia celular, dentre os quais, a clonagem. Técnicas de *data mining* podem ser utilizadas para detectar hábitos dos usuários de celulares. Quando um telefonema for feito e considerado pelo sistema como uma excessão, o programa faz uma chamada para confirmar se foi ou não uma tentativa de fraude.

Administração em Alto Nível

Depois do final da segunda guerra mundial a Pesquisa Operacional (P0) apareceu como ferramenta fundamental para a vitória das tropas contra as potências do eixo. Com a pesquisa operacional foi possível resolver matematicamente o problema de alocação ótima de recursos e isto vem sendo utilizado com grande sucesso em altos níveis de decisão até o presente momento. Cerca de cinquenta anos depois, apareceu o *data mining.* Suas potencialidades estão longe de serem imaginadas e não seria ousado esperar que no mundo globalizado possa vir a dar seus frutos como a PO deu no passado.

Medicina

Atualmente as técnicas de *data mining* são pouco usadas em medicina. No momento, o ponto que está emperrando o uso de *data mining* é o fato de que *data mining*

48

sendo uma nova concepção dirigida para pesquisa ainda é quase completamente desconhecida da comunidade médica. Ora, se existem dados clínicos abundantes, estes dados são frequentemente adequados a um estudo de *data mining* por não conterem dados que aparentemente são inúteis mas que são exatamente os que o pesquisador de *data mining* procura.

Vestibular PUC-RJ

Utilizando as técnicas de *data mining,* 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, então não efetiva matrícula.* Estranho, ninguém havia pensado nisso... 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 excessões: 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!