2
Descoberta de Conhecimento em Bases de Dados

2.1.
Introdução

De acordo com [FAYY96], o conceito de descoberta de conhecimento em bases de dados pode ser resumido como o processo não-trivial de identificar padrões novos, válidos, potencialmente úteis e, principalmente, compreensíveis em meio às observações presentes em uma base de dados.

O objetivo último da descoberta do conhecimento em bases de dados não é o de simplesmente encontrar padrões e relações em meio à imensa quantidade de informação disponível em bases de dados, e sim a extração de conhecimento inteligível e imediatamente utilizável para o apoio às decisões.

O processo de descoberta do conhecimento [BERR00] [FAYY96] [PYLE99] [KELL95] é composto por várias etapas. A Figura 2-1 ilustra o ciclo de descoberta do conhecimento em bases de dados e suas etapas.

A origem diversa dos dados que serão utilizados, coletados em diferentes instantes de tempo em lugares distintos, cria um esforço inicial de consolidação e agrupamento de toda a informação que irá servir de base para o processo. A compreensão do negócio e do ambiente no qual os dados estão inseridos é crítica para o entendimento dos mesmos. Dada essa diversidade e heterogeneidade dos dados, esforços de pré-processamento e limpeza dos mesmos são cruciais na geração de dados que possam vir a ser trabalhados em busca de conhecimento útil. É essencial que seja realizada a investigação de inconsistências e problemas devido a diferenças de escalas, assim como o tratamento de valores fora da normalidade (*outliers*) e observações errôneas. Realizadas essas tarefas iniciais, que tornam os dados tratáveis e homogêneos, a mineração dos dados pode ser iniciada, na busca por padrões e relações que façam sentido e sejam úteis para o problema a ser resolvido ou objetivo a ser alcançado. Finalmente, a interpretação, compreensão e aplicação dos resultados encontrados é o passo que torna o conhecimento adquirido através de bases de dados um real insumo para o apoio às decisões [BERR00] [BOZD03] [FAYY96] [KLOS02].

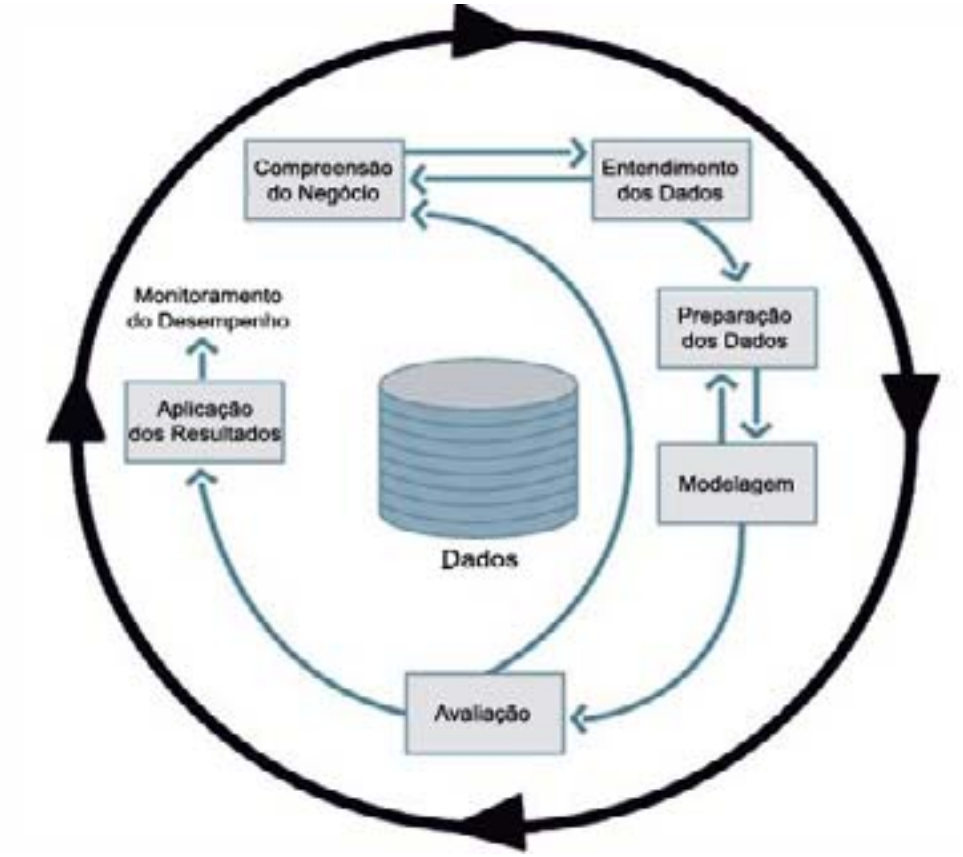

Figura 2.1 – Ciclo da Descoberta do Conhecimento em Bases de Dados

2.2. Dados

Para que qualquer conhecimento seja gerado a partir de dados, o primeiro passo é que tais dados existam e estejam disponíveis em algum lugar [BOZD03] [KLOS02]. Empresas e outras organizações fazem uso de uma grande infra-estrutura de tecnologia da informação para garantir a disponibilidade e o uso adequado da informação no apoio à decisão [BERR00] [PYLE99].

2.2.1. Data Warehouses

Uma das formas de organização, consolidação e disponibilidade de dados são os chamados “armazéns de dados” (*data warehouses*) [BALL99] [KELL95]. Um *data warehouse* é um grande repositório de bases de dados alimentado por muitos sistemas operacionais. *Datas warehouses* nascem da necessidade de empresas e organizações possuírem uma visão de dados e operações centralizadas em um único ponto, ao invés de ter seus dados espalhados por

diversos locais ou departamentos sem muita coesão e por vezes incomunicáveis entre si. Um dos objetivos primordiais de *data warehouses* é garantir a integridade e consistência de todos os dados coletados dos sistemas operacionais de uma empresa, além do acesso eqüânime de todas as partes a esses dados. Em geral, ao serem encaminhados para um *data warehouse*, dados operacionais são limpos e transformados em uma primeira instância, principalmente para garantir que dados de diferentes fontes e formatos passem então a possuir as mesmas definições e obedeçam às mesmas regras.

Por vezes algumas das transformações padrões realizadas nos dados para se encaixarem no *data warehouse* podem danificar ou até mesmo destruir informação que poderia vir a ser valiosa no processo de descoberta de conhecimento. Normalização, agregação e sumarização dos dados são algumas dessas transformações que podem vir a atrapalhar a análise e mineração dos dados. Por exemplo, um *data warehouse* de uma companhia de cartões de crédito pode possuir meses de balanços mensais de seus clientes, mas pode não incluir a informação sobre cada transação realizada, o que resulta em uma grande perda de informação sob o prisma de mineração de dados.

Na medida que o design de *data warehouses* evolui, eles focam cada vez mais em antever as necessidades de análise e evitar a perda de informação, mesmo sendo por vezes inevitável, devido às imensas quantidades de dados e problemas de espaço de armazenamento [KLOS02].

2.2.2. OLAP, Data Marts e Bases de Dados Multidimensionais

Outro tipo comum de organização de bases de dados para seu uso no processo de apoio à decisão são *data marts* desenvolvidos para processamento analítico on-line, ou OLAP (*Online Analytical Processing*). Bases de dados OLAP possuem desempenho superior tanto em velocidade quanto em qualidade de resposta, quando comparados a *data warehouses* em tarefas de apoio à decisão. A razão para isso é o fato delas apresentarem uma única visão dos dados, em geral relacionada ao departamento específico que possui os dados. Bases de dados OLAP são organizadas levando-se em consideração as diferentes dimensões do negócio, como tempo, tipo de produto e geografia, permitindo que o analista em busca de conhecimento se mova facilmente através das dimensões de interesse.

Quando uma base de dados multidimensional é armazenada em bases de dados relacionais, a organização das tabelas é feita no esquema estrela - uma

tabela central com muitas tabelas chave ao redor. Tal modelo é particularmente apropriado para bases de dados que sirvam aos interesses de departamentos específicos, como vendas ou marketing. É mais fácil chegar-se a uma conclusão sobre que fatos centrais serão acompanhados e que dimensões subjacentes são importantes em um nível departamental, do que fazer o mesmo tendo-se em vista todo o negócio. Essas bases de dados criadas especialmente com o intuito de servirem a processos de apoio à decisão, são conhecidas como *data marts* [BERR00].

Alguns *data warehouses* são na verdade coleções de vários *data marts* [BOZD03].

2.3. Preparação dos dados

O processo de preparação dos dados é de grande relevância na busca do conhecimento em bases de dados. Mesmo já existindo *data warehouses* ou até mesmo *data marts* com os dados disponíveis para análise e já pré-processados, é essencial criar-se uma representação dos dados que satisfaça os objetivos da análise de dados a ser realizada e que se encaixe de forma ótima na resolução do problema enfrentado. O conceito de preparação de dados engloba: consultas (*queries*) iniciais a *data warehouses* ou outros repositórios de dados em busca dos dados procurados; consolidação de toda a informação de interesse em um local ou base única; limpeza e tratamento de erros, valores aberrantes, valores *missing* e inconsistências; transformações nos dados como normalização, diferença temporal, e a criação de variáveis derivadas das variáveis originais que possuam mais informação; transposição dos dados para o nível de agregação correto, isto é, por exemplo, escolher trabalhar com a visão de cliente ou a visão de chamada em uma base de telefonia; aplicação de métodos de seleção de variáveis [CONT02] [BLUM97] [KWAK03] [YI97].

De acordo com [PYLE99], o processo de preparação dos dados tem uma razão muito maior para existir do que simplesmente fornecer uma base limpa e sem erros para os modelos a serem utilizados: ao preparar os dados, o responsável pela análise também é “preparado” pelos dados. Ao despender esforço para obter a melhor representação possível para os dados, o analista convive com os dados e aprende suas nuances e detalhes, ao mesmo tempo em que compreende melhor o problema que está sendo estudado, o que contribui

em etapas futuras para um melhor desempenho, tanto do analista quanto dos dados.

2.4. Mineração de dados

O termo mineração de dados (*data mining*) define a exploração e análise de grandes quantidades de dados com o objetivo de encontrar padrões, regras e relações interessantes e significativas para algum fim [BERR00]. Essa definição engloba as mais variadas áreas do conhecimento e é válida, por exemplo, tanto na compreensão de clientes de uma empresa quanto no desenvolvimento de uma nova vacina para alguma doença.

Existem dois tipos de mineração de dados: a direta e a indireta. A mineração de dados direta tenta explicar ou categorizar uma variável alvo definida, como receita proveniente de algum esforço de vendas ou a resposta a uma campanha de marketing. Em geral toma a forma de modelagem preditiva, onde se sabe o que se quer prever. Mineração de dados indireta, por sua vez, procura encontrar padrões ou similaridades entre grupos de registros de uma base de dados sem o uso de um alvo particular ou de alguma coleção de classes pré-definida. Ambas as abordagens não são mutuamente exclusivas e, na verdade, freqüentemente as tarefas de mineração de dados envolvem as duas em conjunto.

Muitas técnicas, algoritmos e tipos de modelos são englobados pelo conceito de mineração de dados, sempre com o objetivo de buscar informação útil em meio a dados. Dado isso, a mineração de dados geralmente envolve uma ou mais de uma das seguintes atividades:

- Classificação
- Estimação
- Predição
- Agrupamento por afinidade ou regras de associação
- Clustering
- Descrição e visualização

Destas, as três primeiras tarefas (classificação, estimação e predição) consistem em exemplos de mineração de dados direta. As outras três são exemplos de atividades da mineração de dados indireta.

Selecionando uma destas atividades, por exemplo classificação, várias famílias de modelos e algoritmos podem ser utilizados na tarefa, dependendo de

peculiaridades do problema estudado e das características dos dados. Entre estes modelos estão, por exemplo, os classificadores bayesianos [DUDA00], as redes neurais [BISH96] [MOZE00] [ZHEN04], as árvores de decisão [QUIN87] [BERR00], os sistemas neuro-fuzzy [CHUN00] [GONÇ01] [SOUZ99] e as máquinas de vetor de suporte [ARCH04] [ZHEN04] [DUDA00]. Tais modelos serão estudados em maior detalhe no capítulo seguinte.

2.5. Interpretação e utilização do conhecimento gerado

Uma etapa essencial da busca do conhecimento em bases de dados é exatamente a interpretação dos resultados e a transformação deles em uma real base para decisões. Após realizado todo o esforço de se coletar dados, prepará-los e minerá-los, os resultados finais apresentados por qualquer modelo ou técnica que tenha sido utilizada requerem uma avaliação cuidadosa sob o prisma da questão a ser resolvida ou do objetivo a ser alcançado. Somente assim o conhecimento gerado realmente se torna útil no apoio à decisão.

2.6. Aplicações em CRM

Em qualquer indústria, empresas estão se movendo na busca de uma compreensão individualizada de seus clientes e usando tal conhecimento para garantir sua satisfação e sua fidelidade a seu negócio. Essas mesmas empresas estão aprendendo a olhar para o valor de cada cliente e entender quais deles são merecedores de dinheiro e esforço para serem mantidos e quais podem ser excluídos. Essa mudança no foco empresarial, de grandes segmentos de mercado para consumidores individuais, requer mudanças significativas em todo o empreendimento, principalmente nas áreas de marketing, vendas e suporte ao cliente.

Construir um negócio voltado para o gerenciamento da relação com o cliente, ou CRM (*Customer Relationship Management*), é um passo revolucionário para grande parte das empresas. Bancos sempre estiveram tradicionalmente interessados em ter certeza que a diferença entre as taxas pelas quais o dinheiro entra e sai eram suficientemente grandes para garantir seus lucros. Empresas de telefonia concentravam seus esforços em garantir a conexão de chamadas na sua rede [MATT01]. Companhias de seguro focavam seu trabalho em processar sinistros e gerenciar investimentos. Mudar o foco de

uma organização de uma visão centrada em produto para uma visão focada no cliente é uma tarefa complicada [KOTL99].

Em qualquer negócio pequeno existe uma estreita relação entre o cliente e a empresa, formando naturalmente um elo de aprendizado. Com o tempo a empresa aprende mais e mais sobre seus consumidores, usando esse conhecimento para servi-los melhor. O resultado dessa relação são clientes satisfeitos, leais e um negócio lucrativo. Em empresas maiores, com centenas de milhares ou milhões de clientes, não existe o luxo do contato pessoal com cada cliente. Essas firmas necessitam então encontrar outros meios para criar uma íntima relação com seus clientes. Em particular, elas precisam aprender a tirar completa vantagem de algo que possuem em abundância: os dados produzidos por praticamente qualquer interação da empresa com cada cliente.

2.6.1. O que se entende por CRM

O gerenciamento da relação com clientes (CRM), está mais em voga do que nunca atualmente. Essa terminologia veio para englobar conceitos como marketing individualizado, customização de ofertas e automatização e direcionamento da força de vendas. Um CRM de qualidade requer o entendimento profundo de quem são os clientes e do que eles gostam ou não. Significa antecipar suas necessidades e anseios, agindo sobre eles de forma pró-ativa. Significa reconhecer quais clientes estão insatisfeitos e tomar alguma atitude antes que eles abandonem a empresa em busca de um concorrente [BERR00] [NOGU04] [ROSS03].

2.6.2. Entendimento do cliente

A descoberta do conhecimento em bases de dados possui um papel fundamental em todas as facetas do CRM. Somente através da aplicação de técnicas de tratamento, mineração e visualização de dados, uma empresa pode pensar em transformar a grande quantidade de informação presente em inúmeras bases de dados [PYLE99], operacionais ou não, no conhecimento inteligível e necessário para compreender seus clientes. O processo de descoberta de conhecimento em bases de dados permite que uma empresa aprenda sobre o comportamento de cada cliente através de todas as

observações presentes em bases de dados que registram cada passo da relação entre consumidor e firma [YAN04].

2.6.3. Geração de inteligência empresarial

Em um empreendimento focado no consumidor, a geração de inteligência empresarial útil sobre o cliente, a partir dos dados disponíveis, é o objetivo final. Em um sentido amplo, isso significa que decisões de negócio devem ser baseadas no aprendizado, que decisões bem informadas são melhores do que decisões sem bases e que medir resultados e utilizar toda a informação disponível sobre o cliente é benéfico para o negócio. De forma a criar uma relação de aprendizado com seus clientes, uma empresa deve ser capaz de:

- Perceber o que o seu cliente faz;
- Lembrar o que ela e seus clientes fizeram ao longo do tempo;
- Aprender a partir do que se lembrou;
- Agir sobre o que aprendeu para tornar os clientes mais lucrativos.

Para tanto, é necessário existir [BERR00] [BOZD03] [KLOS02]: sistemas de processamento de transações capazes de monitorar e capturar todas as interações com os clientes; *data warehouses* que armazenem o histórico de informação comportamental dos clientes; processos de mineração de dados que traduzam o histórico em planos para futuras ações e uma estratégia de relacionamento com o cliente que coloque tudo isso em prática.

2.7. Resumo

Neste capítulo foram apresentados os conceitos básicos que constituem a descoberta de conhecimento em bases de dados, tendo sido dada ênfase especial para suas aplicações em CRM.

O próximo capítulo detalha as etapas descritas aqui, apresentando os métodos e modelos utilizados na busca do conhecimento em bases de dados.