Mineração de Dados

Solange Oliveira Rezende

Resumo

Atualmente, em termos mundiais, o volume de dados armazenado é gigantesco e continua crescendo rapidamente. Infelizmente, devido à incapacidade do ser humano de interpretar tamanha quantidade de dados, muita informação e conhecimento, possivelmente úteis, podem estar sendo desperdiçados, ficando ocultos dentro das Bases de Dados espalhadas pelo mundo. Em conseqüência disso, a necessidade de se desenvolver novas ferramentas e técnicas de extração de conhecimento a partir de dados armazenados também vem crescendo e se mostrando cada vez mais indispensável. Nesse capítulo é apresentado o processo de Mineração de Dados, cujo objetivo é a obtenção de conhecimento útil e interessante, a partir de dados, para a utilização em um processo de tomada de decisão.

1.1. Introdução

O armazenamento de dados em sistemas computacionais já se tornou uma prática comum e essencial nos dias de hoje, principalmente devido à queda no custo do armazenamento dos dados e a rápida automatização das empresas, que faz com que a quantidade de dados armazenados em bases de dados cresça a uma velocidade muito alta. Essas grandes quantidades de dados, armazenadas em Bases de Dados, *Data Warehouses* e outros tipos de repositórios de dados, de maneira centralizada ou distribuída, existem em muitos domínios, com financeiro, médico, produção e manufatura, comercial e científico.

O grande volume de dados armazenados, principalmente em empresas e instituições acadêmicas, vem sendo muito valorizado e analisado, pois muitas informações realmente novas e interessantes estão "embutidas" nessas Bases de Dados, como perfis de clientes no uso de cartão de crédito (que podem ser usados para combater fraudes), padrões de pacientes que desenvolveram doenças (que podem ser úteis na tentativa de prever diagnósticos e antecipar tratamentos), perfis de compra de clientes (para usar em futuras grandes promoções). Conforme citado em T. Y. Lin (1997), as Bases de Dados das grandes empresas contêm uma potencial mina de ouro de informações valiosas, porém, de acordo com S. Mitra (2002), estes dados raramente são obtidos de forma direta. Usualmente, estas informações não estão disponíveis devido à falta de ferramentas apropriadas

para a sua extração; está além da capacidade do ser humano analisar tamanha quantidade de dados e extrair relações significativas entre eles.

A área de Mineração de Dados (Data Mining) surgiu no final da década de oitenta, e focaliza a extração de conhecimento a partir de grandes volumes de dados usando computador. Devido à sua natureza interdisciplinar, a pesquisa e desenvolvimento da área de Mineração de Dados têm estreitas relações com as contribuições oferecidas por diversas áreas como Banco de Dados, Aprendizado de Máquina, Estatística, Recuperação de Informação, Computação Paralela e Distribuída. Como apontado em Zhou (2003), as áreas de Banco de Dados, com poderosas técnicas de gerenciamento de dados, Aprendizado de Máquina, com técnicas práticas de análise de dados e a Estatística, com uma sólida fundamentação teórica, são as áreas de conhecimento e pesquisa que estão contribuindo mais efetivamente para o desenvolvimento e o estabelecimento da área de Mineração de Dados.

No artigo Zhou (2003), o autor analisa as perspectivas destas três áreas, Banco de Dados, Aprendizado de Máquina e Estatística, e enfatiza os diferentes aspectos de Mineração de Dados abordados por cada uma. De acordo com Zhou (2003), a perspectiva de Banco de Dados enfatiza a eficiência, uma vez que focaliza o processo de descoberta como um todo, em um volume de dados imenso. A perspectiva de Aprendizado de Máquina focaliza a efetividade, dado que essa perspectiva é fortemente influenciada por heurísticas efetivas para a análise de dados. A perspectiva da Estatística focaliza validade, dado que enfatiza o rigor matemático que subsidia os métodos de mineração.

Como não poderia deixar de ser, dadas as diferentes perspectivas com as quais a área de Mineração de Dados pode ser abordada, na literatura podem ser encontradas diversas caracterizações da área. No artigo Zhou (2003), o autor evidencia a caracterização da área sob as perspectivas tratadas em três livros sobre Mineração de Dados avaliados por ele, sendo um de cada uma das três áreas. Sob a perspectiva da área de Banco de Dados, citada em J. Han (2001), a Mineração de Dados é "o processo de descoberta de conhecimento interessante em grandes quantidades de dados armazenados em Bases de Dados, *Data Warehouses* ou outros repositórios de dados"; sob a perspectiva da área de Aprendizado de Máquina, conforme apontada em Witten & Frank (1999), é caracterizada como a "extração de conhecimento implícito, previamente desconhecido e potencialmente útil a partir de dados"; e sob a perspectiva da área de Estatística, conforme citado em D. Hand (2001), é "a análise de conjuntos de dados supervisionados, normalmente em grandes quantidades, para encontrar relacionamentos inesperados e resumir os dados em novas formas que são compreensíveis e úteis para o proprietário dos dados".

Na caracterização da área de Mineração de Dados é importante, também, discutir a caracterização do que na literatura é chamado de KDD (*Knowledge Discovery in Databases*). De acordo com W. Frawley (1992), KDD é a "extração de conhecimento previamente desconhecida, implícita e potencialmente útil, a partir de dados". Na literatura existente atualmente as opiniões divergem a respeito dos termos Mineração de Dados e KDD. Existem autores que consideram os termos sinônimos (Fayyad, Piatetsky-Shapiro, & Smyth 1996b; Mitchell 1999; Wei 2003) enquanto outros consideram a Mineração de Dados apenas um dos passos do processo de KDD, embora seja o passo principal de todo o processo (S. Mitra 2002; I. Sarafis 2002).

Mineração de Dados é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como Inteligência Artificial, especialmente Aprendizado de Máquina (AM) (Monard & Baranauskas 2003), Base de Dados e Estatística. Por isso, as técnicas utilizadas em MD não devem ser vistas como substitutas de outras formas de análises (por

exemplo, OLAP), mas, como práticas para melhorar os resultados das explorações feitas com as ferramentas atualmente utilizadas.

O foco central de Mineração de Dados é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados sem o uso de qualquer técnica, entretanto, existe também muito conhecimento que está de certa forma "embutido" na Base de Dados, na forma de relações existentes entre itens de dados que, para ser extraído, é necessário o desenvolvimento de técnicas especiais. Assim na próxima seção são definidos os conceitos básicos dados, informação e conhecimento para em seguida abordar o processo de Mineração de Dados como um todo.

1.2. Dados, Informação e Conhecimento

Os conceitos de dados, informação e conhecimento estão interligados. Na Figura 1.1 é mostrada uma representação gráfica do relacionamento entre dados, informação e conhecimento, em função da capacidade de entendimento e da independência de contexto que cada um destes conceitos implica (Kock Jr., McQueen, & Baker 1996; Kock Jr., McQueen, & Corner 1997).

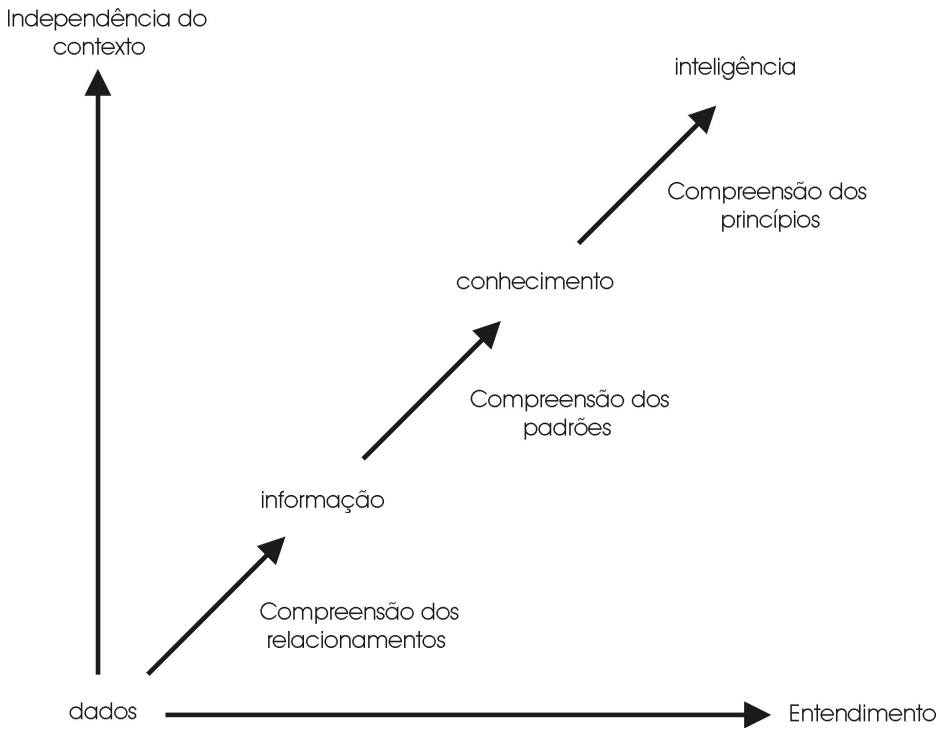

Figura 1.1: Dado, informação e conhecimento (Kock Jr., McQueen, & Baker 1996)

Antes de estabelecer qualquer ligação desses conceitos com as diferentes tecnologias para seu registro e processamento, faz-se necessária uma breve discussão sobre a distinção entre dado, informação e conhecimento.

O dado é um elemento puro, quantificável sobre um determinado evento. Dados são fatos, números, texto ou qualquer mídia que possa ser processada pelo computador. Hoje em dia, as organizações estão acumulando vastas e crescentes quantidades de dados em diferentes formatos e em diferentes tipos de repositórios. Ressalta-se que o dado, por si só, não oferece embasamento para o entendimento da situação. Entre os dados armazenados atualmente estão:

- dados operacionais ou transacionais como vendas, custos, inventários, folhas de pagamento ou contas correntes;
- dados não operacionais como previsões de mercado, vendas ao nível industrial, e dados macro-econômicos;
- metadados, ou dados descrevendo a estrutura dos dados armazenados, tais como projetos lógicos de Bancos de Dados ou dicionários de dados;
- mídia contento imagens, sons ou uma combinatória de ambos, que além de ser armazenada, pode ser catalogada eletronicamente.

A informação é o dado analisado e contextualizado. Envolve a interpretação de um conjunto de dados, ou seja, a informação é constituída por padrões, associações ou relações que todos aqueles dados acumulados podem proporcionar. Por exemplo, a análise de pontos de equilíbrio no mercado pode fornecer informação acerca de quais produtos estão sendo vendidos e a freqüência de tais operações.

A informação pode gerar conhecimento que ajude na análise de padrões históricos para conseguir uma previsão dos fatos futuros (pelo menos no contexto das variáveis que estão sendo envolvidas na análise). Por exemplo, a informação dos dados sumarizados nas vendas de um determinado ambiente comercial pode ser analisada com a finalidade de fornecer informações relacionadas com a natureza dos clientes.

Enquanto que a informação é descritiva, o conhecimento é utilizado fundamentalmente para fornecer uma base de previsão com um determinado grau de certeza. O conhecimento refere-se à habilidade de criar um modelo mental que descreva o objeto e indique as ações a implementar e as decisões a tomar. Uma decisão é o uso explícito de um conhecimento. O conhecimento pode ser representado como uma combinação de estruturas de dados e procedimentos interpretáveis que levam a um comportamento conhecido. Este comportamento fornece informações que podem, então, serem utilizadas para planejar e decidir.

A compreensão, análise e síntese, necessárias para a tomada de decisões inteligentes, são realizadas a partir do nível do conhecimento. Assim, é fundamental que se mantenha a coerência dos dados que estão armazenados nos diferentes repositórios e das informações nos diferentes níveis.

Assim, nesse contexto, o desafio dos anos de 1980 foi migrar os dados para as informações, por meio do desenvolvimento dos Sistemas de Informação, que tinham por finalidade analisar dados e organizar a informação para melhorar o processo decisório empresarial. A partir da década de 1990, o desafio era criar sistemas que fossem capazes de representar e processar conhecimento, em resposta às diferentes necessidades de indivíduos, grupos e culturas.

1.2.1. Algumas Áreas que Trabalham com Dados, Informação e Conhecimento

Atualmente, os analistas de negócios precisam usar ferramentas capazes de responder a perguntas complexas como: “qual produto de alta lucratividade venderia mais com a promoção de um item de baixa lucratividade, analisando os dados dos últimos dez anos de vendas?”. Esse tipo de informação pode ser fundamental para oferecer vantagens competitivas às empresas. Dessa maneira, ferramentas que apoiam o processo de obtenção e análise das informações e extração de conhecimento devem ser usadas no processo decisório.

Gestão do Conhecimento

A sociedade do conhecimento está impondo uma competitividade cada vez maior entre países e entre empresas, o que leva a uma necessidade de mudança e reflexão contínuas. É preciso inovar e adquirir sucessivamente novos conhecimentos (Cavalcanti, Gomes, & Pereira 2001).

A gestão do conhecimento *(Knowledge Management)* tem o objetivo de estabelecer meios, de maneira integrada e colaborativa, para capturar, criar, organizar e usar todos os ativos de informação de uma corporação.

A gestão do conhecimento é o primeiro passo na criação de uma estrutura lógica para manipular as informações que uma determinada entidade possui e gerenciar tanto as entradas quanto os resultados da mesma. Em outras palavras, a gestão do conhecimento é responsável pela recuperação e organização das práticas, documentos, políticas, experiências de funcionários, entre outras fontes, de onde é possível obter conhecimento explícito de uma organização.

O foco principal é o conhecimento organizacional que é inerente a todas as empresas e é definido como "a capacidade de executar coletivamente tarefas que as pessoas não conseguem fazer atuando de forma isolada, tarefas essas projetadas para criar valor para as partes interessas na organização" (Garvin, Nayak, Maira, & Bragar 1998). Há uma diferença entre o conhecimento estar embutido em estruturas, regras e processos de trabalho em grupo – conhecimento explícito – e estar embutido em trabalhos individuais – conhecimento tácito. Garvin, Nayak, Maira, & Bragar (1998) acreditam que o conhecimento organizacional deve ser explícito e tácito, pois o conhecimento tácito, que inclui o discernimento , o instinto e a compreensão individual, é fundamental para tornar o conhecimento explícito útil.

A utilização da Tecnologia de Informação (TI) para a gestão do conhecimento tem seus primórdios nos anos 70, quando esta passa de um foco computacional voltado ao processamento de dados para um foco mais voltado ao processamento da informação, como nos sistema de suporte a decisão gerencial (DSS - *Decision Support System*) e nos sistemas de informação gerencial (MIS - *Management Information System*).

Nos anos 80, o processamento do conhecimento passa a estar cada vez mais presente nos recursos oferecidos, com os sistemas baseados em conhecimento (KBS's - *Knowledge-based Systems*). Talvez a face mais visível e conhecida desses sistemas seja os sistemas especialistas (*Expert Systems*).

Atualmente os DSS e os MIS desdobram-se em inúmeras linhas de atuação, desde a utilização de sistemas baseados em inteligência artificial e modelos matemáticos e estatísticos, passando pela criação de conhecimento a partir de dados e informações presentes em Bases de Dados (Mineração de Dados e *Data Warehousing*), até a representação do conhecimento em sistemas especialistas e redes neurais que procuram automatizar a tomada de decisões.

Entretanto, é necessário destacar que antes de investimentos ou decisões em TI, a empresa já deveria saber identificar, desdobrar e atribuir o devido valor aos conhecimentos disponíveis, respeitar e motivar o trabalho em rede dos grupos voluntariamente formados dentro da organização e destes com os parceiros da empresa, e saber aprender com a experiência (no sentido de refletir e aprender a incorporar as mudanças e atualizações, conseqüência da natureza dinâmica do conhecimento).

Além disso, a organização precisa estar ciente de que muitos dos recursos da TI requerem significativos e continuados esforços para se construir, manter e atualizar, tanto

em seus aspectos funcionais quanto nos conteúdos envolvidos, bem como é necessário conquistar a aceitação e confiança do usuário corporativo.

Inteligência de Negócios

Num mercado cada vez mais concorrido e exigente é grande a demanda por soluções capazes de oferecer vantagens competitivas às empresas. Por isso coorporações dos mais diversos setores buscam ferramentas estratégicas para:

- endenter melhor o nincho de atuação no mercado;
- promover melhoramentos na competência essencial da empresa;
- identificar oportunidades;
- responder adequada e eficientemente às mudanças do mercado;
- melhorar o relacionamento com clientes e fornecedores;
- reduzir custos operacionais.

O conceito de Inteligência de Negócios (*Business Intelligence* (BI)), de forma mais ampla, pode ser entendido como a utilização de variadas fontes de informação para definir estratégias de competitividade nos negócios.

Existe uma grande problemática ao nível empresarial e de mercado: uma grande quantidade de dados está disponível, provocando muitas dificuldades na extração de informações a partir deles e a enorme quantidade de informações dificulta o processo de tomada de decisão na medida em que a gerência se sente impotente no processo de recuperação e análise.

As informações vitais para tomadas de decisões estratégicas estão escondidas em milhares de tabelas e arquivos, ligadas por relacionamentos de correlações transacionais, em uma organização inadequada para o estabelecimento de decisões.

O objetivo maior das técnicas de BI, neste contexto, está exatamente na definição de regras e técnicas para a formatação adequada destes volumes de dados, com a finalidade de transformá-los em depósitos estruturados de informações, independentemente da sua origem.

O grande desafio das empresas hoje, não é apenas organizar seus Sistemas de Gestão do Conhecimento, mas estabelecer uma ponte entre ela e a Inteligência de Negócios. Enquanto o BI transforma dados em informação, produzindo as visões *(views)*, relatórios, etc., a Gestão do Conhecimento realiza as devidas combinações, compilações, subscrições e a distribuição, inserindo pontos de discussão, com o objetivo de transformar informações em conhecimento.

Data Warehousing

Data Warehousing é um processo, não um produto, para montar e gerenciar repositórios de dados a partir de várias fontes com o propósito de ter uma visão detalhada e singular de parte ou do todo de um negócio. O produto principal obtido de um projeto de *Data Warehousing* é o seu *Data Warehouse (DW).*

A realização de *Data Warehousing* (Gardner 1998) é considerada um dos primeiros passos para tornar factível a análise de grande quantidade de dados no apoio ao processo decisório. O objetivo básico é criar um repositório, conhecido por *Data Warehouse,* que contenha dados limpos, agregados e consolidados podendo ser analisados por ferramentas OLAP (*On-Line Analytical Processing*). Tais ferramentas apresentam facilidades para a realização de consultas complexas em Bases de Dados multidimensionais.

Como citado, o processo de construção, acesso e manutenção de um *Data Warehouse* (DW) é denominado de *Data Warehousing*. Este processo objetiva integrar e gerenciar dados extraídos de diversas fontes, com o propósito de ganhar uma visão detalhada de parte ou do todo de um negócio (Chaudhuri & Dayal 1997; Gupta 1997).

Um *Data Warehouse* é, então, um Banco de Dados cuja função é proporcionar aos seus usuários uma única fonte de informação a respeito dos seus negócios, servindo também, como ferramenta de apoio ao processo de extração de conhecimento. É uma coleção de dados orientado por assuntos, integrado, variante no tempo e não volátil, que tem como objetivo básico satisfazer as necessidades dos usuários quanto ao armazenamento dos dados que servirão para realizar consultas e análises necessárias para o apoio à tomada de decisão (Inmon 1997; Poe, Klauber, & Brobst 1998). O termo orientado por assuntos está relacionado ao conceito de *Data Mart*, ou seja, um *Data Mart* implementa um determinado assunto, enquanto um *Data Warehouse* implementa e integra vários assuntos de um organização.

Uma seqüência de passos, que pode servir de guia para o projeto de repositórios de dados (Barquini 1996; Kimball 1997; Inmon 1997; Poe, Klauber, & Brobst 1998). A metodologia é composta por seis fases, onde as cinco últimas devem se repetir para cada nova área de negócio a ser considerada no projeto. A primeira fase corresponde a justificativa de um projeto de DW. O objetivo é procurar identificar quais as vantagens que um sistema de DW trará à organização. Caso o projeto se justifique, então deve tomar lugar a fase de planejamento. Essa fase objetiva a definição de uma visão corporativa do DW, definição da arquitetura e topologia do sistema, e definição da área de negócio a ser enfatizada. Em seguida vem a fase de análise, cuja tarefa principal é a modelagem conceitual dos dados para uma área de negócio selecionada. A próxima fase corresponde ao projeto do sistema, sendo que a mesma enfatiza o detalhamento dos resultados adquiridos na fase de análise, bem como, o projeto das consultas OLAP. A quinta fase é a implementação, na qual são criados os objetos físicos, é feito o povoamento do *Data Warehouse* e a implementação das consultas OLAP. A última fase corresponde à revisão, onde são verificados os resultados obtidos com a implementação do sistema. Um documento final deve registrar todo conhecimento obtido durante o projeto da área de negócio selecionada, servindo como modelo para as próximas iterações do projeto. Um outro tópico que deve ser considerado no projeto de um *Data Warehouse* é a sua granularidade, a qual se refere ao nível de detalhe na qual as unidades de dados são mantidas (Barquini 1996; Gardner 1998).

As técnicas de análise de dados do *Data Warehouse* geralmente não extrapolam a realização de consultas SQL (*Structured Query Language*) simples, a utilização de ferramentas OLAP ou os mecanismos de Visualização de Dados. Por meio desta forma de análise de dados, algumas questões importantes para tomada de decisão não podem ser expressas, como:

- Quais são os usuários potenciais para praticar fraude?
- Quais clientes gostariam de comprar o novo produto X?

As ferramentas utilizadas para analisar um *Data Warehouse*, normalmente, são orientadas às consultas, ou seja, são dirigidas pelos usuários, os quais possuem hipóteses que gostariam de comprovar, ou simplesmente, executam consultas aleatórias.

Essa abordagem dependente do usuário pode impedir que padrões escondidos nos dados sejam encontrados de forma “inteligente”, uma vez que o usuário não terá condições de imaginar todas as possíveis relações e associações existentes em um grande volume de dados. Por isso, faz-se necessária a utilização de técnicas de análise dirigidas por computador que possibilitem a extração automática (ou semi-automática) de novos conhecimentos a partir de um grande repositório de dados (Bradley, Fayyad, & Man-

gasarian 1998). A extração automática de conhecimento a partir de dados, denominada Mineração de Dados, será abordada na Seção 1.3.

Apesar de não ser obrigatória, a construção de um *Data Warehouse* pode reduzir drasticamente a complexidade e a duração do processo de Mineração de Dados. Vale ressaltar a diferença entre os resultados das consultas OLAP e da Mineração de Dados. As consultas OLAP geram informações obtidas a partir do DW enquanto conhecimento pode ser extraído utilizando as técnicas de MD aplicadas nos dados do DW ou diretamente nos dados da Base de Dados, como mostrado na Figura 1.2.

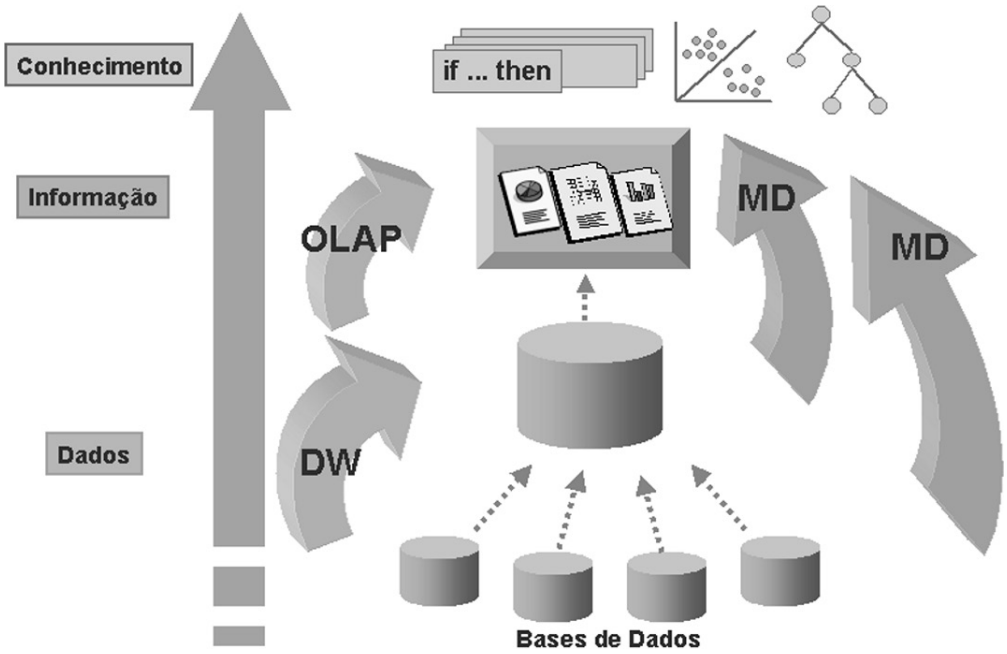

Figura 1.2: Relação entre Base de Dados, *Data Warehouse* e Mineração de Dados

A obtenção do conhecimento é um processo que envolve o uso de múltiplas técnicas e métodos que evoluiram à medida que os requisitos de informação tornaram-se prioridade dos ambientes de negócios. Enquanto as tecnologias de processamento da informação evoluíram separadamente dos sistemas analíticos, o processo de Mineração de Dados pode fornecer um elo entre ambos, apoiando a Gestão do Conhecimento e Inteligência de Negócios.

O objetivo do processo de Mineração de Dados é a extração do conhecimento implícito por meio da descoberta de padrões e da criação de modelos de maneira automática a partir dos dados.

1.3. Mineração de Dados

A área de Inteligência Artificial têm propiciado aos pesquisadores a possibilidade de utilizar diferentes técnicas para o reconhecimento/extração de padrões. Essa extração está acompanhada de técnicas de manipulação de dados e de análises posteriores. Todo esse conjunto de diferentes técnicas está encaixado em um processo denominado Mineração de Dados.

A área de Visualização de Informação, por sua vez, têm contribuído com técnicas de visualização que permitem observar as informações em diferentes níveis e estruturas. Com o apoio dessas técnicas nasceu uma outra técnica, denominada Mineração Visual de Dados. A Mineração Visual de Dados é a combinação entre a Visualização de Informação com os padrões extraídos da Mineração de Dados.

1.3.1. Definição

A definição de Mineração de Dados aceita por diversos pesquisadores foi elaborada por Fayyad, Piatetsky-Shapiro, & Smyth (1996a) como sendo: "Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados". Para compreender melhor o conteúdo dessa definição, deve-se olhar individualmente cada componente da mesma:

Dados Conjunto de fatos ou casos em um repositório de dados. Por exemplo, os dados correspondem aos valores dos campos de um registro de vendas em uma Base de Dados qualquer;

Padrões Denota alguma abstração de um subconjunto dos dados em alguma linguagem descritiva de conceitos;

Processo A Extração de Conhecimento de Base de Dados envolve diversas etapas como a preparação dos dados, busca por padrões e avaliação do conhecimento;

Válidos Os padrões descobertos devem possuir algum grau de certeza, ou seja, devem satisfazer funções ou limiares que garantam que os exemplos cobertos e os casos relacionados ao padrão encontrado sejam aceitáveis;

Novos Um padrão encontrado deve fornecer novas informações sobre os dados. O grau de novidade serve para determinar quão novo ou inédito é um padrão. Pode ser medido por meio de comparações entre as mudanças ocorridas nos dados ou no conhecimento anterior;

Úteis Os padrões descobertos devem ser incorporados para serem utilizados;

Compreensíveis Um dos objetivos de realizar MD é encontrar padrões descritos em alguma linguagem que pode ser compreendida pelos usuários permitindo uma análise mais profunda dos dados;

Conhecimento O conhecimento é definido em termos dependentes do domínio, relacionados fortemente com medidas de utilidade, originalidade e compreensão.

O processo de Extração de Conhecimento de Bases de Dados tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório. Portanto, um requisito importante é que esse conhecimento descoberto seja compreensível a humanos, além de útil e interessante para os usuários finais do processo, que geralmente são tomadores de decisão, de forma que ele forneça um suporte a esses usuários no processo de decisório (Fayyad, Piatetsky-Shapiro, & Smyth 1996a; Freitas 1998b).

Todo o processo de MD é orientado em função de seu domínio de aplicação e dos repositórios de dados inerentes aos mesmos. Para usar os dados é necessário que estes estejam estruturados de forma a serem consultados e analisados adequadamente.

Os sistemas de aplicações, conhecidos por OLTP (*On-Line Transaction Processing*), processam dados armazenados em Base de Dados relacionais usadas para armazenar, consultar e alterar informações do negócio. Normalmente, não é possível aplicar as técnicas de MD diretamente a estas bases, pois isso poderia resultar numa sobrecarga de consultas nas mesmas podendo literalmente "travar" um sistema, impossibilitando qualquer outro tipo de operação transacional.

Assim, é recomendável que os dados a serem utilizados na descoberta de conhecimento estejam separados da Base de Dados operacional. Nesses casos, utilização de DW é indicada, pois possibilita o armazenamento de grande quantidade de dados históricos de uma organização e viabiliza acessos otimizados aos dados previamente consolidados, reduzindo consideravelmente o tempo de realização do processo de MD.

1.3.2. O Processo de Mineração de Dados

A extração de conhecimento a partir de grande quantidade de dados é vista como um processo interativo e iterativo, e não como um sistema de análise automática, sendo centrado na interação entre usuários, especialistas do domínio e responsáveis pela aplicação. Dessa maneira, não se pode esperar que a extração de conhecimento seja útil simplesmente submetendo um conjunto de dados a uma "caixa preta" (Mannila 1997).

Existem diversas abordagens para a divisão das etapas do processo de Extração de Conhecimento de Bases de Dados. Inicialmente, foi proposto por Fayyad, Piatetsky-Shapiro, & Smyth (1996c) uma divisão do processo em nove etapas. Já em Weiss & Indurkhya (1998), essa divisão é composta por apenas quatro etapas. No entanto, neste capítulo é considerada a divisão do processo em três grandes etapas — pré-processamento, extração de padrões e pós-processamento. São incluídas nessa divisão uma fase anterior ao processo de Mineração de Dados, que se refere ao conhecimento do domínio e identificação do problema, e uma fase posterior ao processo, que se refere à utilização do conhecimento obtido. A Figura 1.3 ilustra essas etapas.

Assim, o processo de Extração de Conhecimento de Bases de Dados, ou Mineração de Dados, inicia-se com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados (Base de Dados da qual se pretende extrair conhecimento). Em seguida, é realizada uma seleção de dados a partir dessas fontes, de acordo com os objetivos do processo. Os conjuntos de dados resultantes dessa seleção são, então, pré-processados, ou seja, recebem um tratamento para poderem ser submetidos aos métodos e ferramentas na etapa de extração de padrões.

A etapa de extração de padrões tem o objetivo de encontrar modelos (conhecimento) a partir de dados. Normalmente, como resultado dessa etapa, tem-se preditores que, para um dado problema, fornecem uma decisão. Após a etapa de extração de padrões, vem a de pós-processamento, na qual o conhecimento é avaliado quanto a sua qualidade e/ou utilidade para que, em caso positivo, seja utilizado para apoio a algum processo de tomada de decisão.

É importante notar que, por ser um processo eminentemente iterativo, as etapas da Mineração de Dados não são estanques, ou seja, a correlação entre as técnicas e métodos utilizados nas várias etapas é considerável, a ponto da ocorrência de pequenas mudanças em uma delas afetar substancialmente o sucesso de todo o processo. Portanto, os resultados de uma determinada etapa podem acarretar mudanças a quaisquer das etapas posteriores ou, ainda, o recomeço de todo o processo (Fayyad, Piatetsky-Shapiro, & Smyth 1996b).

O processo de MD é centrado na interação entre as diversas classes de usuários, e o seu sucesso depende, em parte, dessa interação. Os usuários do processo podem ser divididos em três classes: especialista do domínio que deve possuir amplo conhecimento do domínio da aplicação e deve fornecer apoio para a execução do processo; analista que é o usuário especialista no processo de Extração de Conhecimento e responsável por sua execução devendo conhecer profundamente as etapas que compõem o processo e; usuário final que representa a classe de usuários que utiliza o conhecimento extraído no processo para auxiliá-lo em um processo de tomada de decisão.

É importante ressaltar que pode haver situações em que o especialista do domínio também é o usuário final, ou que este auxilie ou execute funções pertinentes ao analista. Entretanto, é pouco provável que o analista encontre conhecimento útil a partir dos dados sem a opinião do especialista sobre o que é considerado interessante em um domínio específico.

Figura 1.3: Etapas do processo de Mineração de Dados(Rezende, Pugliesi, Melanda, & Paula 2003)

Identificação do Problema

O estudo do domínio da aplicação e a definição de objetivos e metas a serem alcançadas no processo de Mineração de Dados são identificados nesta fase.

O sucesso do processo de Extração de Conhecimento depende, em parte, da participação dos especialistas do domínio da aplicação no fornecimento de conhecimento sobre o domínio e apoio aos analistas em sua tarefa de encontrar os padrões. Assim, antes do início das tarefas do processo é imprescindível a realização de um estudo a fim de adquirir um conhecimento inicial do domínio (Fayyad, Piatetsky-Shapiro, & Smyth 1996b).

Algumas questões importantes devem ser respondidas nessa fase de identificação do problema, como:

- Quais são as principais metas do processo?
- Quais critérios de desempenho são importantes?
- O conhecimento extraído deve ser compreensível a seres humanos ou um modelo do tipo caixa-preta é apropriado?
- Qual deve ser a relação entre simplicidade e precisão do conhecimento extraído?

Além dessa análise inicial para definição das principais metas, objetivos e restrições, o conhecimento sobre o domínio fornece um subsídio para todas as etapas do processo de Extração de Conhecimento. Mais especificamente, na etapa de pré-processamento, esse conhecimento pode ajudar os analistas na escolha do melhor conjunto de dados para se realizar a extração de padrões, saber quais valores são válidos para os atributos, os critérios de preferência entre os possíveis atributos, as restrições de relacionamento ou informações para geração de novos atributos.

Na etapa de extração de padrões, o conhecimento sobre o domínio pode ajudar os analistas na escolha de um critério de preferência entre os modelos gerados, no ajuste dos parâmetros do processo de indução, ou mesmo na geração de um conhecimento inicial a ser fornecido como entrada do algoritmo de mineração para aumentar a eficiência no aprendizado dos conceitos e melhorar a precisão ou a compreensibilidade do modelo final.

Na etapa de pós-processamento, o conhecimento extraído pelos algoritmos de extração de padrões deve ser avaliado. Alguns critérios de avaliação utilizam o conhecimento do especialista para saber, por exemplo, se o conhecimento extraído é interessante ao usuário (Piatetsky-Shapiro & Matheus 1994; Liu & Hsu 1996).

Entender o domínio dos dados é naturalmente um pré-requisito para extrair algo útil: o usuário final do sistema deve ter algum grau de entendimento sobre a área de aplicação antes de qualquer informação valiosa ser obtida. Por outro lado, se existem especialistas com profundo conhecimento sobre domínio, a obtenção de novas informações com o uso de ferramentas semi-automáticas é difícil. Esse pode ser o caso em domínios bastante estáveis, no qual os seres humanos tiveram tempo para adquirir o conhecimento especializado em detalhes. Um exemplo ocorre em áreas de venda em que os produtos e os clientes são os mesmos por um longo período de tempo. A Mineração de Dados parece ter melhores resultados em áreas nas quais as propriedades reais dos dados mudam, mas existem especialistas com conhecimentos genéricos abrangentes. Esse é o caso da área de telecomunicações, na qual os operadores das redes têm uma idéia geral das características dos sistemas, mas mudanças em equipamentos e softwares implicam que especialistas em detalhes dos dados sejam difíceis de serem encontrados.

Pré-Processamento

Normalmente, os dados disponíveis para análise não estão em um formato adequado para a Extração de Conhecimento. Além disso, em razão de limitações de memória ou tempo de processamento, muitas vezes não é possível a aplicação direta dos algoritmos de extração de padrões aos dados. Dessa maneira, torna-se necessária a aplicação de métodos para tratamento, limpeza e redução do volume de dados antes de iniciar a etapa de Extração de Padrões. É importante salientar que a execução das transformações deve ser guiada pelos objetivos do processo de extração a fim de que o conjunto de dados gerado apresente as características necessárias para que os objetivos sejam cumpridos.

Diversas transformações nos dados podem ser executadas na etapa de pré-processamento dos dados, entre elas: extração e integração, transformação, limpeza, seleção e redução de dados.

Na extração e integração os dados disponíveis podem ser encontrados em diferentes fontes, como arquivos-texto, arquivos no formato de planilhas, Banco de Dados ou *Data Warehouse*. Assim, é necessária a obtenção desses dados e sua unificação, formando uma única fonte de dados no formato atributo-valor que será utilizada como entrada para o algoritmo de extração de padrões.

Após a extração e integração dos dados, estes devem ser adequados para serem utilizados nos algoritmos de extração de padrões. Algumas transformações comuns que podem ser aplicadas aos dados são: resumo, por exemplo, quando dados sobre vendas são agrupados para formar resumos diários; transformação de tipo, por exemplo, quando um atributo do tipo data é transformado em um outro tipo para que o algoritmo de extração de padrões possa utilizá-lo mais adequadamente; normalização de atributos contínuos, colocando seus valores em intervalos definidos, por exemplo, entre 0 e 1. As transformações de dados são extremamente importantes em alguns domínios, por exemplo, em aplicações que envolvem séries temporais como predições no mercado financeiro.

Os dados disponíveis para aplicação dos algoritmos de extração de padrões podem apresentar problemas advindos do processo de coleta. Estes problemas podem ser erros de digitação ou erro na leitura dos dados pelos sensores. Como o resultado do processo de extração possivelmente será utilizado em um processo de tomada de decisão, a qualidade

dos dados é um fator extremamente importante. Por isso, técnicas de limpeza devem ser aplicadas aos dados a fim de garantir sua qualidade.

A limpeza dos dados pode ser realizada utilizando o conhecimento do domínio. Por exemplo, pode-se encontrar registros com valor inválido em algum atributo, granularidade incorreta ou exemplos errôneos. Pode-se também efetuar alguma limpeza independente de domínio, como decisão da estratégia de tratamento de atributos incompletos, remoção de ruído e tratamento de conjunto de exemplos não balanceados (Batista, Carvalho, & Monard 2000).

Em virtude das restrições de espaço em memória ou tempo de processamento, o número de exemplos e de atributos disponíveis para análise pode inviabilizar a utilização de algoritmos de extração de padrões. Como solução para este problema, pode ser necessária a aplicação de métodos para redução dos dados antes de iniciar a busca pelos padrões. Esta redução pode ser feita de três maneiras: reduzindo o número de exemplos, de atributos e de valores de um atributo (Weiss & Indurkhya 1998).

A redução do número de exemplos deve ser feita a fim de manter as características do conjunto de dados original, isto é, por meio da geração de amostras representativas dos dados (Glymour, Madigan, Pregibon, & Smyth 1997). A abordagem mais utilizada para redução do número de exemplos é a amostragem aleatória (Weiss & Indurkhya 1998), pois este método tende a produzir amostras representativas.

Se a amostra não for representativa, ou se a quantidade de exemplos for insuficiente para caracterizar os padrões embutidos nos dados, os modelos encontrados podem não representar a realidade, não tendo, portanto, valor. Além disso, com uma quantidade relativamente pequena de exemplos, pode ocorrer *overfitting*, isto é, o modelo gerado pode "decorar" os dados do conjunto de treinamento, não se adequando para utilização com novos exemplos (Fayyad, Piatetsky-Shapiro, & Smyth 1996c).

A redução do número de atributos pode ser utilizada para reduzir o espaço de busca pela solução. O objetivo é selecionar um subconjunto dos atributos existentes de forma que isto não tenha grande impacto na qualidade da solução final. Esta redução pode ser realizada com o apoio do especialista do domínio, uma vez que, ao remover um atributo potencialmente útil para o modelo final, a qualidade do conhecimento extraído pode diminuir consideravelmente. Além disso, por não se saber inicialmente quais atributos serão importantes para atingir os objetivos, deve-se remover somente aqueles atributos que, com certeza, não têm nenhuma importância para o modelo final.

Outra maneira de reduzir o número de atributos é por meio da indução construtiva, em que um novo atributo é criado a partir do valor de outros. Caso os atributos originais utilizados na construção do novo atributo não estejam presentes em um novo modelo, eles podem ser descartados, reduzindo assim o número de atributos. A utilização de indução construtiva pode aumentar consideravelmente a qualidade do conhecimento extraído (Lee 2000).

A terceira forma de redução dos dados consiste na redução do número de valores de um atributo. Isso é feito, geralmente, por discretização ou suavização dos valores de um atributo contínuo.

Discretização de um atributo consiste na substituição de um atributo contínuo (inteiro ou real) por um atributo discreto, por meio do agrupamento de seus valores. Essencialmente, um algoritmo de discretização aceita como entrada os valores de um atributo contínuo e gera como saída uma pequena lista de intervalos ordenados. Cada intervalo é representado na forma $[V_{inferior} : V_{superior}]$, de modo que $V_{inferior}$ e $V_{superior}$ são, respectivamente, os limites inferior e superior do intervalo. Os métodos de discretização

podem ser classificados em supervisionados ou não-supervisionados, locais ou globais, e parametrizados ou não-parametrizados (Félix, Rezende, Monard, & Caulkins 2000).

Na suavização dos valores de um atributo, o objetivo é diminuir o número de valores do mesmo sem discretizá-lo. Nesse método, os valores de um determinado atributo são agrupados, mas, ao contrário da discretização, cada grupo de valores é substituído por um valor numérico que o represente. Esse novo valor pode ser a média, a mediana ou mesmo os valores de borda de cada grupo (Weiss & Indurkhya 1998).

As transformações descritas devem ser realizadas criteriosamente e com o devido cuidado, uma vez que é fundamental garantir que as informações presentes nos dados brutos continuem presentes nas amostras geradas, para que os modelos finais sejam representativos da realidade expressa nos dados brutos.

A etapa de pré-processamento é realizada antes da extração de padrões. Mas, como Mineração de Dados é um processo iterativo, algumas atividades de pré-processamento podem ser realizadas novamente após a análise dos padrões encontrados na etapa de extração de padrões. Por exemplo, pode-se desejar acrescentar algum atributo, reduzir ou aumentar o volume de dados, fazer uma outra transformação no tipo de algum atributo, para que o indutor possa utilizá-lo de forma mais eficiente, e melhorar a qualidade do conhecimento extraído.

Extração de Padrões

A etapa de extração de padrões é direcionada ao cumprimento dos objetivos definidos na Identificação do Problema. Assim como todo o processo de MD, essa etapa também é um processo iterativo e pode ser necessário sua execução diversas vezes para ajustar o conjunto de parâmetros visando à obtenção de resultados mais adequados aos objetivos preestabelecidos. Ajustes podem ser necessários, por exemplo, para a melhoria da precisão ou da compreensibilidade do conhecimento extraído.

A etapa de extração de padrões compreende a escolha da tarefa de Mineração de Dados a ser empregada, a escolha do algoritmo e a extração dos padrões propriamente dita.

Escolha da tarefa

A escolha da tarefa é feita de acordo com os objetivos desejáveis para a solução a ser encontrada. As tarefas possíveis de um algoritmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas.

Atividades de predição, ou Mineração de Dados preditivo, consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo. Os dois principais tipos de tarefas para predição são classificação e regressão. A classificação consiste na predição de um valor categórico como, por exemplo, predizer se o cliente é bom ou mau pagador. Na regressão, o atributo a ser predito consiste em um valor contínuo como, por exemplo, predizer o lucro ou a perda em um empréstimo (Weiss & Indurkhya 1998).

Atividades de descrição, ou Mineração de Dados descritivo, consistem na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada. Algumas das tarefas de descrição são *clustering*, regras de associação e sumarização.

Uma vez eleita a tarefa a ser empregada, existe uma variedade de algoritmos para executá-la. A definição do algoritmo de extração e a posterior configuração de seus parâmetros também são realizadas nesta etapa.

Escolha do algoritmo

A escolha do algoritmo é realizada de forma subordinada à linguagem de representação dos padrões a serem encontrados. Podem-se utilizar algoritmos indutores de árvores de decisão ou regra de produção, por exemplo, se o objetivo é realizar uma classificação. Entre os tipos mais freqüentes de representação de padrões, destacam-se: árvores de decisão, regras de produção, modelos lineares, modelos não-lineares (Redes Neurais Artificiais), modelos baseados em exemplos (KNN — *K-Nearest Neighbor*, Raciocínio Baseado em Casos) e modelos de dependência probabilística (Redes Bayesianas).

Um aspecto que merece atenção é a complexidade da solução encontrada pelo algoritmo de extração. A complexidade da solução está diretamente relacionada com a capacidade de representação do conceito embutido nos dados. O problema é que, quando os parâmetros do algoritmo estão ajustados para encontrar soluções mais complexas que o conceito efetivamente existente nos dados, o modelo resultante poderá ter uma precisão boa para o conjunto de treinamento, mas um desempenho ruim para novos exemplos. Diz-se então que o modelo é específico para o conjunto de treinamento, ocorrendo *overfitting* (Monard & Baranauskas 2003).

Por outro lado, se a solução que está sendo buscada não for suficiente para adequar o conceito representado nos dados, por exemplo, ao se definir um número insuficiente de neurônios para uma Rede Neural ou um fator de poda muito alto na geração de uma árvore de decisão, o modelo induzido pode não ser representativo. Isto é o que se chama de *underfitting* (Monard & Baranauskas 2003). Nesse caso, é provável que o modelo encontrado não tenha bom desempenho tanto nos exemplos disponíveis para o treinamento como para novos exemplos.

Portanto, a configuração dos parâmetros do algoritmo deve ser feita de forma bastante criteriosa. Kearns & Vazirani (1994) apresenta uma sugestão para escolher uma determinada função: o modelo mais apropriado é aquele mais simples que seja consistente com todas as observações. Normalmente, soluções mais complexas são preferidas por pesquisadores, enquanto os práticos tendem a preferir modelos mais simples em virtude de sua fácil interpretação (Fayyad, Piatetsky-Shapiro, & Smyth 1996c).

Kohavi, Sommerfield, & Dougherty (1996) mostra experimentalmente que não existe um único bom algoritmo para todas as tarefas de Mineração de Dados. Por isso, a escolha de vários algoritmos para realizar a tarefa desejada pode ser feita, levando a obtenção de diversos modelos que, na etapa de pós-processamento, são tratados para fornecer o conjunto de padrões mais adequado ao usuário final.

Extração de padrões

A extração de padrões consiste da aplicação dos algoritmos de mineração escolhidos para a extração dos padrões embutidos nos dados. É importante ressaltar que dependendo da função escolhida pode ser necessária a execução dos algoritmos de extração de padrões diversas vezes.

Técnicas de combinação de preditores têm sido pesquisadas com o objetivo de

construir um preditor mais preciso pela combinação de vários outros. O resultado dessa combinação é chamado de *ensemble*. A utilização de *ensembles* tem obtido melhores resultados que a utilização de um único preditor (Dietterich 2000; Breiman 2000).

Outro aspecto importante é que com o objetivo de fazer uma avaliação mais precisa da taxa de erro de um classificador, métodos de *resampling* têm sido normalmente utilizados na extração de padrões. Mais detalhes sobre o cálculo da taxa de erro e geração do modelo quando se utiliza *resampling* podem ser encontrados em Monard & Baranauskas (2003).

A disponibilização do conjunto de padrões extraídos nesta etapa ao usuário ou a sua incorporação a um Sistema Inteligente ocorre após a análise e/ou o processamento dos padrões na etapa de pós-processamento.

Pós-Processamento

A obtenção do conhecimento não é o passo final do processo de Extração de Conhecimento de Bases de Dados. O conhecimento extraído pode ser utilizado na resolução de problemas da vida real, seja por meio de um Sistema Inteligente ou de um ser humano como apoio a algum processo de tomada de decisão. Para isso é importante que algumas questões sejam respondidas aos usuários (Liu & Hsu 1996):

- O conhecimento extraído representa o conhecimento do especialista?
- De que maneira o conhecimento do especialista difere do conhecimento extraído?
- Em que parte o conhecimento do especialista está correto?

No entanto, geralmente, não é fácil responder essas questões, já que os algoritmos de extração de padrões podem gerar uma quantidade enorme de padrões, muitos dos quais podem não ser importantes, relevantes ou interessantes para o usuário. Sabe-se também que fornecer ao usuário uma grande quantidade de padrões descobertos não é produtivo pois, normalmente, ele procura uma pequena lista de padrões interessantes. Portanto, é de vital importância desenvolver algumas técnicas de apoio no sentido de fornecer aos usuários apenas os padrões mais interessantes (Silberschatz & Tuzhilin 1995).

Diversas medidas para avaliação de conhecimento têm sido pesquisadas com a finalidade de auxiliar o usuário no entendimento e na utilização do conhecimento adquirido. Estas medidas podem ser divididas entre medidas de desempenho e medidas de qualidade.

Algumas medidas de desempenho são precisão, erro, confiança negativa, sensitividade, especificidade, cobertura, suporte, satisfação, velocidade e tempo de aprendizado (Lavrac, Flach, & Zupan 1999).

As medidas de qualidade são necessárias pois um dos objetivos do processo de Extração de Conhecimento é que o usuário possa compreender e utilizar o conhecimento descoberto. Entretanto, podem ocorrer casos em que os modelos são muito complexos ou não fazem sentido para os especialistas (Pazzani 2000; Pazzani, Mani, & Shankle 1997). Assim, a compreensibilidade do conhecimento extraído é um aspecto bastante importante para o processo de Extração de Conhecimento.

A compreensibilidade de um dado conjunto de regras está relacionada com a facilidade de interpretação dessas regras por um ser humano. A compreensibilidade de um modelo pode ser estimada, por exemplo, pelo número de regras e número de condições por regra. Nesse caso, quanto menor a quantidade de regras de um dado modelo e menor o número de condições por regra, maior será a compreensibilidade das regras descobertas (Fertig, Freitas, Arruda, & Kaestner 1999). Em (Pazzani 2000; Pazzani, Mani, &

Shankle 1997) é discutido que outros fatores, além do tamanho do modelo, são importantes na determinação da compreensibilidade de um conhecimento. Um fator citado é que os usuários especialistas possuem tendência a compreender melhor modelos que não contradizem seu conhecimento prévio.

A interessabilidade é uma maneira de avaliar a qualidade tentando estimar o quanto de conhecimento interessante (ou inesperado) existe e deve combinar fatores numa medida que reflita como o especialista julga o padrão (Piatetsky-Shapiro & Matheus 1994). As medidas de interessabilidade estão baseadas em vários aspectos, principalmente na utilidade que as regras representam para o usuário final do processo de Extração de Conhecimento (Dong & Li 1998). Estas medidas podem ser divididas em objetivas e subjetivas (Silberschatz & Tuzhilin 1995; Piatetsky-Shapiro & Matheus 1994; Freitas 1998a).

Medidas objetivas são aquelas que estão relacionadas somente com a estrutura dos padrões e do conjunto de dados de teste. Elas não levam em consideração fatores específicos do usuário nem do conhecimento do domínio para avaliar um padrão. Algumas medidas objetivas de interessabilidade são: modelos de regras, cobertura de regras mínimas, custo da classificação incorreta e tamanho do disjunto (Horst 1999).

Como diferentes usuários finais do processo de Extração de Conhecimento podem ter diferentes graus de interesse para um determinado padrão, medidas subjetivas são necessárias. Estas medidas consideram que fatores específicos do conhecimento do domínio e de interesse do usuário devem ser tratados ao selecionar um conjunto de regras interessantes ao usuário. Algumas medidas subjetivas são inesperabilidade e utilidade (Silberschatz & Tuzhilin 1995).

Em um ambiente para avaliação de conhecimento, aspectos objetivos de interessabilidade podem ser utilizados como um primeiro filtro para selecionar regras potencialmente interessantes. Os aspectos subjetivos podem ser utilizados como um filtro final para selecionar regras realmente interessantes.

Após a análise do conhecimento, caso este não seja de interesse do usuário final ou não cumpra com os objetivos propostos, o processo de extração pode ser repetido ajustando-se os parâmetros ou melhorando o processo de escolha dos dados para a obtenção de resultados melhores numa próxima iteração.

1.3.3. Principais Tarefas de Mineração de Dados

Com o grande número de sistemas de Mineração de Dados desenvolvidos para os mais diferentes domínios, a variedade de tarefas para MD vem se tornando cada vez mais diversificada. Essas tarefas podem extrair diferentes tipos de conhecimento, sendo necessário decidir já no início do processo de MD qual o tipo de conhecimento que o algoritmo deve extrair.

Como discutido anteriormente, atividades de predição envolvem o uso dos atributos de um conjunto de dados para prever o valor futuro do atributo-meta, ou seja, essas atividades visam principalmente à tomada de decisões. Já as atividades de descrição procuram padrões interpretáveis pelos humanos que descrevem os dados antes de realizar a previsão. Essa tarefa visa o suporte à decisão.

Classificação

A tarefa de classificação é uma função de aprendizado que mapeia dados de entrada, ou conjuntos de dados de entrada, em um número finito de categorias. Nela, cada exemplo pertence a uma classe, entre um conjunto predefinido de classes. Os exemplos consistem

de um conjunto de atributos e um atributo-meta discreto. O objetivo de um algoritmo de classificação é encontrar algum relacionamento entre os atributos e uma classe, de modo que o processo de classificação possa usar esse relacionamento para predizer a classe de um exemplo novo e desconhecido.

Assim, a classificação consiste em obter um modelo baseado em um conjunto de exemplos que descrevem uma função não-conhecida. Esse modelo é então utilizado para predizer o valor do atributo-meta de novos exemplos.

Regressão

A tarefa de regressão é conceitualmente similar à de classificação. A principal diferença é que o atributo a ser predito é contínuo ao invés de discreto.

Os métodos de regressão já são estudados pela comunidade estatística há bastante tempo. Porém, nas áreas de Aprendizado de Máquina e Mineração de Dados, a maioria das pesquisas é voltada para problemas de classificação, que são mais comumente encontrados na vida real do que problemas de regressão (Weiss & Indurkhya 1995).

O objetivo da tarefa regressão é encontrar uma relação entre um conjunto de atributos de entrada (variáveis de entrada ou variáveis preditoras) e um atributo-meta contínuo. Sejam $X = \{x_1, ..., x_d\}$ os atributos de entrada e y o atributo-meta, o objetivo é encontrar uma mapeamento da seguinte forma (Apte & Weiss 1997):

$$y = f(x_1, x_2, ..., x_d)$$

Regressão também é conhecida por predição funcional, predição de valor real, função de aproximação, ou ainda, aprendizado de classes contínuas (Uysal & Güvenir 1999).

Os métodos de regressão geram modelos[1] em diferentes formatos de representação. Como esses modelos expressam o conhecimento obtido durante o processo de Mineração de Dados, então cada método pode expressar o conhecimento de uma forma diferente. Dependendo dos objetivos a serem alcançados ao final do processo de Mineração de Dados, a compreensibilidade dos modelos gerados é considerada muito importante, tanto para uma tarefa de regressão quanto para uma tarefa de classificação. A necessidade de modelos capazes de fornecerem soluções interpretáveis leva à utilização dos métodos de aprendizado simbólico. Regras e árvores são tipos de modelos gerados por esses métodos. Elas são bastante semelhantes, diferenciando-se, por exemplo, pelo fato de que as árvores são mutuamente exclusivas e as regras nem sempre.

A grande maioria dos problemas de regressão faz a predição de um único atributo-meta. Porém, regressão também pode ser realizada predizendo mais que um atributo-meta. Por exemplo, as redes neurais artificiais (Braga, Carvalho, & Ludermir 2000).

Regras de Associação

Uma regra de associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma Base de Dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (Agrawal & Srikant 1994). Desse modo, o objetivo das regras de associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados. Por exemplo, observando os dados de vendas de um supermercado sabe-se que 80% dos clientes que compram o produto Q também

[1] Os modelos de regressão são também conhecidos como regressores.

adquirem, na mesma compra, o produto W. Nessa regra, 80% corresponde a sua confiabilidade.

O formato de uma regra de associação pode ser representado como uma implicação $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens. As regras de associação podem ser definidas como descrito a seguir (Agrawal & Srikant 1994):

> Seja D uma Base de Dados composta por um conjunto de itens $A = \{a_1, ..., a_m\}$ ordenados lexicograficamente e por um conjunto de transações $T = \{t_1, ..., t_n\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens tal que $t_i \subseteq A$.
> A Regra de Associação é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que ocorre LHS ocorre também RHS. A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações em D ocorre $LHS \cup RHS$.

O valor do suporte mede a força da associação entre LHS e RHS, e não relaciona possíveis dependências de RHS com LHS. Por outro lado, a confiança mede a força da implicação lógica descrita pela regra.

1.3.4. Técnicas e Funcionalidades de Ferramentas para Mineração de Dados

Uma parte importante do processo de MD são as técnicas e os algoritmos que podem ser empregados. São várias as áreas que podem apoiar esse processo, entre elas estão Análise Estatística, Aprendizado de Máquina, Algoritmos Genéticos e Redes Neurais.

É importante notar que as técnicas descrevem um paradigma de extração de conhecimento e vários algoritmos podem seguir esse paradigma. Por exemplo, o aprendizado simbólico que gera regras de decisão, muito utilizadas para extração de conhecimento, pode ser realizado utilizando diferentes algoritmos, como C4.5-rules e CN2.

Entre as técnicas mais utilizadas em Mineração de Dados estão as regras e árvores de decisão, as Redes Neurais que apesar de não gerarem conhecimento explícito são bastante empregadas para diversas tarefas de Mineração de Dados, aplicações de Algoritmos Genéticos que fazem parte da computação evolutiva, e Lógica *Fuzzy*.

O progresso da área de Mineração de Dados e sua utilização nos mais variados domínios e pelas mais diversas organizações têm motivado o desenvolvimento de várias ferramentas comerciais além da elaboração de muitos protótipos de pesquisa. O processo de Extração de Conhecimento de Bases de Dados é facilitado consideravelmente se for usada uma ferramenta que ofereça suporte para uma variedade de técnicas, com diferentes algoritmos disponíveis e voltadas para várias tarefas de Mineração de Dados.

O desenvolvimento de ferramentas comerciais de Mineração de Dados tem como objetivo principal fornecer aos tomadores de decisão das organizações, que são usuários geralmente não especialistas em Mineração de Dados, ferramentas intuitivas e amigáveis. É interessante que estas ferramentas ofereçam suporte às várias etapas do processo de Mineração de Dados assim como disponibilizem apoio para diversas técnicas e tarefas.

Inúmeras universidades e laboratórios de pesquisa têm direcionado seus projetos para o desenvolvimento de protótipos de ferramentas de MD com funcionalidades inovadoras. De acordo com essas funcionalidades esses protótipos podem ser agrupados nas seguintes categorias (Thuraisingham 1999):

Novos Modelos Funcionais Essencialmente, na maioria destes projetos, o objetivo é integrar métodos de Mineração de Dados com o gerenciamento de Banco de Dados. A idéia principal desses trabalhos é desenvolver novas técnicas para otimizar as consultas e permitir um melhor suporte aos métodos de Mineração de Dados.

Tratamento de Novos Tipos de Dados O tratamento de diferentes tipos de dados, como dados multimídia, é o objetivo dos protótipos pertencentes a esta categoria. Além de ferramentas que tratam dados multimídia, também pertencem a esta categoria as ferramentas que apóiam o processo de Mineração de Textos (Baeza-Yates & Ribeiro-Neto 1999; Habn & Mani 2000; Vasileios, Gravano, & Maganti 2000).

Escalabilidade Com o desenvolvimento da tecnologia de armazenamento, o volume de dados disponível para análise tem crescido exponencialmente. Dessa maneira, diversos projetos de pesquisa têm sido desenvolvidos com o objetivo de tratar grandes Bases de Dados. Alguns trabalhos que tratam da escalabilidade dos algoritmos de extração de conhecimento estão sendo desenvolvidos.

Compreensibilidade dos Resultados Um outro problema identificado no final do processo de Extração de Conhecimento de Bases de Dados é a compreensibilidade do resultado obtido. Os padrões extraídos pelos algoritmos podem ser muito complexos ou não fazerem sentido para os usuários do processo de extração de conhecimento, dificultando sua compreensão. A visualização dos padrões extraídos tem sido utilizada em muitos projetos como principal técnica para auxiliar a sua compreensão.

Além das técnicas, tarefas e ferramentas, observa-se que essencialmente há dois estilos para se fazer MD: *top-down* e *bottom-up*, ilustrados na Figura 1.4. No estilo *top-down*, o processo é iniciado com alguma hipótese a ser verificada. Nesse caso, em geral, é desenvolvido um modelo, e este é então avaliado para determinar se a hipótese é válida ou não. No estilo *bottom-up*, não é especificada uma hipótese para validação, apenas são extraidos padrões dos dados.

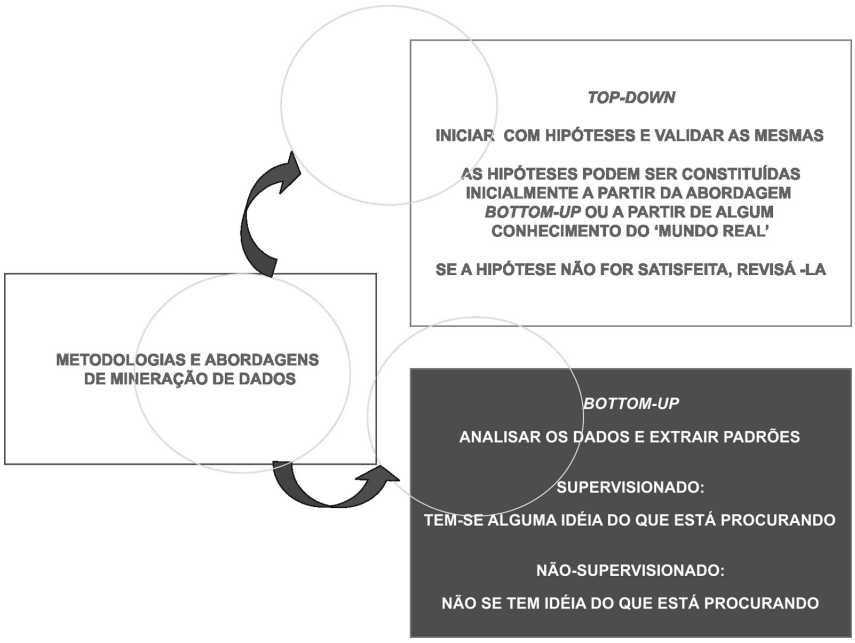

Figura 1.4: Estilos para Mineração de Dados (Thuraisingham 1999)

1.3.5. Mineração Visual de Dados

Visualização de Informação

Representações gráficas têm sido utilizadas, largamente, como instrumentos de comunicação desde os primórdios da humanidade. Com o advento da ciência, as representações gráficas passaram a embutir significados por convenções, como gráficos matemáticos e cartográficos (Branco 2003). Essas representações, normalmente, têm o propósito de comunicar uma idéia existente em um conjunto de valores. Contudo, a fim de aproveitar as características da percepção visual humana, uma segunda abordagem consiste em utilizar as representações gráficas para criar ou descobrir uma idéia que esteja embutida nos valores. Essa segunda abordagem tem crescido consideravelmente, devido a evolução dos computadores para a geração de representações significativas e pelos mesmos permitirem a evolução da interatividade humano-computador.

Visualização é a área em que as representações gráficas produzem um significado que permitam aos usuários desenvolver suas próprias idéias ou de confirmar suas expectativas. Visualização é o processo de mapeamento de dados e informações em um formato gráfico, baseando-se em representações visuais e em mecanismos de interação, fazendo uso de suporte computacional e objetivando a ampliação da cognição (Card, Mackinlay, & Shneiderman 1999).

Na Figura 1.5 pode ser observado um modelo de referência para visualização. A visualização pode ser observada como sendo uma seqüência de mapeamentos “ajustáveis” de dados para uma representação visual, por meio da interação do usuário.

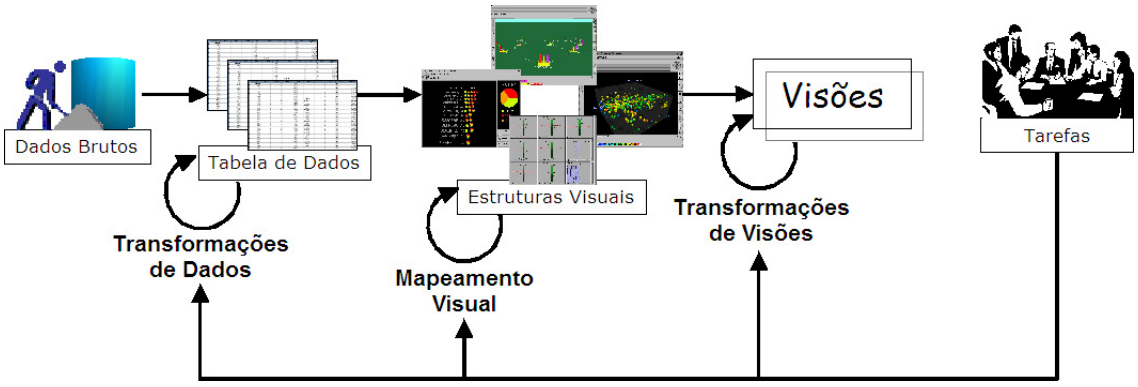

Figura 1.5: Modelo de referência de visualização (Card, Mackinlay, & Shneiderman 1999)

Por meio desse modelo de referência, pode-se obter duas vertentes: visualização científica e visualização de informação.

A visualização científica baseia-se em dados produzidos por fenômenos naturais, do mundo físico, com a finalidade de visualizar objetos a partir dos dados. Por exemplo, para representar as concentrações e a dinâmica de massas de ar na atmosfera.

A visualização de informação baseia-se em analisar uma grande quantidade de dados não-físicos, tais como coleções de documentos e dados financeiros. Esse tipo de avaliação possui uma complexidade maior devido a tornar visíveis as características inerentes desses dados.

Em (Keim 2002) é apresentado um *survey* das diferentes técnicas de visualização, as quais são classificadas segundo 3 critérios: dados a serem visualizados (eixo Y da Fi-

gura 1.6), técnicas de interação e distorção (eixo X da Figura 1.6); e técnica de visualização de informação relacionada (eixo Z da Figura 1.6).

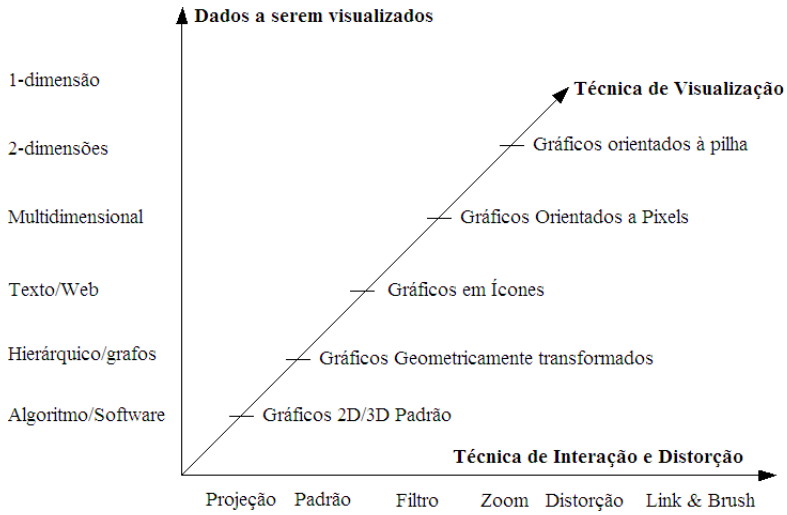

Figura 1.6: Classificação das técnicas de visualização (Keim 2002)

A seguir são discutidos esses três critérios.

Tipos de dados a serem visualizados

Os dados para a visualização de informação possuem, normalmente, uma grande quantidade de registros, sendo que cada registro consiste em uma observação, medida, transação, entre outros e possui um número de variáveis[2] ou dimensões. Para que os dados possam ser visualizados corretamente é necessário identificar se são: unidimensionais, bidimensionais, multidimensionais, texto/web, hierárquicos/grafos ou algorítmicos/software.

Um exemplo de dado unidimensional é o temporal, isto é, para cada unidade de tempo um ou mais valores relacionados ao tempo são associados. Um dado bidimensional é aquele que possui 2 dimensões distintas, por exemplo, longitude e latitude de uma área geográfica. Um dado multimensional consiste de três ou mais variáveis para serem visualizados, por exemplo, dados retirados de uma tabela relacional.

Dados textuais ou retirados da WWW possuem uma grande complexidade, pois, normalmente são não-estruturados, não podendo ser descritos em forma de números ou letras facilmente (Nowell, Havre, Hetzler, & Whitney 2002). Dados hierárquicos/grafos são aqueles que possuem algum nível de relacionamento entre eles (Battista, Eades, Tamassia, & Tollis 1999). Dados providos de algoritmos/software são aqueles que permitem depurar códigos e visualizar a estrutura para detectar ou encontrar erros[3].

Técnicas de visualização

Além das técnicas comuns de visualização 2D/3D como gráficos em barra, gráficos lineares, gráficos em pizza, existem técnicas de visualização mais sofisticadas que encontram-se em diferentes classificações, descritas a seguir.

[2]Para uma Base de Dados as variáveis ou dimensões são os atributos.

[3]Software Visualization – `http://wwwbroy.informatik.tu-muenchen.de/~trilk/sv.html`

- Gráficos Geometricamente Transformados (*Geometrically Transformed Displays*) são técnicas que permitem transformações em dados multidimensionais para verificar se os mesmos são interessantes (*interestingness*). Esta classe inclui técnicas estatísticas exploratórias como *scatterplots*. Porém, a técnica que se destaca é a Coordenadas Paralelas (*Parallel Coordinate*), a qual um espaço de dimensão *k* é mapeado em um espaço visual bidimensional, usando *k* eixos eqüidistantes e paralelos aos eixos principais (X e Y) (Inselber & Dimsdale 1990);
- Gráficos em Ícones (*Iconic Displays*) são técnicas que mapeiam os valores de um dado multidimensional como características de um ícone. Um exemplo de gráfico em ícones é o *Stick Figures* (Pickett & Grinstein 1988);
- Gráficos Orientados à Pixels (*Dense Pixel Displays*) possuem um princípio básico que consiste em mapear cada valor de uma dimensão para um pixel, conseqüentemente com uma cor, agrupando-os ao longo de cada dimensão. Além disso, é necessário determinar um arranjo desses pixels para diferentes propósitos. Em (Keim 2000) podem ser encontradas as diferentes aplicações dos algoritmos desta técnica em forma de um *survey*;
- Gráficos Orientados à Pilha (*Stacked Displays*) são técnicas hierárquicas para visualização, nas quais o espaço n-dimensional dos dados é dividido em subespaços que estão organizados e exibidos na forma hierárquica, projetando ou embutindo esses espaços uns dentro dos outros. Um exemplo de gráfico orientado à pilha é o *Dimensional Stacking* (Keim & Kriegel 1996).

Técnicas de interação e distorção

Em adição as técnicas de visualização vistas anteriormente, técnicas de interação e distorção devem ser combinadas com os usuários para que seja feita uma exploração efetiva. As técnicas de interação permitem aos usuários interagir diretamente com as visualizações e mudá-las dinamicamente dependendo dos objetivos de exploração. Já as técnicas de distorção consistem em mostrar porções dos dados com um alto nível de detalhe.

As técnicas de interação dividem-se em:

- Projeções Dinâmicas (*Dynamic Projections*) consiste em trocar dinamicamente as projeções para explorar a multidimensionalidade do conjunto de dados. O sistema XGobi (Swayne, Cook, & Buja 1992) utiliza esse tipo de técnica de projeção;
- Filtragem Interativa (*Interactive Filtering*) possui duas características: *browsing* e *querying*. *Browsing* seria a seleção direta de um subconjunto dos dados disponíveis e *querying* seria a seleção de um subconjunto dos dados por meio de especificações de propriedades inerentes aos dados. Um exemplo de sistema que utiliza este tipo de técnica é o Polaris (Stolte, Tang, & Hanrahan 2002);
- Zoom Interativo (*Interactive Zooming*) é uma extensão de uma técnica conhecida como *zooming*, que é capaz de apresentar, em diferentes níveis, valores de uma variável em diferentes resoluções. Os dados, por exemplo, podem ser representados como simples pixels em um baixo *zoom*, como ícones em um médio *zoom* e como objetos em um alto *zoom*. Uma interessante técnica de *interactive zooming* pode ser encontrada em (Rao & Card 1994).
- Distorção Interativa (*Interactive Distortion*) é uma técnica que mostra porções dos dados em um alto nível de detalhamento, enquanto outros podem ser mostrados com um baixo nível de detalhamento. Técnicas utilizando grafos, *fisheye* (Furnas 1986) e *perspective wall* (Mackinlay, Robertson, & Card 1991) pertencem a esta classe;

- *Interactive Linking and Brushing* é uma técnica que tenta combinar diferentes métodos de visualização. Pode-se combinar *scatterplots* de diferentes projeções, colorindo somente os pontos que aparecem em todas as projeções, obtendo-se assim, visualizações de dependências ou correlações entre os dados. Em (Stolte, Tang, & Hanrahan 2002) e (Swayne, Cook, & Buja 1992) são detalhados usos dessa técnica.

Visualização de Informação e Mineração Visual de Dados

Como foi descrito anteriormente, a Mineração de Dados têm-se tornado um fenômeno de uso por várias áreas do conhecimento. Esse fenômeno é motivado pela abrangência das técnicas para uso com grandes Bases de Dados a fim de obter padrões inteligíveis para os usuários, sejam especialistas, analistas ou usuários finais. Porém, somente padrões em forma de árvores, regras, redes ou qualquer linguagem de descrição, normalmente, não são suficientes. A visualização dos padrões encontrados têm-se tornado uma necessidade, pois para uma grande Base de Dados é gerada uma grande quantidade de padrões. Essa grande quantidade pode ser filtrada, por exemplo, observando-se padrões ou comportamentos gráficos em uma visualização.

Segundo (Ganesh, Han, Kumar, Shekhar, & Srivastava 1996; Keim & Kriegel 1996), a Mineração Visual de Dados (VDM – *Visual Data Mining*) consiste de uma parte da área de Visualização de Informação concentrando-se as técnicas de visualização para dados multidimensionais e combinando-as com a Mineração de Dados.

Esta junção de técnicas de Visualização com Mineração de Dados têm-se tornado freqüente, pois a visualização é um valioso instrumento de apoio ao processo de Mineração de Dados (Rezende 2003). Apesar da visualização para Mineração de Dados ser considerada exploratória (Rezende, Oliveira, Félix, & Rocha 1998), (Keim 2001) considera que a "exploração visual pode ser vista como um processo de geração de hipóteses, segundo o qual a visualização dos dados permite ao usuário adquirir percepções dos dados, podendo provocar o surgimento de novas hipóteses, que, por sua vez, podem também ser confirmadas ou rejeitadas com o uso da exploração visual".

Em relação a combinação da Visualização de Informação e Mineração de Dados, Wong (1999) descreveu duas formas de integração:

- **Acoplamento forte**, na qual a visualização é combinada com a Mineração de Dados de modo a aproveitar as características de cada área;
- **Acoplamento fraco**, na qual as áreas estão simplesmente intercaladas, possibilitando um pouco aproveitamento do potencial de cada área.

A forma de integração também pode auxiliar os usuários do processo a obter um maior ganho de interatividade, pois um acoplamento forte entre a visualização e a mineração permite ao usuário verificar ou analisar hipóteses já sugeridas pelo especialista, por exemplo. Uma visualização com acoplamento fraco é desejável quando se possui uma idéia ou um entendimento de um subconjunto de um domínio, ou seja, é necessário somente visualizar algum padrão que o especialista deseja comprovar.

Em (Ankerst 2000) a Mineração Visual de Dados está intimamente ligada com a fase de extração de padrões, e dividida em 3 categorias ilustradas na Figura 1.7: visualização anterior, visualização posterior e visualização fortemente integrada.

A visualização anterior (Figura 1.7(a)) ou visualização dos dados é uma categoria de visualização, na qual os dados são visualizados sem a execução dos algoritmos de extração de padrões. No processo de Mineração de Dados esse tipo de visualização é,

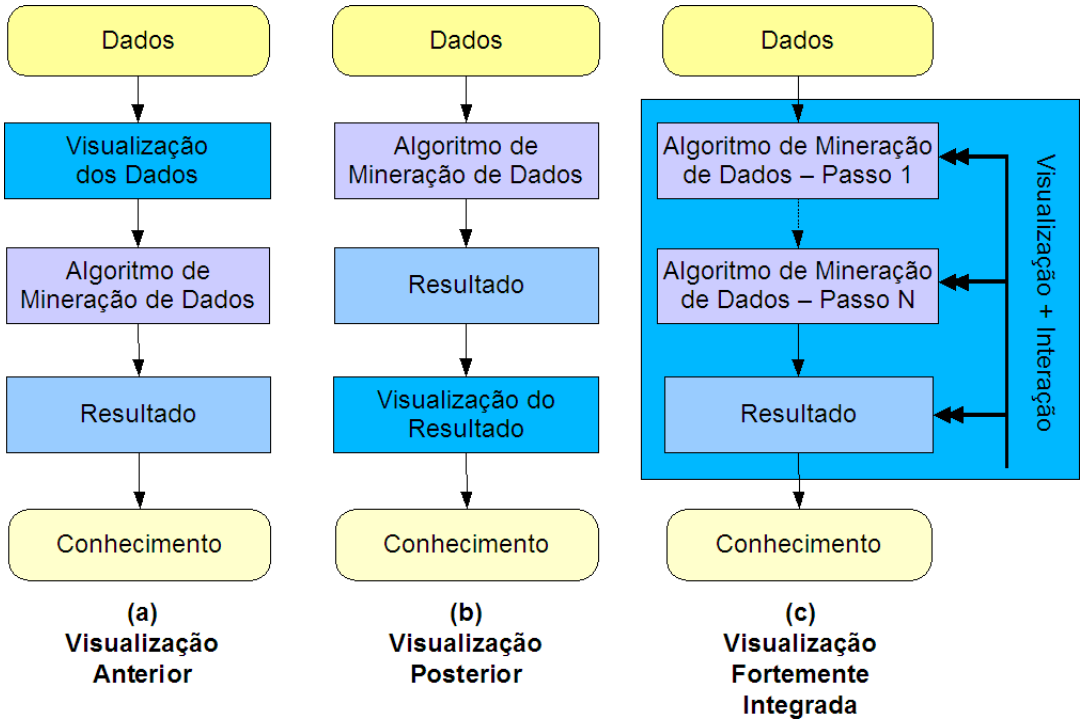

Figura 1.7: Categorias de Mineração Visual de Dados (Ankerst 2000)

comumente, utilizado nas etapas de identificação do problema e pré-processamento dos dados (Figura 1.3).

A visualização do resultado da Mineração de Dados (Figura 1.7(b)) corresponde a visualização dos padrões obtidos pela Mineração de Dados, ou seja, dependendo da linguagem de descrição do algoritmo é utilizada uma determinada visualização. Um exemplo de visualização de árvores de decisão no sistema MineSet^{TM} da SGI Inc. durante o processo de Mineração de Dados pode ser encontrado em Rezende, Oliveira, Félix, & Rocha (1998).

A visualização fortemente integrada é uma categoria, na qual, podem ser observados resultados intermediários de uma mineração. Essa categoria de visualização ocorre quando os algoritmos que executam uma análise dos dados não produzem padrões finais, mas sim, padrões intermediários. Com isso, o usuário pode encontrar padrões de interesse na visualização, visto que essa forma promove um conhecimento do domínio. Essa abordagem torna-se particularmente importante quando se observa que não há algoritmos genéricos para Mineração de Dados, e que pode fornecer uma forma eficiente de avaliar e validar o andamento do processo (Branco 2003).

1.3.6. Mineração de Dados Textuais

O processo de descoberta de conhecimento em textos[4] pode ser visto como uma associação de técnicas para preparação de textos ao processo de Mineração de Dados. Pela natureza dos dados[5] e pelas aplicações a que está freqüentemente relacionado, o processo de Mineração de Textos combina muitas das técnicas de Recuperação de Informação (RI) e pode também contar com a utilização de algumas estratégias de Processamento de Língua Natural (PLN), ao longo de suas etapas.

Tendo em vista o grande volume de informações disponíveis, principalmente na

[4]As palavras texto e documento são usadas indistintamente, nesse trabalho.
[5]Textos são dados não estruturados.

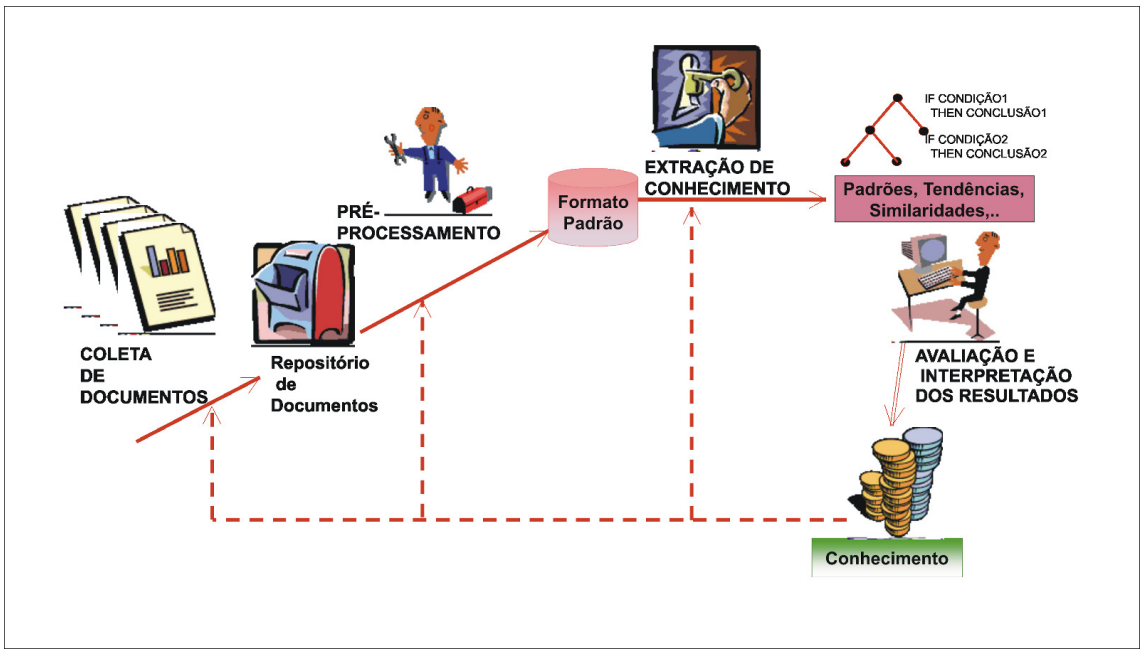

Figura 1.8: Etapas do processo de Mineração de Textos

WWW, o que se pode perceber é que grande parte das aplicações de Mineração de Textos estão voltadas à construção de ferramentas que propiciem melhorias na recuperação de documentos, de forma a tornar esse processo mais preciso. Dessa forma, muitas pesquisas associam a realização desse processo à atividade de classificação de documentos. Independente de qual seja a finalidade do processo de Mineração de Textos, alguns conceitos em mais alto nível sobre suas etapas podem ser considerados os mesmos.

O Processo de Mineração de Dados Textuais

O processo de Mineração de Textos, como ilustrado na Figura 1.8, pode ser dividido em quatro etapas fundamentais: coleta de documentos, pré-processamento, extração de conhecimento e avaliação e interpretação dos resultados.

A coleta de documentos consiste no primeiro passo do processo e tem como função recuperar os documentos que possam ser relevantes para alcançar o objetivo almejado.

Como os documentos coletados podem estar em diferentes formatos, o pré-processamento pode envolver a padronização dos mesmos para um formato único. Além disso, essa etapa é responsável por obter uma estrutura, tal qual uma tabela atributo-valor, que represente o conteúdo de uma coleção de documentos.

Com a representação dos documentos na forma atributo-valor, é possível aplicar as técnicas de extração de padrões. Para isso, na etapa de extração de conhecimento pode ser feita a utilização de sistemas de aprendizado com a finalidade de encontrar padrões/tendências/similaridades de acordo com o objetivo e requisitos do usuário e/ou domínio da aplicação.

Na etapa de avaliação e interpretação dos resultados, assim como no processo de MD, os padrões encontrados podem ser analisados junto ao usuário final, ao especialista do domínio e ao analista de dados, que cumprem diferentes atividades importantes e não são necessariamente pessoas distintas.

Mediante algumas circunstâncias, como na obtenção de um resultado pouco sig-

nificativo ao usuário, pode ser necessário que o processo seja refeito, adequando-se algumas de suas etapas, seja para contar com uma gama mais informativa de documentos, ou mesmo com alguma estratégia que possa melhorar o desempenho do processo. E por essa razão, o processo pode contar com algumas iterações, como mostrado na Figura 1.8.

Coleta de Documentos

Na etapa de coleta de documentos é almejada a recuperação de documentos com descrições textuais que sejam relevantes ao domínio de aplicação do conhecimento a ser extraído. Mediante esse objetivo, podem ser consideradas diversas fontes, tais como, livros (cujas páginas possuem cópias eletrônicas ou pelo uso de um *scanner*), e documentos provenientes da WWW.

A WWW é um dos principais meios para disponibilizar e acessar uma grande gama de documentos. Para facilitar o acesso a esses, muitas ferramentas de apoio têm sido construídas, geralmente fazendo uso de diversas técnicas de RI. A busca de documentos nesse meio pode ser realizada usando cinco abordagens diferentes: *Robotic Internet Search Engines, Mega-Indexes, Simultaneous Mega-Indexes, Subject Directories e Robotic Specialized Search Engines*.

1. Os *Robotic Internet Search Engines* usam uma arquitetura *Web* robô, também chamado de *spider*, *Web wanderers* ou *Web worms*, e percorrem todo o domínio da WWW enviando páginas novas ou adaptadas ao servidor a que eles estão indexados (Baeza-Yates 1998).
   Procedendo dessa forma, esses *search engines* indexam toda a WWW como um completo Banco de Dados de textos. A recuperação de um número significativo de documentos da WWW, geralmente, é dada por meio da coincidência entre as palavras da consulta e as palavras das páginas WWW indexadas.
   Dentro dessa categoria, fazem parte os seguintes sistemas de busca: Altavista[6], Excite[7], Lycos[8], The Open Text Index[9], InfoSeek Guide[10], InfoSeek Ultra[11], WebCrawler[12], ALIWEB[13], HotBot[14], entre outros.
2. A abordagem *Mega-Indexes* é caracterizada por não possuir seu próprio Banco de Dados. Ao invés disso, mantém *links* para os *Robotic Search Engines*. De acordo com esse conceito, páginas pessoais com indicação de *links* para sistemas de busca podem ser consideradas *Mega-Indexes*. Além dessas páginas, alguns sistemas de busca que podem ser considerados *Mega-Indexes* são: Galaxy[15] e Magellan[16].
3. Os *Simultaneous Mega-Indexes* fazem uso de diversos *Robotic Search Engines* paralelamente, que colaboram para apresentar um pacote correspondente a um resultado unificado. Como exemplo dessa abordagem, pode ser citado o MetaCrawler[17].
4. Os *Subject Directories*, também chamados de catálogos, mantém a indexação das páginas de acordo com uma organização conceitual, que é mantida geralmente

[6] http://www.altavista.com
[7] http://www.excite.com
[8] http://www.lycos.com
[9] http://index.opentext.net
[10] http://guide.infoseek.com
[11] http://ultra.infoseek.com
[12] http://webcrawler.com
[13] http://aliweb.emnet.co.uk
[14] http://www.hotbot.com
[15] http://www.galaxy.com
[16] http://www.mckinley.com/magellan/
[17] http://www.metacrawler.com/index.html

de forma manual. A arquitetura de diretórios pode ser vista como uma árvore de taxonomias hierárquicas, que classifica o conhecimento humano. O exemplo mais conhecido para essa categoria é o Yahoo![18].

5. Os *Robotic Specialized Search Engines* são *Robotic Search Engines* que cobrem uma pequena ou especializada porção da *Web*, que é o caso de listas de discussão, *Yellow Pages*, *White Pages*, etc.

Na realidade, essas cinco abordagens podem ser, de forma geral, classificadas em apenas dois grandes grupos: os *search engines* (tais como *Robotic Search Engines* e suas variações: *Mega-Indexes*, *Simultaneous Mega-Indexes* e *Robotic Specialized Search Engines*) e os WWW *directories* (*Subject Directories*).

O principal problema relacionado aos *search engines* é a manutenção de páginas novas e modificadas, pois em função da natureza altamente dinâmica da WWW os *links* de comunicação dos servidores de páginas podem se tornar saturados. Já a principal desvantagem dos WWW *directories* está relacionada ao processo de classificar as páginas na hierarquia, que freqüentemente é realizada por um número limitado de pessoas, e por esse motivo não conseguem classificar todas as páginas *Web*, cujo volume de documentos tem crescido muito (Baeza-Yates 1998). Problemas como esses podem contribuir para a busca resultar em uma grande lista de páginas que muitas vezes não são do interesse do usuário.

Para tornar as buscas mais eficientes, algumas técnicas de outras áreas também têm sido empregadas, como as de PLN e AM. As técnicas de Aprendizado de Máquina têm sido usadas, sobretudo, para prover buscas inteligentes através do mapeamento do perfil do interesse do usuário. Como exemplo, pode-se citar o *WebWatcher* (Joachims, Freitag, & Mitchell 1997), que acompanha o usuário página a página sugerindo *hyperlinks* apropriados e aprende pelas experiências passadas para melhorar os conselhos a serem dados.

Embora o uso da combinação de técnicas de AM e de RI para determinação de padrões do perfil do usuário seja bastante relevante e interessante para melhorar a coleta de documentos.

Fazendo uso de alguma ferramenta de suporte à recuperação de documentos, a primeira etapa do processo de Mineração de Textos pode ser então cumprida. Diante disso, deve-se dar início a segunda etapa do processo, na qual os documentos recuperados são pré-processados para gerar a possível representação a ser utilizada pelos algoritmos de extração de conhecimento.

Pré-processamento de Textos

Como mencionado, a etapa de pré-processamento é responsável por converter os textos em uma representação atributo-valor que possa ser manipulada pelos métodos de extração de conhecimento. A obtenção de tal representação pode ser feita através da realização de algumas tarefas como identificação dos atributos, atribuição de pesos e redução da representação, como mostrado na Figura 1.9.

Existem diferentes abordagens para determinar os termos e os pesos que estarão presentes na representação. Freqüentemente as operações realizadas para a escolha dos termos são:

- substituição de marcadores HTML por símbolos especiais, quando aplicável;

[18] http://www.yahoo.com

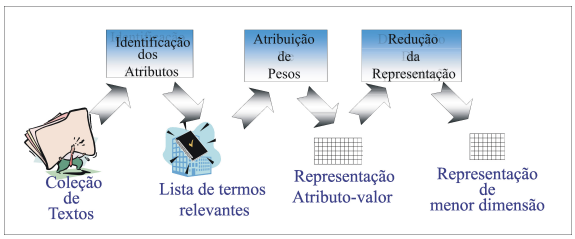

Figura 1.9: Algumas atividades realizadas no pré-processamento de textos

- reconhecimento de palavras individuais ou compostas que estejam presentes no texto;
- uso de uma lista de palavras a serem desconsideradas, como artigos, preposições, etc. Essa lista é conhecida como *stop list* ou lista de *stopwords*;
- remoção do sufixo das palavras para mapeá-las na sua forma canônica;
- organização do valor semântico das palavras por meio do uso de dicionários eletrônicos e mapas de sinônimos, como *thesaurus*.

As estratégias empregadas na identificação dos atributos, como mostrado na Figura 1.9 são, em geral, dependentes do idioma. Isso porque as palavras que compõem a *stop list* e os métodos para remover os sufixos dos termos podem variar de acordo com o idioma dos textos.

Para completar a representação dos textos, além das estratégias para identificação dos atributos, deve-se escolher quais serão as abordagens utilizadas para o cálculo do valor do peso de cada atributo, que geralmente pode ser booleano ou numérico (Weiss & Indurkhya 1998). Os valores booleanos são empregados para indicar a presença ou ausência do termo em cada documento, enquanto que os numéricos são calculados por meio de medidas estatísticas baseadas na freqüência dos termos nos documentos.

Em alguns casos, a representação originalmente obtida possui muitos atributos tornando sua dimensão relativamente grande a ponto de exceder a capacidade de processamento dos algoritmos usados para extração de conhecimento. Desta forma, são empregados métodos para redução da dimensão.

Extração de Conhecimento

A descoberta de conhecimento na mineração de texto pode ser interpretada como a classificação automática de documentos, ou a descoberta de associações entre textos e autores, ou ainda a determinação de tendências, entre outras atividades. Essas tarefas podem ser realizadas por meio de diversos algoritmos, sendo que grande parte deles também são utilizados em Mineração de Dados.

Em Mineração de Dados, os algoritmos freqüentemente empregados são de Aprendizado de Máquina. Além dos algoritmos de AM, como muitas das aplicações da extração de conhecimento em textos são dirigidas às tarefas da área de Recuperação de Informação, algoritmos clássicos, como Rochio (Hull 1994) e mais recentemente *Support Vector Machine* (SVM) (Vapnik 1995), vêm sendo utilizados.

Avaliação e Interpretação do Conhecimento

Assim como no processo de Mineração de Dados, os resultados obtidos durante a etapa de extração de conhecimento ainda não devem ser imediatamente utilizados nas aplicações de interesse do usuário final. É necessário que se faça a verificação dos padrões neles

contidos. Essa verificação pode ser realizada pelos participantes do processo e visa determinar se os resultados obtidos condizem com o objetivo a ser alcançado por meio do processo de Mineração de Textos.

A questão que diz respeito à aplicabilidade não é uma tarefa fácil de ser definida, uma vez que está diretamente relacionada com a noção de quão compreensível, válido, útil e novo é o conhecimento. Em muitos casos, um modelo é considerado compreensível de acordo com a sua simplicidade, porém a análise do que deve ser considerado um modelo complexo ou simples, também depende de muitos fatores e contextos.

Com intuito de tornar mais simples, útil, e até mesmo de facilitar a avaliação e interpretação dos padrões extraídos, muitas abordagens estatísticas e ferramentas de visualização são usadas para identificar e remover padrões redundantes, e pouco ou sem utilidade para a aplicação final.

O fator chave desta etapa, além dessas técnicas e ferramentas de apoio, é a interação de todos os participantes do processo, uma vez que pode ser detectado se houve erros em alguma parte do processo, ou mesmo se a aplicação de outros métodos poderia fornecer melhores resultados. Se fatos como estes forem constatados, algumas das etapas anteriores devem ser refeitas, como mostrado na Figura 1.8 (ilustrada na página 26), caso contrário os resultados são validados e podem ser utilizados pelo usuário final em sua aplicação.

Como pode ser observado pela descrição das medidas e por todos os conceitos abordados até o momento neste capítulo, o processo de Mineração de Textos está voltado principalmente para aplicações de classificação de textos. Recentemente, atividades relacionadas à inteligência competitiva também têm sido almejadas. Isso vem comprovar que a importância do processo vêm aumentando como um todo. De fato, em virtude dessa importância, muitas ferramentas têm sido desenvolvidas para dar apoio à realização do processo de descoberta de conhecimento em textos.

1.3.7. Alguns Desafios em Mineração de Dados

Como MD é uma área relativamente recente, diversas dificuldades e desafios estão surgindo continuamente, os quais devem ser superados. Dessa forma, esta seção pretende exemplificar alguns dos principais problemas e direções futuras para o processo de MD.

Alguns desafios importantes que os pesquisadores de MD poderão enfrentar são:

Interação com Usuário Muitas das tecnologias de Mineração de Dados existentes não são realmente interativas e não conseguem incorporar facilmente o conhecimento prévio a respeito de um domínio de aplicação. O uso do conhecimento *a priori* relevante é de grande importância para o processo de MD (Pazzani & Kibler 1992). Pode-se tentar incorporar um Sistema Baseado em Conhecimento para tentar colher o conhecimento de especialistas em uma Base de Conhecimento.

Integração com Outros Sistemas Um sistema de descobrimento isolado pode não possuir muita utilidade se ele não puder ser integrado a outros sistemas, como gerenciadores de Bases de Dados, ferramentas analíticas e de visualização.

Suporte a Novas Tecnologias de Base de Dados Com a evolução da tecnologia de armazenamento, os dados armazenados passarão a conter além de textos e números, objetos gráficos, multimídia, bem como dados não-estacionários (alteração contínua), temporais, entre outros. Gerenciadores de Base de Dados orientados a objetos podem tratar este tipo de problema de armazenamento, facilitando a geração de metadados.

Outro ponto que merece atenção é o crescimento das Bases de Dados multimídia que incluem dados estruturados, semi-estruturados e não-estruturados como áudio,

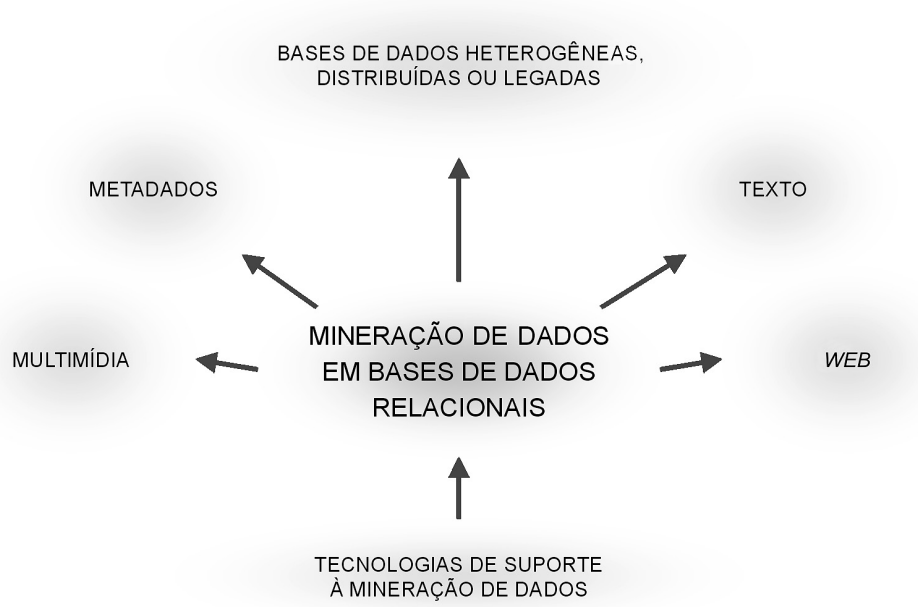

Figura 1.10: Tendências e perspectivas futuras

vídeo, texto e imagens. Uma solução refere-se à extração de dados estruturados das Bases de Dados multimídia e então minerá-los com as técnicas tradicionais. Outra solução é o desenvolvimento de ferramentas que operem diretamente nos dados multimídia.

MD em um Ambiente de Rede e Distribuído O rápido crescimento de recursos disponíveis na internet demanda uma grande necessidade por pesquisas para o desenvolvimento de ferramentas, técnicas e sistemas que possam permitir a realização do processo de MD nesse ambiente conectado e distribuído. Ainda, a tendência da área de MD é guiar-se para um descobrimento de conhecimento "colaborativo", envolvendo uma equipe de analistas e especialistas do domínio que utilizarão Bases de Dados distribuídas pela rede. As pesquisas atuais de agentes inteligentes, é um começo para atingir os desafios impostos à área de MD pelas novas tecnologias de WWW e Base de Dados multimídia.

As discussões apresentadas assumem que os dados minerados estão representados no formato atributo-valor, ou seja, para dados oriundos de diversas fontes uma integração é necessária. Porém, muitas aplicações de dados podem ser distribuídas e gerenciadas por um sistema distribuído necessitando assim de extração de conhecimento em dados distribuídos e heterogêneos que ainda tem recebido pouca atenção da comunidade. Um modelo alternativo para este caso é a implementação de ferramentas no sistema distribuído.

Como qualquer área em desenvolvimento, novos desafios devem ser continuamente superados. Uma síntese dessas tendências em termos de tipos de dados pode ser visualizada na Figura 1.10.

1.4. Considerações Finais

Recentemente, com o desenvolvimento da tecnologia de armazenamento, a quantidade de dados disponível nas Bases de Dados das organizações tem crescido demasiadamente. Este crescimento inviabilizou a utilização de técnicas manuais para análise desses dados. Ao mesmo tempo, as organizações têm visto a possibilidade de utilizar seus dados a fim de

obter conhecimento sobre seus clientes, produtos e parceiros, e assim adquirir vantagem competitiva perante seus concorrentes.

Nesse contexto, os pesquisadores da área de Inteligência Artificial têm direcionado estudos para o processo de Mineração de Dados, que visa automatizar a tarefa de extrair conhecimento útil a partir de grandes volumes de dados.

Este capítulo teve o objetivo de apresentar o processo de Mineração de Dados, detalhando as atividades que devem ser realizadas desde a compreensão do domínio, pré-processamento dos dados, até a avaliação do conhecimento extraído. Também foram descritas as principais tarefas, algumas particularidades da Mineração Visual de Dados e Mineração de Textos assim como alguns desafios para a Área de Mineração de Dados.

Conforme apresentado, pode-se observar que o processo de Extração de Conhecimento é bastante complexo e trabalhoso, pois envolve a execução de muitas tarefas, configuração de diversos parâmetros e grande interação com os usuários. Contudo, o sucesso do processo pode trazer uma recompensa valiosa para as organizações.

Referências

Agrawal, R. & R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Databases*, pp. 487–499. `http://www.almaden.ibm.com/u/ragrawal/pubs.html` (último acesso: 27/01/2004).

Ankerst, M. (2000). *Visual Data Mining*. Tese de Doutorado, Faculty of Mathematics and Computer Science, University of Munich.

Apte, C. & S. Weiss (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems 13*(2–3), 197–210.

Baeza-Yates, R. A. (1998). Searching the www: Challenges and possible solutions. In *Procedings of IBERAMIA'98 – 6th Iberoamerican Conference on Artificial Intelligence*, Lisboa, Portugal, pp. 39–51.

Baeza-Yates, R. A. & B. Ribeiro-Neto (1999). *Modern Information Retrieval*. Addison-Wesley.

Barquini, R. (1996). *Planning and Designing the Warehouse*. New Jersey: Prentice-Hall.

Batista, G. E. A. P. A., A. C. P. L. F. Carvalho, & M. C. Monard (2000). Applying one-sided selection to unbalanced datasets. In *Proceedings of the Mexican Congress on Artificial Inteligence MICAI, Lecture Notes in Artificial Inteligence*, pp. 315–325.

Battista, G. D., P. Eades, R. Tamassia, & L. G. Tollis (1999). *Graph Drawing*. Prentice-Hall.

Bradley, P., U. Fayyad, & O. Mangasarian (1998). Data mining: Overview and optimization opportunities. Technical Report MSR-TR-98-04, Microsoft Research Report, Redmond, WA.

Braga, A. P., A. C. P. L. F. Carvalho, & T. B. Ludermir (2000). *Redes Neurais Artificiais: Teoria e Aplicações*. Rio de Janeiro, Brasil: LTC Press.

Branco, V. M. A. (2003). Visualização como suporte à exploração de uma base de dados pluviométricos. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP, Brasil. Capítulo 2.

Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 577, University of California, Berkeley.

Card, S. K., J. D. Mackinlay, & B. Shneiderman (1999). Information visualization. In S. Card, J. Mackinlay, & B. Shneiderman (Eds.), *Readings in Information Visualization - Using Visualization to Think*, San Francisco, pp. 1–34. Morgan Kaufmann Publ.

Cavalcanti, M., E. Gomes, & A. Pereira (2001). *Gestão de Empresas na Sociedade do Conhecimento*. Editora Campus.

Chaudhuri, S. & U. Dayal (1997). An overview of data warehousing and OLAP technology. *SIGMOD Record 26*(1), 65–74.

D. Hand, H. Mannila, P. S. (2001). *Principles of Data Mining*. Cambridge, CA: MIT Press.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli

(Eds.), *Multiple Classifier Systems, Lecture Notes in Computer Science*, Volume 1857, pp. 1–15.

Dong, G. & J. Li (1998). Interestingness of discovered association rules in terms os neighborhood-based unexpectedness. *Lecture Notes in Artificial Intelligence, 1394*, 72–86.

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth (1996a). From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery & Data Mining*, pp. 1–34.

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth (1996b). Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 82–88.

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth (1996c). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM 39*(11), 27–34.

Fertig, C. S., A. A. Freitas, L. V. R. Arruda, & C. Kaestner (1999). A fuzzy beam-search rule induction algorithm. In *Proceedings of the Third European Conference (PKDD-99) Lecture Notes in Artificial Intelligence 1704*, pp. 341–347.

Félix, L. C. M., S. O. Rezende, M. C. Monard, & C. W. Caulkins (2000). Transforming a regression problem into a classification problem using hybrid discretization. *Computación y Sistemas 4*, 44–52.

Freitas, A. A. (1998a). A multi-criteria approach for the evaluation of rule interestingness. In *Proceedings of the International Conference on Data Mining*, pp. 7–20.

Freitas, A. A. (1998b). On objective measures of rule surprisingness. In *Proceedings of the 2nd European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD-98)*, Volume 1510, pp. 1–9.

Furnas, G. (1986). Generalized fisheye views. In *Proceedings of Human Factors in Computing Systems (CHI'86)*, pp. 18–23.

Ganesh, M., E.-H. Han, V. Kumar, S. Shekhar, & J. Srivastava (1996). Visual data mining: Framework and algorithm development. Technical Report TR-96-2001, Department of Computer Science, University of Minnesota, Minneapolis.

Gardner, S. R. (1998). Building the data warehouse. *Communications of the ACM 41*(9), 52–60.

Garvin, D. A., P. R. Nayak, A. N. Maira, & J. L. Bragar (1998). *Aprender a Aprender*. HSM Management.

Glymour, C., D. Madigan, D. Pregibon, & P. Smyth (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery 1*, 11–28.

Gupta, V. R. (1997). An introduction to data warehousing. Technical report, System Services Corporation, Chicago, Illinois.

Habn, U. & I. Mani (2000). The challenges of automatic summarization. *IEEE Computer 33*(11), 29–36.

Horst, P. S. (1999). Avaliação do conhecimento adquirido por algoritmos de aprendizado de máquina utilizando exemplos. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP, Brasil.

Hull, D. A. (1994). Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of SIGIR'94 – 17th ACM International Conference on Research and Development in Information Retrieval*, pp. 282–289.

I. Sarafis, A. M. S. Zalzala, P. W. T. (2002). A genetic rule-based data clustering toolkit. In *Congress on Evolutionary Computation (CEC)*, Honolulu, USA.

Inmon, W. H. (1997). *Como construir o Data Warehouse*. Rio de Janeiro: Editora Campus.

Inselber, A. & B. Dimsdale (1990). Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *Proceedings of Visualization'90*, pp. 361–370.

J. Han, M. K. (2001). *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann.

Joachims, T., D. Freitag, & T. Mitchell (1997). WebWatcher: A tour guide for the world wide web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'97)*, pp. 770–777.

Kearns, M. J. & U. V. Vazirani (1994). *An introduction to computational learning theory*. Ellis Horwood.

Keim, D. A. (2000). Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics 6*(5), 59–78.

Keim, D. A. (2001). Visual exploration of large data sets. *Communications of the ACM 44*(8), 38–44.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*(1), 1–8.

Keim, D. A. & H. P. Kriegel (1996). Visualization techniques for mining large databases: A comparison. *IEEE Transacations on Knowledge and Data Engineering 8*(6), 923–938.

Kimball, R. (1997). *Data Warehouse Toolkit*. São Paulo: Makron Books.

Kock Jr., N. F., R. J. McQueen, & M. Baker (1996). Learning and process improvement in knowledge organisations: A critical analysis of four contemporary myths. *The Learning Organization*, 31–40.

Kock Jr., N. F., R. J. McQueen, & J. L. Corner (1997). The nature of data, information and knowledge exchanges in business processes: Implications for process improvement and organizational learning. *The Learning Organization 4*(2), 70–80.

Kohavi, R., D. Sommerfield, & J. Dougherty (1996). Data mining using $\mathcal{MLC}^{++}$: A machine learning library in C^{++}. In *Tools with Artificial Intelligence*, pp. 234–245. IEEE Computer Society Press.

Lavrac, N., P. Flach, & B. Zupan (1999). Rule evaluation measures: A unifying view. In S. Džeroski & P. Flach (Eds.), *Ninth International Workshop on Inductive Logic Programming (ILP'99)*, Volume 1634 of *Lecture Notes in Artificial Intelligence*, pp. 174–185. Springer-Verlag. `http://link.springer.de/link/service/series/0558/papers/1634/16340174.p%df` (último acesso: 27/01/2004).

Lee, H. D. (2000). Seleção e construção de features relevantes para o aprendizado de máquina. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP, Brasil.

Liu, B. & W. Hsu (1996). Post-analysis of learned rules. *AAAI 1*, 828–834.

Mackinlay, J. D., G. G. Robertson, & S. K. Card (1991). The perspective wall: Detail and context smoothly integrated. In *Proceedings of Human Factors in Computing Systems (CHI'91)*, pp. 173–179.

Mannila, H. (1997). Data mining: Machine learning, statistic and databases. In *Proceedings of the 8th International Conference on Scientific and Statistical Database Management*, pp. 1–8.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM 42*(11).

Monard, M. C. & J. A. Baranauskas (2003). Conceitos sobre aprendizado de máquina. In S. O. Rezende (Ed.), *Sistemas Inteligentes: Fundamentos e Aplicações*, pp. 89–114. Manole.

Nowell, L., S. Havre, B. Hetzler, & P. Whitney (2002). Themeriver: Visualizing the-

matic changes in large document collections. *IEEE Transaction Visualization and Computer Graphics 8*(1), 9–20.

Pazzani, M. & D. Kibler (1992). The utility of knowledge in inductive learning. *Machine Learning, 9 9*, 57–94.

Pazzani, M., S. Mani, & W. Shankle (1997). Comprehensible knowledge discovery in databases. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum*, pp. 596–601.

Pazzani, M. J. (2000). Knowledge discovery from data? *IEEE Intelligent Systems 15*(2), 10–13.

Piatetsky-Shapiro, G. & C. J. Matheus (1994). The interestingness of deviations. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD-94)*, pp. 23–36.

Pickett, R. M. & G. G. Grinstein (1988). Iconographic displays for visualizing multidimensional data. In *Proceedings of IEEE Conference on Systems, Man and Cybernetics'88, Piscataway, NJ*, pp. 361–370.

Poe, V., P. Klauber, & S. Brobst (1998). *Building a Data Warehouse for Decision Support*. New Jersey: Prentice-Hall.

Rao, R. & S. K. Card (1994). The table lens: Merging graphical and symbolic representation in an interactive focus+context visualization for tabular information. *Human Factors in Computing System – CHI'94*, 318–322.

Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações* (1 ed.). Barueri, SP: Manole.

Rezende, S. O., R. B. T. Oliveira, L. C. M. Félix, & C. A. J. Rocha (1998). Visualization for knowledge discovery in database. In *Ebecken, N.F.F. (ed.) Data Mining. WIT Press, England*, pp. 81–95.

Rezende, S. O., J. B. Pugliesi, E. A. Melanda, & M. F. Paula (2003). Mineração de dados. In S. O. Rezende (Ed.), *Sistemas Inteligentes – Fundamentos e Aplicações*, pp. 307–335. Editora Manole.

S. Mitra, S. K. Pal, P. M. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks 13*(1), 3–14.

Silberschatz, A. & A. Tuzhilin (1995). On subjective measures of interestingness in knowledge discovery. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining 1*, 275–281.

Stolte, C., D. Tang, & P. Hanrahan (2002). Polaris: A system for query, analysis and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics 8*(1), 52–65.

Swayne, D. F., D. Cook, & A. Buja (1992). User's manual for XGobi: A dynamic graphics program for data analysis. Bellcore technical memorandum, Bell Labs.

T. Y. Lin, N. C. (1997). *Rough Sets and Data Mining*. Norwell, Massachusetts: Kluwer Academic Publishers.

Thuraisingham, B. (1999). *Data Mining: Technologies, Techniques, Tools, and Trends*. CRC Press.

Uysal, I. & H. A. Güvenir (1999). An overview of regression techniques for knowledge discovery. *The Knowledge Engineering Review 14*(4), 319–340.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vasileios, H., L. Gravano, & A. Maganti (2000). An investigation of linguistic features and clusters algorithms for topical document clustering. In *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 224–231.

W. Frawley, G. Piatetsky-Shapiro, C. M. (1992). Knowledge discovery in databases: An overview. *AI Magazine*.

Wei, J. M. (2003). Rough set based approach to selection of node. *International Journal of Computational Cognition 1*(2), 25–40.

Weiss, S. M. & N. Indurkhya (1995). Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research 3*, 383–403.

Weiss, S. M. & N. Indurkhya (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Witten, I. H. & E. Frank (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

Wong, P. (1999). Visual data mining. *IEEE Computer Graphics and Applications 19*(5), 20–21.

Zhou, Z. H. (2003). Three perspectives of data mining. *Artificial Intelligence jounal 143*(1), 139–146.