ANÁLISE PREDITIVA E DATA MINING

CREDENCIADA JUNTO AO MEC PELA PORTARIA
N 3.455 DO DIA 19/11/2003

SUMÁRIO

NOSSA HISTÓRIA

A nossa história inicia com a realização do sonho de um grupo de empresários, em atender à crescente demanda de alunos para cursos de Graduação e Pós-Graduação. Com isso foi criado a nossa instituição, como entidade oferecendo serviços educacionais em nível superior.

A instituição tem por objetivo formar diplomados nas diferentes áreas de conhecimento, aptos para a inserção em setores profissionais e para a participação no desenvolvimento da sociedade brasileira, e colaborar na sua formação contínua. Além de promover a divulgação de conhecimentos culturais, científicos e técnicos que constituem patrimônio da humanidade e comunicar o saber através do ensino, de publicação ou outras normas de comunicação.

A nossa missão é oferecer qualidade em conhecimento e cultura de forma confiável e eficiente para que o aluno tenha oportunidade de construir uma base profissional e ética. Dessa forma, conquistando o espaço de uma das instituições modelo no país na oferta de cursos, primando sempre pela inovação tecnológica, excelência no atendimento e valor do serviço oferecido.

1. Introdução

Os avanços obtidos nas áreas de software e hardware possibilitaram a criação de aplicações comerciais e científicas capazes de processar grandes volumes de dados. Por exemplo, o sistema que é usado por uma grande empresa do setor petroquímico para processar as compras de suprimentos diversos processa milhões de transações diariamente, produzindo um volume de dados que pode chegar a mais de uma dezena de Gigabytes.

De acordo Send e Jacob (1998), cada vez mais, as empresas vêm fazendo grandes investimentos em aplicativos e equipamentos usados para o armazenamento, integração, análise e gerenciamento dos seus dados. Isto se deve a uma mudança de filosofia, pois, atualmente, as bases de dados não são mais consideradas simples repositórios de informações, mas sim, um importante patrimônio da organização.

Os dados gerados pelas organizações de médio e grande porte superam a capacidade humana de interpretar, analisar e compreender tanta informação. Por isso, são necessárias novas ferramentas e técnicas capazes de analisar automaticamente o volume de dados produzidos, fornecendo o conhecimento necessário para auxiliar nos processos decisórios.

O custo de uma má qualidade desses dados pode ser decisivo para o sucesso de uma empresa. Com isso, é essencial o estudo da Administração de Dados para garantir a qualidade dos dados que são inseridos e recuperados em uma base de dados. A administração de dados consiste no desenvolvimento e execução de estratégias, práticas e procedimentos para o processo de gerência do ciclo de vida completo dos dados em uma empresa [DAMA International, 2006].

A área conhecida por Extração de Conhecimento de Base de Dados ou Knowledge Discovery in Databases (KDD) surgiu para auxiliar a análise de grande volume de dados. Os trabalhos neste segmento objetivam o estudo da aplicação de novas metodologias, ferramentas e técnicas capazes de extrair conhecimento contido em grandes volumes de dados.

1.1 Processo decisório e seu histórico

Por tratar-se de tema longo e muito abrangente, o surgimento do processo decisório e sua evolução serão apresentados de forma resumida com base em BISPO e CAZARINI (1998) destacando os mais importantes pontos de sua evolução.

Segundo Bispo e Cazarini (1998), o homem sempre procurou alguma ajuda para seu processo decisório desde o começo da civilização. Considerava-se que pessoas com “místicos poderes” teriam livre e direto contato com os seres considerados divinos e que todas as orientações dadas por essas pessoas eram, também, divinas. Dessa forma, as decisões eram consideradas sábias e se, no entanto, os resultados não fossem os esperados, tais erros significavam que as divindades estariam insatisfeitas. Nesta época, as entidades divinas e as pessoas que as representavam tinham forte influência nas decisões.

Segundo Pereira e Fonseca (1997), no início do século XX, os critérios usados para tomada de decisão se concentravam no executivo maior, que muitas vezes além de ser o dono do negócio, possuía o privilégio da escolha que acreditasse ser a melhor para a empresa e seus trabalhadores. Isso devido o entendimento existente na época de que os trabalhadores eram pessoas sem capacidade e não estavam preparados para tomarem decisões, sendo avaliados por sua produção e descartados quando não produziam o esperado pela empresa. Acreditava-se que apenas os executivos de alto escalão tivessem capacidade para sábias decisões devido ao amplo conhecimento a eles atribuído sobre todas as alternativas possíveis e suas consequências.

Somente no início dos anos 60, essa perspectiva mudou com o surgimento do movimento conhecido como Escola de Relações Humanas, oriundo da contribuição da Psicologia Social à Teoria da Administração. A partir desse movimento os trabalhadores passam a ser reconhecidos como alguém capaz de pensar, de decidir e de ser motivado (PEREIRA e FONSECA, 1997), ou seja, não mais se restringia ao alto escalão a capacidade de decidir.

1.2 Evolução dos estudos

Segundo Simon (1986), o estudo do processo decisório, principalmente após a Segunda Guerra, ganhou muita força, mas especificamente centrada no modelo racional, seguindo uma teoria prescritiva.

Na Teoria da Administração, o processo decisório foi esquecido até por volta da metade do século passado devido a ciência administrativa ter nascido tendo como base um conjunto de valores funcionais e mecanicistas, e as organizações foram concebidas somente como instrumentos técnicos, tendo como objetivo principal a maximização dos lucros e dos resultados.

Os modelos de tomada de decisão e sua classificação surgem pela divisão do estudo e/ou abordagem do processo decisório pelas diferentes escolas de Administração. A teoria da decisão, hoje, assume um privilegiado lugar no pensamento administrativo, contemplando os níveis estratégico, tático e operacional. A partir de Simon, a teoria da decisão vem conquistando sua relevância e sua especificidade, deixando, ao longo do tempo, a abordagem simplesmente quantitativa e adaptando-se a nova realidade decorrente das complexas mudanças pelas quais vêm passando, nas últimas décadas, as organizações.

Hall (2004) afirma que o processo decisório está envolvido de pressões imediatas sobre o tomador das decisões, a análise do tipo do problema e de suas dimensões básicas, da busca de soluções variadas e do exame minucioso de suas consequências, inclusive a antecipação dos diversos tipos de conflito pós-decisório e a escolha final.

Procura-se, através de uma curta retrospectiva histórica, apresentar com a sistematização das características principais das diferentes etapas ou fases da evolução do processo decisório e a forma como foi discutido ao longo dos anos, alguns destaques pinçados pelas escolas e pelos ideólogos em busca do aperfeiçoamento da gestão empresarial.

Estes destaques podem ser identificados em alguns dos principais estudiosos que se destacaram em abordagens gerenciais, tais como: Friedrich Wislow Taylor, Peter Drucker, Earnest Archer, Joseph Newman e Herbert Simon, que, certamente, encontram-se entre os precursores dessa sistêmica abordagem.

Fases	Método Científico	Earnest Archer	Peter Drucker	Herbert Simon	Joseph Newman	Abordagem Sistêmica
1	Observação	Monitoração do ambiente decisório		Inteligência (busca de condições que pedem por solução)		Escolha do problema
2	Formulação do problema	Conceituação de problemas ou situações	Definição de problema		Reconhecimento do sistema que requer ação de decisão	Definição e quantificação do problema
3	Estabelecimento dos objetivos	Objetivos de decisão	Definição de expectativas			
4	Determinação das causas	Diagnóstico do problema ou situação				Determinação de relações causais entre os fatos para soluções
5	Formulação de hipóteses	Desenvolvimento de Soluções alternativas	Desenvolvimento de soluções alternativas	Invenção, desenvolvimento e análise de curso de ação	Identificação e desenvolvimento de caminhos alternativos de ação	Determinação de tentativas opcionais de solução
6	Metodologia	Definição de metodologia ou critério para avaliar alternativas				
7	Teste de hipóteses	Avaliação das soluções alternativas			Avaliação de alternativas	Teste das possíveis soluções
8	Formulação de conclusões	Escolha da melhor alternativa		Seleção de um caminho de ação	Escolha de uma das alternativas	
9	Comunicação de Resultados	Implementação da melhor alternativa	Saber o que fazer com a decisão	Implementação do caminho de ação selecionado	Implementação do caminho de ação selecionado	Documentação dos procedimentos

Figura 1.2 – Evolução do processo decisório

1.3 A tomada de decisão e os sistemas de apoio

A sobrevivência das empresas e a situação das pessoas que direta ou indiretamente estão a ela ligadas, empregados, fornecedores, clientes ou acionistas, são afetadas diretamente pelas decisões gerenciais. Assim, o tomador de decisões é atingido por vários fatores de influência, inclusive por cobranças das pessoas atingidas para obtenção de um resultado de sucesso. Cada uma dessas pessoas solicita soluções diferenciadas e, possivelmente antagônicas, como solução de um problema, e é preciso que prioridades sejam estabelecidas quando estamos diante de posições e objetivos diferentes, antagônicos ou disputas de informações e recursos. É preciso transformar os objetivos da organização em objetivos gerais para todos os

membros da empresa, buscando o compartilhamento da participação e da visão do futuro, buscando a satisfação dos usuários e clientes, não se descuidando, no entanto, dos demais grupos de interesses - acionistas e empregados.

Segundo Pereira e Fonseca (1997), no dia a dia, a viabilização desse processo envolto em conflitos de interesses, exige liderança, habilidade de negociação permanente, objetivos compartilhados e comunicação efetiva.

Nesse contexto, segundo Gates (1997), a informação é desejada e existe quem esteja disposto a pagar por ela, não é mensurável em tangível, porém é um valioso produto no mundo moderno porque proporciona poder.

É pela informação que se tem a possibilidade de suportar melhor o processo decisório, sendo função das diversas ferramentas que darão esse suporte ao processo, obter as informações necessárias de forma confiável, rápida e mostrá-las de maneira compreensível.

Segundo Power (2002), a conceituação de suporte computacional à decisão surge com o desenvolvimento de duas vertentes de pesquisa: estudos teóricos sobre Processo de Tomada de Decisão Organizacional, feitos durantes as décadas de 50 e 60 no Carnegie Institute of Technology e os trabalhos feitos com Sistemas Computacionais Interativos, durante a década de 60 no Massachuttes Institute of Technology.

1.4 Avanço das análises no decorrer dos tempos

Segundo Fisher (2006), Costa (1997) e Pearson e Shim (1995), os pioneiros SADs - Sistemas de Apoio à Decisão - surgiram em 60 e 70 para suporte aos tomadores de decisão em soluções de problemas gerenciais que não fossem estruturados.

Durante esse período, os sistemas de computador que davam suporte à decisão eram desenvolvidos, primeiramente, para auxílio na resolução de problemas gerenciais específicos e, depois, aperfeiçoados a fim de incorporar outros problemas gerenciais. No entanto, não houve possibilidade de que, com um sistema desses, se chegasse a um bom suporte ao processo de tomada de decisão por tratar-se de um processo dinâmico, onde o fornecimento das informações tem que ocorrer no momento certo.

Apenas nos anos 80, com o crescimento na utilização dos Sistemas de Gerenciamento de Banco de Dados - SGDB - é que foi possível um acesso melhor aos dados disponíveis, à formatação desses dados e a construção de consultas e relatórios de maneira mais rápida, prática e barata. No entanto, quando se fazia necessária uma análise mais profunda dos dados essas eram feitas fora de um sistema computacional, ou seja, ainda faltava o desenvolvimento de uma ferramenta que auxiliasse realmente os tomadores de decisão.

Apesar dos avanços obtidos, o grande problema era que a modelagem dos dados se baseava na estrutura de processos quando deveria se basear na estrutura de negócios. Começam então a surgir os primeiros sistemas desenvolvidos especialmente para gerentes: os Sistemas de Informações para Executivos - Executive Information Systems (EIS) - , mas as empresas e os negócios evoluem mais rápido do que esses sistemas.

Ao longo do tempo, com o crescimento das empresas e o aumento dos negócios, o volume de dados armazenados também aumentou e também surgiu a necessidade de aumento do número de gerentes ou de divisão de tarefas em diversos níveis gerenciais. Com isso ocorreu a necessidade de crescimento da análise de dados, de respostas rápidas, confiáveis e adaptáveis as novas formas de gerenciamento das empresas e negócios. Novos métodos de gestão empresarial foram elaborados, tais como: Reengenharia (HAMMER, 1994) e o Gerenciamento pela Qualidade Total (DEMING e SCHERKENBACH, 1992).

Segundo Fisher (1998), quando as necessidades de progresso tecnológico e as necessidades de mercado convergem, estes realizam mudanças primordiais na prática dos negócios e a evolução das Tecnologias da Informação possibilitou muitas empresas a encarar um ambiente cada vez mais competitivo.

Segundo Bispo e Cazarini (1998), nos anos 90, diversos sistemas para apoio e suporte às decisões nas empresas foram desenvolvidos. Dentre essas novas ferramentas está o ERP (Enterprise Resource Planning), como ferramenta de gestão integrada utilizada para gerenciamento no ambiente operacional e, também, uma nova geração de Sistemas: o Data Warehouse, o OLAP e o Data Mining que vêm sendo utilizadas para o gerenciamento no ambiente gerencial.

1.5 Cenário atual

Com as ferramentas Data Warehouse e OLAP, os relatórios e as consultas passam a ser feitos pelos próprios usuários dos sistemas sem que haja a necessidade de um profundo conhecimento em tecnologias computacionais, sendo sua confecção barata, rápida, confiável e adaptável aos modelos diversos de negócios. Ao usarem essas ferramentas os gerentes gastam um tempo bem menor manipulando os dados e construindo modelos conforme suas necessidades, usando melhor o tempo para as necessárias análise e soluções de problemas.

O surgimento dessa nova geração de Sistemas de Apoio a Decisão não inutiliza nem substitui os tradicionais e antigos sistemas. Na maioria das vezes os antigos e os novos sistemas atuam em conjunto no auxílio a gerência dos negócios, na solução de problemas e na elaboração de estratégias novas. Informações obtidas através do Data Mining ou do OLAP podem alimentar qualquer sistema que trabalhe em otimização ou na linha de pesquisa operacional, como o do próximo tópico a ser estudado.

1.6 Data Warehouse

Vários pesquisadores das áreas de inteligência artificial, machine learning (WEISS e KULIKOWSKI, 1991), estatística, base de dados espaciais, aquisição de conhecimentos, visualização de dados, entre outras, consideram a possibilidade de obtenção de informações valiosas e extração de conhecimentos geradas por grandes massas de dados, como sendo um ponto chave de pesquisa, e devido a essa importância, têm demonstrado grande interesse do assunto, que é conhecido como Data Mining.

Visando facilitar o trabalho de Mineração de Dados, aponta-se como fundamental uma análise criteriosa dos dados armazenados nas várias bases.

Além disso, as empresas de grande porte possuem um enorme volume de dados que estão espalhados em vários sistemas diferentes, não possibilitando a busca de informações que permitam a tomada de decisão baseada em um histórico dos dados, o que possibilitaria a identificação de tendências e o posicionamento das empresas estrategicamente para a competitividade e para a maximização dos lucros.

Dessa forma, foi introduzido no mercado um novo conceito que permite reagrupar esses dados espalhados pelos diversos sistemas e reorganizá-los estrategicamente: o Data Warehouse.

Data Warehouse é uma organização de banco de dados para análises e business intelligence, surgiu como um conceito acadêmico, criado na década de 1980. Sua arquitetura e desenho é voltado para processamento e armazenamento de altos volume de dados.

Um conceito que define um Banco de Dados com capacidade de armazenar e organizar um grande volume de dados; responsável por criar e organizar relatórios por meio de históricos, que podem ajudar uma empresa obter insights e auxílio na tomada de decisões importantes. Traduzindo diretamente ao português temos “Armazem de Dados”

Segundo Singh (2001), “um ambiente de suporte a decisão que alavanca os dados armazenados em diferentes fontes e os organiza e entrega aos tomadores de decisões da empresa, independente de plataforma que utilizam ou de seu nível de qualificação técnica”.

Enfim, um conceito que consiste em organizar dados corporativos da melhor forma, a fim de subsidiar os gerentes e diretores das empresas com informações para a tomada de decisão, num banco de dados paralelo aos sistemas operacionais.

1.6.1 Objetivo

O objetivo do Data Warehouse é centralizar os dados retirados de diversas fontes e facilitar a consulta. Os dados podem ser extraídos de:

- Planilhas
- ERP’s
- CRM’s, etc

Com diferentes formatos:

- Banco de dados
- XLS
- TXT
- CSV

- JSON, etc

Após a extração, os dados normalmente são acomodados na Staging Area, que é uma área destinada aos processos de qualidade e padronização dos dados. Posteriormente podem ser direcionados ao Enterprise Data Warehouse (EDW) ou aos Data Marts diretamente. Com isso, é possível buscar todas as informações importantes em um único lugar – organizado e atualizado, criado com foco em facilitar a consulta.

COMPLEMENTO

https://www.youtube.com/watch?v=Kt_Q__gO47U

https://www.youtube.com/watch?v=YJiqvR6n_vA

2. DADOS E SUA RELEVÂNCIA

Dados se constituem como a matéria-prima para que processos de mineração ocorram. Essa matéria-prima pode se manifestar, basicamente, de duas formas: estruturada e não estruturada. A forma como os dados estão disponíveis para a realização da mineração é importante para determinar o tipo de tarefa de mineração que é possível resolver, o tipo de conhecimento factível de ser descoberto e o tipo de técnica de mineração aplicável. Além disso, a quantidade e a qualidade dos dados disponíveis também são determinantes para o sucesso da mineração de dados.

2.1 Dados estruturados

Tipicamente, uma base de dados usada em sistemas informatizados convencionais é organizada de forma que se tenham dados armazenados em estruturas tabulares, em que as linhas armazenam uma ocorrência de um evento caracterizado por um conjunto de colunas que representam características que descrevem um exemplar (instância) daquele evento. Na maioria dos casos, os dados estruturados são resultantes de processos de geração de dados inerentes a sistemas transacionais (Seção 1.5) ou resultantes de observações e processos de medição. Esses dados geralmente são armazenados em um conjunto de tabelas relacionadas entre si. Exemplos de processos de geração de dados são:

- Um sistema bancário que armazena informações sobre seus clientes, sobre suas respectivas contas bancárias e sobre as transações executadas sobre tais contas.
- O trabalho de funcionários de um departamento de recursos humanos de uma empresa que observam o comportamento de candidatos a uma vaga de emprego, fazendo anotações sobre suas atitudes durante a realização de atividades de um processo de seleção.
- Uma empresa que deseja lançar um produto novo e, para isso, realiza pesquisas sobre sua aceitação com um conjunto específico de pessoas que representa o público-alvo.
- Uma corrida de Fórmula 1 na qual se medem os tempos que cada piloto gasta para chegar em diferentes trechos de um circuito.

Considere uma base de dados na qual são armazenados dados sobre candidatos a vagas de emprego em um restaurante. Nela são armazenados dados que descrevem a unidade observacional: o candidato. Uma forma de estruturar os dados referentes a cada candidato é apresentada na Figura 2.1, em que cada uma de todas as *n* observações possui um identificador (ID), seis atributos descritivos e um atributo de rótulo.

	atributos						
ID	*descritivos*						*rótulo*
01	João	masculino	viúvo	65	garçom	12.000,00	SIM
02	Maria	feminino	solteiro	32	cozinheiro	4.200,00	SIM
...	...	...	...	...	...	...	...
n	Pedro	masculino	casado	18	entregador	1.650,00	NÃO

Figura 2.1 - Base de dados de observações documentadas a partir de um processo seletivo

À unidade observacional e seus descritores é possível e útil associar significado, dando a ela o caráter informacional. Uma representação resumida para essa associação é:

Candidato = {Matrícula, Nome, Sexo, Estado civil, Idade, Especialidade, Pretensão salarial, Experiência}

Podendo a base de dados ser semanticamente representada pela Figura 2.1.1.

	atributos						
ID	*descritivos*						*rótulo*
Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão salarial	Experiência
01	João	masculino	viúvo	65	garçom	R$ 12.000,00	SIM
02	Maria	feminino	solteiro	32	cozinheiro	R$ 4.200,00	SIM
...	...	...	...	...	...	...	...
n	Pedro	masculino	casado	18	entregador	R$ 1.650,00	NÃO

Figura 2.1.1 - Base de dados sobre candidatos com associação de significado

2.2 Dados não estruturados

Muitos dos dados disponíveis para análise e extração de informação e conhecimento estão apresentados de forma não estruturada, a exemplo de textos, imagens, vídeos e sons.

Uma coleção de textos (ou documentos) compõe uma base textual, que pode ser usada como entrada em um processo de mineração de dados. Um exemplo de dados não estruturados com dois exemplares (dois dados, dois documentos) em uma base textual é mostrado na Figura 2.2. Os textos versam sobre postagens extraídas da rede social de um restaurante, que podem ser analisadas em um processo de mineração de dados (por exemplo, ser classificadas em polaridades como positiva ou negativa; ou agrupadas por similaridade de assuntos).

```
Ambiente agradável e tranquilo. Comida e música com qualidade. Adoramos o Filé
à Parmegiana.

O Filé à Parmegiana da cidade. Ambiente agradável e qualidade no atendimento.

O Filé à Parmegiana com fritas é uma delícia.
```

Figura 2.2 – Exemplo de textos (dados não estruturados)

Os dados presentes em uma imagem se referem a valores relacionados com um sistema de cores e associados a uma estrutura matricial no arquivo da imagem. A informação se torna presente a partir das relações de vizinhança e disposição das cores que, a partir de uma interpretação, manifestam-se em formas e texturas que representam conceitos abstratos ou concretos do mundo real. Um vídeo, de forma simplificada, pode ser visto como uma sequência de imagens que, em associação à dimensão “tempo”, manifestam também conceitos do mundo real. De forma similar, um arquivo de som é uma composição de ondas que se manifestam em uma informação audível.

Para fins de mineração de dados, dados não estruturados precisam passar por uma etapa de pré-processamento, de forma que uma representação adequada lhes seja produzida.

2.3 Relevância

Quando aplicada em uma empresa, a mineração de dados melhora a interação entre empresa e cliente, aumenta vendas e dirige as estratégias de marketing. A mineração de dados, porém, pode ser aplicada a qualquer massa de dados, sejam eles oriundos da Medicina, Economia, Astronomia, Geologia, entre outras áreas de estudo. A relevância deste trabalho fundamenta-se na importância da adoção de técnicas de mineração de dados para melhorar a qualidade de dados em um SGBD, como parte do trabalho de Administração de Dados.

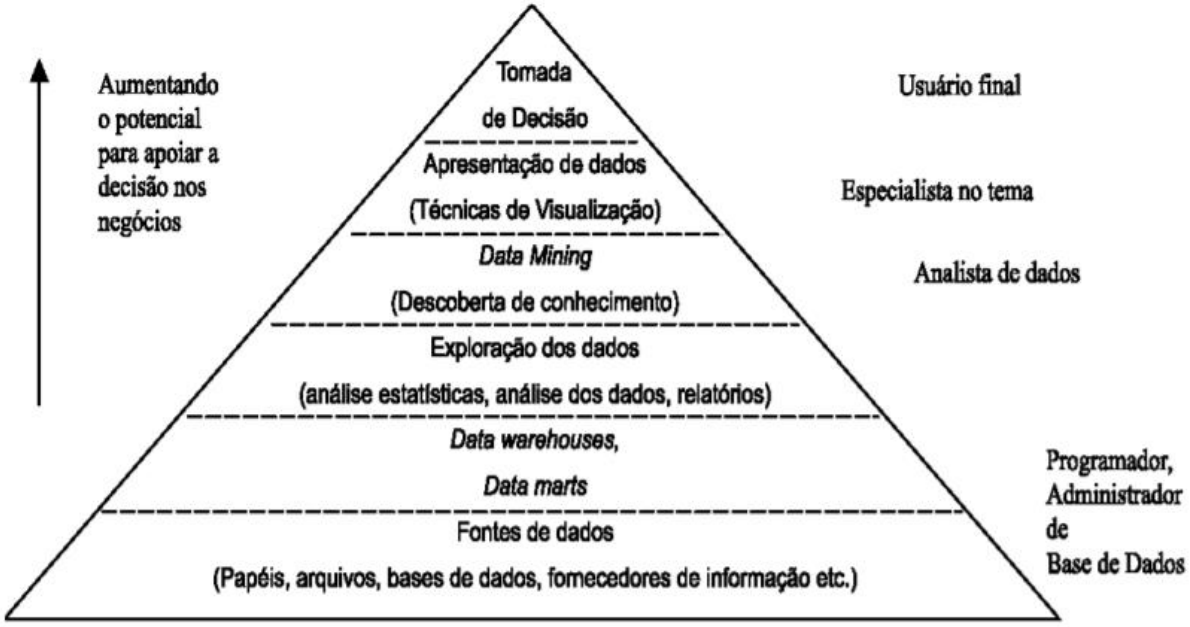

Figura 2.3 – Evolução do valor estratégico de bases de dados

A Figura 2.3 mostra o posicionamento lógico de diferentes fases da tomada de decisão com seu valor potencial para as dimensões tática e estratégica de uma organização. Em geral, o valor da informação para apoiar a tomada de decisão aumenta a partir da base da pirâmide. Uma decisão baseada em dados nas camadas mais baixas, onde há tipicamente milhões de registros de dados, não possui muito valor agregado; já aquela apoiada em dados altamente resumidos nas camadas superiores da pirâmide tem probabilidade de alto valor estratégico.

Da mesma forma, encontram-se diferentes usuários nas diferentes camadas. Um administrador, por exemplo, no nível operacional, trabalha primariamente com informações diárias e operações de rotina, encontradas em arquivos e bases de dados, na base da pirâmide informacional. Esses criam dados. Enquanto analistas de negócios e executivos, responsáveis por indicarem direções, formulam estratégias e táticas, supervisionando a sua execução, e estes necessitam de informações de maior qualidade. Preocupam-se com tendências, padrões, ameaças, pontos fortes e fracos, oportunidades, informação de mercado, entre outros. Necessitam de informações internas e externas. São os que demandam dados analisados com alto valor agregado, as do topo da pirâmide.

2.4 Processo de descoberta de dados – Processo KDD

Nas últimas décadas, todo o mundo tem armazenado uma considerável quantidade de dados, superando consideravelmente as nossas habilidades de interpretação, gerando uma necessidade de criação de técnicas e ferramentas que automatizem e analisem a base de dados de maneira inteligente (FAYYAD, 1996). Essas ferramentas e técnicas que procuram transformar os dados armazenados em conhecimento, são o objetivo do chamado Knowledge Discovery in Databases - KDD (descoberta de conhecimento em bases de dados).

Um número crescente de publicações vem se dedicando ao tema.

Segundo Fayyad (1996), o termo Knowledge Discovery in Databases ou KDD, foi criado em 1989 como referência ao amplo processo para encontrar conhecimento nos dados e enfatizar uma aplicação em especial - o método Data Mining (Mineração de Dados). KDD é todo processo de descoberta de conhecimento útil nos dados, enquanto Data Mining refere-se à aplicação de algoritmos para extração de modelos dos dados. No entanto, os termos KDD e Data Mining foram considerados como sinônimos por muitos autores até 1995.

Dessa forma, cabe ressaltar que o processo KDD é dependente de uma nova geração de técnicas e ferramentas de análise de dados envolvendo diversas etapas. A principal etapa desse processo chama-se Data Mining ou Mineração de Dados, também conhecida como reconhecimento de padrões ou processo de arqueologia de dados (CHEN; HAN; YU, 1996).

A relação existente entre KDD e Data Mining esta representado na Figura 2.4 abaixo:

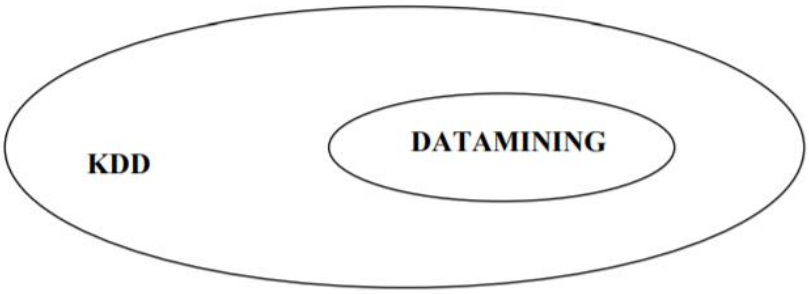

Figura 2.4 – Diferença entre KDD e Data Mining

Assim, o processo global para achar e interpretar modelos extraídos de dados é chamado de processo KDD, tipicamente iterativo e interativo, que envolve aplicações específicas repetidas de métodos ou algoritmos Data Mining e a interpretação dos padrões gerados por estes algoritmos (FAYYAD, 1996).

2.5 As etapas do processo

KDD é um processo de descoberta de conhecimento em bases de dados que envolvem uma diversificada abrangência, como: estatística, banco de dados, matemática, visualização de dados, inteligência artificial e reconhecimento de padrões. Este processo utiliza técnicas, métodos e algoritmos com origem dessas áreas, em que o principal objetivo é a extração do conhecimento partindo de grandes bases de dados.

Sendo o processo de KDD um conjunto de atividades contínuas para o compartilhamento do conhecimento descoberto a partir de bases de dados, segundo FAYYAD (1996), esse conjunto é composto de 5 (cinco) etapas:

- Seleção dos dados;
- Pré-processamento e limpeza dos dados;
- Transformação dos dados;
- Data Mining;
- Interpretação e avaliação dos resultados.

Etapas que podem ser visualizadas através da Figura 2.5, a seguir:

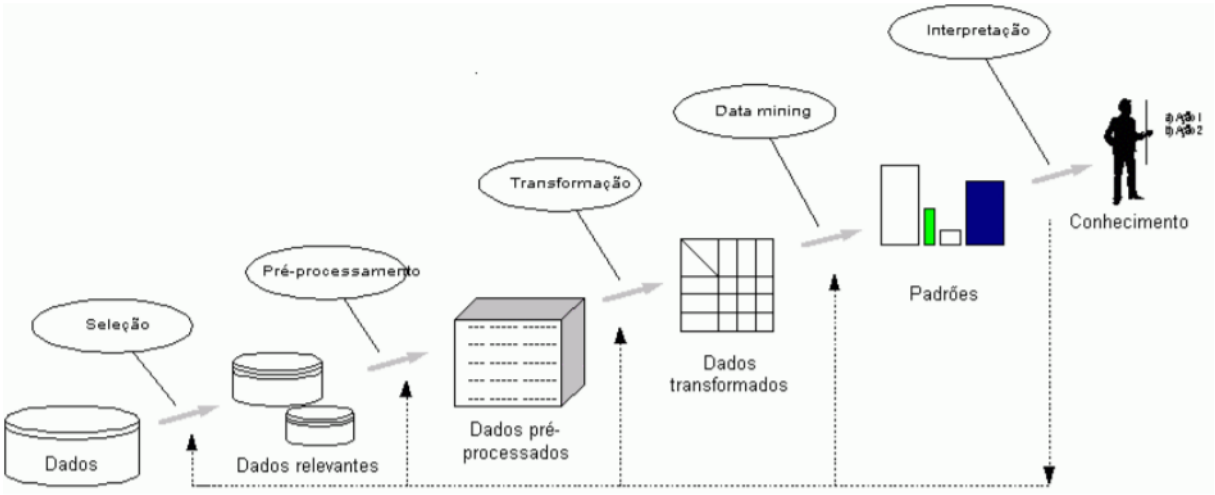

Figura 2.5 – Processo KDD

O processo de KDD inicia-se, obviamente, com o entendimento do domínio da aplicação e dos resultados finais a serem atingidos. A seguir, é feito um agrupamento de uma massa de dados organizadamente, alvo da prospecção. A etapa seguinte, limpeza dos dados (data cleaning), é realizada através de um pré-processamento dos dados, com vistas a adequá-los aos algoritmos. Isso se faz com a integração de dados heterogêneos, com a eliminação de dados incompletos e outras. Essa etapa é uma das mais demoradas podendo tomar até 80% de todo o tempo necessário para o processo completo, devido às muito conhecidas dificuldades de integração de bases de dados heterogêneas (MANNILA, 1996).

Os dados pré-processados devem ainda sofrer uma transformação com o consequente armazenamento adequado, visando facilitar o uso das técnicas de Data Mining.

É aí que o uso de Data Warehouse (Armazenamento de Dados) se torna significativo, pois com essa tecnologia as informações estarão armazenadas de maneira bastante eficiente.

Segundo Inmon (2005), o Data Warehouse é um conjunto de dados, integrado, não volátil e variável em relação ao tempo, dando apoio às decisões gerenciais.

Dando continuidade ao processo, chega-se à etapa de Data Mining especificamente, que se inicia com a escolha das ferramentas (algoritmos) que serão utilizadas. Essa escolha depende basicamente do objetivo do processo de KDD: classificação, agrupamento, regras associativas, ou outra. De maneira geral, na fase de Data Mining, as ferramentas especializadas buscam padrões nos dados. Essa pesquisa pode ser efetuada pelo sistema automaticamente, de forma livre (roams - percorrer/vasculhar o banco de dados) ou interativamente com um analista responsável pela geração de hipóteses, chamada análise direcionada (directed analysis) ou também chamada aprendizado supervisionado (supervised learning), onde temos como que um "professor" que "ensina" o sistema indicando, por exemplo, quando uma premissa foi ou não correta.

2.6 Áreas relacionadas ao KDD

O processo KDD é interdisciplinar e envolve áreas relativas a estatística, banco de dados, matemática, visualização de dados, inteligência artificial, aprendizado de máquina e sistemas especialistas. Este processo utiliza métodos, técnicas e algoritmos oriundos destas áreas, com o principal objetivo de extrair conhecimento a partir de grandes bases de dados.

Na figura 2.6 visualizamos a relação das áreas no processo KDD:

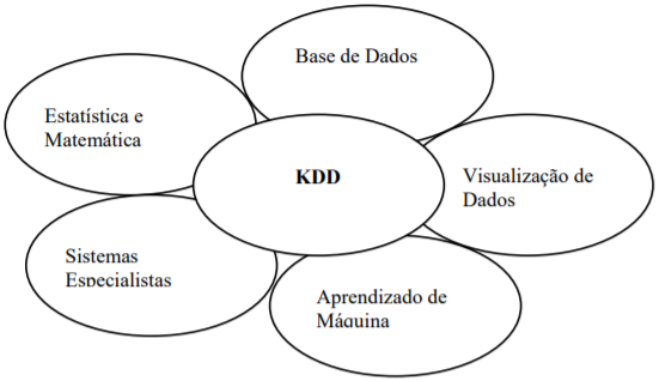

Figura 2.6 – KDD é um campo multidisciplinar

2.7 Tarefas de mineração de dados

Na literatura especializada são encontradas diferentes taxonomias para caracterizar as tarefas de mineração de dados. Fayyad et al. (1996) apresentam uma taxonomia em dois níveis. No primeiro nível, as tarefas de mineração de dados são

divididas em preditivas e descritivas. Tarefas preditivas usam os valores dos atributos descritivos para predizer valores futuros ou desconhecidos de outros atributos de interesse. Já as tarefas descritivas têm o objetivo de encontrar padrões que descrevem os dados de maneira que o ser humano possa interpretar. No segundo nível, as tarefas preditivas e descritivas são especializadas. No conjunto de tarefas preditivas, os autores inserem classificação e regressão. Já no conjunto de tarefas descritivas, as especializações agrupamento, sumarização, modelagem de dependências e detecção de desvios são incluídas pelos autores.

Han, Kamber e Pei (2011) também seguem o primeiro nível da taxonomia já apresentada; porém, o segundo nível da taxonomia seguida por esses autores difere levemente da taxonomia de Fayyad et al. (1996). Para Han, Kamber e Pei (2011), as tarefas de mineração de dados no segundo nível se dividem em: classificação e regressão; mineração de padrões frequentes, associações e correlações (que correspondem a um dos objetivos da “modelagem de dependências” na primeira taxonomia); análise de grupos (equivalente a “agrupamento”); e análise de outliers (similar à detecção de desvios). Ainda, no segundo nível da taxonomia, para o caso de tarefas descritivas, Rokach e Maimon (2008) enumeram duas tarefas adicionais: resumo linguístico e visualização. Interessante notar que esses autores, na realidade, apresentam uma taxonomia em três níveis para “paradigmas de mineração de dados”, em que um nível ainda mais alto é definido, dividido em paradigmas de verificação e de descoberta. No primeiro, tarefas de teste de hipótese, análise de variância e teste de goodness-of-fit são incluídas. Abaixo do paradigma de descoberta, desenvolvem-se os dois níveis de taxonomia já discutidos.

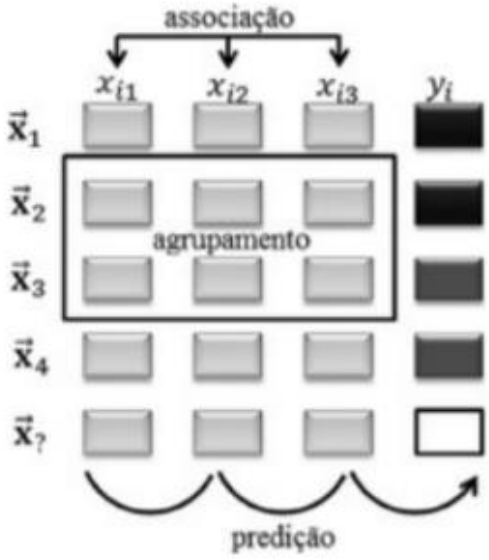

Figura 2.7 – Tarefas de mineração de dados

2.7.1 Tarefas de Predição e Classificação

Tarefas de predição consistem na análise de um conjunto de dados nos quais estão presentes os dados, descritos por atributos, e seus rótulos associados. O objetivo nessa tarefa é descobrir um modelo capaz de mapear corretamente cada um dos dados x→1, x→2, x→3 e x→4 aos seus rótulos (y) (Figura 2.7). Esse objetivo é alcançado por meio de técnicas chamadas supervisionadas, ou seja, capazes de encontrar o modelo de mapeamento a partir de procedimentos que associam um dado a um rótulo e corrigem tal associação quando ela não corresponde ao rótulo esperado (aquele associado ao dado no conjunto de dados). A análise preditiva pode ser dividida em duas subtarefas: análise preditiva categórica, também chamada de tarefa de classificação; e análise preditiva numérica, também chamada de tarefa de regressão. A primeira subtarefa se manifesta quando os rótulos associados aos dados pertencem a um conjunto discreto e finito de categorias. Já a segunda se faz presente quando os rótulos associados aos dados são numéricos e pertencentes a um conjunto de valores contínuos.

Para exemplificar situações nas quais a resolução de tarefas de predição pode ser útil, considere o contexto do restaurante. Imagine que o dono do restaurante desejasse oferecer um serviço de harmonização aos seus clientes, ou seja, quisesse classificar os pratos (x→i) servidos no restaurante em relação ao tipo de vinho (y) que

deveria ser servido como acompanhamento. Tendo a descrição de cada um dos pratos (xi1 , xi2 e xi3 para cada x→i) em termos de ingredientes (temperos principalmente), determinado tipo de vinho (branco seco, branco meio seco, tinto seco ou tinto meio seco) deve ser associado

Para aprender o modelo que associa um tipo de vinho a determinado prato, o dono do restaurante convidaria um sommelier para ajudá-lo. O sommelier iria ao restaurante e, a cada prato, diante dos ingredientes que os compõe, associaria o tipo de vinho correto; e aquele “funcionário”, como um bom aprendiz (ou um bom algoritmo), observaria e entenderia o comportamento da função de mapeamento (tipo de prato para tipo de vinho). Suponha que alguns meses depois, o cozinheiro do restaurante inserisse um novo prato (x→?) no cardápio. Diante dessa situação, para associar o novo prato ao vinho correto (o y desconhecido representado pelo retângulo em branco na Figura 2.7), o “funcionário” analisaria os ingredientes usados e aplicaria a função de mapeamento que aprendeu, concluindo o tipo de vinho associado ao prato.

Uma das tarefas mais comuns, a Classificação, visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de ’aprender’ como classificar um novo registro (aprendizado supervisionado). Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa: Perfil Técnico, Perfil Negocial e Perfil Gerencial. O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa. A tarefa de classificação pode ser usada por exemplo para:

- Determinar quando uma transação de cartão de crédito pode ser uma fraude;
- Identificar em uma escola, qual a turma mais indicada para um determinado aluno;
- Diagnosticar onde uma determinada doença pode estar presente;
- Identificar quando uma pessoa pode ser uma ameaça para a segurança.

2.7.2 Subtarefa de Estimação ou Regressão

Para exemplificar a subtarefa de regressão, considere que o dono do restaurante deseja saber a relação entre o número de clientes que frequenta o estabelecimento (x) e seu faturamento mensal (y). A partir daí, ele poderia descobrir, por exemplo, qual seria seu faturamento se o número de clientes fosse o dobro, ou ainda, qual o impacto no faturamento com a redução do número de clientes em 10%. Assim, a empresa pode modelar uma tarefa de regressão e induzir a função capaz de fazer o mapeamento entre uma perspectiva de quantidade de clientes (x?) e o faturamento mensal relacionado (o y desconhecido representado pelo retângulo em branco na Figura 2.7).

A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor. A tarefa de estimação pode ser usada por exemplo para:

- Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas;
- Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal

2.7.3 Tarefa de Agrupamento de Dados

A tarefa de agrupamento de dados consiste na análise de conjuntos de dados em que estão presentes apenas as descrições dos dados. Não há, nesse caso, a necessidade do uso da informação sobre qualquer tipo de rotulação dos dados. O objetivo na resolução dessa tarefa é descobrir relações entre os dados por meio de suas similaridades e fornecer, como resposta, a indicação de quais dados são similares entre si, oferecendo um modelo de agrupamento ou perfis para grupos de dados. Os algoritmos aplicados na resolução dessa tarefa executam procedimentos que organizam os dados em grupos, de forma que a similaridade entre os dados de um grupo seja máxima (ou seja, devem ser colocados em um grupo os mais similares

entre si) e a similaridade entre dados colocados em grupos diferentes seja mínima (ou seja, devem ser separados em grupos diferentes os dados não similares entre si).

Como exemplo para a tarefa de agrupamento, considere um restaurante que possui vários ambientes. É interessante que clientes com características similares sejam direcionados ao mesmo ambiente, enquanto clientes com características diferentes sejam direcionados a ambientes distintos. Por exemplo, clientes jovens, que preferem ambientes mais movimentados e dinâmicos, devem ser colocados no grupo de clientes que irão para o ambiente no qual há, por exemplo, música eletrônica e uma pequena pista de dança. Já clientes com suas famílias e com crianças devem ir para o grupo direcionado para o ambiente mais próximo ao playground. Embora, nesse exemplo, a tomada de decisão muito provavelmente não será feita por um software (a menos que, em um cenário futurista, haja um funcionário recepcionista robô – o que seria muito interessante); o funcionário que tomará a decisão está analisando as características de cada cliente e os agrupando em ambientes adequados ao seu perfil. Observando a Figura 2.7, pode-se associar cada cliente a um dado x→i e, seguindo o esquema gráfico da figura, descobrir que os clientes x→2 e x→3 são mais similares entre si que em relação aos clientes x→1 e x→4. Note que o agrupamento realizado sobre os dados independe de qualquer informação de rótulo.

A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares. Exemplos:

- Segmentação de mercado para um nicho de produtos;
- Para auditoria, separando comportamentos suspeitos;
- Reduzir para um conjunto de atributos similares registros com centenas de atributos

2.7.4 Tarefa de Associação

A tarefa de associação é definida como a busca por ocorrências frequentes e simultâneas entre elementos de um contexto. Os algoritmos que resolvem essa tarefa analisam conjuntos de dados que representam eventos ou transações (x→1 , x→2 , x→3 e x→4), procurando por itens (xi1 , xi2 e xi3) (Figura 2.7) frequentemente envolvidos nos mesmos eventos ou que apresentam algum tipo de correlação em seus comportamentos em tais eventos.

Nesse tipo de tarefa, é comum a descoberta de padrões triviais, mas o que se espera, no entanto, é que padrões inesperados sejam revelados. No contexto do restaurante, considere que cada prato seja um evento e que os itens que aparecem nos eventos sejam os ingredientes. Uma análise sobre a base de dados (as receitas de cada prato) pode revelar que “sempre que cebola é usada no preparo de um prato, alho também é usado”. Esse é um exemplo de uma descoberta sobre uma relação óbvia entre dois itens, a cebola e o alho. Porém, regras como “pratos nos quais se usa queijo brie também se usa geleia de pimenta” ou “quanto menor o teor de álcool na cerveja usada em molhos, menor a necessidade do uso de açúcar” podem representar conhecimento novo, inesperado e interessante, a depender, é claro, do nível de conhecimento sobre gastronomia de quem recebe a resposta da resolução da tarefa de associação.

2.8 Aplicações da Mineração de Dados

Aplicações A Mineração de Dados pode ser definida como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano. De fato, muitas são as técnicas utilizadas, porém a mineração de dados ainda é mais uma arte do que uma ciência. O sentimento do especialista não pode ser dispensado, mesmo que as mais sofisticadas técnicas sejam utilizadas. Ainda que as técnicas da Mineração de Dados sejam antigas, foi apenas nos últimos anos que passaram a ser usadas como exploração de dados, por vários motivos [Carvalho, 2005]:

- **O volume de dados disponível atualmente é enorme** – Mineração de Dados é uma técnica que só se aplica a grandes massas de dados, pois

necessita disto para calibrar seus algoritmos e extrair dos dados conclusões confiáveis. Empresas de telefonia, cartões de crédito, bancos, televisão por assinatura, comércio eletrônico, entre outras, vem gerando a cada dia uma grande quantidade de dados sobre seus serviços e clientes. Estes dados são passíveis de análise por mineração;

- **Os dados estão sendo organizados** - Com a tecnologia do data warehouse, os dados de várias fontes estão sendo organizados e padronizados de forma a possibilitar sua organização dirigida para o auxílio à decisão. As técnicas de mineração de dados necessitam de bancos de dados limpos, padronizados e organizados;
- **Os recursos computacionais estão cada vez mais potentes** - A mineração de dados necessita de muitos recursos computacionais para operar seus algoritmos sobre grandes quantidades de dados. O aumento da potência computacional, devido ao avanço tecnológico e à queda dos preços dos computadores, facilita o uso da mineração de dados atualmente. O avanço da área de banco de dados, construindo bancos de dados distribuídos, também auxiliou em muito à mineração de dados;
- **A competição empresarial exige técnicas mais modernas de decisão** - As empresas da área de finanças, telecomunicações e seguro experimentam a cada dia mais competição. Como estas empresas sempre detiveram em seus bancos de dados uma enorme quantidade de informação, é natural que a mineração de dados tenha se iniciado dentro de seus limites. Atualmente, outras empresas buscam adquirir dados para analisar melhor seus caminhos futuros através dos sistemas de apoio à decisão. Para empresas de serviços, a aquisição de dados é importante, pois precisam saber que serviço oferecer a quem. Para outras empresas, até a venda das informações pode ser um produto;
- **Programas comerciais de mineração de dados já podem ser adquiridos** - As técnicas de mineração de dados são antigas conhecidas da Inteligência Artificial, porém somente recentemente saíram dos laboratórios para as empresas. Alguns pacotes já podem ser encontrados no comércio, contendo algumas destas técnicas. As técnicas mais recentes, no entanto, ainda se encontram no campo

acadêmico, sendo necessário que a empresa se dirija a uma universidade que realize pesquisa para obter ajuda.

2.9 Fases da Mineração de Dados

Em 1996, um conjunto de três empresas especializadas no então jovem e imaturo mercado de data mining, desenvolveram um modelo de processos genéricos, com o intuito de padronizar as etapas do processo de mineração de dados, dando início ao denominado projeto CRISP-DM (CRoss Industry Standard Process for Data Mining) [The CRISP-DM Consortium, 2000].

Este projeto desenvolveu um modelo de processo de mineração de dados industrial e livre de ferramenta. Começando pelos embrionários processos de descoberta de conhecimento usados nos primeiros projetos de mineração de dados e respondendo diretamente aos requerimentos do usuário, esse projeto definiu e validou um processo de mineração de dados que é aplicável em diversos setores da indústria. Essa metodologia torna projetos de mineração de dados de larga escala mais rápidos, mais baratos, mais confiáveis e mais gerenciáveis. Até mesmo projetos de mineração de dados de pequena escala se beneficiam com o uso do CRISP-DM. O modelo CRISP, atualmente, é uma referência para que seja desenvolvido um plano de integração para a descoberta de conhecimento.

O atual processo para mineração de dados propõe uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases correspondentes de um projeto, suas respectivas tarefas e relacionamentos entre essas tarefas.

Na Figura 3.1 abaixo é mostrado o ciclo de vida de um projeto de mineração de dados, que consiste de 6 (seis) fases. A sequência de fases não é obrigatória, ocorrendo a transição para diferentes fases, dependendo do resultado de cada fase, e que etapa particular de cada fase precisa ser executada em seguida. As setas indicam as mais importantes e mais frequentes dependências entre as fases. O ciclo externo na figura simboliza o ciclo natural da mineração de dados. Um processo de mineração de dados continua após a solução ter sido desenvolvida. As lições aprendidas durante o processo podem provocar perguntas novas, frequentemente mais pertinentes ao negócio. Processos subsequentes se beneficiarão das experiências de processos anteriores

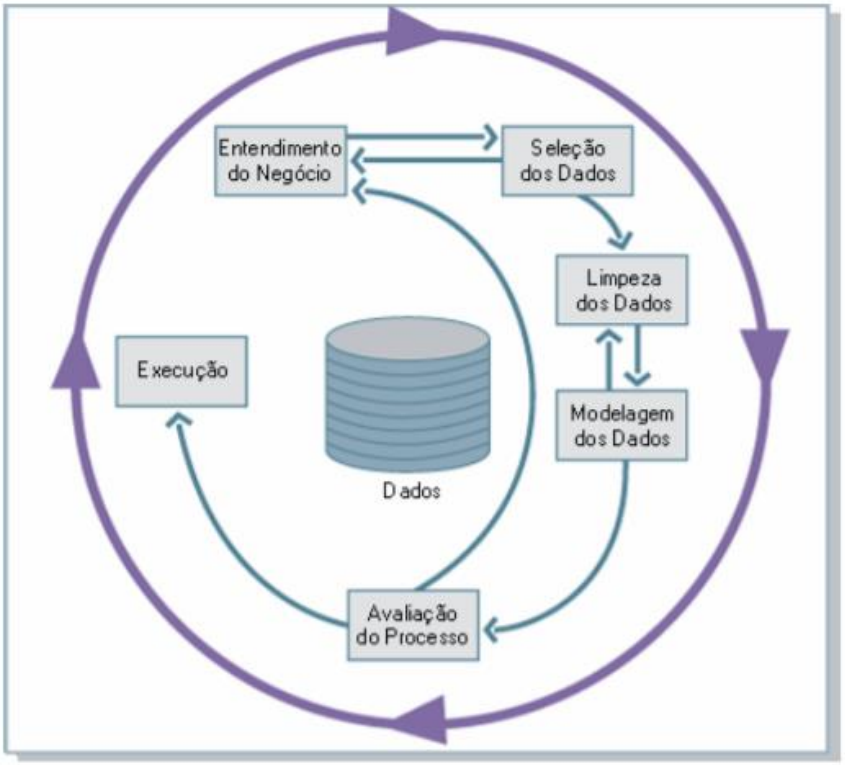

Figura 3.1 - Fases do CRISP-DM Process Model (Baseado em [The CRISP-DM Consortium, 2000])

Segue abaixo uma síntese das etapas pertencentes ao modelo.

2.9.1 Entendimento do Negócio (Business Understanding)

Essa fase inicial tem o foco no entendimento do negócio que visa obter conhecimento sobre os objetivos do negócio e seus requisitos, e então converter esse conhecimento em uma definição de um problema de mineração de dados, e um plano preliminar designado para alcançar esses objetivos.

2.9.2 Seleção dos Dados (Data Understanding)

Consiste no entendimento dos dados, que visa à familiarização com o banco de dados pelo grupo de projeto, utilizando-se de conjuntos de dados "modelo". Uma vez definido o domínio sobre o qual se pretende executar o processo de descoberta, o próximo passo é selecionar e coletar o conjunto de dados ou variáveis necessárias.

Essa fase se inicia com uma coleta inicial de dados, e com procedimentos e atividades visando a familiarização com os dados, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formar hipóteses.

2.9.3 Limpeza dos Dados (Data Preparation)

A fase de preparação de dados consiste na preparação dos dados que visa a limpeza, transformação, integração e formatação dos dados da etapa anterior. É a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados. Esta fase abrange todas as atividades para construir o conjunto de dados final (dados que serão alimentados nas ferramentas de mineração), a partir do conjunto de dados inicial.

A utilização de Data Warehouses facilita muito esta etapa do processo de mineração de dados, que costuma ser a fase que exige mais esforço, correspondendo geralmente a mais de 50% do trabalho. Por isso, é muito importante para uma organização, que ela possua em seus processos habituais boas práticas da administração de dados, como o Data Cleansing, que é uma parte fundamental da cadeia da administração da informação, responsável pelas etapas de detecção, validação e correção de erros em bases de dados [Chapman, 2005].

2.9.4 Modelagem dos Dados (Modeling)

Fase que consiste na modelagem dos dados, a qual visa a aplicação de técnicas de modelagem sobre o conjunto de dados preparado na etapa anterior.

Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para se obter valores otimizados. Geralmente, existem várias técnicas para o mesmo tipo de problema de mineração. Algumas técnicas possuem requerimentos específicos na forma dos dados. Consequentemente, voltar para a etapa de preparação de dados é frequentemente necessário.

A maioria das técnicas de mineração de dados são baseadas em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação e clusterização.

2.9.5 Avaliação do processo (Evaluation)

A avaliação do processo visa garantir que o modelo gerado atenda às expectativas da organização. Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer um dos estágios anteriores do processo de mineração. Nesta etapa se construiu um modelo que parece de alta qualidade, de uma perspectiva da análise de dados. Antes de prosseguir, é importante avaliar mais detalhadamente o modelo, e rever as etapas executadas para construir o modelo, para se certificar de que ele conseguirá alcançar os objetivos de negócio.

Deve se determinar se houve algum importante objetivo do negócio que não foi suficientemente alcançado. No fim desta fase, uma decisão sobre o uso dos resultados da mineração deve ser tomada

2.9.6 Execução (Deployment)

Esta fase consiste na definição das fases de implantação do projeto de Mineração de Dados.

A criação do modelo não é o fim do projeto. Mesmo se a finalidade do modelo for apenas aumentar o conhecimento dos dados, o conhecimento ganho necessitará ser organizado e apresentado em uma maneira que o cliente possa usar. Dependendo das exigências, a fase de execução pode ser tão simples quanto a geração de um relatório, ou tão complexo quanto executar processos de mineração de dados repetidamente.

Em muitos casos será o cliente, não o analista dos dados, que realizará as etapas da execução. Entretanto, mesmo se o analista não se encarregar da execução é importante que ele faça o cliente compreender que medidas deverão ser tomadas a fim de empregar efetivamente os modelos criados.

3.0 Limitações

Apesar da grande potencialidade oferecida pela Mineração de Dados, alguns fatores devem ser analisados. Wang et all. [85] discutem como alguns desses fatores podem prejudicar as técnicas de mineração:

- As relações entre os atributos precisam ser muito bem definidas, caso contrário os resultados podem ser mal interpretados;
- Permitir que o processo de treinamento execute por muito tempo, até que se consiga obter indícios que possam levar à conclusões factíveis;
- Gerar subsídios para uma conclusão errada tornando-a mais plausível. Porém, uma interpretação falha pode disfarçar as falhas nos dados;
- Usar um grande número de variáveis

Referências Bibliográficas

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. Artificial Intelligence Magazine, v. 17, n. 3, p. 37-54, 1996a.

AGRAWAL, R. & SRIKANT, R. Fast algorithms for mining association rules. Proc. of the 20th Int'l Conference on Very Large Databases. Santiago, Chile, 1994.

CARVALHO, Deborah et al. Mineração de dados aplicada à fisioterapia. Fisioterapia e Movimento, Curitiba, v. 25, n. 3, p. 595-605, jul./set. 2012.

KURETZKI, Carlos Henrique. Técnicas de mineração de dados aplicadas em bases de dados da saúde a partir de protocolos eletrônicos. 2009. 98f. Dissertação (Mestrado) - Universidade Federal do Paraná, Curitiba.

WITTEN, Ian; FRANK, Eibe; HALL, Mark. Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2011.