PROCESSO DE EXTRAÇÃO DE CONHECIMENTO DE BASES DE DADOS – ELEMENTOS DE APOIO E PRINCIPAIS PROBLEMAS

*Cláudio Alex Jorge da Rocha**

RESUMO: O interesse cada vez maior das empresas em acompanhar novas tecnologias de processamento e armazenamento de dados, além de visualizar a informação como seu maior patrimônio, tem direcionado várias pesquisas para o estudo do processo de transformação dos dados em conhecimento, o que proporciona um auxílio efetivamente inteligente à tomada de decisão. Neste contexto, o processo de Extração de Conhecimento de Bases de Dados (KDD - *Knowledge Discovery in Database*) desponta como uma tecnologia capaz de cooperar, amplamente, com a busca do conhecimento embutido nos dados. Este trabalho envolve a investigação dos conceitos, técnicas, métodos e ferramentas para o processo de extração de conhecimento a partir de grandes volumes de dados, considerando os elementos que apoiam esse processo e os principais problemas envolvidos.

1. INTRODUÇÃO

Na última década, três fatores importantes nortearam o crescimento de nossa capacidade de gerar e colecionar dados: primeiro, a disponibilidade de tecnologias que oferecem um grande poder de armazenamento e processamento de dados a baixo custo; segundo, pelo acúmulo de dados a uma razão crescente; e terceiro, a novos métodos e ferramentas desenvolvidos pela comunidade de informática para o processamento de dados [6].

O interesse cada vez maior das empresas públicas e privadas em acompanhar essas novas tecnologias de processamento e armazenamento, além de visualizar a informação como seu maior patrimônio, tem direcionado várias pesquisas para o estudo do processo de transformação de dados em conhecimento. Tradicionalmente, essa transformação tem sido realizada via processos manuais de análise e interpretação de dados, o que torna o processo de extração de conhecimento de dados muitas vezes caro, lento e altamente subjetivo, além de inviável em se tratando de grandes volumes de dados.

O interesse em automatizar o processo de análise de grandes volumes de dados tem fomentado várias pesquisas em um campo emergente chamado Extração de Conhecimento de Bases de Dados (KDD - *Knowledge Discovery in Database*) [9].

KDD refere-se ao processo de extração de conhecimento de grandes volumes de dados com o objetivo de obter significado e conseqüente entendimento dos dados, bem como adquirir novos conhecimentos. Esse processo é bastante complexo, consistindo da combinação de métodos e ferramentas de Estatística, Inteligência Artificial, Visualização e Banco de Dados para encontrar padrões e regularidades nos dados [6, 9, 16].

Na literatura são encontradas várias nomenclaturas para o processo KDD, entre elas destacam-se: Data Mining, Processamento de Padrões de Dados, Arqueologia de Dados, Garimpagem de Dados e *Siftware* [6]. Porém, segundo Fayyad, Data Mining[1] pode ser visto como parte de um processo maior de extração de conhecimento de bases de dados [9].

* *Mestre em Ciência da Computação pela USP, professor do Curso de Ciência da Computação – UNAMA.*

Este trabalho está dividido da seguinte forma. Na próxima seção são apresentados os conceitos relacionados à extração de conhecimento de bases de dados. Na seção 3, são focalizadas as etapas do processo KDD. Na seção 4, são mostrados alguns dos elementos que apoiam esse processo. Na seção 5, são sublinhados os principais problemas relacionados ao processo de Extração de Conhecimento de Bases de Dados. Na seção 6, são feitas algumas consideração finais a respeito deste trabalho. Finalmente, na seção 7, são apresentadas as Referências Bibliográficas que foram utilizadas para que a realização deste trabalho fosse possível.

2. EXTRAÇÃO DE CONHECIMENTO

De uma forma geral, no processo KDD, os dados podem ser vistos como a matéria-prima bruta. No momento em que o usuário atribui algum significado especial aos dados, estes passam a ser entendidos como uma informação. Quando os especialistas de um domínio de aplicação estabelecem, por exemplo, uma regra, a interpretação de confronto entre essas informações e essa regra constitui um conhecimento a respeito dos dados. A relação hierárquica entre os dados, a informação e o conhecimento existentes em uma base de dados, considerando o volume e o valor que os usuários de níveis decisórios atribuem a cada um dos elementos dessa hierarquia, pode se visualizado na Figura 1.

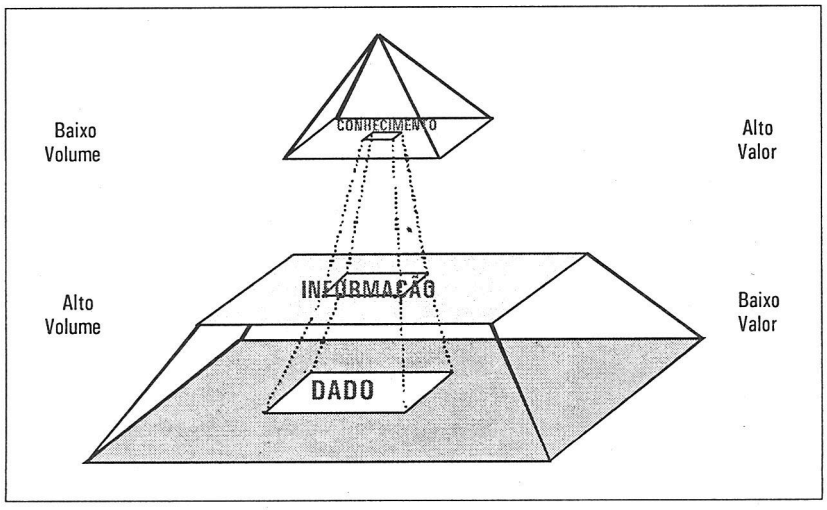

Figura 1. Pirâmide do Processo de Conhecimento.

Dessa forma, o conhecimento pode ser visto como uma abstração ou um nível de informação acima dos dados. Segundo Carbonel, conhecimento pode ser definido como uma informação interpretada, categorizada, aplicada, revisada e que possui um determinado valor para o usuário [2].

A aquisição de conhecimento representa um dos principais problemas relacionados ao desenvolvimento de sistemas inteligentes. Isso se deve, principalmente, à grande quantidade de procedimentos, relações e técnicas que interagem nesse processo. Além disso, o conhecimento adquirido nem sempre é apropriado para uso efetivo, visto que o mesmo pode possuir propriedades indesejáveis como, por exemplo, grande volume e difícil representação [21].

O processo de aquisição de conhecimento pode ser realizado de forma *explícita*, em que a extração do conhecimento é feita por meio do engenheiro de conhecimento, utilizando técnicas convencionais como, por exemplo, estudo de casos e entrevista a especialistas do domínio [20]; ou *implícita*, que é geralmente feita utilizando sistemas de aprendizado de máquina para extrair conhecimento a partir de dados.

O processo KDD desponta como um conjunto de técnicas e ferramentas capazes de contribuir amplamente para o problema de aquisição de conhecimento implícito em grandes volumes de dados. A seguir serão apresentadas as etapas que compõem esse processo.

3. ETAPAS DO PROCESSO DE EXTRAÇÃO DE CONHECIMENTO DE BASES DE DADOS

A extração de conhecimento a partir de grande volumes de dados deve ser vista como um processo interativo e iterativo, e não como um sistema de análise automática. Dessa forma, não se pode esperar a extração de conhecimento útil simplesmente submetendo um conjunto de dados a uma "caixa preta" [16].

[1] ***Neste trabalho Data Mining é visto como parte do processo KDD.***

A interatividade do processo KDD é pautada no amplo entendimento, por parte dos usuários desse processo, sobre o domínio da aplicação. Esse entendimento envolve, por exemplo, a seleção de um subconjunto representativo dos dados e bons critérios para avaliar o conhecimento extraído. Para uma melhor compreensão das funções dos usuários que utilizam o processo KDD, neste trabalho, eles são divididos, em três classes: *especialista do domínio*, que deve possuir amplo entendimento do domínio da aplicação; *analista*, que executa o processo KDD e, portanto, deve possuir amplo conhecimento das etapas que compõem esse processo; e o *usuário final*, que é freqüentemente aquele que utiliza o conhecimento extraído para auxiliá-lo em algum processo de tomada de decisão.

Vale ressaltar que os usuários, geralmente, não possuem funções disjuntas no processo KDD, portanto, pode haver situações, por exemplo, em que o especialista também é o usuário final ou que o especialista auxilie e/ou execute funções do analista.

O sucesso do processo KDD depende, em grande parte, da interação entre esses usuários. É pouco provável que o analista encontre conhecimento útil a partir dos dados sem o aval do especialista sobre o que seria útil para um domínio específico. Além disso, a interatividade do processo requer a participação efetiva do usuário final e do especialista nas escolhas e decisões ao longo do processo [9].

A extração de conhecimento a partir dos dados é entendida como um processo contendo, pelo menos, as seguintes grandes etapas: (1) Compreensão do Domínio da Aplicação; (2) Seleção e Preparação dos Dados; (3) Data Mining; (4) Avaliação do Conhecimento Extraído e (5) Consolidação e Utilização do Conhecimento Extraído.

Um esquema representativo contendo todos essas etapas é ilustrado na Fig. 2. O processo KDD inicia com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados (bases de dados de onde se pretende extrair conhecimento). Em seguida, uma amostra representativa (e. g. utilizando técnicas estatísticas) é retirada da base de dados, pré-processada e submetida aos métodos e ferramentas da etapa de Data Mining[2] com o objetivo de encontrar padrões/modelos (conhecimento) a partir dos dados. Depois esse 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. Vale acentuar que, apesar das ferramentas de visualização serem mais utilizadas

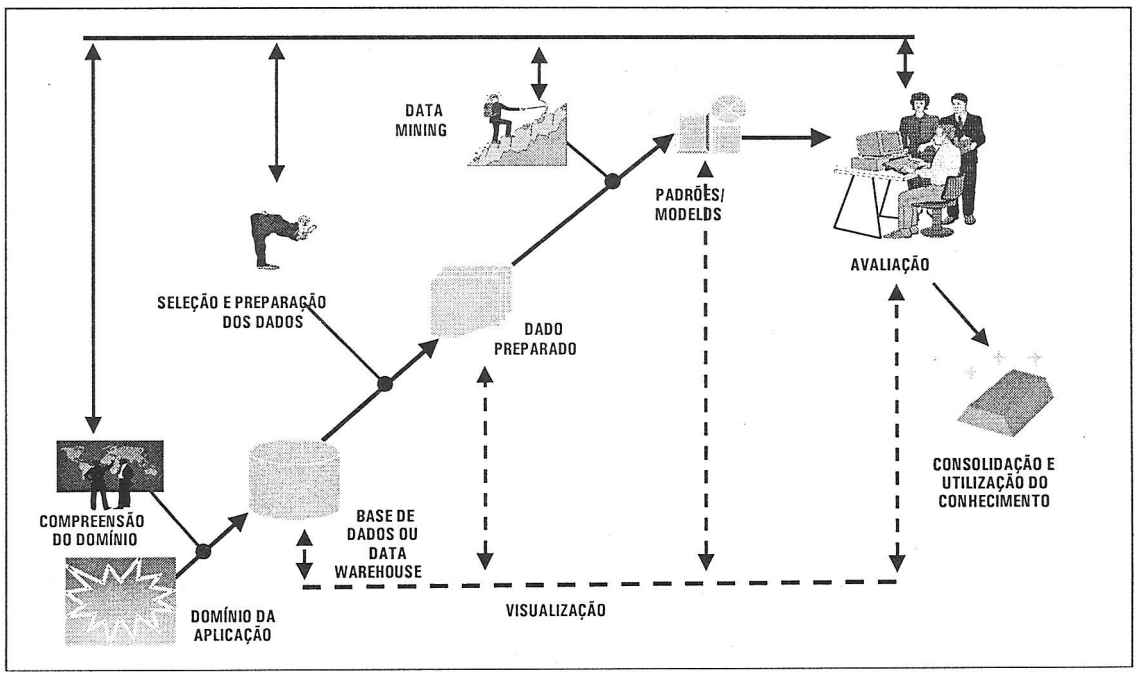

Figura 2. Etapas do Processo de Extração de Conhecimento de Bases de Dados.

[2] *Esses métodos e ferramentas de Data Mining serão chamados de algoritmo ao longo deste trabalho.*

É importante notar que, por ser um processo eminentemente iterativo, as etapas do processo KDD 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. Desta forma, os resultados de um determinada etapa podem acarretar mudanças a quaisquer das etapas anteriores ou ainda, o recomeço de todo o processo [8]. Um detalhamento maior das metas de cada etapa será apresentado a seguir.

3.1. COMPREENSÃO DO DOMÍNIO DA APLICAÇÃO

Nesta etapa, procura-se o completo entendimento do domínio da aplicação, considerando aspectos como:

- Estudo de viabilidade e custos da aplicação do processo.
- Verificação da quantidade e do tipo de conhecimento disponível antes do processo KDD iniciar.
- Condições e metas do usuário final.

A sinergia usuário final/especialista do domínio/ analista deve ser bastante acentuada nessa etapa. O produto dessa interação deve ser uma documentação completa do domínio, considerando entre outros fatores:

- Identificação das fontes dos dados: internas (e.g. bases de dados da empresa) e externas (e.g. *internet*).
- Identificação dos gargalos do domínio, considerando, entre outros aspectos:
  - bases de dados com valores de atributos ausentes;
  - bases de dados com um grande volume de dados;
  - tipos de dados armazenados nas bases de dados, como por exemplo: dados multimídia, gráficos, combinação de dados discretos e contínuos, etc.
- Estabelecimento de critérios para avaliação dos resultados do processo, verificando a relação entre a facilidade de entendimento com a qualidade (utilidade) do conhecimento a ser extraído.
- Especificação do modo como o conhecimento extraído deverá ser utilizado. Por exemplo, para classificação, visualização, exploração, etc.

3.2. SELEÇÃO E PREPARAÇÃO DOS DADOS

A extração direta de conhecimento a partir de grandes volumes de dados pode se tornar uma tarefa inviável. Grandes volumes de dados podem gerar um espaço de busca de padrões combinatorialmente explosivo. Além disso, a maioria dos algoritmos para extração desses padrões possuem limitações quanto ao volume de dados que podem manipular. A busca de conhecimento em grandes bases de dados pode ocasionar, ainda, o aumento das chances de encontrar-se padrões pouco significativos e até mesmo espúrios. Uma possível solução para esse problema envolve a tentativa de selecionar uma amostra significativa da base de dados da aplicação.

Segundo Mannila, a seleção e preparação dos dados consome cerca de 80% de todo o tempo gasto no processo KDD [16]. Esta etapa envolve a seleção de uma amostra representativa da base de dados e o pré-processamento e redução dessa amostra a fim de adequá-la aos padrões de entrada dos algoritmos para Data Mining. Para um melhor entendimento, essa etapa pode ser dividida em seleção e, preparação e redução da amostra, que serão apresentadas a seguir.

3.2.1. SELEÇÃO DA AMOSTRA

A escolha de uma amostra que reflita com a maior fidelidade possível a base de dados é de suma importância para as demais etapas do processo KDD. Seleções de amostras pouco significativas podem produzir resultados (padrões extraídos) imprecisos ou sem valia. Além disso, pequenas amostras podem levar a conclusões incorretas, por outro lado, grandes quantidades de dados podem tornar o processamento lento e até mesmo inviável devido às limitações dos algoritmos para Data Mining. Portanto, entre outros aspectos, devem ser considerados:

- Tamanho da amostra.
- Estratégias para obtenção da amostra (técnicas estatísticas podem ser utilizadas).
- Homogeneidade dos dados.
- Dinâmica dos dados (e.g. mudanças de valores de atributos ao longo do tempo).

3.2.2. PREPARAÇÃO E REDUÇÃO DA AMOSTRA

Após a etapa de seleção, é necessário o pré-processamento da amostra de dados a fim de atender as exigências e as limitações dos formatos de entrada dos algoritmos para Data Mining. Esse pré-processamento inclui a preparação e a redução dos dados, observando, entre outros fatores:

- Características da base de dados, como, por exemplo, tipos de dados e padronização do conteúdo dos registros (e.g. no caso do sexo das pessoas, é comum que seja especificado através de valores M ou F, 0 ou 1 e Mas ou Fem).
- Eliminar registros duplicados, lixo nos dados produzidos pelas migrações, etc.
- Tratamento de ruídos nos dados.
- Manipulação de valores de atributos ausentes (dados incompletos).
- Representação dos dados de acordo com os objetivos da tarefa.
- Redução efetiva do número de variáveis a serem consideradas.

É importante destacar que se a base de dados estiver em um *data warehouse*, problemas como padronização e limpeza nos dados podem ser em grande parte resolvidos, pois o *data warehouse* provê métodos que, entre outros, permitem a integração, a padronização e a sumarização de dados [13].

Observados esses fatores, a amostra, freqüentemente chamada de conjunto de dados, pode ser submetida aos algoritmos para Data Mining.

3.3. DATA MINING

Data Mining (DM) envolve a criação e/ou a utilização de modelos apropriados de representação dos padrões e relações identificados a partir dos dados. O resultado desses modelos, depois de avaliados pelo especialista e/ou usuário final, são empregados para predizer novas situações baseadas em novos dados [9].

Os modelos gerados por DM seguem geralmente os padrões estatísticos, neurais ou simbólicos. Um modelo estatístico típico é gerado pelo método de regressão (e.g. regressão linear) e é representado por um sistema de equações. Um modelo neural é representado como uma arquitetura (e.g. rede *feedforward*) de nós e conexões (com pesos) entre eles, além de uma função de aprendizado (e.g. regra delta). Já os modelos simbólicos são geralmente representados por regras do tipo IF...THEN ou árvores de decisão. Existe uma vasta gama de algoritmos referenciados na literatura que seguem esses modelos [12, 17].

Em particular, algoritmos para DM consistem da combinação de três componentes básicos: Modelo, Critério de Preferência e Algoritmo de Busca [9]. Dessa forma, um algoritmo é freqüentemente uma instanciação de um Modelo/Critério de Preferência/Algoritmo de busca. Para um melhor entendimento dos principais componentes de um algoritmo para DM, eles serão descritos a seguir.

3.3.1 MODELO

O modelo de um algoritmo para DM considera dois fatores: a função do modelo e a representação do modelo.

As *funções do modelo*, também conhecidas como técnicas de DM, especificam o modo como o conhecimento extraído deverá ser utilizado. Entre as funções mais comuns e aceitas pela comunidade KDD, destacam-se a classificação e as regras de associação [7, 9].

A classificação visa o mapeamento de um determinado caso (registro do conjunto de dados) dentro de uma das várias classes pré-definidas (e.g. regras de classificação a respeito de doenças podem ser extraídas de um conjunto de casos conhecidos e usadas para fazer diagnóstico em novos pacientes baseados em seus sintomas).

As regras de associação determinam as relações entre os campos de uma base de dados (e.g. regras de associação podem descrever que itens são comumente

comprados juntamente com outros em um supermercado).

Para que essas técnicas sejam aplicadas é necessário que os algoritmos para DM possuam uma linguagem para representar os conceitos (padrões) obtidos a partir dos dados. Essa linguagem, chamada de *representação do modelo*, descreve o estado interno desse algoritmo. Esta representação geralmente determina a flexibilidade do modelo em representar os dados e a facilidade de compreensão do modelo em termos humanos. Na literatura pode-se encontrar vários tipos de representação, entre as quais destacam-se: árvores e regras de decisão [17], modelos não lineares (e.g. redes neurais) [12], *instance-based* (e.g raciocínio baseado em casos [1]) e modelos de dependência gráfica probabilística (e.g. redes bayesianas) [11].

3.3.2. CRITÉRIO DE PREFERÊNCIA

O critério de preferência verifica a qualidade do modelo e seus parâmetros, considerando, por exemplo, mecanismos para evitar *overfitting*, ou seja, evitar que o algoritmo "decore". Há dois critérios a serem considerados: o explícito, embutido nos algoritmos de busca, como por exemplo os critérios de máxima verossimilhança de encontrar os padrões que melhor representem o conjunto dados observado; e, o implícito, utilizado principalmente na etapa de seleção e preparação dos dados, e que refletem os critérios subjetivos do analista em termos de quais modelos devem ser considerados [9].

3.3.3. ALGORITMO DE BUSCA

O algoritmo de busca especifica um método para encontrar modelos e parâmetros a partir dos dados, isto é, extrair conhecimento a partir desses dados. Entre os algoritmos de busca para DM encontrados na literatura, destacam-se: CN2 [4], C4.5 [17], Naive Bayes [15], Autoclass [3] e BKD [18]. Vale ressaltar que a maioria dos algoritmos utilizados na fase de Data Mining, também chamados de sistemas de aprendizado, foram desenvolvidos e são largamente aceitos e utilizados pela comunidade de Aprendizado de Máquina.

A escolha do melhor algoritmo para DM é freqüentemente crítica, pois é sabido que nenhum deles tem desempenho ótimo em todos os domínios de aplicação [17]. A seleção desses algoritmos é realizada pelo analista e deve ser pautada nas restrições do domínio e/ou nas preferências do usuário final (e/ou especialista no domínio). Considerando essas restrições, o analista pode selecionar o algoritmo baseado em alguns parâmetros como, por exemplo, o tipo de aprendizado, paradigmas de aprendizado, linguagens de descrição e como novos exemplos são integrados [9].

Vale ressaltar que, além da observação desses parâmetros, as avaliações experimentais desempenham um papel fundamental na seleção de um algoritmo, uma vez que não existem métodos formais para decidir qual o melhor algoritmo para um determinado domínio de aplicação [14].

3.4. AVALIAÇÃO DO CONHECIMENTO EXTRAÍDO

O processo KDD não termina quando os padrões nos dados de entrada são extraídos. É preciso que o usuário entenda e possa julgar a utilidade do conhecimento extraído, contrastando-o com o conhecimento do especialista do domínio. Essa interação pode facilitar a busca das causas de possíveis erros ocorridos ao longo de todo esse processo.

A avaliação do modelo é uma tarefa bastante difícil que envolve, entre outros aspectos, a utilização de métodos (geralmente estatísticos) para "filtrar" o conhecimento extraído, removendo padrões redundantes e/ou irrelevantes [8]. Esses métodos devem ser acompanhados de técnicas de visualização para auxiliar os usuários na filtragem dos padrões, bem como na decisão sobre a utilidade do conhecimento extraído. Além disso, devem ser observados os critérios de desempenho do processo, considerando fatores como a precisão e a representação do conhecimento extraído. Por exemplo, um algoritmo para DM pode possuir uma elevada capacidade preditiva mas pouca capacidade descritiva (simplicidade de representação do conhecimento).

3.5. CONSOLIDAÇÃO E UTILIZAÇÃO DO CONHECIMENTO EXTRAÍDO

A consolidação do conhecimento extraído pressupõe a verificação e a solução de potenciais conflitos com o conhecimento previamente extraído antes do processo ser iniciado.

O conhecimento pode, então, ser organizado pelo analista dentro de um modelo, usado para refinar o modelo existente na aplicação ou simplesmente documentado e informado ao usuário. Na próxima seção são apresentados alguns dos elementos que apoiam o processo KDD.

4. ELEMENTOS DE APOIO AO PROCESSO DE EXTRAÇÃO DE CONHECIMENTO DE BASES DE DADOS

Nesta seção, serão abordados três dos principais elemento de apoio ao processo KDD: *data warehouse*, técnicas estatísticas e visualização de dados. Esse elementos podem ser considerados ferramentas de auxílio para a extração de conhecimento devido a otimização dos recursos e tempo gastos nesse processo, bem como o maior controle dos dados, no que concerne ao armazenamento e recuperação. Além disso, a compreensão do domínio é acentuadamente facilitada, posto que as técnicas estatísticas, em conjunto com as ferramentas de visualização, têm um papel fundamental em todas as etapas do processo KDD, sobretudo nas etapas de seleção e preparação dos dados, Data Mining e avaliação do conhecimento extraído.

4.1. DATA WAREHOUSE

Data Warehouse (DW) congrega várias tecnologias, tais como, plataformas de hardware, ferramentas de gerenciamento de dados, bases de dados voltadas para consultas complexas e ferramentas inteligentes de análise de dados. Seu uso como modelo de infra-estrutura para o suporte à tomada de decisão tem como objetivos fundamentais [13]:

- Prover um ambiente de informação bem administrado e protegido contra acessos indiscrimidados de usuários, haja vista o grande ativo que representam essas informações para uma empresa.
- Construir um modelo de dados corporativo que permita uma padronização da manipulação das informações, tanto nos sistemas de produção quanto nos de tomada de decisão.
- Manter a independência entre os processos dos usuários e os da administração.

Criar um DW não é uma simples questão de tecnologia de bases de dados ou processadores paralelos, envolve planejamento e modelagem (aspectos muitas vezes deixados em segundo plano), integração de vários softwares e uma contínua atualização e refinamento.

Uma solução bem projetada de DW visa à satisfação das necessidades de análise de informações dos usuários, como monitorar o histórico das operações, além de prever situações futuras. Ao transformar, consolidar e racionalizar as informações dispersas em diferentes bases de dados e plataformas, um DW permite que sejam realizadas análises estratégicas bastantes eficazes em informações antes inacessíveis ou subaproveitadas. Um dos métodos mais populares para análise de DW é conhecido como OLAP (On-Line Analytical Processing) [5]. OLAP focaliza a manipulação e análise de dados por meio de métodos multidimensionais, com o objetivo de suprir as limitações impostas pelas linguagens de consultas (e.g. SQL) e pelos esquemas de bancos de dados relacionais para armazenamento e acesso a dados. A partir de todos estes elementos envolvidos no DW, o analista do processo KDD, em interação com o especialista do domínio, pode aproveitar essas bases de dados já padronizadas e a facilidade de recuperação dos dados para, por exemplo, selecionar atributos mais significativos para uma conseqüente extração de conhecimento, via algoritmos para Data Mining.

4.2. FERRAMENTAS DE VISUALIZAÇÃO

As ferramentas de visualização de dados estão se tornando cada vez mais importantes no processo KDD, pois permitem o aumento da capacidade de análise e de interpretação dos resultados obtidos [19].

Geralmente, a visualização de dados pode ser utilizada como uma ferramenta exploratória na análise

desses dados. Um especialista em visualização pode construir diferentes tipos de gráficos onde o especialista do domínio e o analista do processo KDD podem verificar tendências a partir dos dados, determinar os atributos mais significativos para um determinado padrão descoberto e, principalmente, mostrar ao usuário final ou ao próprio especialista do domínio, de maneira mais clara e concisa, o conhecimento extraído.

Os principais tipos de ferramentas usadas para realizar as aplicações de visualização são: as linguagens de programação especializadas e as ferramentas *Graphic User Interface* (GUI). As linguagens de programação exigem geralmente habilidades de programação necessárias para criação de gráficos (e.g. C++). Essa linguagens são voltadas, particularmente, para os casos em que se exigem gráficos especializados, não disponíveis no mercado. As ferramentas GUI são apropriadas às situações em que as necessidades de visualização dos dados não excede a própria capacidade dessas ferramentas.

Vale ressaltar que é freqüente o uso de ferramentas avançadas de visualização de dados no processo KDD, principalmente as que envolvem análise estatística. Nesse caso são também necessários conhecimentos estatísticos para uma plena utilização dessas ferramentas.

4.3. TÉCNICAS ESTATÍSTICAS

A estatística tem desenvolvido uma vasta infra-estrutura (teoria) como suporte de seus próprios métodos e uma linguagem (cálculo probabilístico) para descrever suas abordagens para quantificar incerteza associada às inferências a partir dos dados. Estes métodos permitem descrever relações entre variáveis para predição, quantificando efeitos, ou sugerindo caminhos [10].

A relação entre o processo KDD e estatística é bastante estreita. Embora com enfoques diferentes, ambas as áreas objetivam a localização de padrões e regularidade nos dados. Em geral, o processo KDD, particularmente na etapa de Data Mining, enfatiza mais a facilidade de entendimento do conhecimento adquirido que basicamente a precisão. Além disso, a maioria dos algoritmos para DM estão mais voltados à produção de conjuntos de declarações sobre dependências locais entre variáveis de interesse (e.g. na forma de regras) do que propriamente na construção de modelos globais que incluam todas as variáveis de interesse do problema [10].

Além da etapa de DM, as técnicas estatísticas têm um papel fundamental em mais três etapas do processo KDD. Na etapa de seleção dos dados, tentando extrair amostras enxutas e representativas dos dados, na etapa de preparação e redução da amostra, por exemplo no tratamento de dados com ruído, e na avaliação do conhecimento extraído, especialmente no que concerne a utilização de técnicas de visualização [8].

Além disso, técnicas estatísticas, juntamente com técnicas de Inteligência Artificial, especialmente as que manipulam incerteza [11], provêem mecanismos para evitar *overfitting*, tratar ruídos, manipular conjunto de dados incompletos (*missing values*), e combinar conhecimento de fundo (do domínio ou a priori) com os dados. Na próxima serão destacados alguns dos principais problemas relacionados ao processo KDD.

5. PROBLEMAS RELACIONADOS AO PROCESSO DE EXTRAÇÃO DE CONHECIMENTO DE BASES DE DADOS

Nesta seção serão abordados alguns dos principais problemas relacionados ao processo KDD e algumas das principais medidas que podem ser adotadas para evitá-los.

5.1. DEFINIÇÃO DOS OBJETIVOS DA APLICAÇÃO

A clareza das metas a serem alcançadas influencia sobremaneira o êxito ou fracasso do processo KDD. Dessa forma, faz-se necessário ter um profundo conhecimento da base de dados do domínio de aplicação para dentre outras coisas, determinar a forma com que o conhecimento pode ser representado, de que maneira os resultados (conhecimento) vão ser empregados para evitar que o especialista, ao longo do processo, especifique muitos outros objetivos (não previsto na etapa de Compreensão do Domínio), o que pode acarretar um desperdício muito grande de tempo e

recursos na tentativa de encontrar "às cegas" o que se pretende obter por intermédio do processo KDD.

É necessário contar com a capacidade de gerência do analista do processo KDD para controlar essa "ansiedade" do especialista do domínio. Uma possível solução é apresentar o conhecimento extraído (e.g. utilizando alguma ferramenta de visualização) delimitando essa extração ao que foi previamente planejado na etapa de Compreensão do Domínio do processo KDD [7]. Por exemplo, se foi previamente especificado apenas a extração de padrões de adimplência e inadimplência a partir dos dados de um grupo de clientes, não proceder na busca de outras singularidades entre esses dados.

5.2. RUÍDO NOS DADOS

Assumir que todos os valores de um conjunto de dados estão totalmente corretos não representa uma hipótese sustentável, em se tratando de dados reais. A maioria dos algoritmos para Data Mining trabalham com a suposição de que os registros de uma base de dados podem incluir valores de atributos baseados nas medidas subjetivas e/ou de juízo, o que pode significar que esses valores representam erros. Os erros não sistemáticos deste tipo, em determinados valores dos atributos, são normalmente chamados de ruído.

O tratamento de ruído se dá geralmente pela consideração ou não da parte dos dados que apresentam ruídos. Os ruídos nos dados são geralmente identificados por intermédio da excepcionalidade (ou variação acentuada) de valores do conjunto de dados, mediante as técnicas estatísticas. Vale ressaltar que alguns pesquisadores, como Quinlan, consideram que não vale a pena destinar esforços para eliminar ruídos do conjunto de dados se é bastante provável que o conhecimento seja aplicado, na prática, em outros conjuntos de dados com ruído [17].

5.3. DADOS INCOMPLETOS

Em bases de dados reais é muito freqüente a ausência de alguns valores de atributos (dados incompletos), especialmente em bases de dados comerciais. Esse problema ocorre por muitos motivos, entre os quais, destacam-se [9, 18]:

- Armazenamento de dados impuros.
- Perda de dados (e.g. causada por problemas físicos de armazenamento).
- Revisão dos dados armazenados (e.g. novos atributos são considerados, quando há alguns meses atrás não o foram).
- Falta de observação de atributos preditivos ou que, apesar de aparentemente irrelevantes, em conjunto com outros atributos tenham elevado poder preditivo.
- Falhas na medição dos valores dos atributos.

Os métodos de tratamento de dados incompletos (*missing values*) em bases de dados têm recebido atenção especial em pesquisas relacionadas a KDD, especialmente aqueles que envolvem soluções estatísticas (e.g. redes bayesianas) [18, 22].

6. CONSIDERAÇÕES FINAIS

O constante avanço dos mecanismos de coleta e armazenamento de dados, bem como o inexorável processo de automação do mundo das ciências, negócios e governo, tem gerado a necessidade do uso de novas técnicas e ferramentas capazes de automatizar o processo de análise e entendimento dos dados. Portanto, sem uma forte ênfase em pesquisas de técnicas de extração de conhecimento de bases de dados, corre-se o risco de estar privado do verdadeiro "valor" de grande parte dos dados armazenados nessas bases.

Como forma de solucionar esse problema, KDD desponta como uma tecnologia capaz de cooperar, amplamente, na busca do conhecimento embutido nos dados. Desta forma, o principal objetivo do processo KDD é encontrar padrões válidos e potencialmente úteis nos dados.

Neste trabalho, foram focalizadas as principais etapas do processo KDD, bem como a importância da interação entre os usuários para o êxito desse processo. Além disso, foram apresentados os elementos que dão suporte à realização das tarefas pertinente a essas etapas, bem como os principais problemas encontrados para realizá-las.

7. BIBLIOGRAFIA CONSULTADA

1. AADMODT, A. , PLAZA, E. **Case-Based Reasoning**: Foundational Issues, Methodological Variations and System Approaches. AI Communications. S.l., v.7, n.1, p. 39-59, 1994.
2. BATISTA, G. E. A. P. A. **Um Ambiente de Avaliação de Algoritmos de Aprendizado de Máquina Utilizando Exemplos.** São Paulo: ICMC-USP, 1997. Dissertação de Mestrado
3. CARBONEL, J. G., LANGLEY, P. **Machine Learnin - Encyclopedia of Artificial Intelligence**. S.l., John Wiley & Sons, 1987. p. 464-488.
4. CHEESEMAN, P., STUTZ, J. **AutoClass**: A Bayesian Classification System. Readings in Machine Learning, Morgan Kauffman. S.l., p. 431-441, 1990.
5. CLARK, P., BOSWELL, R. **Rule Induction with CN2:** Some Recent Improvements. In Proc. 5th European Conference (EWSL 91).S.l., p. 151-163, 1991.
6. CODD, E. F. **Providing OLAP (On-line Analytical Processing) to User-Analyst**: An IT Mandate. S.l., E. F. Codd and Assoc., 1993.
7. DECKER, K. M., FOCARDI, S. **Technology Overview**: A Report on Data Mining. CSCS-ETH, S.l., Swiss Scientific Computer Center, 1995.
8. FÉLIX, L. C. M. **Data Mining no Processo de Extração de Conhecimento de Bases de Dados**. São Paulo: ICMC-USP, 1998. Dissertação de Mestrado.
9. FAYYAD, U. **Data Mining and Knowledge Discovery**: Making Sense Out of Data. IEEE Expert. S.l., v. 11, n.5, p. 20-25, 1996.
10. FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. **The KDD Process for Extracting Useful Knowledge from Volumes of Data.** In: Communication of the ACM. S.l., v.39, n.11, p. 27-34, November, 1996.
11. GLYMOUR, C. et al. **Statistic Themes and Lessons for Data Mining**. Data Mining and Knowledge Discovery. S.l., Kluwer Academic Publishers, v. 1, 1997.
12. HECKERMAN, D. **Bayesian Networks for Data Mining**. Data Mining and Knowledge Discovery. Kluwer Academic Publishers. S.l., v.l. , p. 79-119, 1997.
13. HAYKIN, S. **Neural Networks – A Comprehensive Foundation**. S.l., Macmillan College Publising Company, Inc., *1994.*
14. INMON, W. H. **The Data Warehouse and Data Mining**. In: Communication of the ACM. S.l., v.39, n. 11, p. 49-50, nov., 1996.
15. KLIBER, D., LANGLEY, P. **Machine Learning as a Experimental Science**. Machine Learning, S.l.. v.3, n.1, p. 5-8, 1988.
16. LANGLEY, P., IBA, W., THOMPSON, K. **An Analysis of Bayesian Classifiers**. In Proc. 10th National Conference on Artificial Intelligence. AAAI Press and MIT Press, p. 223-228, 1992.
17. MANNILA, H. **Data Mining**: Machine Learning, Statistic and Databases. Helsinki: Department of Computer Science/ University of Helsinki, 1997. URL: http://www.cs.helsinki.fi/~mannila/
18. QUINLAN, J. R. **C4.5**: Programs for Machine Learning. S.l., Morgan Kaufmann Publishers,1993.
19. RAMONI, M., SEBASTIANI, P. **Discovering Bayesian Networks in Incomplete Databases**. Knowledge Media Institute, The Open University, 1997. Kmi Technical Report, 46.
20. REZENDE, S. O. **Visualization for Knowledge Discovery in Database.** In: Proc. of International Conference on Data Mining, Rio de Janeiro, p. 81-96, set. ,1998.
21. REZENDE, S. O., PUGLIESI, J. B. **Aquisição de Conhecimento Explícito ou Manual**. São Paulo: ICMC/USP, n. 37, mar. 1998. (Notas do ICMC).
22. RICH, E., KNIGHT, K. **Inteligência Artificial**. 2. ed. São Paulo: Makron Books, 1993.
23. ROCHA, Cláudio Alex J. da. **Processo de Extração de Conhecimento de Bases de Dados, Considerando a Incorporação de conhecimento de Fundo e Tratamento de Dados Incompletos.** São Paulo: ICMC-USP, 1999. Dissertação de Mestrado.