Descoberta de Conhecimento em Bases de Dados

Célia Ramos | Prof.ª Adjunta na ESGHT - UALG
Fernando Lobo | Prof. Auxiliar na FCT - UALG

A sociedade actual é caracterizada por uma inundação de dados provenientes de diversas fontes, existentes numa vasta gama de áreas económicas, sociais, científicas, etc. A análise dos dados, armazenados em Bases de Dados, é cada vez mais pertinente para garantir a competitividade e o sucesso das organizações. Este artigo apresenta a nova área de conhecimento "Descoberta de Conhecimento em Bases de Dados" que permite efectuar análises dos dados e extrair conhecimento útil e adequado para o processo de tomada de decisão. O processo de Descoberta de Conhecimento em Bases de Dados é constituído por um conjunto de etapas, dentro das quais se encontra o *Data Mining*: um conjunto de técnicas que efectuam a extracção de conhecimento.

1. Introdução

Nos últimos anos, tem ocorrido um crescimento muito acentuado na dimensão das Bases de Dados referentes a dados de organizações empresariais, do governo ou de experiências científicas.

A existência e utilização de Bases de Dados de elevadas dimensões só é pertinente se existirem ferramentas adequadas para efectuar análises automáticas e inteligentes dos dados. A necessidade destas ferramentas torna-se cada vez mais evidente, e a sua relevância é cada vez mais perceptível se, por exemplo, nos apercebermos dos elevados volumes de transacções financeiras que são efectuados todos os dias ou do número de medicamentos prescritos, em cada dia, aos doentes nos centros de saúde.

Alguns dos tipos de suportes que podem ser utilizados para armazenar e extrair dados de elevadas dimensões são: Bases de Dados Relacionais (*Relational Databases*), *Data Warehouses*[1] (permitem a análise de dados cuja origem provem de diferentes fontes), Base de Dados de operações comerciais ou financeiras (*Transactions Databases*).

As ferramentas acima referidas estão inseridas numa nova área de conhecimento, designada por *Knowledge Discovery in Databases* (KDD), a qual está a surgir como uma das formas para resolver os problemas inerentes à análise de dados em Bases de Dados com grandes volumes de informação.

2. Knowledge Discovery in Databases

À área de conhecimento *Knowledge Discovery in Databases* (KDD), que em português tem a designação de "Descoberta de Conhecimento em Bases de Dados (DCBD)", tem sido atribuída várias designações como

por exemplo: *Data Mining* (DM), *Knowledge Extraction*, *Extraction of Knowledge from Databases* (EKDB), entre outros.

O termo *Data Mining* tem sido mais utilizado pelos estatísticos, analistas de dados e pela comunidade dos gestores dos Sistemas de Informação. Enquanto que, o termo *Knowledge Discovery in Databases* tem sido mais utilizado pelos investigadores da área de Inteligência Artificial e de *Machine Learning*.

No entanto, o termo *Knowledge Discovery in Databases* refere-se a todo o processo de descoberta de conhecimento útil realizado a partir dos dados, enquanto que o termo *Data Mining* refere-se apenas à aplicação de algoritmos para a procura de padrões extraídos a partir dos dados.

2.1 Definição de KDD

Para os autores desta área de conhecimento (Fayyad, 1996: 6), a definição mais correcta para KDD é "o processo não trivial de identificação válida, original, potencialmente útil, de padrões compreensíveis existentes nos dados". Os termos apresentados são explicados pelo autor como se segue:

- **Dados** (*Data*): é um conjunto de Factos (*F*), isto é, casos em Bases de Dados.

- **Padrão** (*Pattern*): é uma Expressão (*E*) numa Linguagem (*L*) que descreve factos num subconjunto F_E de *F*.

- **Processo** (*Process*): Geralmente os processos de KDD são constituídos por vários passos, que incluem a preparação dos dados, a procura de padrões, a avaliação de conhecimento e o refinamento, envolvendo iterações após respectiva modificação.

- **Validade** (*Validity*): Os padrões descobertos devem ser válidos perante novos dados.

- **Original** (*Novel*): Os padrões são novos (pelo menos para o sistema).

- **Potencialmente útil** (*Potentially Useful*): Potencialmente, os padrões devem conduzir para acções úteis.

- **Compreensível** (*Understandable*): Um dos objectivos de KDD é tornar os padrões compreensíveis para os humanos de forma a facilitar uma melhor compreensão dos dados extraídos.

A especificação das definições tem como objectivo identificar o que um algoritmo de *Data Mining*[2], utilizado num processo KDD, pode entender por conhecimento (*Knowledge*[3]).

A descoberta de conhecimento envolve a avaliação e a possível interpretação dos padrões para efectuar decisões sobre o que constitui ou não o conhecimento.

2.2 Etapas do processo de KDD

O processo KDD é interactivo e iterativo, envolvendo numerosos passos com muitas decisões a serem tomadas pelo utilizador.

A figura 1 apresenta os passos básicos inseridos no processo KDD, os quais são apresentados a seguir (Fayyad, 1996:10):

1. Desenvolver uma compreensão do domínio da aplicação.

2. Criar um conjunto de dados para analisar (*target data*).

3. Limpeza de Dados (*data cleaning*) e Preprocessamento (*preprocessing*).

4. Redução de Dados e Projecção para procurar características relevantes nos dados.

5. Escolher a tarefa ou o objectivo de *Data Mining*.

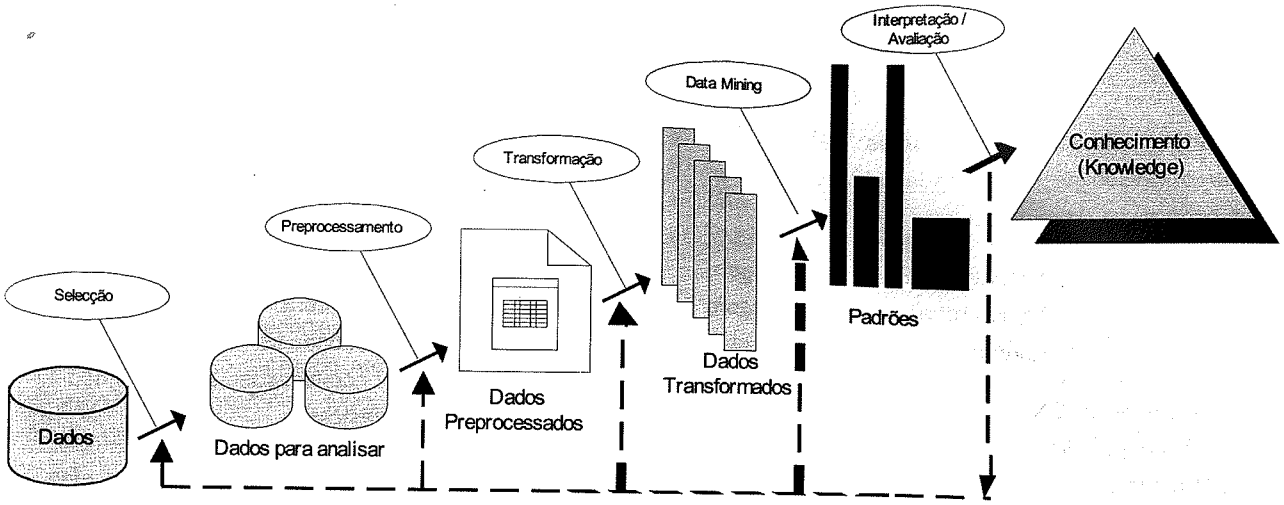
6. Escolher o(s) algoritmo(s) de *Data Mining*.

7. *Data Mining*.

8. Interpretar padrões resultantes da procura e, possivelmente, terá de retornar para um dos passos de 1-7 para nova iteração.

9. Consolidar o conhecimento descoberto.

Figura 1- Passos do Processo KDD. Fonte (Fayyad, 1996:10)



Os nove passos apresentados acima são considerados como os básicos no processo de KDD, no entanto, não apresentam a dimensão de iterações e de ciclos que podem constituir todo o processo de KDD. A grande parte do trabalho num processo de KDD está centrada no passo 7 designado por "*Data Mining*".

3. O que é o Data Mining

A definição apresentada por Usama Fayyad (Fayyad, 1996:11) refere que "*Data Mining* é uma componente do processo de KDD que frequentemente envolve aplicações de um método particular de *Data Mining,* o qual será utilizado repetidamente e de forma iterativa".

Considerando a definição acima proposta, o termo padrão é visto como um caso particular do modelo. Por exemplo, $f(x)=3x^2+x$ é um padrão onde $f(x)=\alpha x^2+\beta x$ é considerado o modelo.

Muitos dos métodos de *Data Mining* são baseados em conceitos de *Machine Learning*, Reconhecimento de Padrões e de Estatística.

3.1. As Tarefas de *Data Mining*

Na prática, os dois níveis principais dos objectivos primários de *Data Mining* tendem a ser de previsão ou de descrição. O tipo de padrões a descobrir depende da tarefa de *Data Mining* empregue:

- **Previsão:** implica a utilização de algumas variáveis ou campos da Base de Dados para prever valores, desconhecidos ou futuros, de outras variáveis relevantes.

- **Descrição:** centra-se principalmente na procura de padrões interpretáveis pelos humanos, os quais descrevem os dados.

A importância relativa da previsão e da descrição, para uma determinada aplicação de *Data Mining,* pode variar consideravelmente. Os objectivos da previsão e da descrição são concretizados recorrendo às tarefas primárias de *Data Mining*:

- **Classificação:** organiza os dados dentro de uma ou mais categorias pré-definidas.

Exemplo: Classificações das tendências nos mercados financeiros.

- **Regressão:** mapeia os dados numa variável de valor real.

Exemplos: a) De acordo com os valores de microondas medidas por sensores remotos, permite prever a quantidade de biomassa presente numa floresta; b) De acordo com os resultados das análises, permite estimar a probabilidade de que um paciente tem de viver; c) De acordo com uma função dos gastos em publicidade, permite prever os pedidos dos consumidores para um novo produto.

- **Agrupamento** ou ***Clustering:*** mapeia os dados dentro de classes, as quais são definidas a partir dos dados.

Exemplo: descobertas de subpopulações de consumidores nas Bases de Dados de Marketing.

- **Sumarização:** descrições compactas de um subconjunto de dados. As técnicas de sumarização são muitas vezes aplicadas às análises de dados

exploradas iterativamente e à geração automática de relatórios.

Exemplo: Análise do total de vendas efectuado mensalmente.

• **Modelização das Dependências:** permite analisar o vínculo de determinada relação entre os campos numa Base de Dados.

Exemplo: desenvolvimento de Sistemas Inteligentes para identificação de probabilidades médicas e na modelação do genoma humano.

• **Alterações e Detecção de Desvios:** também designada por **Sequência de Análise,** define o modelo das sequências de padrões.

3.2. Componentes dos Algoritmos de *Data Mining*

Após a apresentação das tarefas primárias de *Data Mining*, o próximo passo é construir algoritmos para as resolver. Em qualquer algoritmo de *Data Mining*, podem ser identificadas três componentes principais: a representação do modelo, a avaliação do modelo (critério de preferência) e a procura (algoritmo de procura).

• A **representação do modelo** é a linguagem de representação *L* para descrever padrões que se podem descobrir.

• A **avaliação do modelo** estima como um padrão, em particular, se encaixa bem ou não no processo de KDD.

• O **algoritmo de procura** é composto por duas componentes:

1. Procura de Parâmetros, o algoritmo deve procurar pelos parâmetros que optimizam o critério de avaliação do modelo.

2. Procura do modelo, ocorre uma mudança nos parâmetros do método de procura e a representação do modelo é alterada.

Geralmente, um algoritmo de *Data Mining* é uma instância das componentes: modelo, avaliação e procura.

4. Métodos de Data Mining

Existe uma variedade de métodos de *Data Mining*, especializados para determinados tipos de dados e de domínios. A seguir são apresentados alguns dos mais relevantes:

• **Métodos das Árvores de Decisão e das Regras de Indução:** utilizam especificações que têm uma representação formal muito simples e que tornam o modelo deduzido relativamente fácil de compreender.

• **Métodos de Regressão Não Linear e de Classificação:** estes métodos consistem numa família de técnicas para previsão, onde as variáveis de entrada são constituídas por combinações lineares e não-lineares de funções básicas.

• **Métodos Baseados em Exemplos:** são efectuadas previsões sobre novos exemplos, utilizando as propriedades de exemplos semelhantes e cuja previsão já é conhecida.

• **Modelos Gráficos de Dependências Estatísticas:** especificam as dependências probabilísticas para salientar um determinado modelo.

• **Caracterização:** permitem a definição de regras de caracterização, ou seja, o resumo das características gerais dos objectos presentes numa classe.

• **Discriminação:** permitem a definição de regras de discriminação, tendo por base a comparação das características gerais de objectos entre duas classes referenciadas como a classe alvo e a classe de contraste.

• **Análises de Associações:** permitem a descoberta de relações e correlações entre as associações. É a descoberta que é designada por "regra de associação".

• **Classificações:** permitem a organização dos dados em determinadas classes. A classificação é efectuada pela atribuição de etiquetas às classes.

• **Previsão ou predição**[4]**:** permitem a previsão de valores futuros através de análises efectuadas aos dados do passado[5].

• **Agrupamentos:** permitem a organização dos dados em classes.

• **Análises de Destaques (*Outliers*):** permitem a identificação de elementos que não podem ser agrupados numa determinada classe ou num agrupamento.

• **Análises de Séries Temporais**, **Análises de Evolução e de Desvio:** permitem o estudo das alterações ocorridas nos dados, ao longo de um determinado tempo.

• Existem ainda outros como por exemplo a **Lógica Fuzzy** e os **Conceitos de Vizinhança**.

5. Aplicações de KDD

As aplicações que recorrem a técnicas de *Data Mining* existem numa vasta gama de áreas. A seguir são apresentados alguns exemplos elucidativos desta realidade (Ullman, 2000):

• No mundo dos negócios, a aplicação de KDD mais usada e com mais sucesso é designada por "*Database Marketing*" que é um método para analisar as Bases de Dados dos clientes; é utilizada para procurar padrões entre as preferências do cliente e permite utilizar esses padrões para efectuar uma selecção direccionada a um alvo de futuros clientes. Outras utilizações são a análise e selecção de *stocks* e de outros instrumentos financeiros, a detecção e prevenção de fraudes.

• Finanças: tendo por base historiais de empréstimos bancários, são construídas árvores de decisão que permitirão decidir em que situação se concede o empréstimo.

• Recursos Humanos: para procurar padrões de comportamento dos viajantes, de forma a gerir o desconto dos lugares de avião, dos quartos de hotel, entre outros.

• Astronomia: por observação do céu, permite a distinção entre galáxias, estrelas, entre outros astros.

• Biologia molecular: por comparação de genomas, permite a detecção de diabetes.

• Modelação de alterações climáticas: por análise de padrões de espaço - tempo, permite a detecção de ciclones.

O potencial das aplicações de KDD (Brachman, 1996) é a sua utilização em ambientes em constante mutação, onde não existem modelos pré-definidos e em que são requeridas decisões baseadas em conhecimentos específicos do domínio do problema.

Os custos e benefícios de uma aplicação dependem das alternativas, da relevância dos dados, do volume dos casos, da complexidade da aplicação, da qualidade em termos de erros, da acessibilidade dos dados, das alterações nos dados e da existência de um especialista no domínio da aplicação.

5.1 Linhas gerais para seleccionar uma potencial aplicação de KDD

Segundo o autor Usama Fayyad (Fayyad, 1996), o critério para seleccionar aplicações pode ser dividido numa parte prática e numa parte técnica.

O **critério prático** inclui considerações sobre:

• O potencial do impacte significativo de uma aplicação.

• A não existência de boas alternativas e a dificuldade na obtenção de soluções.

• O apoio organizacional.

• O potencial de existirem problemas de invasão de privacidade e legalidade.

O **critério técnico** inclui considerações sobre:

• A disponibilidade de dados suficientes (casos).

• A existência de atributos relevantes.

• Os baixos níveis de ruído (poucos erros nos dados).

• A possibilidade de anexar intervalos de confiança ao conhecimento extraído.

• O conhecimento prévio ou conhecimento do domínio.

5.2 Privacidade e a descoberta de conhecimento

Segundo o mesmo autor (Fayyad, 1996), quando se trabalha com Bases de Dados de informação privada relacionada com pessoas, governos e negócios, tem de se ter muito cuidado de forma a não existir invasão de privacidade. O procedimento que tem sido adoptado é o de sugerir que os dados, sobre aspectos específicos da vida das pessoas, não sejam analisados sem o seu consentimento e apenas devam ser coleccionados para um determinado objectivo.

Em muitos casos (médicos, socio-económicos,...), o objectivo é descobrir padrões de conhecimento sobre

grupos e não sobre pessoas individuais. Enquanto que a descoberta de padrões sobre grupos não parece violar as restrições na extracções de dados pessoais, a combinação engenhosa de vários grupos de padrões, especialmente em pequenos conjuntos de dados, pode permitir a extracção de dados de uma determinada pessoa.

Nesta situação, pode ser efectuado um tratamento prévio aos dados, em que estes são transformados em anónimos, por exemplo, através da eliminação dos atributos que permitem identificar os indivíduos e da atribuição aleatória de um número diferente a cada registo pessoal. O algoritmo de *Data Mining* quando extrai dados, mesmo agrupados de forma engenhosa, não possibilita a detecção da identidade de indivíduos ou de organizações porque esta aparece codificada salvaguardando assim a invasão de propriedade.

5.3 Preocupações dos Investigadores de KDD

A área de investigação de KDD, apesar de ser relativamente recente, coloca algumas preocupações aos seus investigadores que a seguir são apresentadas:

- **Sociais:** Em relação à privacidade de dados, que podem pertencer a pessoas individuais ou a empresas.

- **Instrumento de Interacção Homem-Máquina**: A descoberta de conhecimento só é importante se for perceptível para o utilizador.

- **Metodologias de extracção de conhecimento:** Devido às fontes, a informação tem diversas características que levantam imensas limitações, desde o controlo à dimensão do espaço necessário para cálculo e armazenamento de dados.

- **Desempenho:** Problemas de eficiência e de escalonamento, devido ao elevado número de dados a analisar.

- **Fontes de dados:** Diferentes tipos de dados requerem diferentes tipos de algoritmos e de métodos para os processar.

5.4 Desafios na Investigação e nas aplicações de KDD

A área de investigação de KDD, devido à sua complexidade e a todos os passos que lhe estão associados, coloca vários desafios e respectivas soluções aos seus investigadores que a seguir são apresentados:

- **Bases de Dados de elevadas dimensões:**
Identificar algoritmos eficientes para enumerar todas as associações que excedem um determinado limite de confiança em Bases de Dados de grandes dimensões (Agrawal et al, 1996).

- **Dimensão elevadíssima (muitos campos):**
Incluir métodos para reduzir a dimensão efectiva do problema.

Utilizar a noção de conhecimento, adquirido previamente, para identificar variáveis irrelevantes.

- **Alteração de dados e de conhecimento:**
Incluir métodos incrementais para actualização dos padrões, para tratar as alterações como uma nova oportunidade para descobrir e para interpretar os resultados das novas procuras efectuadas, a partir de padrões que foram alterados (Matheus et al, 1996).

- **Dados perdidos e com ruído:**
Incluir estratégias estatísticas mais sofisticadas para identificar variáveis escondidas e dependências (Heckerman, 1996).

- **Existência de relações complexas entre os dados:**
Desenvolver novas técnicas para detectar relações entre variáveis (Dzeroski, 1996).

- **Integração com outros sistemas**[6]**:**
Incluir integrações com Sistemas de Gestão de Bases de Dados.
Integrar com folhas de cálculo e com ferramentas de visualização.

6. Conclusão

O KDD permite o desenvolvimento de novas ferramentas para a gestão, análise e, eventualmente, o controlo sobre a inundação de dados que caracteriza a sociedade moderna.

O processo de KDD é um conjunto de actividades contínuas, desde a entrada de elevados volumes de dados até à partilha e utilização do conhecimento descoberto, num ambiente de tomada de decisão.

Esta área científica é conduzida pelas fortes necessidades sociais e económicas que impelem o seu crescimento.

Este documento apresentou as várias facetas do processo de KDD. O que ficou por fazer foi uma descrição pormenorizada das diversas técnicas de *Data Mining* que permitem com que o processo de KDD seja posto em prática. Esta descrição pormenorizada será explorada num documento futuro.

Notas:

[1] Um método muito popular utilizado nas análises de *Data Warehouse* é chamado Processamento Analítico *On-line* (*On-line Analytical Processing* - OLAP) (Azevedo, 2001).

[2] *Data Mining* é um passo no processo KDD e consiste na utilização de algoritmos particulares que, perante limitações de eficiência computacionalmente aceitáveis, produz um conjunto de padrões $E_{j, \text{ com } j=1,...,n}$ encontrados em F.

[3] A definição de conhecimento não significa que é absoluta, uma vez que é puramente orientada para o utilizador e determinada pelas funções e pelas fronteiras que o utilizador escolheu.

[4] As técnicas mais utilizadas são: Análises Estatísticas, Algoritmos Genéticos e Redes Neuronais.

[5] Os dados do passado também são designados por "dados históricos" ou "historial".

[6] Exemplos de sistemas KDD integrados são descritos por vários autores em (Simoudis et al, 1996) e (Shen et al, 1996).

Referências Bibliográficas

AGRAWAL, R. , Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. (1996) "Fast Discovery of Association Rules" In Fayyad, U. et al (1996) *Advances in Knowledge Discovery and Data Mining*, pp. 307-328, AAAI Press/MIT Press, California.

AZEVEDO, Paulo J. (2001), *Técnicas Avançadas para Bases de Dados*, Universidade do Minho, Departamento de Informática.

BRACHMAN, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E. (1996) "Mining Business Databases" In *Communications of the ACM*, Vol. 39, Nº. 11, pp. 42-48.

DZEROSKI, S. (1996) "Inductive Logic Programing and Knowledge Discovery in Databases" In Fayyad, U. et al (1996) *Advances in Knowledge Discovery and Data Mining*, pp. 117-152, AAAI Press/MIT Press, California.

FAYYAD, Usama M. et al (1996), *Advances in Knowledge Discovery and Data Mining,* AAAI Press/MIT Press, California.

HAUSSLER, D., Fayyad, U. and Stolorz, P. (1996) "Mining Scientific Data" In *Communications of the ACM*, Vol. 39, Nº. 11, pp. 51-57.

HECKERMAN, D. (1996) "Bayesian Networks for Knowledge Discovery" In Fayyad, U. et al (1996) *Advances in Knowledge Discovery and Data Mining*, pp. 273-305, AAAI Press/MIT Press, California.

MATHEUS, C. J., Piatetsky-Shapiro, G., Perona, P. (1996) "Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data" In Fayyad, U. et al (1996) *Advances in Knowledge Discovery and Data Mining*, pp. 495-515, AAAI Press/MIT Press, California.

SHEN, W., Ong, K., Mitbander, B. and Zaniolo, C. (1996) "Metaqueries for Data Mining" In Fayyad, U. et al (1996) *Advances in Knowledge Discovery and Data Mining*, pp. 375-398, AAAI Press/MIT Press, California.

SIMOUDIS, E., Livezey, B. And Kerber, R. (1996) "Integrating Inductive e Deductive Reasoning" In Fayyad, U. et al (1996) *Advances in Knowledge Discovery and Data Mining*, pp. 353-373, AAAI Press/MIT Press, California.

ULLMAN, J. (2000), http://www-db.stanford.edu/~ullman/mining/mining.html.