



UNIPAC - CENTRO UNIVERSITÁRIO PRESIDENTE ANTÔNIO CARLOS
CAMPUS BARBACENA

Bacharelado em Ciência da Computação



Mineração de dados

Material de Apoio

Parte I – Conceitos iniciais

Prof. Felipe Roncalli de Paula Carneiro
felipecarneiro@unipac.br

2º sem / 2023

Material cedido pela Profª Livia e Profº Osvano

Explosão de dados

- Quantidades gigantescas de dados são coletados e armazenados em empresas, indústrias, pesquisas científicas, comércio eletrônico, clínicas e hospitais, etc.

Explosão de dados

- Dados de comércio eletrônico;
- Dados de navegação na internet;
- Dados de compras de clientes em grandes lojas de departamentos, supermercados;
- Dados de transações bancárias, ou de cartão de crédito.
- Sensores remotos em satélites;
- Telescópios;
- Microarrays gerando dados de expressões de genes; etc.

Explosão de dados

- Dados coletados e armazenados a velocidades enormes;
- Técnicas tradicionais não são apropriadas para analisar tais dados:
 - Ruídos (dados incoerentes) e grande dimensionalidade.

Efeitos da Explosão de Dados

- Estima-se que a quantidade de informação é duplicada a cada *vinete meses*.
- Somos ricos em dados e pobres em informação e conhecimento (número grande de informações não significa conhecimento).
- Necessidade: técnicas para explorar grande volume de dados e transformar esses dados em conhecimento.
- Existem informações escondidas nos dados que não são evidentes – muitos conjuntos de dados nunca são analisados.

Mineração de dados

- A mineração de dados (ou *data mining*) está relacionada à subárea da Estatística chamada análise de dados exploratória, que tem objetivos semelhantes e conta com medidas estatísticas.
- Ela também está intimamente relacionada às subáreas da inteligência artificial chamada descoberta de conhecimento e aprendizado de máquina.

Mineração de dados

- Se dedica a explorar e analisar o gigantesco número de dados produzido pela sociedade, visando encontrar tendências ou padrões nesses dados. Com expectativa de que as ferramentas encontrem esses padrões com a menor interatividade com o usuário.
- Surgiu o termo Big Data, utilizado para descrever esses grandes volumes de dados.

Big data

- Refere-se a um conjunto muito grande de dados gerados e armazenados, no qual os aplicativos de processamento desses dados tradicionais ainda não conseguem atuar em um tempo aceitável.
- Representa uma nova era na sociedade, na qual os dados transformam-se em informações valiosas, mudando a forma como agimos, atuando sobre as tomadas de decisões e impactando, inclusive, na economia e na ciência.

Big data

- Passou a ser essencial para as relações econômicas e sociais, representando evolução nos sistemas de negócio e na ciência e tecnologia. Ferramentas de Big Data são de extrema importância na definição de estratégias de marketing, para o aumento de produtividade, na redução de custos e na tomada de decisões mais inteligentes, gerando valor para os negócios.

Conceitos



- Dado: algo bruto; é a matéria-prima da qual podemos extrair informação.
- Informação: dado processado, com significado e contexto bem definido.
- Conhecimento: uso inteligente da informação; é a informação contextualizada e utilizada na prática.

Dado

- Do lat. *Datus*, elemento ou quantidade conhecida que serve de base à solução de algum problema.
- É um estímulo sensorial: é um fato, uma imagem ou um som.
- Qualquer representação simbólica de fatos ou ideias do qual pode ser extraído informação.

Informação

- Do lat. *Informatio*, significado de um Dado.
- A informação existe quando o cérebro humano recebe um conjunto de dados e os processa neuralmente.
- Informação é a compreensão do dado, é a matéria prima para a atividade cerebral.

Conhecimento

- Do lat. *Cognoscere*, ato ou efeito de compreender as propriedades e as características de alguma coisa; Entendimento; intuição.
- Consiste nos processos de Percepção e Reflexão (a cognição)

Aprender

- Do lat. *Apprehendere*, adquirir conhecimento de alguma coisa; ficar sabendo; instruir-se.
- Habilidade para adaptar-se ao meio ambiente.

Descobrir

- Do lat. *Discooperire*, encontrar o que era desconhecido, que estava escondido, achar.
- Descobrir é também divulgar e revelar.

KDD

- *Knowledge Discovery in Databases* – processo de descoberta de conhecimento em bases de dados.
- É a extração automatizada de novos fatos e relações em dados com informações não óbvias.
- Encontrar e divulgar “alguma coisa” que era desconhecida, compreendendo suas propriedades e características, a partir de um conjunto de dados, proporcionando um significado útil que permita aumentar a capacidade de adaptar-se ao meio ambiente (Luis E. Zárate).

KDD

- No KDD procura-se descobrir o conhecimento existente num conjunto de dados, através da interatividade do engenheiro de conhecimento e as técnicas de data mining.
- Informação incompleta tem valor se colocarmos restrições às nossas conclusões.

Tipos de conhecimento prévio

- Explícito: representado por modelos, algoritmos, documentos, descrição de procedimentos, desenhos, sínteses e principalmente ***Bases de Dados***.
- Ele é frequentemente ***redundante ou incompleto***, definido pelas circunstâncias que o geraram.

Tipos de conhecimento prévio

- Tácito: é de difícil representação, e é adquirido pela prática da *experiência*. Refere-se às habilidades pessoais e às aptidões profissionais

Conhecimento procurado pelo KDD

- O KDD trabalha sobre o conhecimento explícito reforçado pelo conhecimento tácito no objetivo da descoberta organizada de padrões e de conhecimento útil.

(Luis Enrique Zárate)

Domínio = tácito + explícito

- É necessário ter uma boa ideia do domínio do problema, dentro de um determinado contexto, para ser efetivos na extração do conhecimento.
- Os conhecimentos Tácito (do especialista) e Explícito contribuem para uma boa definição do domínio de problema sendo observado.

KDD - Etapas

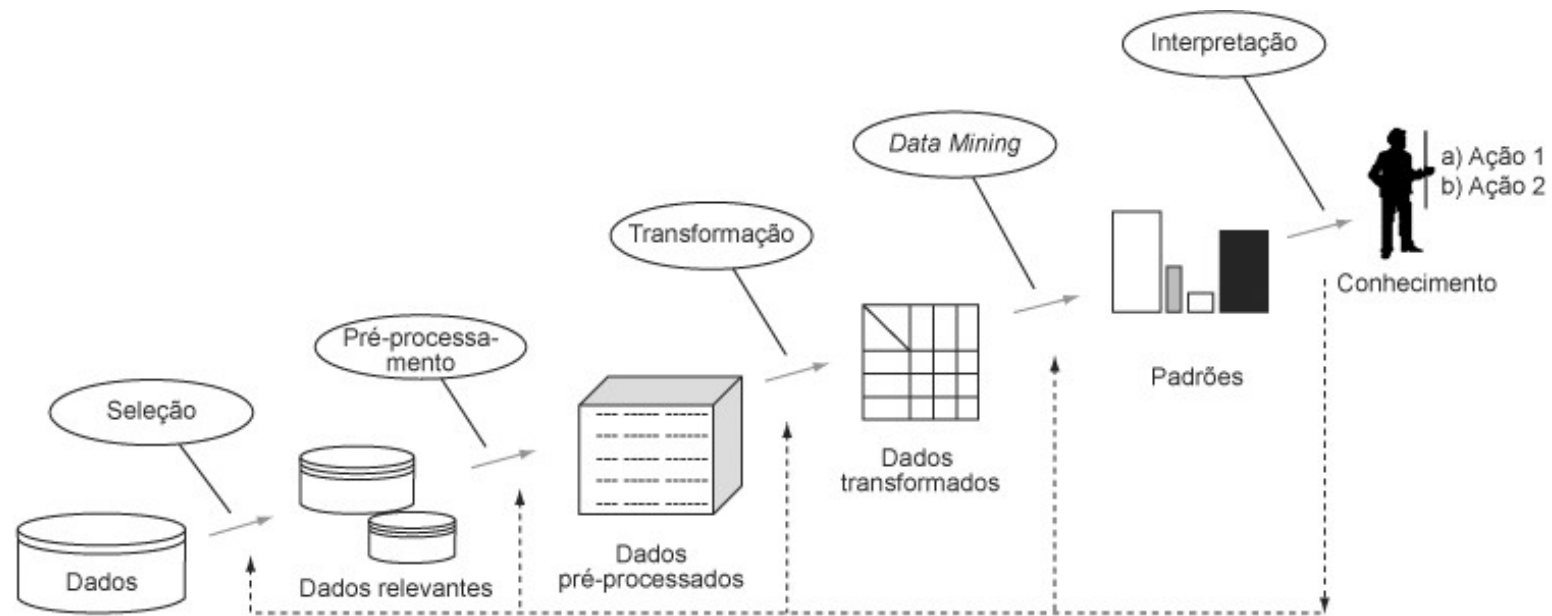


Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

Mineração de dados

- É o processo da descoberta de novas informações a partir de grandes massas de dados.
- Descoberta de padrões não triviais (implícitos e desconhecidos) em grandes bases de dados.
- Processo de extração de modelos uteis de um conjunto de dados.

Mineração de dados

- Técnicas de Mineração podem ajudar analistas:
 - Entender e prever as necessidades de clientes;
 - Descobrir fraudes;
 - Descobrir perfis de comportamento de clientes.
- Técnicas de Mineração podem ajudar cientistas:
 - Classificar e segmentar dados;
 - Formular hipóteses.

Mineração de dados: Exemplos

- Uma consulta à um banco de dados (SQL, por exemplo) -> *NÃO*
- Procurar um nome em uma lista qualquer (telefônica, por exemplo) -> *NÃO*
- Descobrir se existe alguma ligação entre atributos em um banco de dados -> *É*
- Descobrir se certos nomes aparecem com mais frequência em determinadas regiões da cidade -> *É*

Mineração de dados: Exemplos

- Qual o perfil do cliente que consome mais ?
- Que produtos são comprados conjuntamente ? E em sequência ?
- Como as chuvas, variação de temperatura, aplicação de pesticidas afetam as colheitas ?
- Existe uma relação entre o aquecimento global e a frequência e intensidade das perturbações no ecossistema tais como secas, furacões, enchentes?
- Há alguma relação entre determinada doença e padrão de consumo?