UNIVERSIDADE JEAN PIAGET DE MOÇAMBIQUE

- **DISCIPLINA: <u>MINERAÇÃO DE DADOS</u>**

- **DOCENTE: ENGº.** CARLITOS GOVE

MINERAÇÃO DE DADOS

Introdução à Mineração de Dados

Introdução à Mineração de Dados

- **“Mineração:**

1. ação ou efeito de minerar; trabalho de extração do minério.
2. depuração do minério extraído das minas.”

- Podemos usar como base o conceito de mineração encontrado no dicionário, pois seu objetivo com os dados, é encontrar os padrões e as informações que seriam consideradas como “pepitas” na extração e busca pelo ouro.
- Seu objetivo principal é utilizar dos conceitos de estatística e aprendizado de máquina (Machine learning, ML) para gerar resultados, predições e padrões relevantes, sendo que com consultas SQL apenas, seriam inviáveis. Vamos revisar alguns conceitos da área:

Mineração de dados-Cont...

- **Conceito:**
- **Prospecção de dados** ou **mineração de dados** (também conhecida pelo termo inglês ***data mining***) é o processo de explorar dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.
- No campo da administração, a mineração de dados é o uso da tecnologia da informação para descobrir regras, identificar fatores e tendências-chave, descobrir padrões e relacionamentos ocultos em grandes bancos de dados para auxiliar a tomada de decisões sobre estratégia e vantagens competitivas.
- Esse é um tópico recente em ciência da computação, mas utiliza várias técnicas da estatística, recuperação de informação, inteligência artificial e reconhecimento de padrões.

Mineração de dados – Cont..

- A **mineração de dados** é formada por um conjunto de ferramentas e técnicas que através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística, são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. Esse conhecimento pode ser apresentado por essas ferramentas de diversas formas: agrupamentos, hipóteses, regras, árvores de decisão, grafos, ou dendrogramas.
- O ser humano sempre aprendeu observando padrões, formulando hipóteses e testando-as para descobrir regras. A novidade da era do computador é o volume enorme de dados que não pode mais ser examinado à procura de padrões em um prazo razoável. A solução é instrumentalizar o próprio computador para detectar relações que sejam novas e úteis.

Etapas da mineração de dados

- Os passos fundamentais de uma mineração bem sucedida a partir de fontes de dados (bancos de dados, relatórios, logs de acesso, transações, etc.) consistem de uma limpeza (consistência, preenchimento de informações, remoção de ruído e redundâncias, etc.). Disto nascem os repositórios organizados (Data Marts e Data Warehouses).
- É a partir deles que se pode selecionar algumas colunas para atravessarem o processo de mineração. Tipicamente, este processo não é o final da história: de forma interativa e frequentemente usando visualização gráfica, um analista refina e conduz o processo até que os padrões apareçam. Observe que todo esse processo parece indicar uma hierarquia, algo que começa em instâncias elementares (embora volumosas) e terminam em um ponto relativamente concentrado.

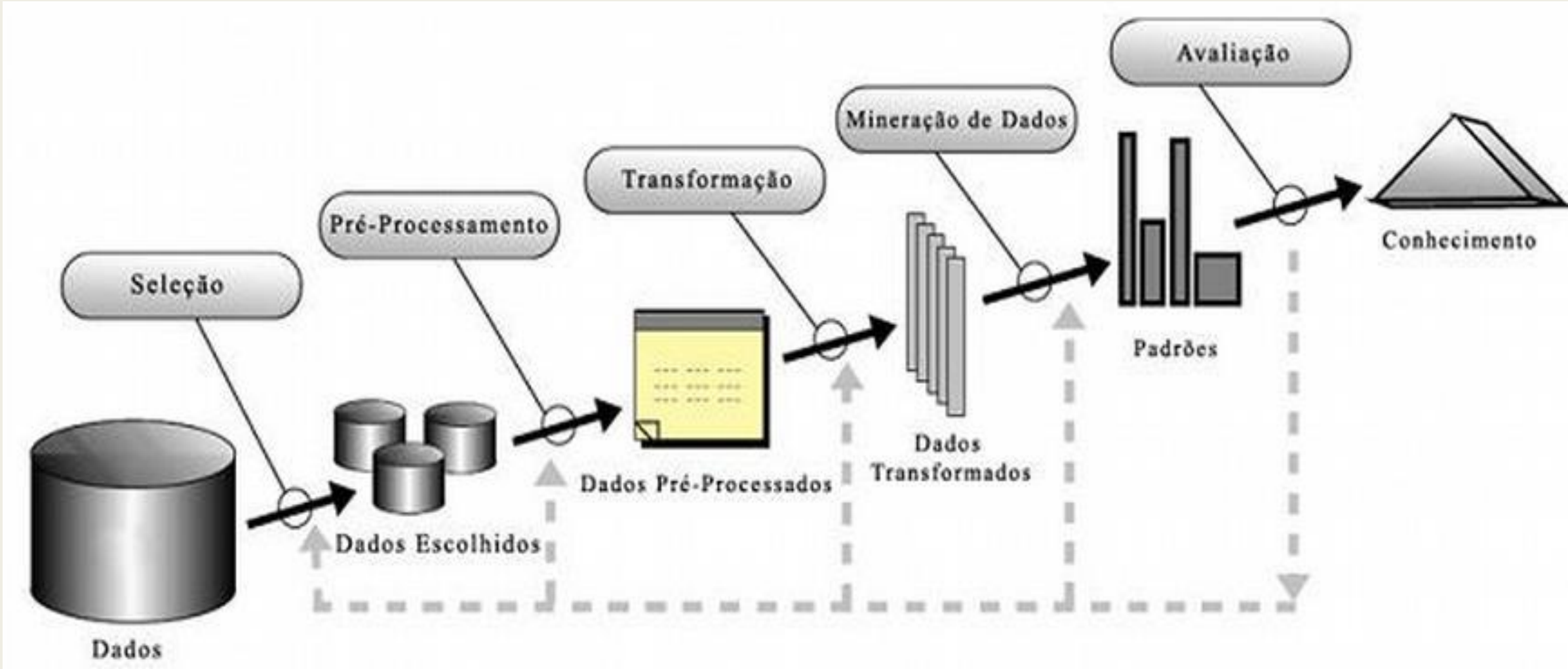
Etapas da mineração de dados – Cont...

- Encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados" de forma a desconsiderar aquilo que é específico e privilegiar e/ou valorizar tudo o que for generalizado. Em um determinado produto uma única data pode apenas significar que esse cliente em particular procurava grande quantidade desse produto naquele exato momento. Mas isso provavelmente não indica nenhuma tendência de mercado.

Mineração de Dados - Cont...

- **Inteligência Artificial:** sistemas que aprendem, e seguem aprendendo conforme surgem novas possibilidades, onde dado um novo cenário, ele responde conforme as probabilidades de resultados para cada movimento.
- **Aprendizado de Máquina**: os programas normalmente recebem dados de entrada, e após algum processamento são retornados os dados alterados na saída. ML utiliza da inteligência artifical, para receber dados, aprender com os mesmos e gerar programas na saída, programas que são criados com base nos padrões da base de dados.
- Então, cabe à mineração de dados utilizar destes componentes para explorar grandes quantidades de dados e facilitar tarefas que seriam extremamente exaustivas para os humanos. Mas e como faz? Um dos processos mais importantes na mineração é o KDD (Knowledge Discovery in Databases), uma série de passos de normalização dos datasets para manipulação e visualização de resultados.

Figura 1: Etapas do processo de descoberta do conhecimento (KDD)



Fonte: Processo de KDD, por Fayyad et. al.

Etapas do processo de descoberta do conhecimento (KDD)-Cont..

- **Seleção**: a primeira etapa consiste da seleção e formação do dataset, que pode incluir subconjuntos de dados de várias fontes, sendo algumas destas, API's, planilhas, dados abertos, sistemas, data warehouses, etc.
- **Pré-Processamento**: esta etapa visa verificar a qualidade dos dados. Muitas das bases normalmente vêm com dados faltantes ou inconsistentes, os quais devem ser ajustados de acordo com os princípios da consulta, para evitar os chamados ruídos. Por exemplo, em uma base que contém dados nulos, estes devem ser corrigidos ou removidos, dependo do objetivo da mineração.
- **Transformação**: etapa de normalização, agregação, inserção de novos atributos, redução e sintetização dos dados. Os dados nesta etapa, são adaptados, dependendo do tipo de algoritmo que será rodado.

Etapas do processo de descoberta do conhecimento (KDD)-Cont..

- **Mineração**: aplicar técnicas e algoritmos dependendo de cada objetivo, como verificar hipóteses ou descobrir padrões de forma autônoma, que sejam úteis e desconhecidas aos analistas. Cabe nesta fase também, verificar quais algoritmos se comportaram melhor para aquela base de dados.
- **Avaliação**: na última etapa, é feita a análise dos dados para que seja apresentado o conhecimento adquirido com aquelas informações e como irão impactar nos processos de decisão, com o propósito de deixar as informações mais simples de serem entendidas e apresentar sua relevância.

Estas etapas podem ser visualizadas como um fluxograma, onde os passos vistos na figura 1, podem ser ajustados dependo dos resultados.

Nada impede, por exemplo, de retornar à etapa de pré-processamento depois de realizar a etapa de mineração, sendo que os dados não foram apresentados da maneira esperada.

Etapas do processo de descoberta do conhecimento (KDD)-Cont..

- Para aplicar a mineração, é fundamental definir bem os problemas que se pode resolver com cada base de dados. Este processo é bem simples, seguindo alguns passos, como a descrição do dataset, da classe, e os atributos da base que vamos utilizar para resolver o problema. Considere uma imobiliária que quer analisar o banco de dados de vendas de propriedades, e que deseja descobrir quais variáveis mais influenciam no preço da venda.

- **Descrição**: Dados sobre o histórico de venda de propriedades de um imobiliária.
- **Classe**: Quais variáveis mais influenciam no preço de venda das propriedades?
- **Atributos**: área da propriedade, quantidade de quartos e preço da venda.

- Após esta descrição é necessário definir em qual tipo de mineração tal problema se encaixa, sendo esta classificação fundamental para a escolha dos algoritmos. Esta análise requer de um estudo aprofundado nos 4 tipos, que são a Associação, Regressão, Classificação e Clusterização.

Tipos de informação obtidos com a Mineração de Dados

- Com o uso da Mineração de dados, é possível descobrir informações relacionadas a associações, sequências,classificação, aglomeração e prognósticos.

- **Associações:** São ocorrências ligadas a um único evento. Por exemplo:um estudos de modelos de compra em supermercados pode revelar que, na compra de salgadinhos de milho, compra-se também um refrigerante tipo cola em 65% das vezes: mas, quando há uma promoção, o refrigerante é comprado em 85% das vezes.Com essas informações, os gerentes podem tomar decisões mais acertadas pois aprenderam a respeito da rentabilidade de uma promoção.
- **Sequências:** Na sequência os eventos estão ligados ao longo do tempo. Pode-se descobrir, por exemplo, que quando se compra uma casa, em 65% as vezes se adquire uma nova geladeira no período de duas semanas; e que em 45% das vezes, um fogão também é comprado um mês após a compra da residência.

Tipos de informação obtidos com a Mineração de Dados – Cont..

- **Classificação:** Reconhece modelos que descrevem o grupo ao qual o item pertence por meio do exame dos itens já classificados e pela inferência de um conjunto de regras. Exemplo: empresas de operadoras de cartões de crédito e companhias telefônicas preocupam-se com a perda de clientes regulares, a classificação pode ajudar a descobrir as características de clientes que provavelmente virão abandona-las e oferecer um modelo para ajudar os gerentes a prever quem são, de modo que se elabore antecipadamente campanhas especiais para reter esses clientes.
- **Aglomeração** (***clustering***): Funciona de maneira semelhante a classificação quando ainda não foram definidos grupos. Uma ferramenta de data mining descobrirá diferentes agrupamentos dentro da massa de dados. Por exemplo ao encontrar grupos de afinidades para cartões bancários ou ao dividir o banco de dados em categorias de clientes com base na demografia e em investimentos pessoais.

Tipos de informação obtidos com a Mineração de Dados – Cont..

- **Prognóstico:** Embora todas essas aplicações envolvam previsões, os prognósticos as utilizam de modo diferente. Partem de uma série de valores existentes para prever quais serão os outros valores. Por exemplo um prognóstico pode descobrir padrões nos dados que ajudam os gerentes a estimar o valor futuro de variáveis com números de vendas.
- Esses sistemas realizam uma análise de alto nível quanto a padrões ou tendências, mas também podem esmiuçar os dados para revelar mais detalhes, se necessário. Existem aplicações de data mining para todas as áreas funcionais da empresa, bem como para o trabalho científico ou governamental. É como usar o data mining para analisar detalhadamente padrões em dados sobre consumidores e, a partir disso, montar campanhas de marketing um-a-um ou identificar clientes lucrativos (LAUDON & LAUDON, 2011, p. 159).

Localizando padrões

- Padrões são unidades de informação que se repetem. A tarefa de localizar padrões não é privilégio da mineração de dados. O **cérebro** dos seres humanos utiliza-se de processos similares, pois muito do conhecimento que temos em nossa mente é, de certa forma, um processo que depende da localização de padrões. Para exemplificar esses conceitos, vamos propor um breve exercício de indução de regras abstratas. Nosso objetivo é tentar obter alguma expressão genérica para a seguinte seqüência:
- **Sequência original:** ABCXYABCZKABDKCABCTUABEWLABCWO
- Observe atentamente essa seqüência de letras e tente encontrar alguma coisa relevante. Veja algumas possibilidades:
- **Passo 1:** A primeira etapa é perceber que existe uma seqüência de letras que se repete bastante. Encontramos as sequências "AB" e "ABC" e observamos que elas ocorrem com freqüência superior à das outras sequências.

Localizando padrões – Cont...

- **Passo 2:** Após determinarmos as sequências "ABC" e "AB", verificamos que elas segmentam o padrão original em diversas unidades independentes:

"ABCXY"

"ABCZK"

"ABDKC"

"ABCTU"

"ABEWL"

"ABCWO"

Localizando padrões – Cont...

- **Passo 3:** Fazem-se agora induções, que geram algumas representações genéricas dessas unidades:

"ABC??" "ABD??" "ABE??" e "AB???",

onde '?' representa qualquer letra

- No final desse processo, toda a seqüência original foi substituída por regras genéricas indutivas, o que simplificou (reduziu) a informação original a algumas expressões simples. Esta explicação é um dos pontos essenciais da mineração de dados, como se pode fazer para extrair certos padrões de dados brutos. Contudo, mais importante do que simplesmente obter essa redução de informação, esse processo nos permite gerar formas de predizer futuras ocorrências de padrões.

Big Data

- **Big data:** são dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior.
- Os três Vs de big data são: **Volume, Velocidade e Variedade.**
- **Processamento de dados:** Processo de colecta, Armazenamento, Montagem e Analise.
- O Big Data fornece novas informações que abrem novas oportunidades e modelos de negócios. (Integrar, Gerenciar e Analisar).
- **Características de analise dos dados com o Big Data:** Volume, Complexidade, Velocidade, Veracidade e Variabilidade, Valor e Adaptabilidade.
- **Casos de uso do Big Data:** Desenvolvimento de produtos, Manutenção Preditiva, Experiencia do cliente, Fraude e Conformidade, Machine Learning, Eficiência Operacional e Impulsiona a inovação.

Data Warehouse(DW)

- **Data warehouse (DW):** é um conjunto de técnicas que aplicadas em conjunto geram um sistema de dados que proporcionam informações para tomada de decisões. Ela funciona tipicamente na arquitectura cliente/servidor.
- **Características:**
- A orientação por assunto;
- A integração dos dados;
- Variante temporal dos dados;
- A não volatilidade dos dados.

Data Warehouse(DW) – Cont...

- **Objectivo:** o objectivo de um data warehouse eh fornecer uma imagem única da realidade do negocio.
- Sistema de data warehouse compreendem um conjunto de programas que extraem dados do ambiente de dados operacionais da empresa, um banco de dados que os mantem, e sistemas que fornecem estes dados aos seus usuários.
- **Implantação de um Data Warehouse:** o data warehouse (DW) apresenta um processo complexo composta por vários itens como:

- Metodologias;
- Técnicas;
- Maquinas;
- Bancos de dados;

Data Warehouse(DW) – Cont...

- Ferramentas de front-end;
- Extração;
- Metadados;
- Refinamento de dados;
- Replicação;
- Recursos humanos.

Referências Bibliográficas

- O'brien, James A. (2005). Sistemas de Informação e as decisões gerenciais na era da internet 2º ed. São Paulo: Saraiva. p. 143. ISBN 9788502044074
- Laudon, Kenneth; Laudon, Jane (2011). Sistemas de Informações Gerenciais: Fundamentos da inteligência de negócios:gestão da informação e de banco de dados. 9º ed. São Paulo: ABDR. p. 159
- O'brien, James A.; Marakas, George M. (2007). Administração de Sistemas de Informação: uma introdução 13º ed. São Paulo: McGraw-Hill. p. 171

FIM

VAMOS A PRÓXIMA SESSÃO