PÓS-GRADUAÇÃO

Integração e fluxo de dados

PÓS-GRADUAÇÃO

Limpeza de dados

Bloco 1

Thiago Salhab Alves

Limpeza de dados

Objetivos

- Compreender as definições e conceitos de limpeza de dados.
- Aprender a remover “ruído”, dados irrelevantes, como também corrigir inconsistências nos dados.
- Aprender as técnicas de remoção de “ruídos” conhecidas como compartimentalização (*binning*), regressão e agrupamento (*clustering*).

Limpeza de dados

Limpeza de dados

- Os dados, segundo Han e Kamber (2006) se originam de diferentes fontes, quase sempre heterogêneas, possuindo:
  - Formatos diferentes.
  - Campos nem sempre preenchidos.
  - Campos duplicados.

Limpeza de dados

- A limpeza de dados, segundo Han e Kamber (2006), é utilizada para remover “ruídos”, dados irrelevantes e corrigir inconsistências nos dados.
- As rotinas de limpeza dos dados tradicionalmente buscam:
  - Preencher valores faltantes.
  - Suavizar os dados, eliminando ruído e detectando *outliers* (dados que se apresentam grande afastamento dos demais dados da série).
  - Corrigir inconsistências presentes nos dados.

Limpeza de dados

Segundo Kimball e Caserta (2009), dados precisos apresentam as seguintes características:

- Corretos: os valores e descrições nos dados descrevem seus objetos de verdade e devem estar corretos.
- Não ambíguos: os valores e descrições nos dados devem ter apenas um significado.
- Consistentes: os valores e descrições nos dados usam uma notação para transmitir seus dados.
- Completos: dados com valores e descrições (não nulos) e número de registros completos.

Limpeza de dados

- Segundo Han e Kamber (2006), algumas técnicas de preenchimento de dados podem ser aplicadas a valores faltantes:
  - Descarte de toda a tupla.
  - Preenchimento manual do valor faltante.
  - Uso de uma constante global para preencher o valor faltante.
  - Usar da média do atributo com o valor a ser substituído.
  - Uso do valor mais provável do atributo como valor a ser preenchido.

Limpeza de dados

- De acordo com Han e Kamber (2006), um ruído (*noise*) é uma variação ou erro aleatório observado em uma variável medida, podendo introduzir erros nos resultados.
- A compartimentalização (*binning*) é uma técnica de remoção de ruído que suaviza dados ordenados a partir dos dados em posições vizinhas.

Limpeza de dados

- Outra técnica para remoção de ruídos, segundo Han e Kamber (2006), a regressão consiste em suavizar os dados, substituindo-os pelo resultado de uma função que os aproxime.
- Essa regressão pode ser:
  - Regressão linear: aproxima os dados por uma reta, plano ou hiperplano (conforme a dimensão dos dados).
  - Regressão não-linear: dados são aproximados por outras funções.

PÓS-GRADUAÇÃO

Limpeza de dados

Bloco 2

Thiago Salhab Alves

Limpeza de dados

- O agrupamento (*clustering*), segundo Han e Kamber (2006), é utilizado principalmente para eliminar *outliers.*
- *Outliers* são valores “espúrios” que não seguem o comportamento geral ou o modelo dos dados. Geralmente, são causados por erros na coleta dos dados.

Limpeza de dados

- Na técnica de agrupamento (*clustering*), os dados são automaticamente divididos em grupos (*clusters*); pontos que não pertencem a qualquer dos grupos são eliminados.
- Segundo Han e Kamber (2006), uma das etapas mais importantes na limpeza dos dados é detectar discrepâncias nos dados.
- Deve-se procurar por erros de códigos, formato e armazenamento.
- Verificar situações em que valores devem ser únicos ou explicitamente faltantes (*null values*).

PÓS-GRADUAÇÃO

Teoria em Prática

Bloco 3

Thiago Salhab Alves

Limpeza de dados

Uma empresa nacional de revenda de cosméticos está enfrentando alguns problemas financeiros. Dado o grande volume de produtos lançados pelo setor de cosméticos, a empresa está tendo dificuldades em acompanhar a demanda por produtos de lançamento o que, muitas vezes, acaba por comprometer o resultado financeiro, por investir em produtos com baixa procura. Outro problema são os produtos que possuem prazo de validade curto, que acabam por vencer e não poderem ser trocados pelos fornecedores.

Limpeza de dados

- Hoje, a empresa conta com um sistema de vendas e controle de estoque, com banco de dados relacional e um processo de marketing pelas redes sociais, porém está tendo dificuldades para a tomada de decisões relacionado ao que o seu público-alvo realmente consome e, assim, evitar gastos desnecessários. Após o processo de extração e controle de dados do sistema de vendas e do marketing das redes social, constatou-se que os dados necessitavam passar por um processo de limpeza. Como podemos auxiliar a organização no processo de limpeza dos dados?

Limpeza de dados

- R: Aplicar técnicas para preencher os valores faltantes, eliminar ruídos utilizando técnicas de compartimentalização (*binning*), regressão ou agrupamento (*clustering*).

PÓS-GRADUAÇÃO

Dica do Professor

Bloco 4

Thiago Salhab Alves

Limpeza de dados

Indicação de leitura do capítulo 4 de livro de KIMBALL e CASERTA, disponível na Biblioteca Virtual:

- KIMBALL, L., R.; CASERTA, J. **The Data Warehouse ETL Toolkit**: Practical Techniques for Extracting, Cleaning, Conforming, and Data Delivering Data. Indianopolis: Wiley Publishing, 2009.

Referências

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. Elsevier, 2006.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit**: Practical Techniques for Extracting, Cleaning, Conforming, and Data Delivering Data. Indianopolis: Wiley Publishing, 2009.