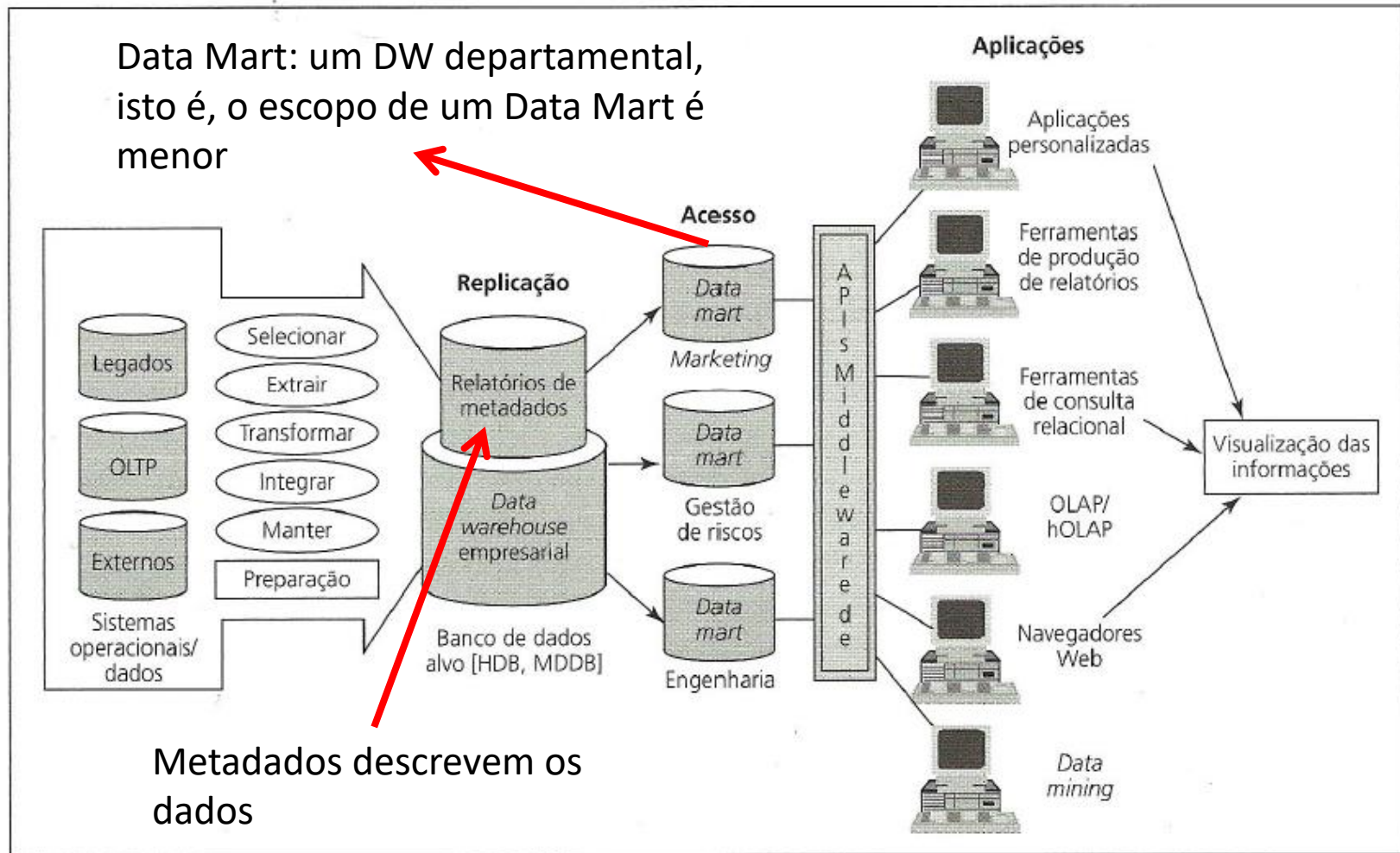


Banco de Dados II

Mineração de Dados – Parte 2

DW - Ambiente



Fonte: TURBAN et al., 2009

Descoberta de Conhecimento em Base de Dados

(Knowledge Discovery Database – KDD)

- Processo **não trivial** de identificação de padrões, a partir de dados válidos, novos, potencialmente úteis e compreensíveis. (FAYYAD, 1996).
- Data Mining é normalmente considerado parte do processo de DCBD.

Processo de DCBD

1. **Definição do problema.** É necessário conhecimento do domínio
2. **Seleção dos dados.** Uma parte dos dados é selecionada.
3. **Limpeza dos dados/Pré-processamento dos dados.**
Inconsistências são corrigidas (o destino por ser um DW).
4. **Transformação dos dados.** Eventualmente é reduzido o número de variáveis ou de registros a serem consideradas no processo de mineração de dados. Exemplo: discretização.
5. **Mineração dos dados/Data Mining.** Envolve escolha de algoritmos de mineração.
6. **Interpretação dos dados.** Os resultados do processo de mineração são interpretados

Resultado: Conhecimento...(ou não...)

Data Mining - Mineração de Dados

- Mineração de dados utiliza técnicas e algoritmos de diferentes áreas do conhecimento:
 - Inteligência artificial (aprendizagem de máquina)
 - Banco de dados (recursos para manipular grandes bases de dados)
 - Estatística (avaliação e validação de resultados)
- Mineração de Dados => Algoritmos

O Valor da Informação

- Informação descoberta deve ser
 - Nova
 - Inesperada
 - Válida (estatisticamente)
- Valor da Informação => Impacto nas decisões

Tarefas de Mineração de Dados

- Indica o tipo de problema que será resolvido
 - Classificação
 - Agrupamento (*Clustering*)
 - Associação
 - Regressão

Associação

- No caso da classe de uma tarefa de mineração não ser determinada como na classificação uma possibilidade é o uso de algoritmos de associação
 - Algoritmo de associação *Apriori* do *Weka*.
 - Itens que ocorrem juntos
 - Exemplo (clássico!!!): Fraldas -> Cerveja
 - Quem compra fraldas compra cerveja...

Exemplificando




- Comércio eletrônico

<http://www.amazon.com>

Amazon.com

Firefox ▾ a Data Mining: Practical Machine Learning... +

www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569/ref=pd_ys_sf_s_283155_a1_4_p



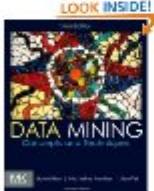
Price for all three: \$121.90

[Add all three to Cart](#) [Add all three to Wish List](#)

[Show availability and shipping details](#)

- ☒ **This item:** Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management) by Ian H. Witten Paperback **\$42.92**
- ☒ Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management) by Jiawei Han Hardcover \$52.62
- ☒ Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites by Matthew A. Russell Paperback \$26.36


Customers Who Bought This Item Also Bought




Data Mining: Concepts and Techniques, Third ...
➤ Jiawei Han
★★★★☆ (14)
Hardcover
\$52.62



Data Mining with R: Learning with Case ...
Luis Torgo
★★★★☆ (7)
Hardcover
\$66.34



Mining the Social Web: Analyzing Data from ...
➤ Matthew A. Russell
★★★★☆ (21)
Paperback
\$26.36



The Elements of Statistical Learning: Data Mining, ...
➤ Trevor Hastie
★★★★☆ (48)
Hardcover
\$72.59

Definição do problema

- Dado um conjunto de transações, encontre regras para prever a ocorrência de um item a partir da ocorrência de outros itens presentes na transação

<i>TID</i>	<i>Items</i>
1	Pão, Leite
2	Pão, Fralda, Cerveja, Ovos
3	Leite, Fralda, Cerveja, Coca
4	Pão, Leite, Fralda, Cerveja
5	Pão, Leite, Fralda, Coca

Terminologia

- *Itemset*
 - Um conjunto de um ou mais itens
Exemplo: {leite}, {leite, pão, fralda}, etc
- *k-itemset*
 - Um *itemset* com k itens
- Contador de suporte
 - Número de transações que contém um *itemset*
- Regra de Associação
 - Uma expressão de implicação no formato $X \rightarrow Y$

Como avaliar uma regra de associação

$$\text{Suporte } (X \rightarrow Y) = \frac{\text{registros_com_X_e_Y}}{\text{total de registros}}$$

- X= fralda, Y=cerveja
- Suporte (X -> Y) = 0,5 significa que em 50% dos registros está registrada a compra conjunta de fralda e cerveja

$$\text{Confiança}(X \rightarrow Y) = \frac{\text{registros_com_X_e_Y}}{\text{registros_com_X}}$$

- X= fralda, Y=cerveja
- Confiança (X->Y) = 0,8 significa que 80% dos que compram fralda compraram também cerveja

Exemplificando

<i>TID</i>	<i>Items</i>
1	Pão, Leite
2	Pão, Fralda, Cerveja, Ovos
3	Leite, Fralda, Cerveja, Coca
4	Pão, Leite, Fralda, Cerveja
5	Pão, Leite, Fralda, Coca

Exemplos de regras:

$\{\text{cerveja}\} \rightarrow \{\text{fralda}\}$

$\{\text{leite, fralda}\} \rightarrow \{\text{cerveja}\}$

$\{\text{leite, cerveja}\} \rightarrow \{\text{fralda}\}$

$\{\text{fralda, cerveja}\} \rightarrow \{\text{leite}\}$

$\{\text{cerveja}\} \rightarrow \{\text{leite, fralda}\}$

$\{\text{fralda}\} \rightarrow \{\text{leite, cerveja}\}$

$\{\text{leite}\} \rightarrow \{\text{fralda, cerveja}\}$

Exemplificando

<i>TID</i>	<i>Items</i>
1	Pão, Leite
2	Pão, Fralda, Cerveja, Ovos
3	Leite, Fralda, Cerveja, Coca
4	Pão, Leite, Fralda, Cerveja
5	Pão, Leite, Fralda, Coca

Exemplos de regras:

$\{\text{cerveja}\} \rightarrow \{\text{fralda}\} \quad (s=0.6, c=1)$

$\{\text{leite, fralda}\} \rightarrow \{\text{cerveja}\} \quad (s=0.4, c=0.67)$

$\{\text{leite, cerveja}\} \rightarrow \{\text{fralda}\} \quad (s=0.4, c=1.0)$

$\{\text{fralda, cerveja}\} \rightarrow \{\text{leite}\} \quad (s=0.4, c=0.67)$

$\{\text{cerveja}\} \rightarrow \{\text{leite, fralda}\} \quad (s=0.4, c=0.67)$

$\{\text{fralda}\} \rightarrow \{\text{leite, cerveja}\} \quad (s=0.4, c=0.5)$

$\{\text{leite}\} \rightarrow \{\text{fralda, cerveja}\} \quad (s=0.4, c=0.5)$

Rodando...