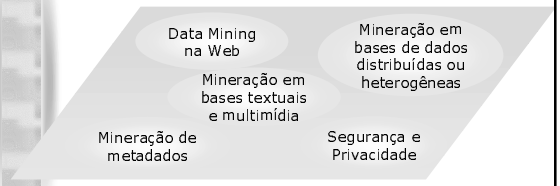
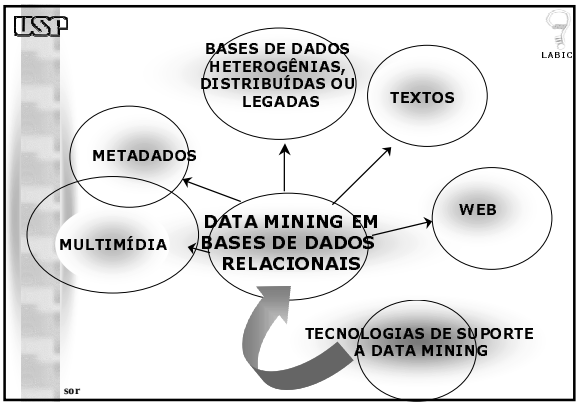
Data Mining (Tendências)

Solange Oliveira Rezende

USP São Carlos / ICMC
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
http://www.icmc.usp.br

Tendências





DM em Textos Text Mining

- **Motivação**
  - ✧Sociedade da informação
  - ✧Conhecimento e capacidade inovadora
  - ✧posição competitiva

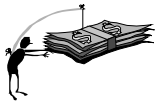
Motivação (cont.)

- **A[illegible]o a in[illegible]ção**
  - ✧Wa[illegible]
  - D[illegible]
  - Docum[illegible]ídos de qualq[illegible]ion[illegible]

Excesso de informação não significa conhecimento

Motivação (cont.)

- **Motivação nos meios acadêmicos com Data Mining que hoje já auxilia a tomada de decisão**

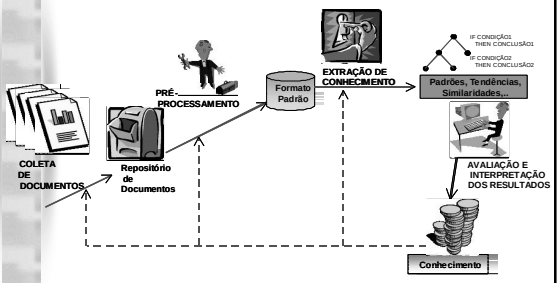


- **Textos - 80 % do conteúdo presente em todas as bases de dados do mundo**

Text Mining

- É um processo, cuja finalidade é a extração de conhecimento de coleções de documentos textuais
- Pode ser considerado uma adaptação do processo DM para dados não estruturados

Etapas de Text Mining



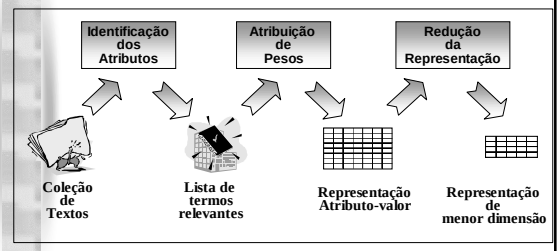
Coleta de Documentos

- Etapa bastante trabalhosa
- Documentos possuem formatos diferentes
- Utilização de técnicas especiais
- Documentos convertidos para um formato único

Pré-Processamento

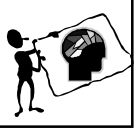
- Transformação dos documentos ➘ formato atributo-valor
  - ✧criação de dicionários
  - ✧atribuição de valores ou pesos às *features*
  - ✧seleção de *features* relevantes

Pré-Processamento (cont.)



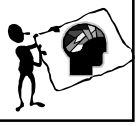
Extração de Conhecimento

- Encontrar padrões, modelos ou classificações presentes na base de dados
- Modelos geralmente utilizados:
  - ✧Estatísticos
  - ✧Neurais
  - ✧Simbólicos



Extração de Conhecimento
(cont.)

- **Tipo de conhecimento almejado:**
  - ✧categorização (classificação)
  - ✧clustering
  - ✧sumarização
  - ✧associação
- **Algoritmo de aprendizado ↘ relacioná-lo à natureza do problema**



Algumas Áreas de Apoio

Automação da aquisição de conhecimento

Melhor entendimento dos mecanismos de aprendizado humano

Aplicação e análise lingüística das features

DM em dados MULTIMÍDIA