

USP

LABIC

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>

© Solange Oliveira Rezende

USP

LABIC

Tendências

Data Mining na Web

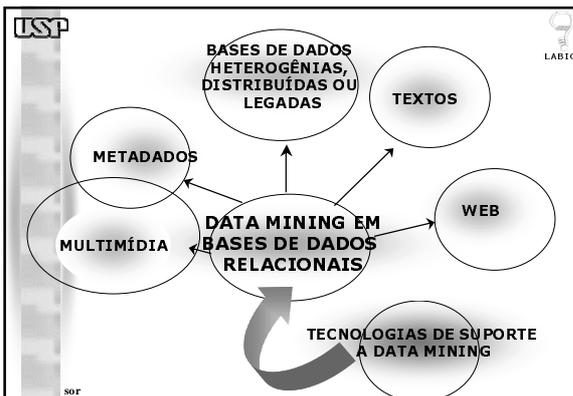
Mineração em bases textuais e multimídia

Mineração de metadados

Mineração em bases de dados distribuídas ou heterogêneas

Segurança e Privacidade

SOF



USP

LABIC

DM em Textos Text Mining

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

SOF

USP

LABIC

Motivação (cont.)

**Excesso de
informação não significa
conhecimento**

SOF

USP

LABIC

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**

SOF

USP LABIO

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

SOF

USP LABIO

Etapas de Text Mining

The diagram illustrates the workflow of Text Mining. It starts with 'COLETA DE DOCUMENTOS' (Document Collection), which leads to a 'Repositório de Documentos' (Document Repository). From there, the process moves to 'PRÉ-PROCESSAMENTO' (Pre-processing), resulting in 'Formato Padrão' (Standard Format). This is followed by 'EXTRAÇÃO DE CONHECIMENTO' (Knowledge Extraction), which produces 'Padrões, Tendências, Similaridades...' (Patterns, Trends, Similarities...). The final stage is 'AVALIAÇÃO E INTERPRETAÇÃO DOS RESULTADOS' (Evaluation and Interpretation of Results), which leads to 'Conhecimento' (Knowledge). A feedback loop is shown from 'Conhecimento' back to 'EXTRAÇÃO DE CONHECIMENTO'.

SOF

USP LABIO

Coleta de Documentos

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

SOF

USP LABIO

Pré-Processamento

- ♦ Transformação dos documentos ↘ formato atributo-valor
 - ✦ criação de dicionários
 - ✦ atribuição de valores ou pesos às *features*
 - ✦ seleção de *features* relevantes

SOF

USP LABIO

Pré-Processamento (cont.)

The flowchart details the pre-processing steps:

- Identificação dos Atributos**: Starting from 'Coleção de Textos' (Text Collection), leading to 'Lista de termos relevantes' (List of relevant terms).
- Atribuição de Pesos**: Leading to 'Representação Atributo-valor' (Attribute-value representation), shown as a grid.
- Redução da Representação**: Leading to 'Representação de menor dimensão' (Lower dimension representation), shown as a smaller grid.

SOF

USP LABIO

Extração de Conhecimento

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

SOF

USP **Extração de Conhecimento** (cont.) **LABIO**

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



SOF

USP **Algumas Áreas de Apoio** **LABIO**

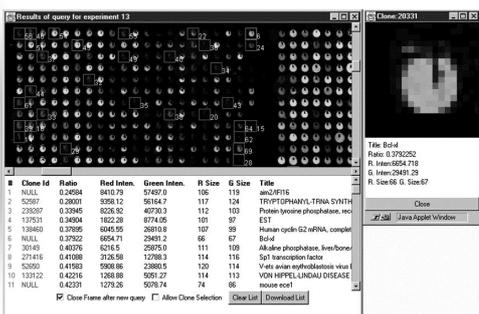
Automação da aquisição de conhecimento

Melhor entendimento dos mecanismos de aprendizado humano

Aplicação e análise lingüística das features

SOF

USP **DM em dados MULTIMÍDIA** **LABIO**



Clone Id	Ratio	Red Inten.	Green Inten.	R Size	G Size	Title
1	NULL	0.24584	8410.79	67457.0	106	119
2	52507	0.28001	8556.12	56164.7	117	124
3	235307	0.33645	8235.32	46795.3	112	103
4	137531	0.34504	1822.28	8774.05	101	97
5	136460	0.37055	4246.05	20816.8	107	98
6	NULL	0.37522	8554.71	29451.2	86	87
7	30149	0.40376	4216.5	25975.0	111	109
8	271416	0.41088	3128.58	12708.3	114	116
9	53593	0.41953	5908.86	23885.5	120	114
10	133132	0.42216	1368.89	5951.27	114	113
11	NULL	0.42331	1279.26	5078.74	74	86

SOF