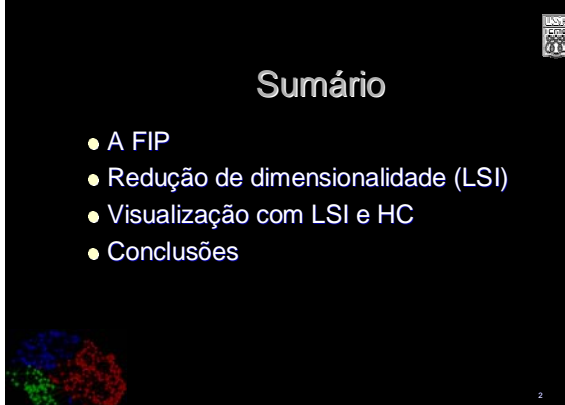




Redução de Dimensionalidade na Visualização de textos

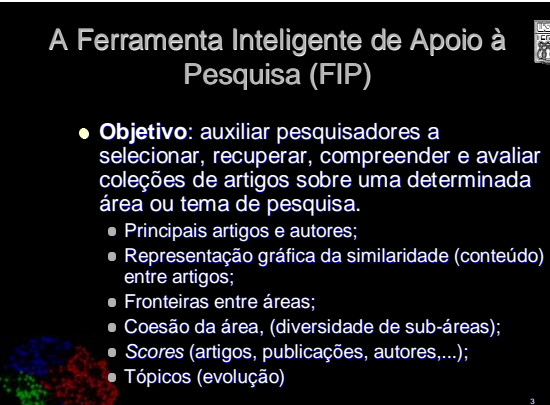
Alneu de Andrade Lopes

Workshop de Mapeamento Visual de Coleções de Documentos – Set 2005
ICMC – Universidade de São Paulo



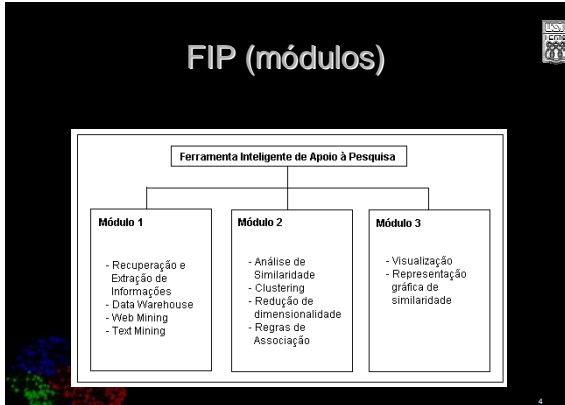
Sumário

- A FIP
- Redução de dimensionalidade (LSI)
- Visualização com LSI e HC
- Conclusões



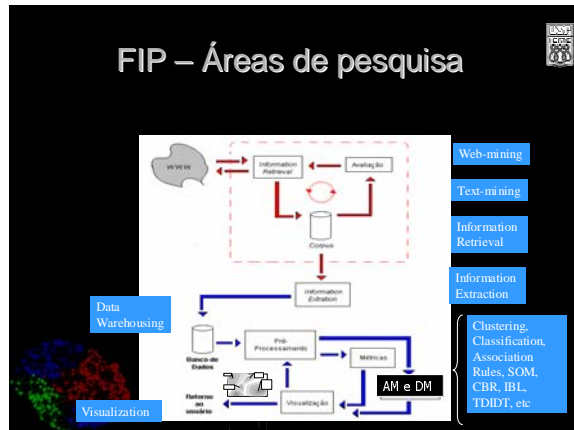
A Ferramenta Inteligente de Apoio à Pesquisa (FIP)

- **Objetivo:** auxiliar pesquisadores a selecionar, recuperar, compreender e avaliar coleções de artigos sobre uma determinada área ou tema de pesquisa.
 - Principais artigos e autores;
 - Representação gráfica da similaridade (conteúdo) entre artigos;
 - Fronteiras entre áreas;
 - Coesão da área, (diversidade de sub-áreas);
 - Scores (artigos, publicações, autores,...);
 - Tópicos (evolução)



FIP (módulos)

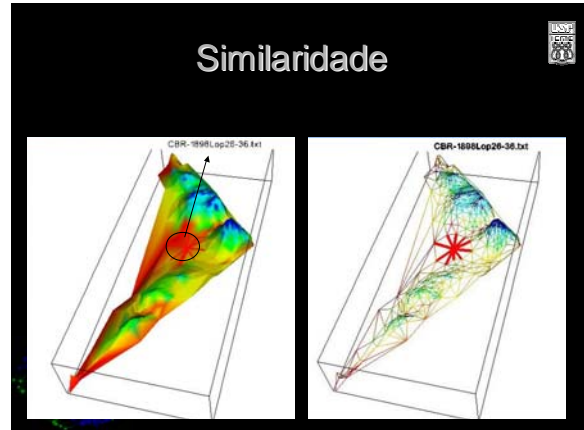
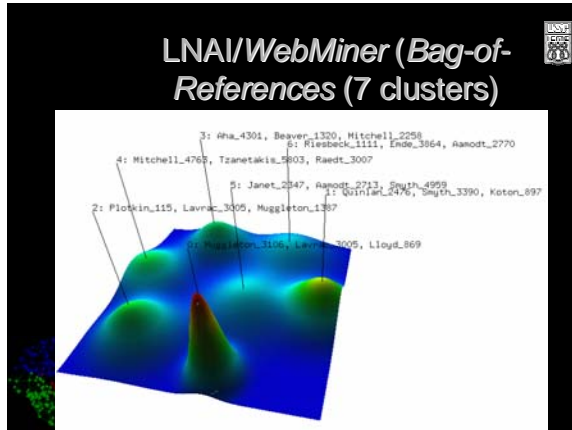
Ferramenta Inteligente de Apoio à Pesquisa		
Módulo 1	Módulo 2	Módulo 3
<ul style="list-style-type: none"> - Recuperação e Extração de Informações - Data Warehouse - Web Mining - Text Mining 	<ul style="list-style-type: none"> - Análise de Similaridade - Clustering - Redução de dimensionalidade - Regras de Associação 	<ul style="list-style-type: none"> - Visualização - Representação gráfica de similaridade



- ## Redução da Dimensionalidade
- Pré-processamento
 - Stopword, stemming, corte de Luhn
 - Representação vetorial (tf, tfidf)
 - LSI
 - Redução de termos à semântica latente

- ## LSI
- Redução de dimensionalidade (explorando co-ocorrência de palavras, projetando-as em uma mesma dimensão).
 - LSI: Aplicação de SVD na matriz de documento-termos em IR.
 - SVD: projeta um vetor n-dimensional (A) em um espaço k-dimensional (\hat{A}), onde $n \gg K$.
 - Minimizando $\Delta = \|A - \hat{A}\|_2$
- Analogia: uma reta é um objeto de uma dimensão, mas pode-se ajustar um conjunto de pontos em um espaço de duas dimensões a uma reta.

- ## SVD
- Computação da projeção
 - Decomposição da Matriz documento-termo $A_{t \times n} = T_{t \times n} S_{n \times n} \text{Transp}(D_{d \times n})$
 - $n = \min(t, d)$
 - $D_{ij} = \text{Transp}(D)_{ji}$
 - T = Matriz de termos por dimensões
 - S = Matriz de valores singulares da decomposição
 - $\hat{A} = S \text{Transp}(D), \quad a = \hat{a}T$
 - Vetor de teste no espaço reduzido $\hat{a} = \text{transp}(T).a$



- ### Conclusões
- Vantagens
 - Recuperação de informação
 - Classificação
 - visualização
 - Desvantagens
 - Custo. Aplicável a corpus limitado a centenas de artigos
 - Dimensão apropriada?