

# Mapeamento Rápido de Textos Utilizando Projeções e Posicionamento de Pontos Baseado em Força<sup>1</sup>

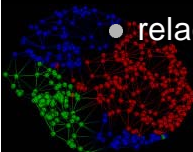
Fernando Vieira Paulovich

<sup>1</sup> baseado em: Minghim, R., Paulovich, F. V., Lopes, A. A. Fast Content-Based Visual Mapping for Interactive Exploration of Document Collection. Relatório técnico Nº 258. Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, 2005.

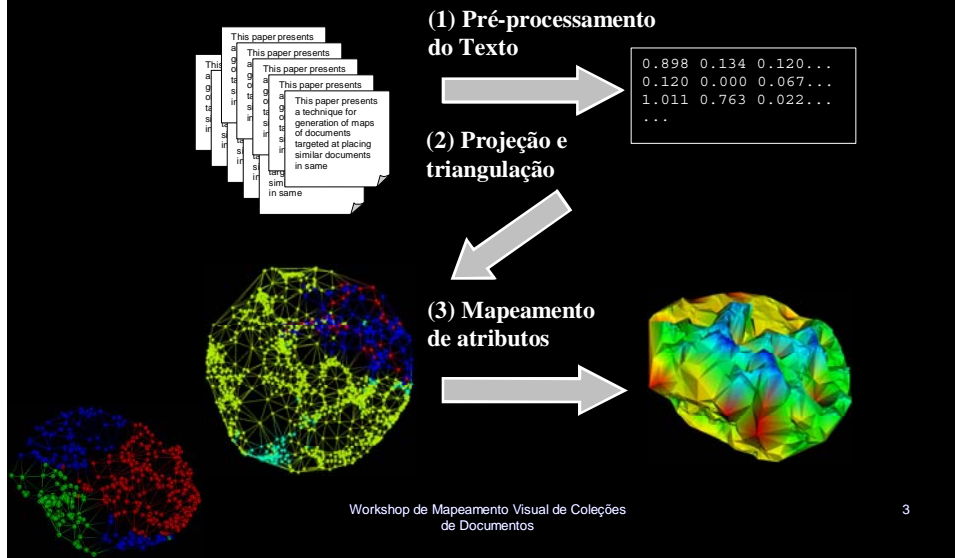
1

## Introdução

- Gerar mapas de documentos utilizando projeções multidimensionais;
- Mapas gerados a partir do conteúdo dos documentos;
- Através desses mapas é possível:
  - identificar estruturas dentro de coleções de documentos;
  - identificar relacionamentos entre tais documentos;
  - relacionamento objetivo: similaridade por vizinhança.

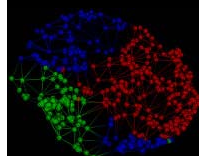


# Visão Geral do Processo

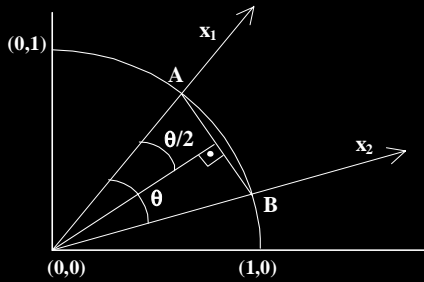


# Técnicas de Projeção

- $X = \{x_1, x_2, x_3, \dots, x_n \mid x_i \in \mathbb{R}^n\}$
- $P = \{x'_1, x'_2, x'_3, \dots, x'_n \mid x'_i \in \mathbb{R}^2\}$
- $d: \mathbb{R}^n \rightarrow \mathbb{R}$
- $d_2: \mathbb{R}^2 \rightarrow \mathbb{R}$
- $\alpha: X \rightarrow P, |d(x_i, x_j) - d_2(\alpha(x_i), \alpha(x_j))| \approx 0$



## Métrica de Distância

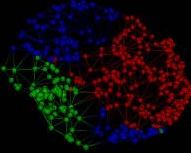


$$d(\vec{x}_1, \vec{x}_2) = 2 * \text{sen}(\theta/2)$$

$$d(\vec{x}_1, \vec{x}_2) = \sqrt{2 * (1 - \cos(\theta))}$$

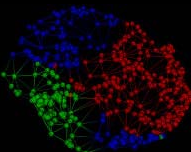
$$\cos(\vec{x}_1, \vec{x}_2) = \frac{\vec{x}_1 \circ \vec{x}_2}{\|\vec{x}_1\|_2 * \|\vec{x}_2\|_2}$$

$$d(\vec{x}_1, \vec{x}_2) = \sqrt{2 * (1 - (\vec{x}_1 \circ \vec{x}_2))}$$



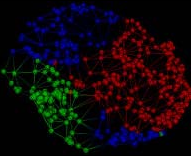
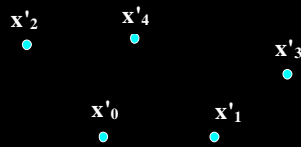
## Técnicas de Projeção

- Duas técnicas de projeção foram usadas:
  - Nearest Neighbor Projection (NNP);
  - Fastmap.



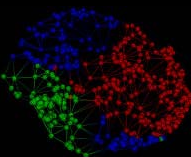
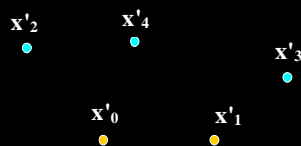
# NNP

- $X = \{x_0, x_1, x_2, x_3, x_4, x_5, \dots, x_m\}$
- $P = \{x'_0, x'_1, x'_2, x'_3, x'_4\}$



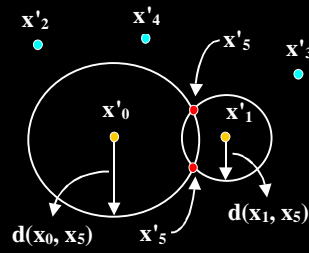
# NNP

- $X = \{x_0, x_1, x_2, x_3, x_4, x_5, \dots, x_m\}$
- $P = \{x'_0, x'_1, x'_2, x'_3, x'_4\}$



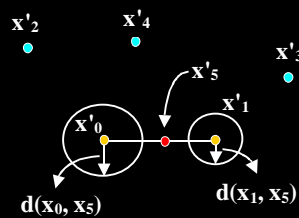
# NNP

- $X = \{x_0, x_1, x_2, x_3, x_4, x_5, \dots, x_m\}$
- $P = \{x'_0, x'_1, x'_2, x'_3, x'_4\}$

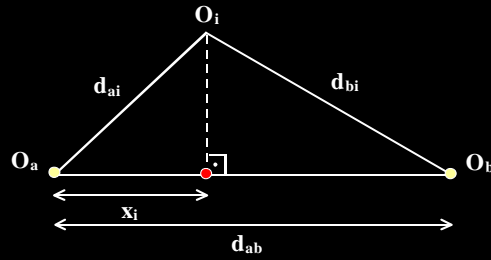


# NNP

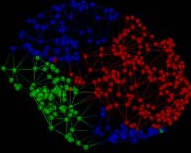
- $X = \{x_0, x_1, x_2, x_3, x_4, x_5, \dots, x_m\}$
- $P = \{x'_0, x'_1, x'_2, x'_3, x'_4\}$



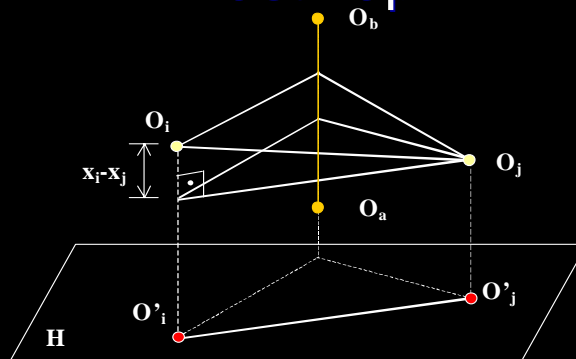
# Fastmap



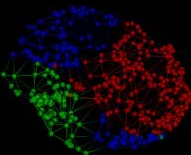
$$x_i = \frac{d_{ai}^2 + d_{ab}^2 - d_{bi}^2}{2d_{ab}}$$



# Fastmap



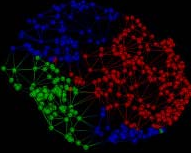
$$d'(O'_i, O'_j) = \sqrt{(d(O_i, O_j))^2 - (x_i - x_j)^2}$$



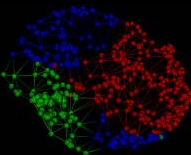
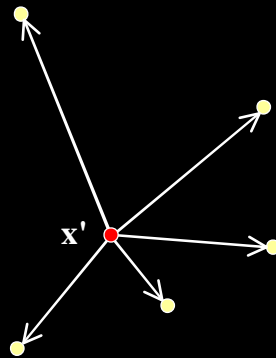
# Melhoria da Projeção

1. Para cada ponto projetado  $x'$ 
  - 1.1. Para cada ponto projetado  $q' \neq x'$ 
    - 1.1.1 Calcular  $\vec{v}$  como sendo o vetor entre  $x'$  e  $q'$
    - 1.1.2 Mover  $q'$  na direção de  $\vec{v}$  uma fração de  $\Delta$
2. Normalizar as coordenadas da projeção entre  $[0,1]$  nas duas dimensões.

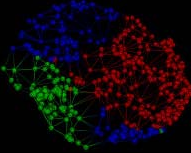
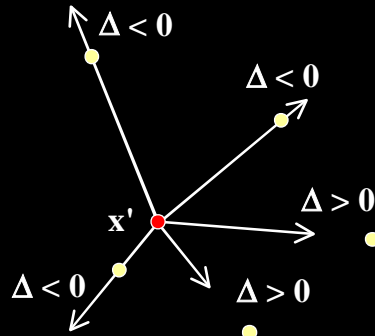
$$\Delta = \frac{d(x, q) - d_{\min} - d_2(x', q')}{d_{\max} - d_{\min}}$$



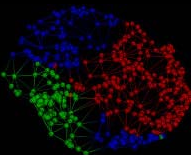
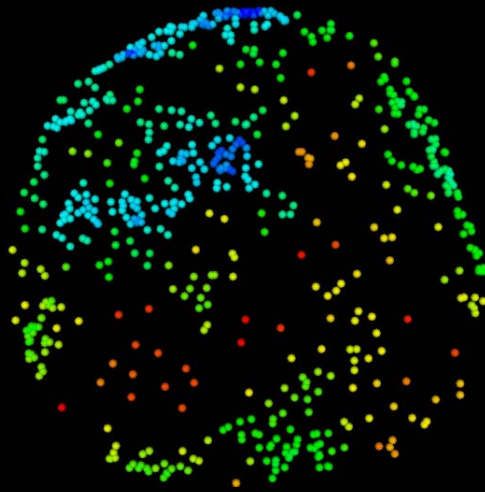
# Force



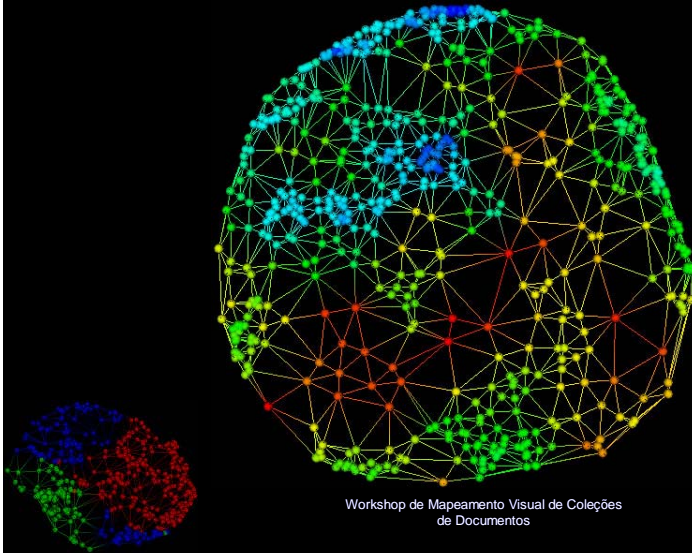
# Force



# Triangulação



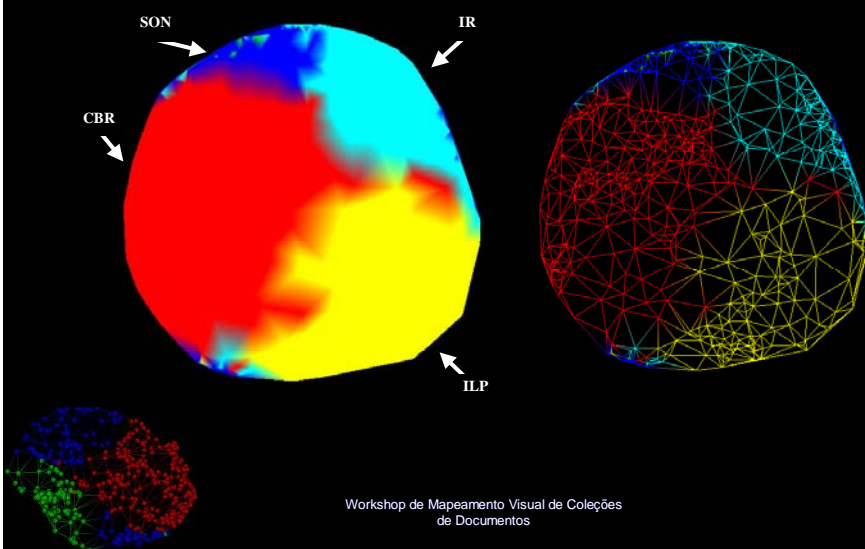
# Triangulação



Workshop de Mapeamento Visual de Coleções de Documentos

17

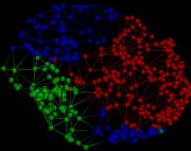
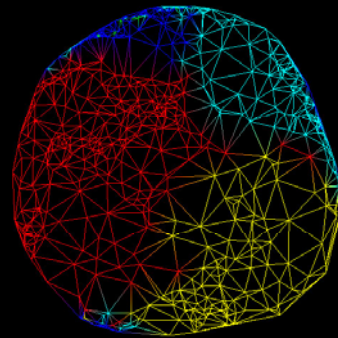
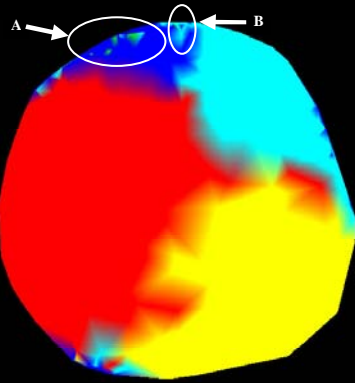
# Resultados: Agrupamentos



Workshop de Mapeamento Visual de Coleções de Documentos

18

# Resultados: Agrupamentos



Workshop de Mapeamento Visual de Coleções de Documentos

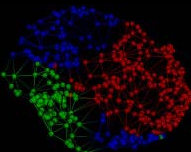
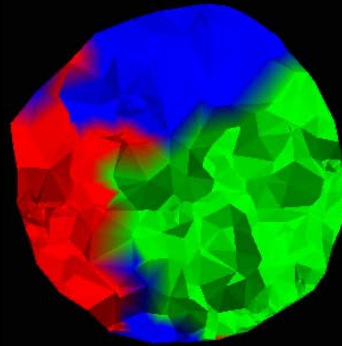
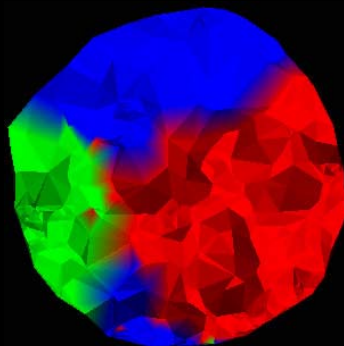
19

# Resultados: Classificação



Pseudo-classe

K-means

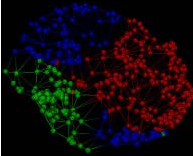


Workshop de Mapeamento Visual de Coleções de Documentos

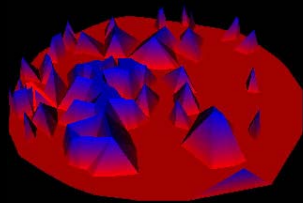
20

# Mapeamento de Atributos

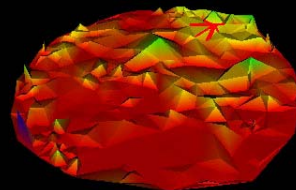
- É possível demonstrar informações adicionais mapeando atributos para a cor ou altura do mapa, tais como:
  - relevância do documento;
  - número de citações;
  - ano de publicação, etc.



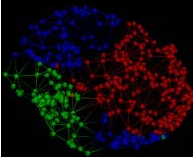
# Mapeamento de Atributos



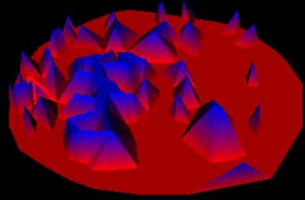
User+interface



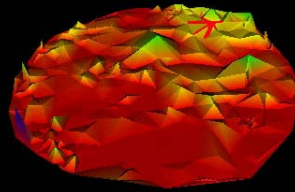
Document+text+visual



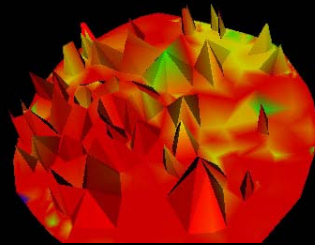
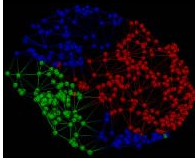
# Mapeamento de Atributos



User+interface



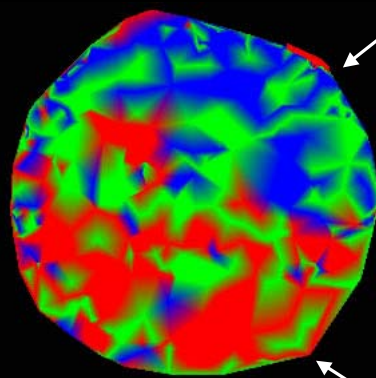
Document+text+visual



es

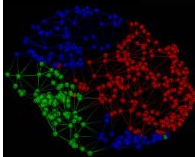
23

# Mapeamento de Atributos



Ferramentas para  
interação e  
tratamento de alta  
dimensionalidade

Técnicas básicas de  
visualização e grafos

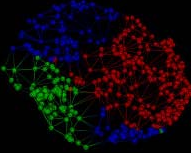
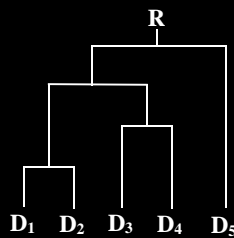


Workshop de Mapeamento Visual de Coleções  
de Documentos

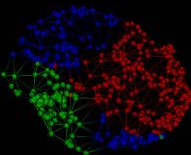
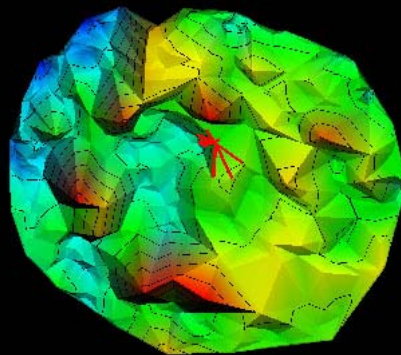
24

# Mapeamento de Atributos

- De forma a mapear para um atributo sub-grupos de documentos, um **Agrupamento Hierárquico (AH)** dos dados projetados pode ser feito:

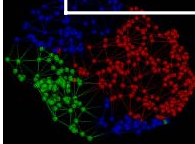


# Mapeamento de Atributos



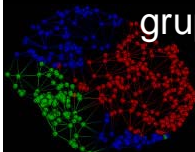
## Tempo de Execução

Corpora	Entradas	Dimensão	Tempo (s)	
			Fastmap	NNP
CBR+ILP+IR	574	5495	11.609	11.484
CBR+ILP+IR+SON	682	6371	18.703	18.781
KDViz	1624	9398	165.844	159.546
InfoVis04	534	4287	8.266	-



## Conclusão e Trabalhos Futuros

- IDMAP (Interactive Document Map);
- Tempo de processamento e separação de (sub)grupos são promissores;
- Permite identificar:
  - documentos similares;
  - (novas) áreas de pesquisa;
  - junção de duas áreas (fronteiras de dois grupos).



# Conclusão e Trabalhos Futuros



- Trabalhos futuros:
  - escalabilidade da técnica;
  - aplicação em diferentes coleções de documentos.
- Problema: não é incremental.

