

Visualização Multi-dimensional:

Mineração Visual de Dados multidimensionais e aplicações Parte II – Variações e Aplicações

Rosane Minghim

2008-2012



1

Basic Concepts

- Text Preprocessing
- Data and text mining
- Projection techniques
- Point Placement Strategies



2

Text Preprocessing

1. Stopwords elimination
2. Extraction of words radicals (stemming)
3. Creation of n-grams
4. Frequency count and Luhn's lower cut (n-grams appearing less than x times are ignored)
5. Weighting process (*term-frequency inverse document-frequency - (tfidf)*)



3

Result is a Vector Model

- Attributes: terms (n-grams)
- Value: term weight
- Table Data



4

Vector Representation – term weighting

- tf – term frequency
- tfidf – tf x idf = tf x inverse document frequency

$$w_{ik} = tf_{ik} \times \log \left(\frac{N}{n_k} \right)$$



5

Vector Representation

	term ₁	term ₂	term ₃	term ₄	...	term _m
Doc ₁	0.92	0.62	0.92	0.10	...	0.67
Doc ₂	0.13	0.11	1.00	0.34	...	0.33
Doc ₃	0.52	0.00	0.00	0.44	...	0.77
...
Doc _n	0.02	0.12	0.22	0.92	...	0.00



6

Vector Representation – Similarity calculation

EUCLIDEAN

$$sim_{i,j} = \sqrt{(w_{i,1} - w_{j,1})^2 + \dots + (w_{i,k} - w_{j,k})^2}$$

MANHATAN

$$sim_{i,j} = |w_{i,1} - w_{j,1}| + \dots + |w_{i,k} - w_{j,k}|$$

COSINE

$$sim_{i,j} = \frac{(w_{i,1} \times w_{j,1}) + \dots + (w_{i,k} \times w_{j,k})}{(w_{i,1}^2 + \dots + w_{i,k}^2) \times (w_{j,1}^2 + \dots + w_{j,k}^2)}$$



7

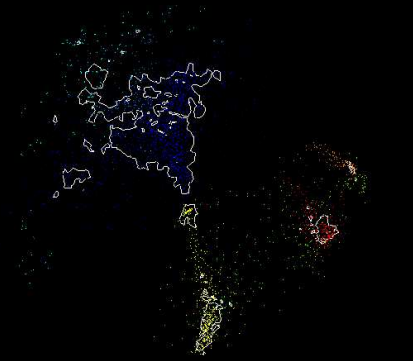
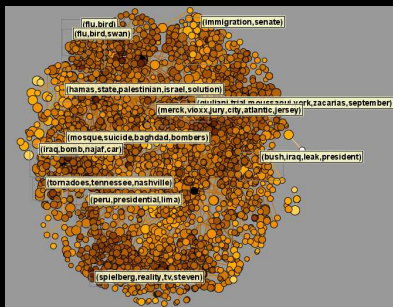
Alternative to Vector Representation

- Similarity Calculation text against text
 - Ex: NCD Normalized Compression Distance
 - Approximation of Kolmogorov Complexity



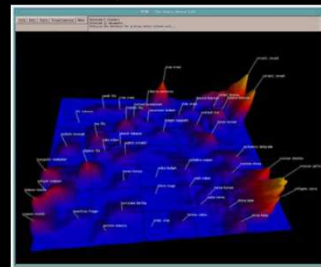
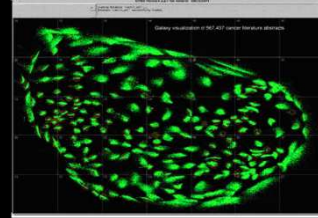
8

Visual representations: graphs, surfaces, volumes, triangulations



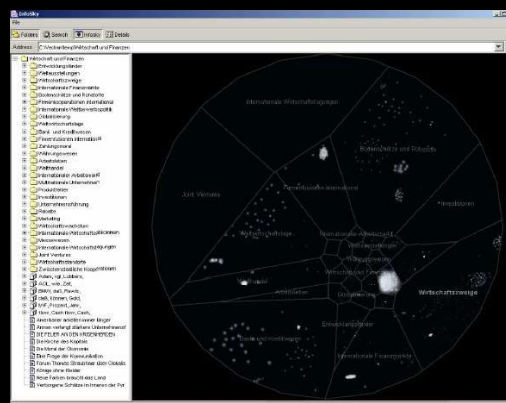
IN-SPIRE

- Spatial Paradigm for Information Retrieval - Pacific Northwest National Laboratories
- Two Visualization Metaphors:
 - Galaxies – dimensional reduction
 - Themescape



InfoSky

Granitzer (Granitzer et al., 2004) also employs galaxy metaphor



The screenshot displays the InfoSky Data Communications interface. On the left is a hierarchical tree view of categories such as Artificial Intelligence, Home Automation, Robotics, and Data Communications. The main area features a network visualization with nodes and edges, highlighting clusters like 'Human Computer Interaction', 'Speech Technology', 'Telephony Digital Wireless', 'Ethernet Education Vendor', 'Education', 'Computers', and 'Composites'. Below the visualization is a table with columns for Name, Size, Modified, and Keywords.

Name	Size	Modified	Keywords
Ethernet	88 documents	Wed Dec 31 21:00:00 ERT 1969	Ethernet, IEEE, networks, links, standards, Site, collection
vendors	241 documents	Wed Dec 31 21:00:00 ERT 1969	vendors, Provides, service, Networks, Cisco, Includes, product
Software	85 documents	Wed Dec 31 21:00:00 ERT 1969	software, applications, communications, services, solutions, Provides, networks
telephony	146 documents	Wed Dec 31 21:00:00 ERT 1969	voice, software, products, category, systems, consumer, solutions
Modems	15 documents	Wed Dec 31 21:00:00 ERT 1969	modems, modem, Modems, Modem, Modems, Telematic, applications
organizations	11 documents	Wed Dec 31 21:00:00 ERT 1969	IEEE, International, development, society, inter-Operability, site, technology
Reference	67 documents	Wed Dec 31 21:00:00 ERT 1969	Information, information, networking, data, communications, Network, Technology
Support	1 documents	Wed Dec 31 21:00:00 ERT 1969	TeleSource, Communications, Manages, custom, voice, data, Internet
Frame Relay	13 documents	Wed Dec 31 21:00:00 ERT 1969	Frame, Relay, service, centers, network, frame, relay
T collections			

http://www.know-center.at/forschung/wissenserschliessung/downloads_demos/infosky_demo¹³

VxInsight

- Sandia National Laboratories, mountain metaphor (Boyack et al., 2002).

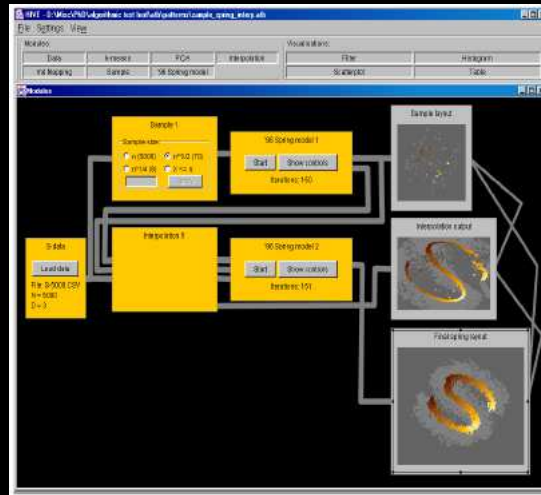
The visualization shows a 3D landscape of research topics represented as mountains. The peaks are labeled with terms and their corresponding document counts:

- SELF-ASSEMBLED INAS (85/62/450)
- SELF-ASSEMBLED INAS (53/49/920)
- FIBER/FIBEROPTIC (62/34/103)
- PAAS SUBSTRATES (54/33/198)
- SEMICONDUCTOR CDSE (68/67/469)
- ONE-DIMENSIONAL EXCITATIONS (28/25/176)
- SILICON POROUS (108/96/295)
- SELF-ASSEMBLED MONOLAYERS (285/221/596)
- CONDUCTANCE/T RANSPORT (27/23/212)

http://www.know-center.at/forschung/wissenserschliessung/downloads_demos/infosky_demo¹³

HIVE (Ross and Chalmers 2003)

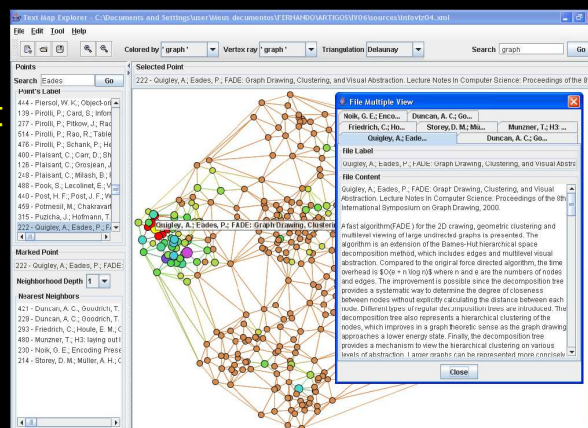
- Interconnected components:
 - Import
 - Transform
 - Render multi-dim data



15

Projection Explorer (PEX)

- Projection and Point placement
- Precision
- Graphs and surfaces (Super Spider)



16

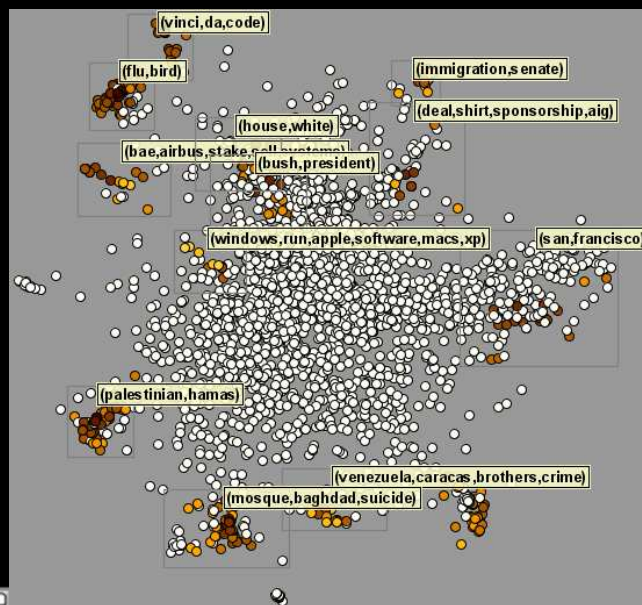


Mapping Text Collections via Projections and Point Placement

- Positioning and labeling

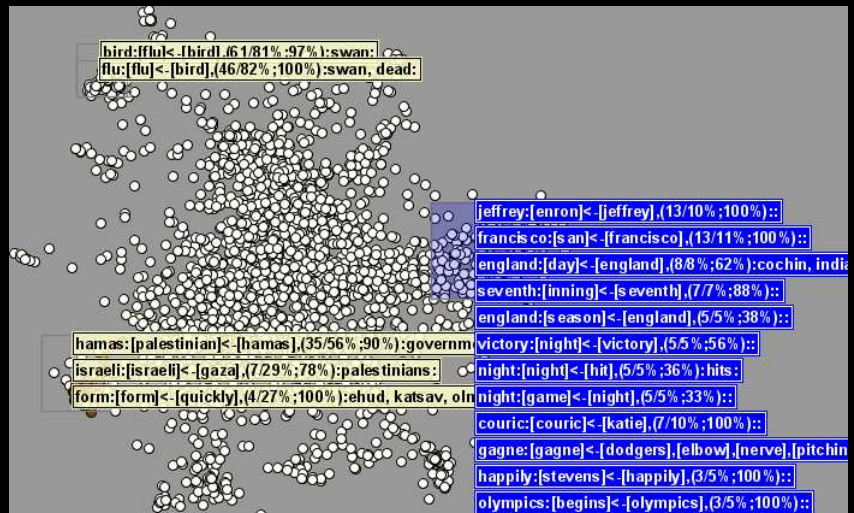


17



18

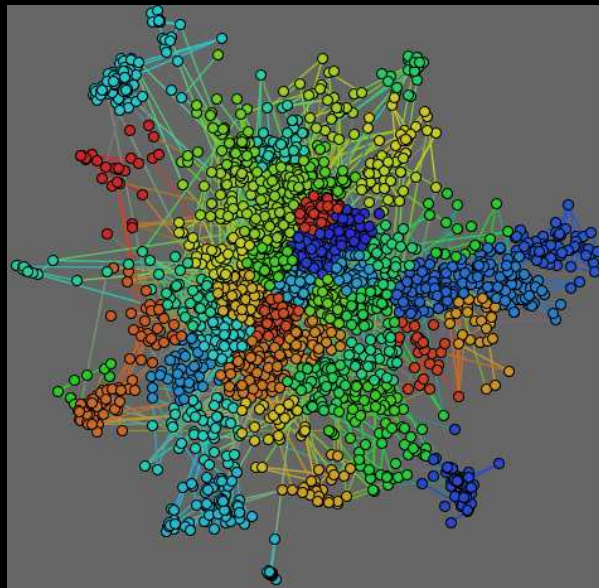
- Detailing topics



- Finding Relationships



21

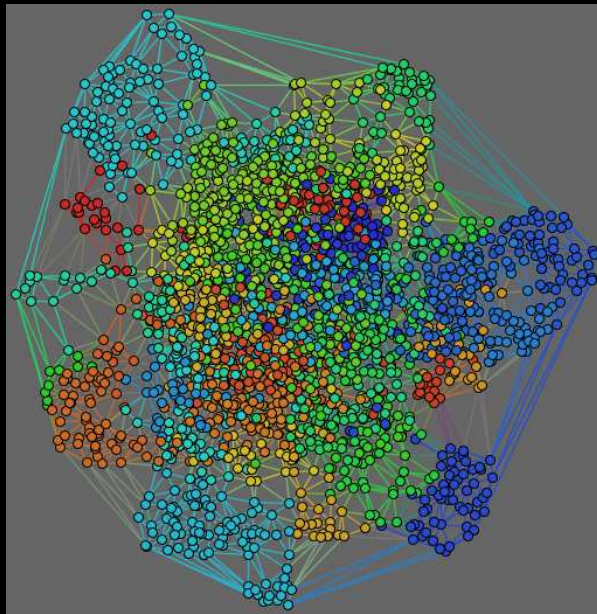


22

- Building a mesh



23

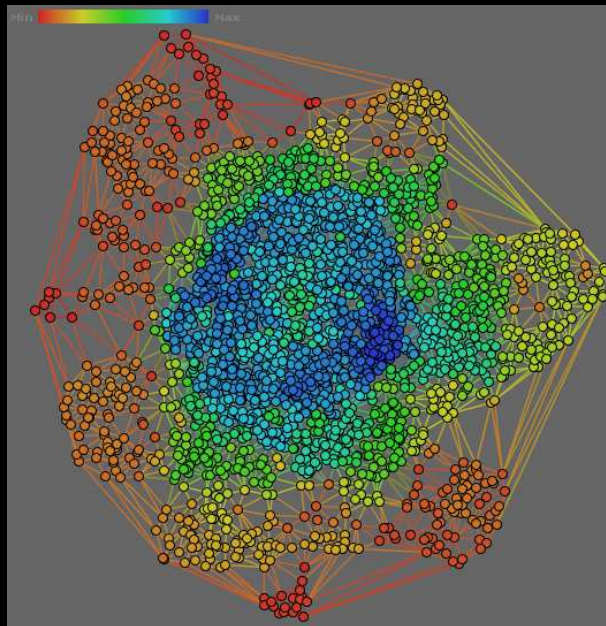


24

- Coloring by degree of proximity



25

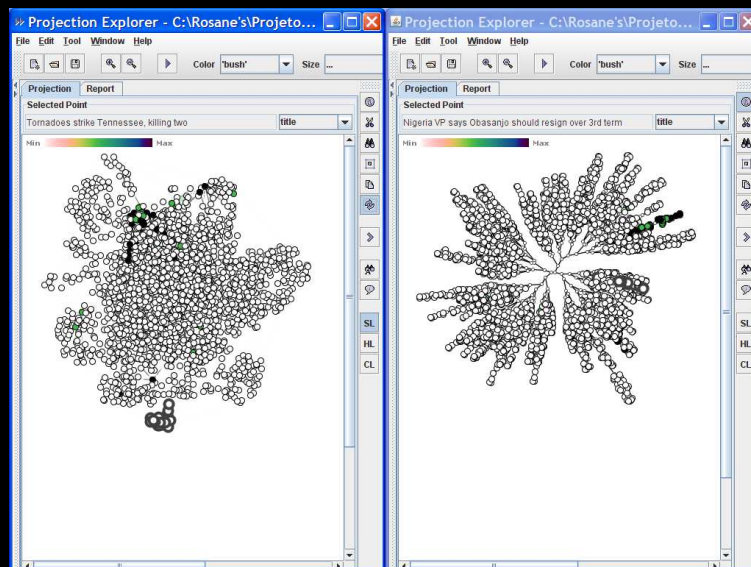


26

- Coordinating



27

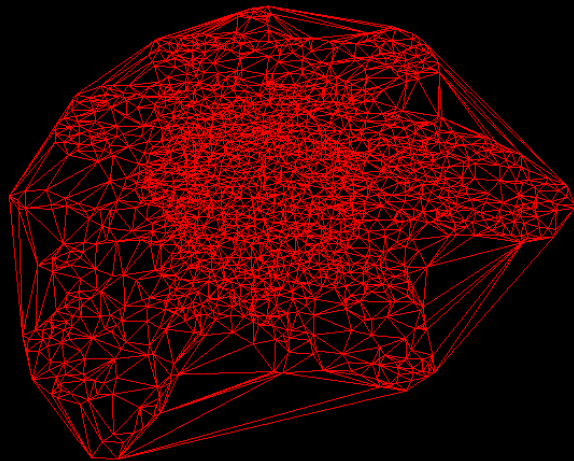
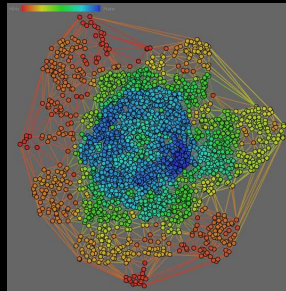


28

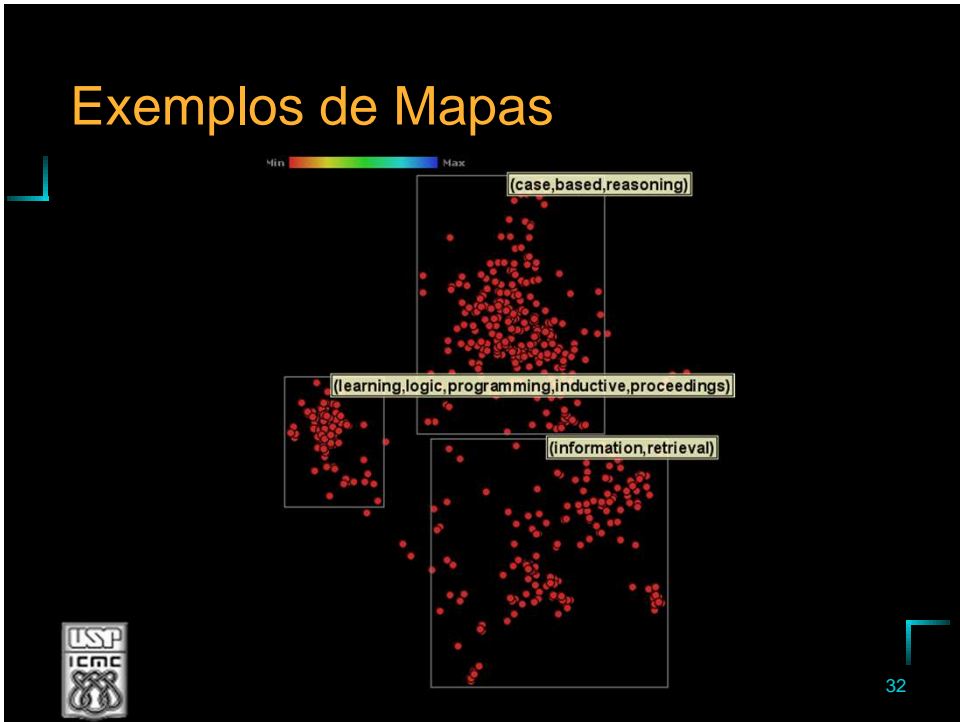
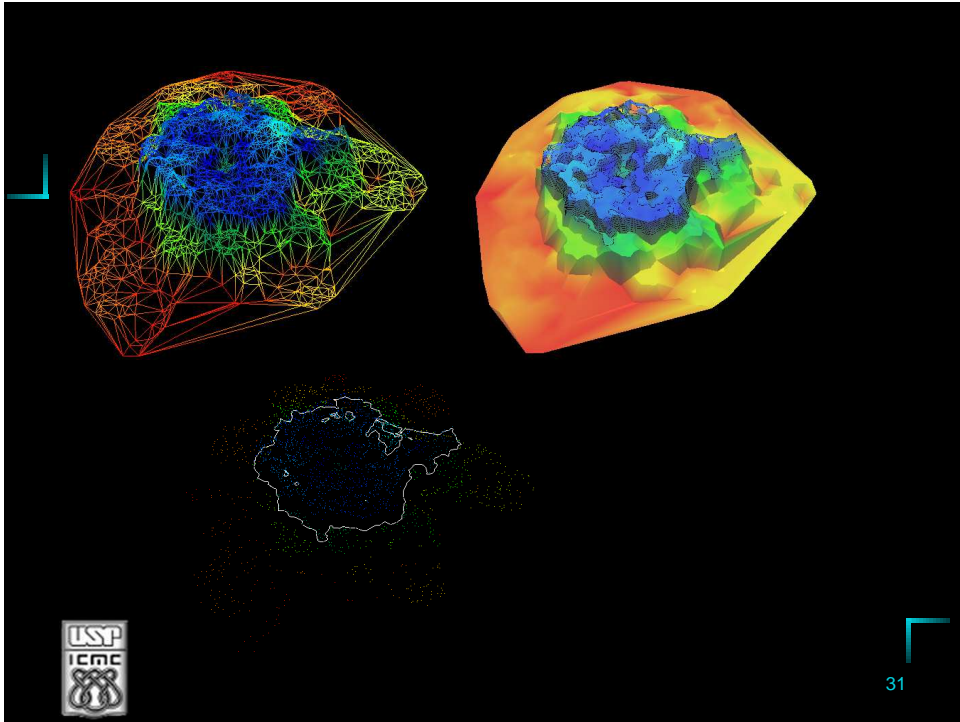
- Building a Surface



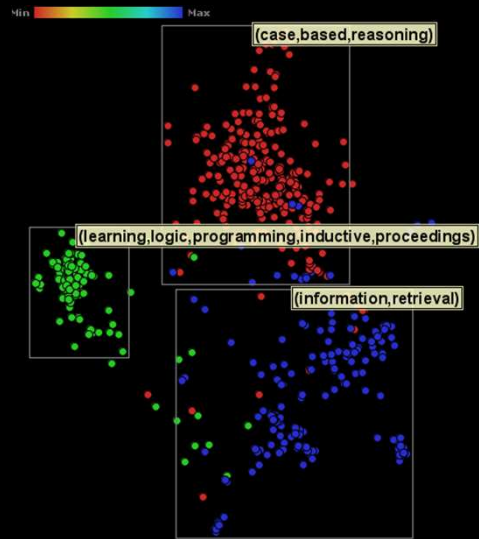
29



30

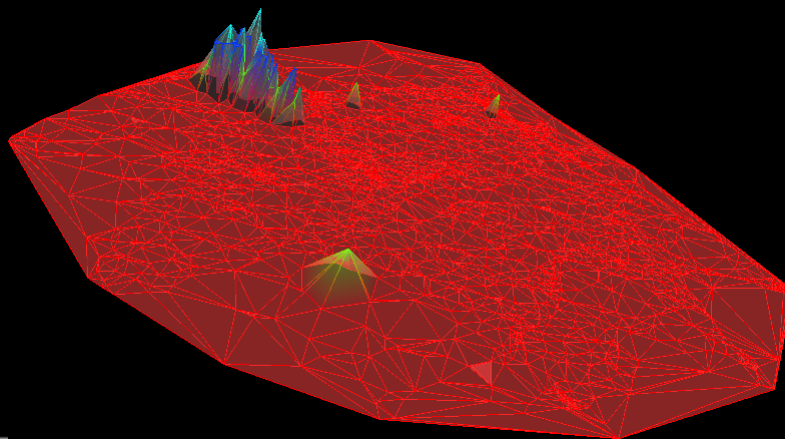


Exemplos de Mapas



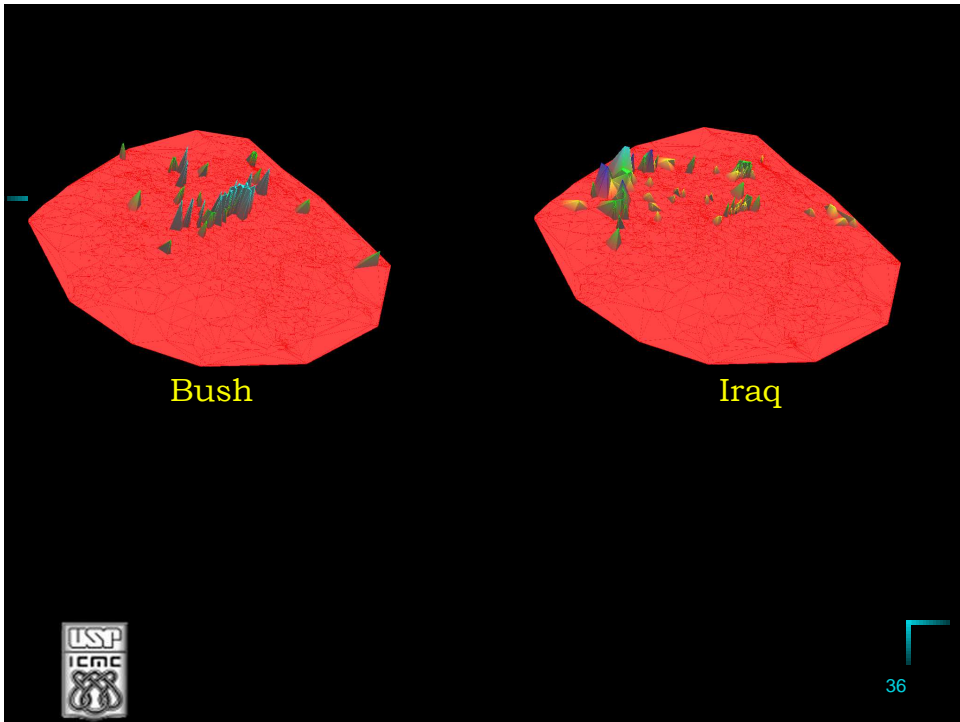
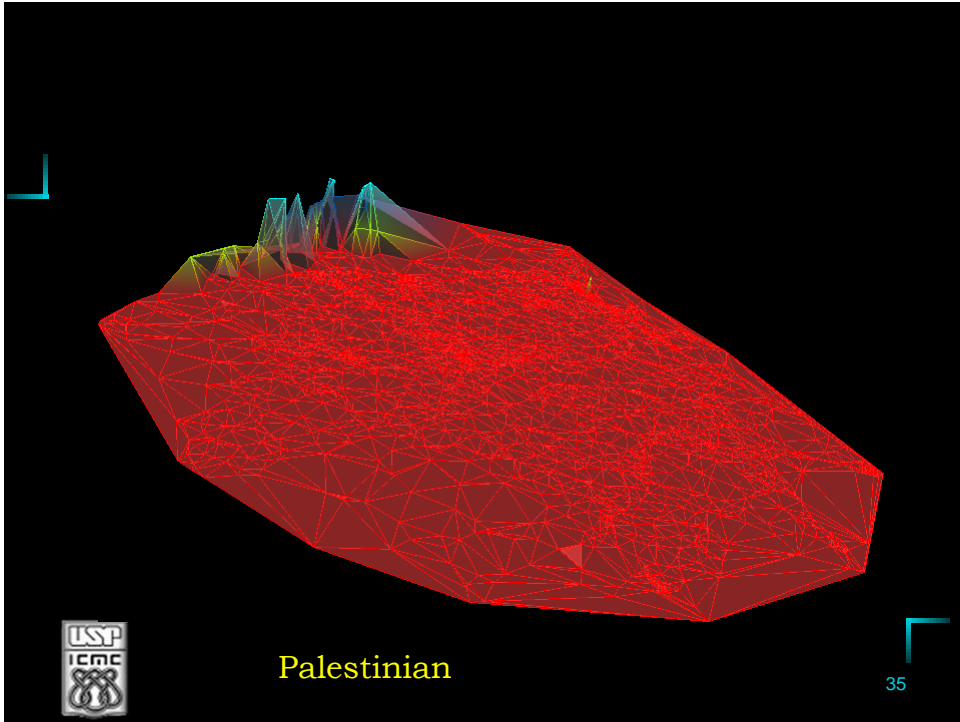
33

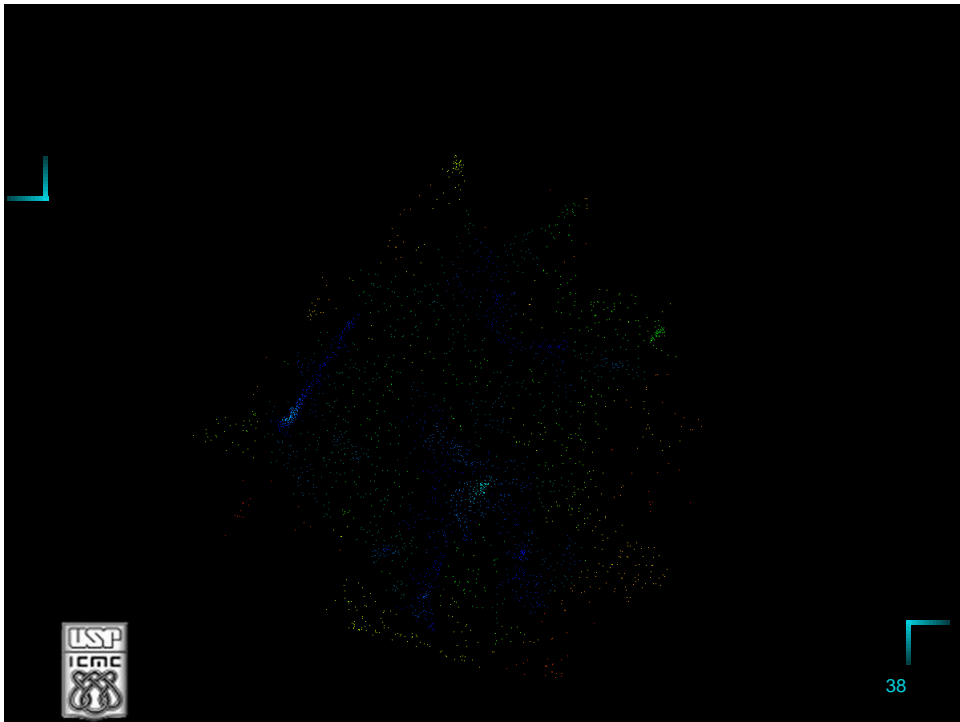
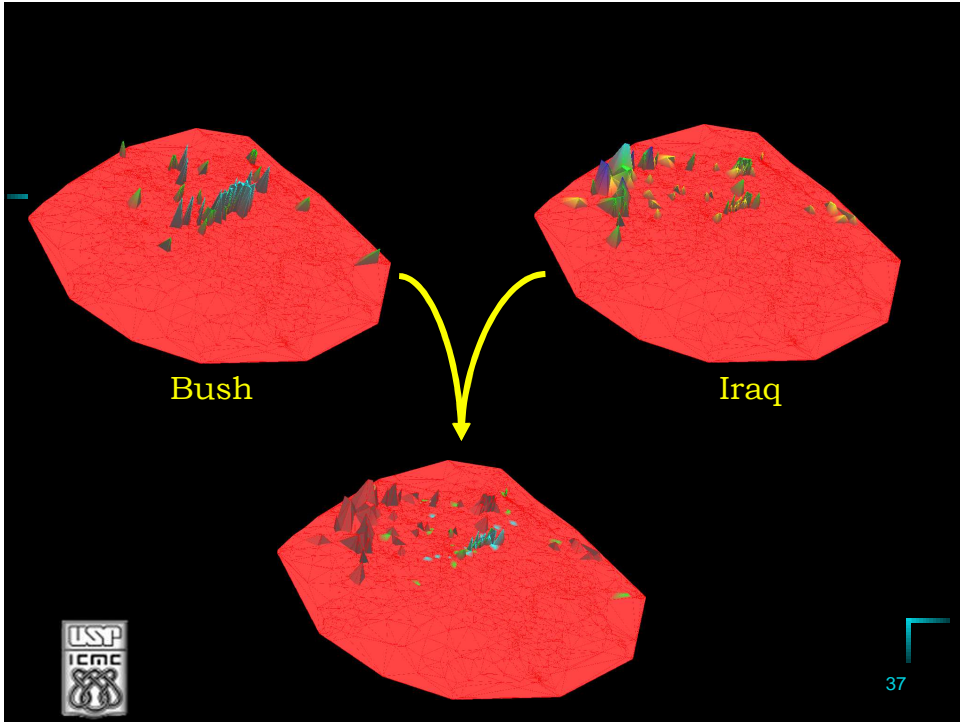
RSS News Flash



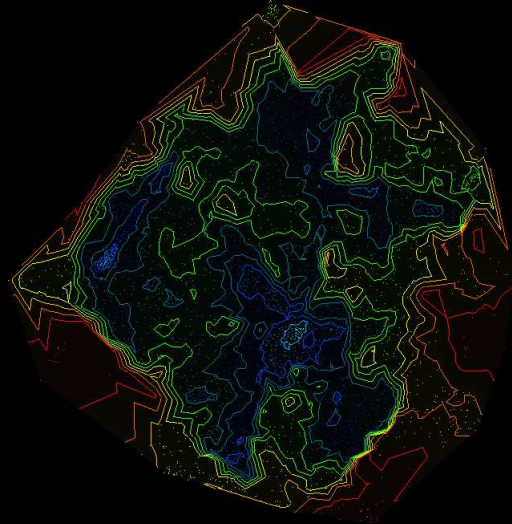
Bird and Flu

34





Curvas de Nível



39

Séries Temporais – Vazão em Hidrelétricas

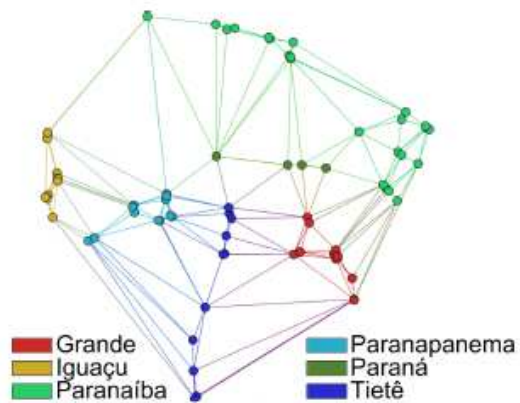
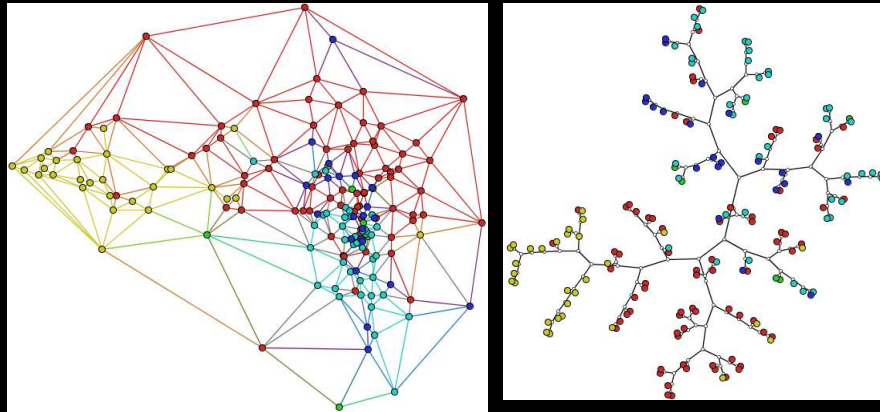


Figure 2. Power plants of the basin Paraná



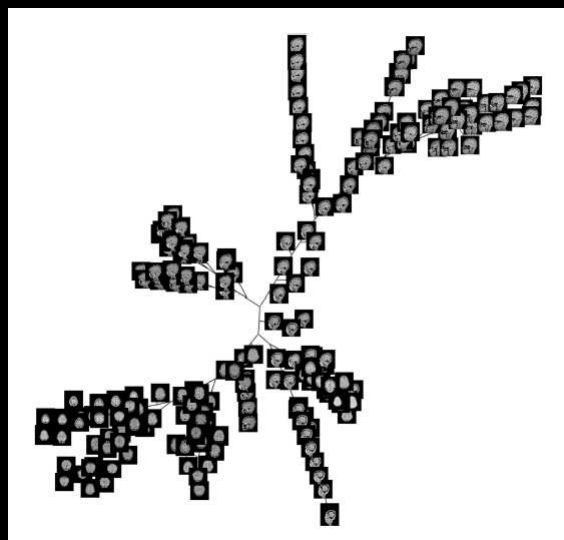
40

Further Example - patents



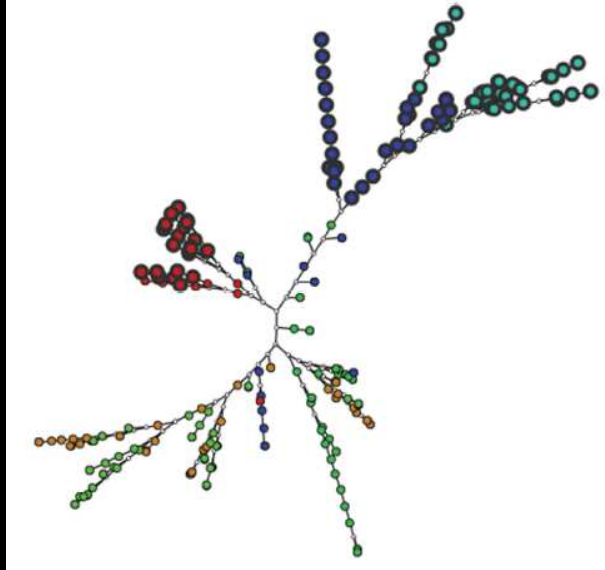
41

Images



42

Imagens



43

Exercícios

- Download Pex
- Usar os conjuntos de dados presentes no site e explorá-los.
- Registrar as conclusões

- Download Pex-Image
- Fazer o mesmo

- Infoserver.lcad.icmc.usp.br/infovis2



44

Referências

- Cuadros, A. M, Paulovich, F. V., Minghim, R., Telles, G. P - Point Placement by Phylogenetic Trees and its Application to Visual Analysis of Document Collections IEEE VAST 2007, Sacramento, CA, USA, IEEE CS Press, pp.99-106.
- Paulovich, F. V., Oliveira, M.C.F., Minghim, R. - The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization, IEEE Sibgrapi 2007, IEEE CS Press, Belo Horizonte, Brazil, pp. 27-34.
- Lopes, A. A., Minghim, R., Melo, V., Paulovich, F.V.; Mapping texts through dimensionality reduction and visualization techniques for interactive exploration of document collections, **SPIE Conference on Visualization and Data Analysis**, San Jose, CA, USA Jan. 2006, 6060T-11.
- Minghim, R., Paulovich, F.V., Lopes, A. A.; Content-based text mapping using multidimensional projections for exploration of document collections, **SPIE Conference on Visualization and Data Analysis**, San Jose, CA, USA Jan. 2006, 6060T-11.



Referências

- Pinho, R. D. ; Oliveira, M. C. F. ; Minghim, R. ; Andrade, M. G. . Voromap: A Voronoi-based Tool for Visual Exploration of Multidimensional Data. In: **10th International Conference on Information Visualization**, 2006, Londres. Proceedings of Information Visualisation 2006, 2006. v. 1. p. 39-44
- Paulovich, F. V. ; Minghim, R. . Text Map Explorer: a Tool to Create and Explore Document Maps. In: Information Visualisation 2006 (IV06) **10th International Conference on Information Visualisation**, 2006, Londres. Proceedings of Information Visualisation 2006, 2006. v. 1. p. 245-251.
- Paulovich, F. V. ; Nonato, L. G. ; MINGHIM, R. ; Levkowitz, H. . Least Square Projection: a fast high precision multidimensional projection technique and its application to document mapping. IEEE Transactions on Visualization and Computer Graphics, 2008.
- Minghim, R. ; Levkowitz, H. ; Nonato, L. G. ; Watanabe, L. S. ; Salvador, V. C. L. ; Lopes, H. ; Pesco, S. ; Tavares, G. . Spider Cursor: A simple versatile interaction tool for data visualization and exploration. In: **ACM GRAPHITE'05** - 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia, 2005, Dunedin. Proceedings of Graphite 2005, 2005. p. 307-314.

