Point Placement by Phylogenetic Trees and its Application to Visual Analysis of Document Collections

Conference Paper · October 2007

DOI: 10.1109/VAST.2007.4389002 · Source: IEEE Xplore

4 authors, including:

Fernando Paulovich
Dalhousie University
87 PUBLICATIONS **1,786** CITATIONS

SEE PROFILE

Rosane Minghim
University of São Paulo
163 PUBLICATIONS **1,983** CITATIONS

SEE PROFILE

University of São Paulo, São Carlos/SP, Brazil
Mathematical and Computer Sciences Institute (ICMC)
Computer Science Department

IEEE Symposium on VAST
Visual Analytics Science and Technology

Point Placement by Phylogenetic Trees and its Application to Visual Analysis of Document Collections

Ana M. Cuadros (anamaria@icmc.usp.br)
Fernando V. Paulovich (paulovic@icmc.usp.br)
Rosane Minghim (rminghim@icmc.usp.br)
Guilherme P. Telles (gpt@icmc.usp.br)

Infovis2/MineVis Project: *http://infoserver.lcad.icmc.usp.br/*

Presentation outline

- Problem Statement and motivation
- Multidimensional Projections drawbacks
- Description of the approach
- Results
- Conclusions

Introduction

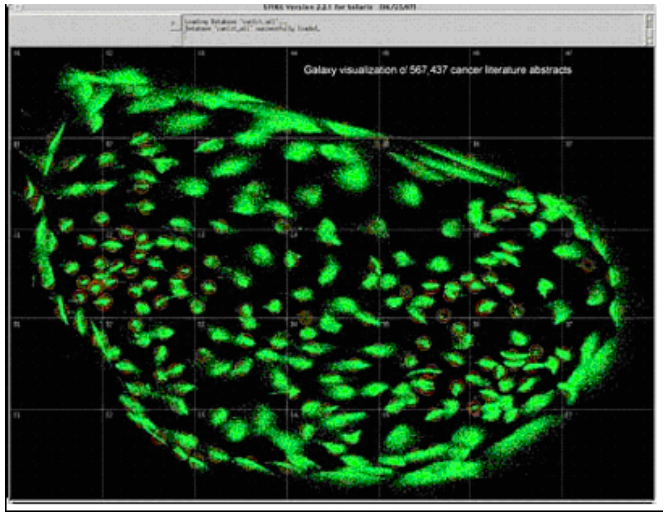
- Problem:
  - Generate a 2D map that reflects relationship among documents
  - Degrees of similarity reflected by proximity in the display
  - Exploratory tasks for document sets based on content
    - Vs. mapping based on metadata

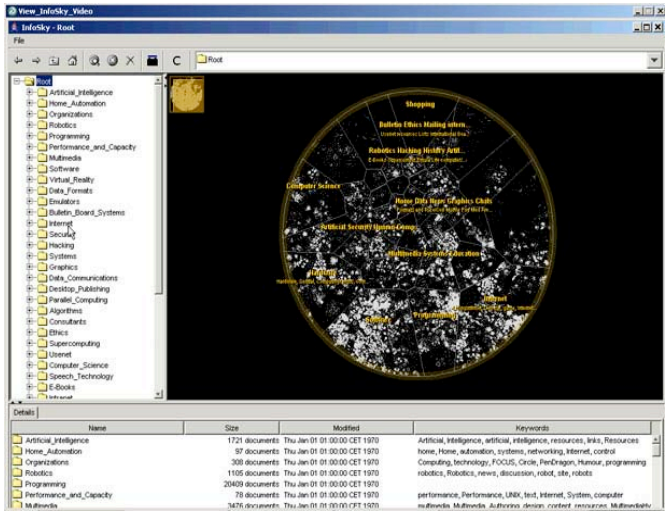
Point placement and multidimensional projections for visualization

- Multidimensional projection technique for data/text analysis:
  - Maps the data into p-dimensional space p={1,2,3}
  - Examples:
    - Principal Component Analysis (PCA) [5]
    - Multidimensional Scaling (MDS) [8]
    - Least-Square Projection (LSP) [11]
  - Projection Explorer (PEx) - http://infoserver.lcad.icmc.usp.br/
- Point Placement Strategies
  - eg. force-directed
- Hybrid Strategies

Maps of documents based on their content

- Ex:
  - IN-SPIRE™ [9]
  - Infosky [1]
- Handle massive amounts of texts and global displays as well as subgroups
- Drawbacks:
  - Documents that should be together get placed in different groups
  - Heterogeneous text sets cause, overlapping regions





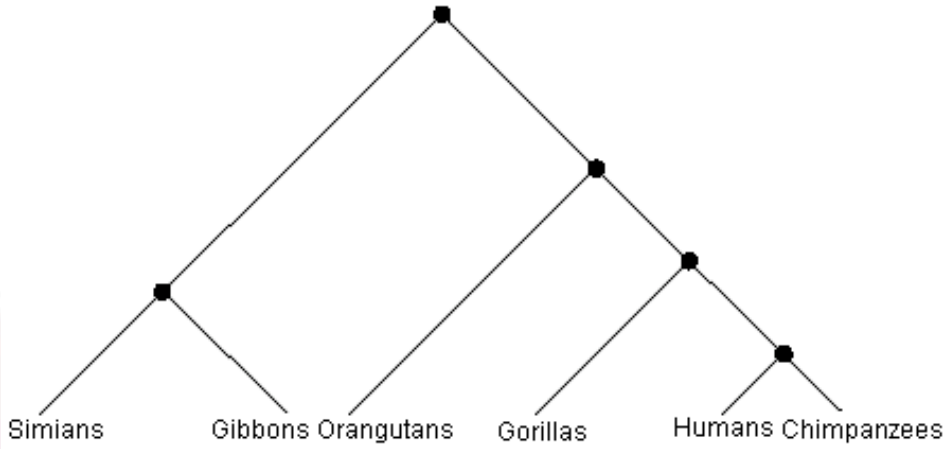
Motivation

- Problem with multidimensional projection techniques
  - Some points are always misplaced
  - Difficulty estimating density of groups
  - Difficulty distinguishing individual points in dense groups

- Alternative:
  - Similarity tree from a distance matrix employing an algorithm for phylogenetic tree reconstruction
    - Reflects the relationships as determined by the similarity measurement

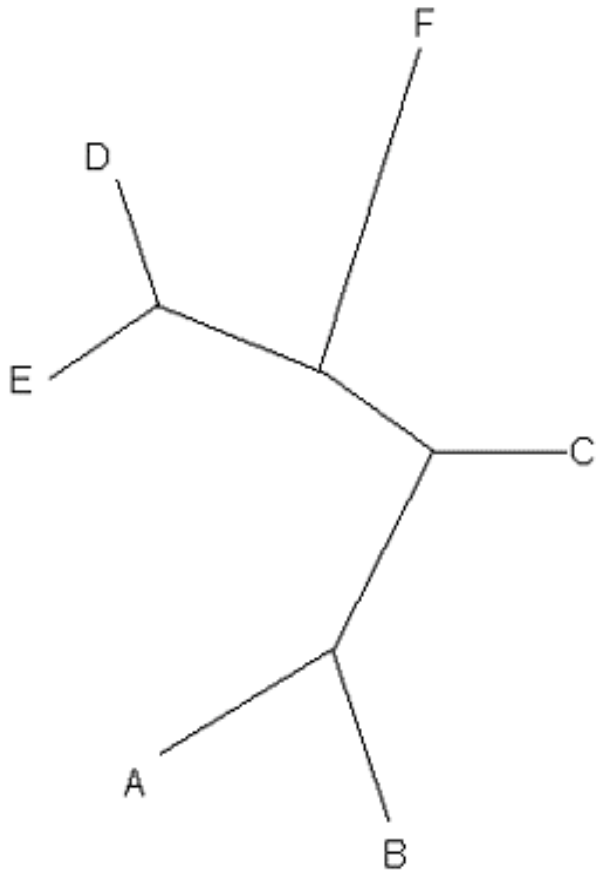
Phylogenetic reconstruction

- Biological problem of building a tree that reflects evolutionary relationship
- Leaves represent species and internal nodes hypothetical ancestors
- Two types of inputs:
  - Character matrix
  - Distance matrix

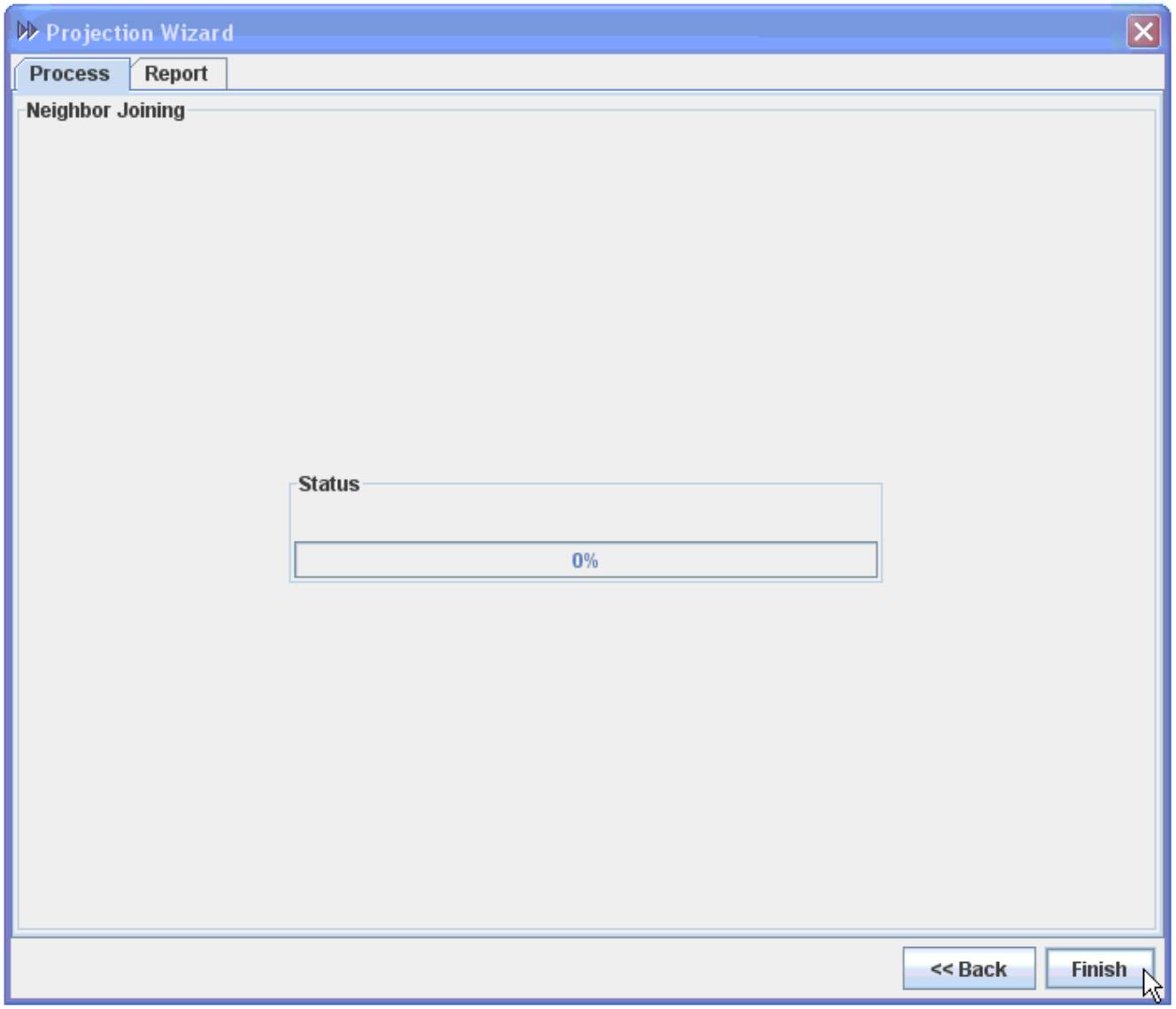


Phylogenetic reconstruction for m-d Vis

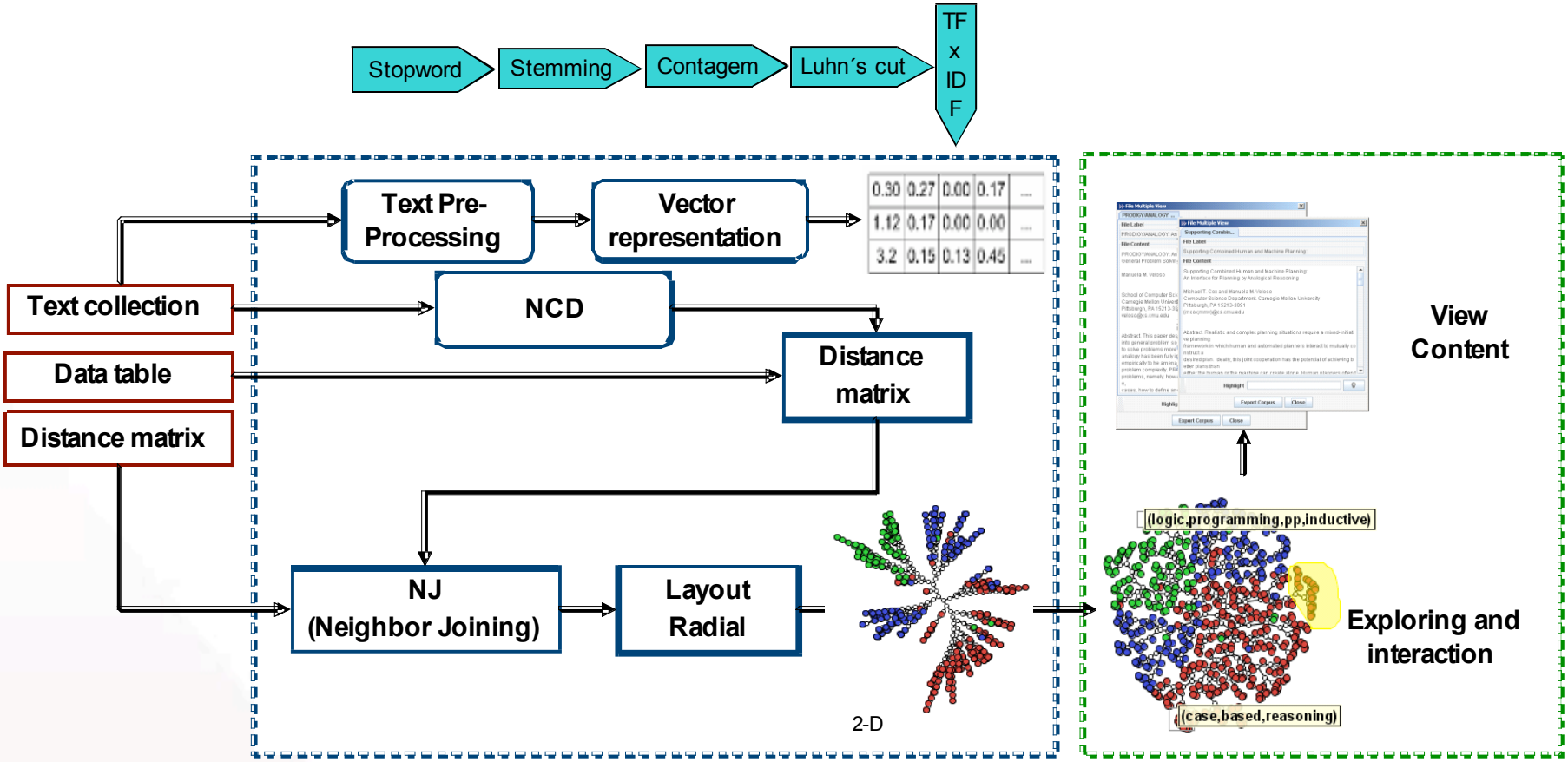
- Generates a distance matrix between all pair of documents
- Builds a similarity tree using a phylogenetic tree reconstruction algorithm (eg. NJ algorithm)
- Employs a lay-out strategy to display the tree
  - Clustered view
- Simplified point placement constrained to branch connections to spread the points around the 2D plane



Building a Tree from a Set of Documents



Process to create the document map

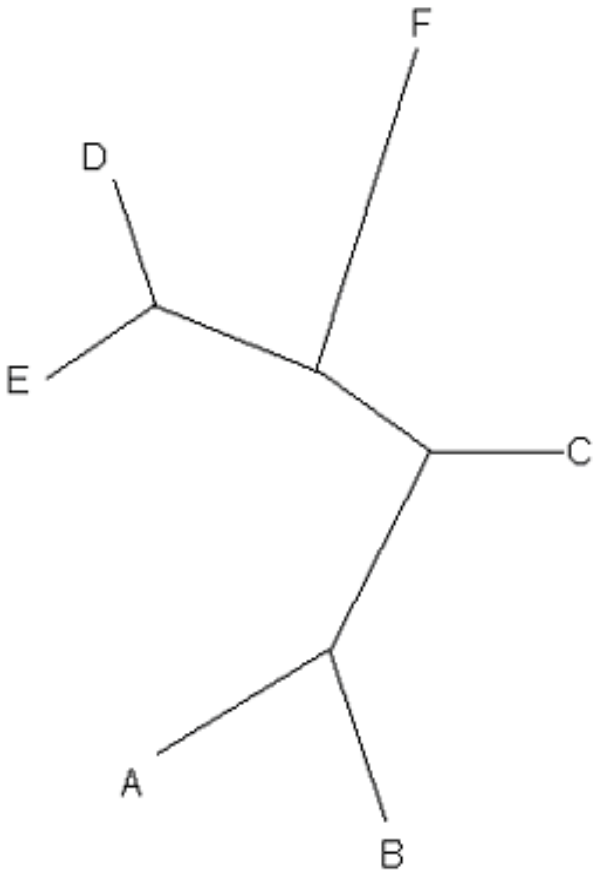


Creating the map NJ

Exploring the map

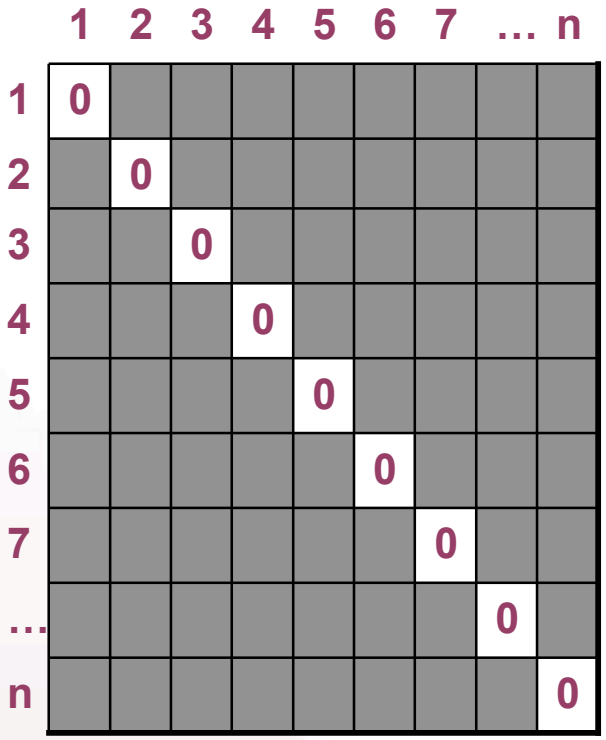
Neighbor-Joining (NJ) [10]

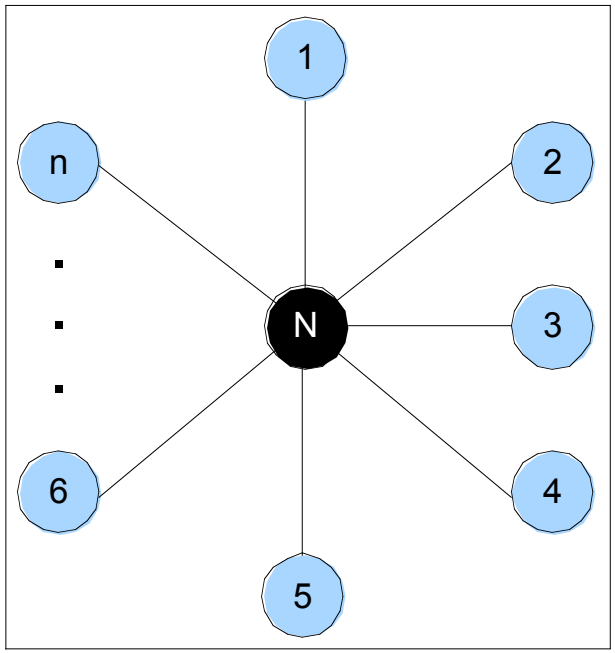
- Neighbor-Joining (NJ) technique [7]:
  - Heuristic algorithm for tree construction
  - Define the tree topology and branches length
  - Builds an unrooted tree
  - Selects the closest pair of documents and joins them into a hypothetical ancestor
  - With text:
    - Leaves: doc.
    - Internal nodes: ancestor hypothetical doc.
    - Edges' lengths: distance between docs.



Neighbor-Joining (NJ) [10]

- Starts with a start-like tree, with n leaves connected to a single internal node



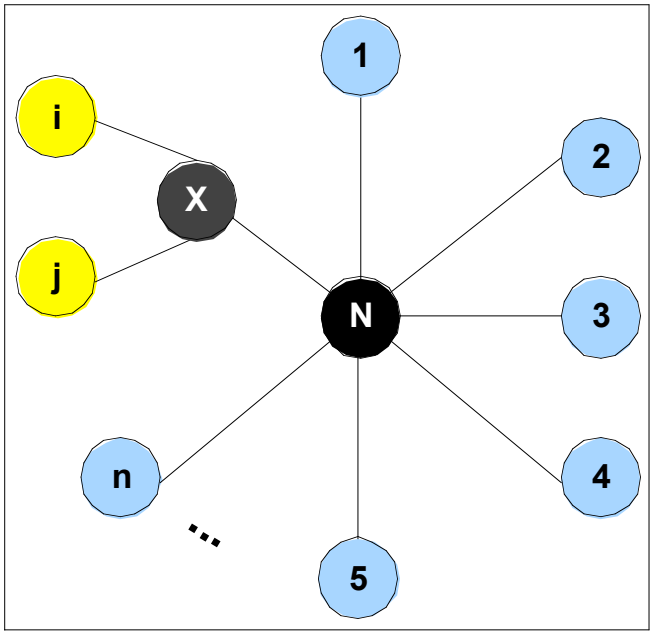


Neighbor-Joining (NJ) [10]

- Selects the smallest sum of branch lengths S_{ij}

$$S_{ij} = \frac{1}{2(n-2)} \sum_{k \neq i,j} (D_{ik} + D_{jk}) + \frac{D_{ij}}{2} + \frac{1}{n-2} \sum_{k,l \neq i,j}^{k<l} D_{kl}$$

- Adds a node x to the tree, with i and j as children and connected to the common ancestor of i and j

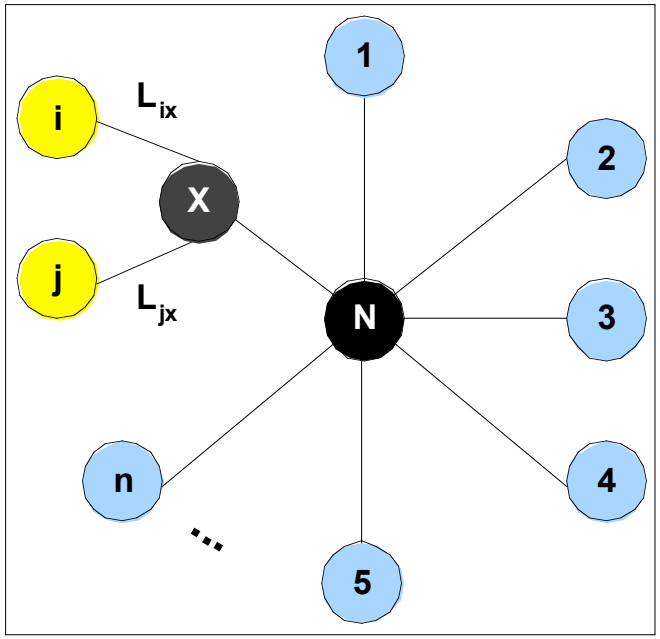


Neighbor-Joining (NJ) [10]

- Evaluates the branch lengths L_{ix} and L_{jx}

$$L_{ix} = \frac{D_{ij} + \frac{\sum_{k \neq j} D_{ik}}{n-2} - \frac{\sum_{k \neq i} D_{jk}}{n-2}}{2}$$

$$L_{jx} = \frac{D_{ij} + \frac{\sum_{k \neq i} D_{jk}}{n-2} - \frac{\sum_{k \neq j} D_{ik}}{n-2}}{2}$$



Neighbor-Joining (NJ) [10]

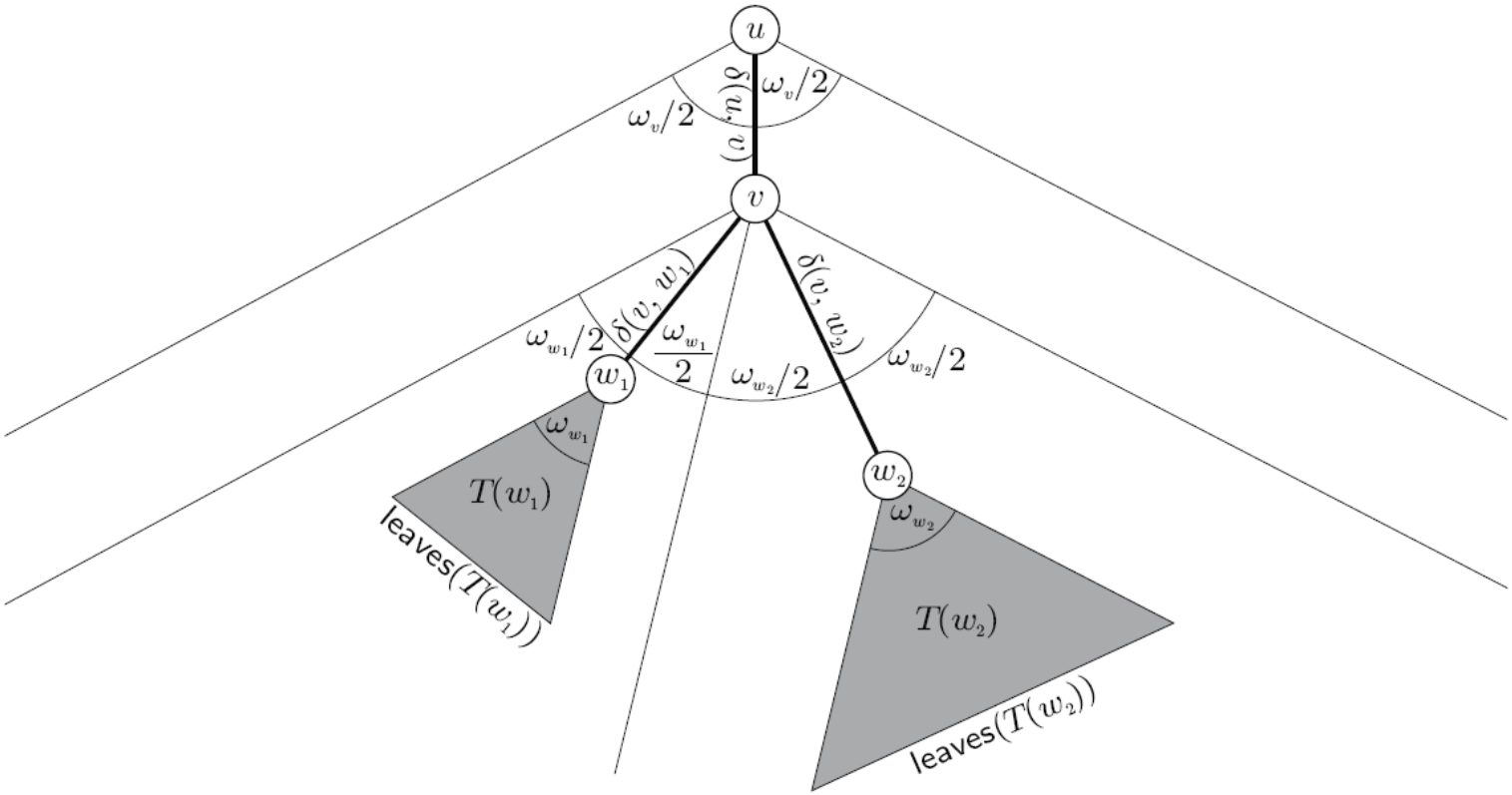
- Replaces i and j by x in the distance matrix evaluating the D_{xy} for every y in the matrix

$$D_{xy} = \frac{D_{iy} + D_{jy}}{2}$$

- Repeats this steps until there is only two nodes remaining in the matrix

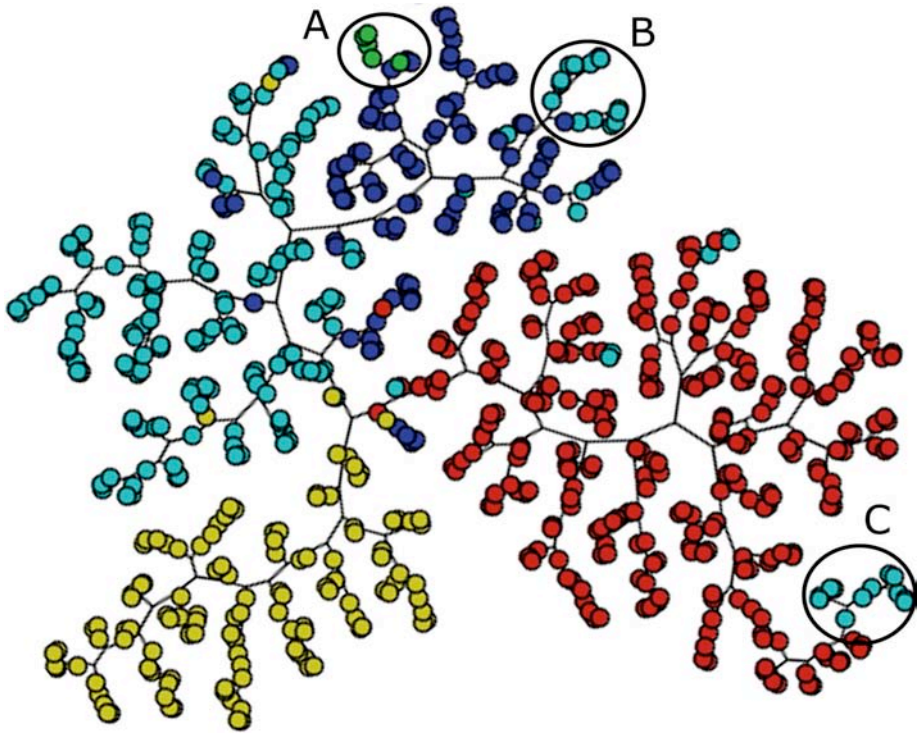
Radial layout [2]

- Preserves edge lengths
- Computed in linear time



Results

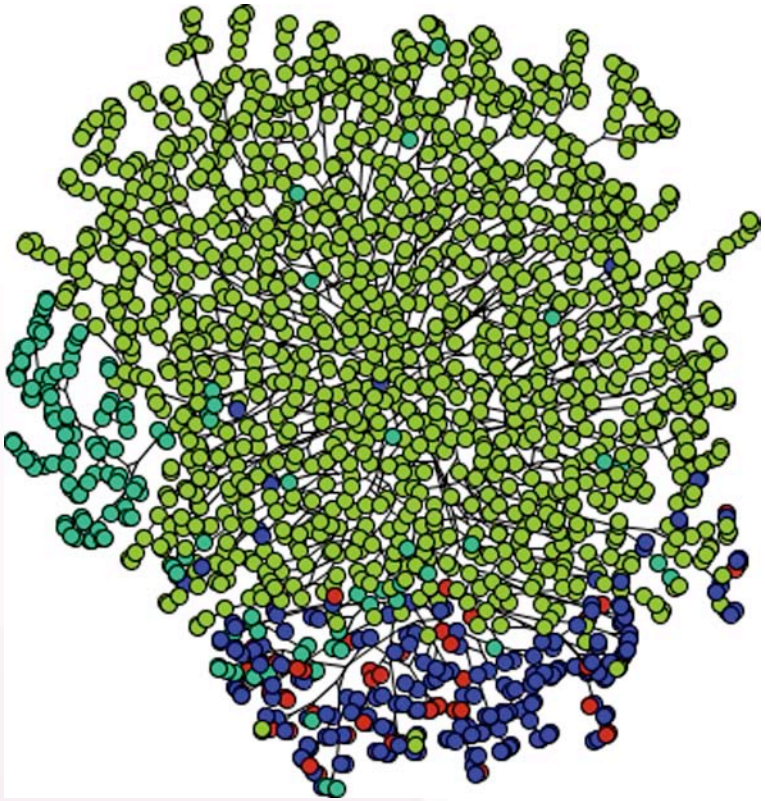
- **Mapping Scientific data sets**
  - CBR+ILP+IR+SON, 680 files, Scientific papers

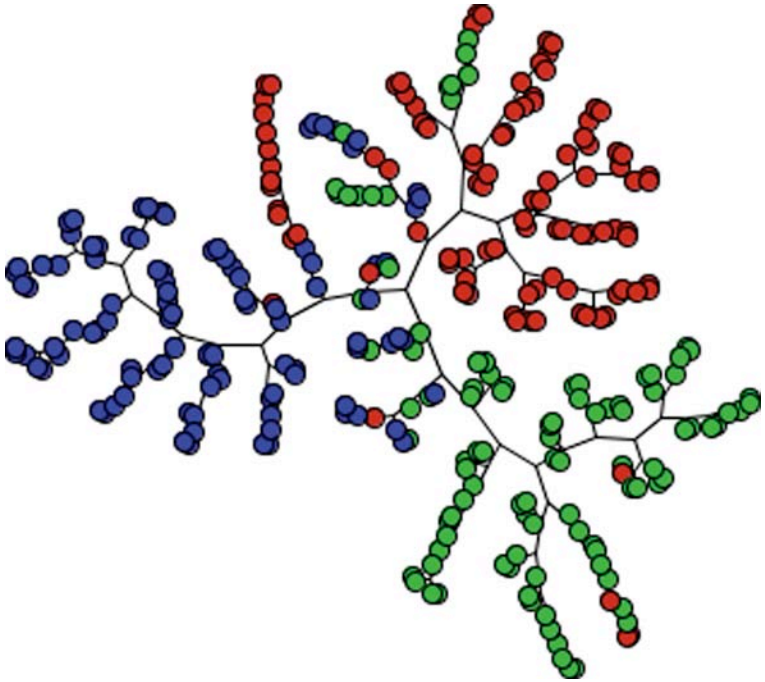


- **Case-based Reasoning**
- **Inductive Logic Programming**
- **Information Retrieval**
- **Sonification**

Results

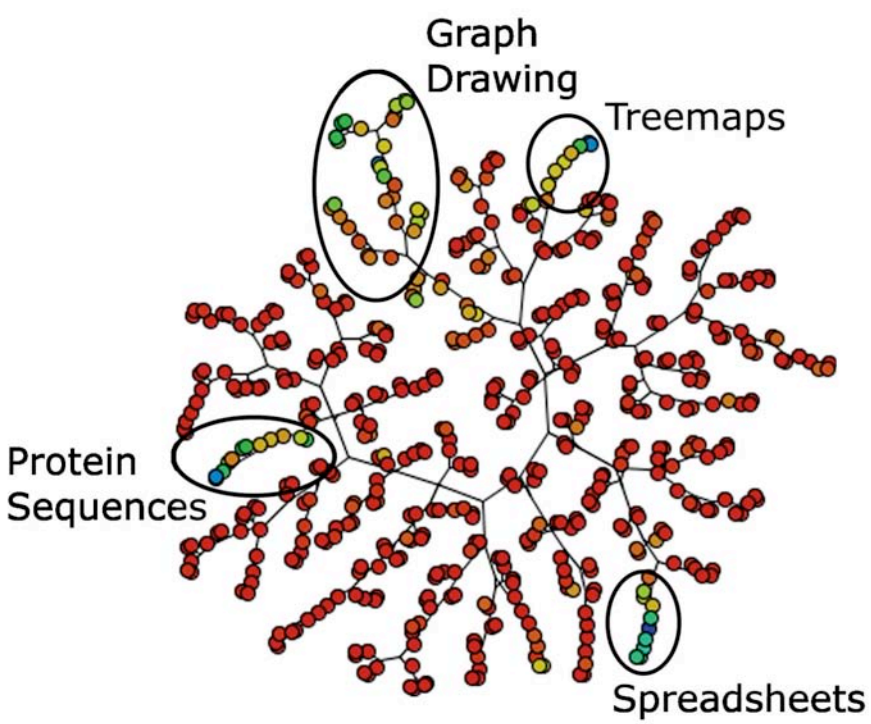
- **Mapping data sets**
  - KDVis, 1,624 files, Scientific papers
  - MESSAGES, 300 files, Discussion groups



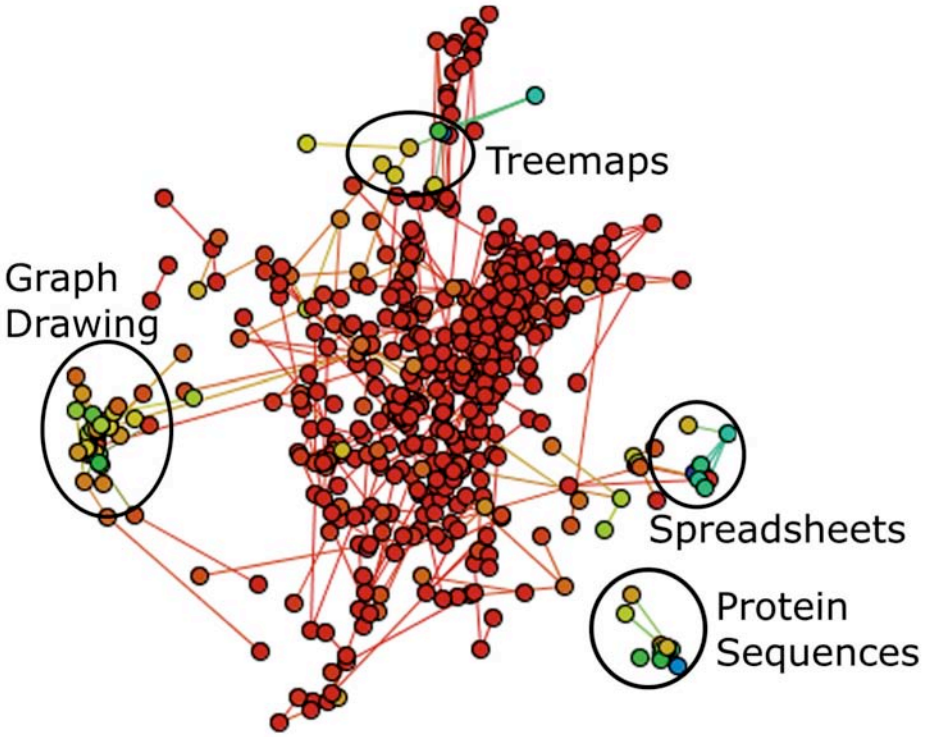


Results

- **Mapping data sets for NJ and projection techniques**
  - 2004 IEEE Infovis Contest, 515 files, Scientific papers



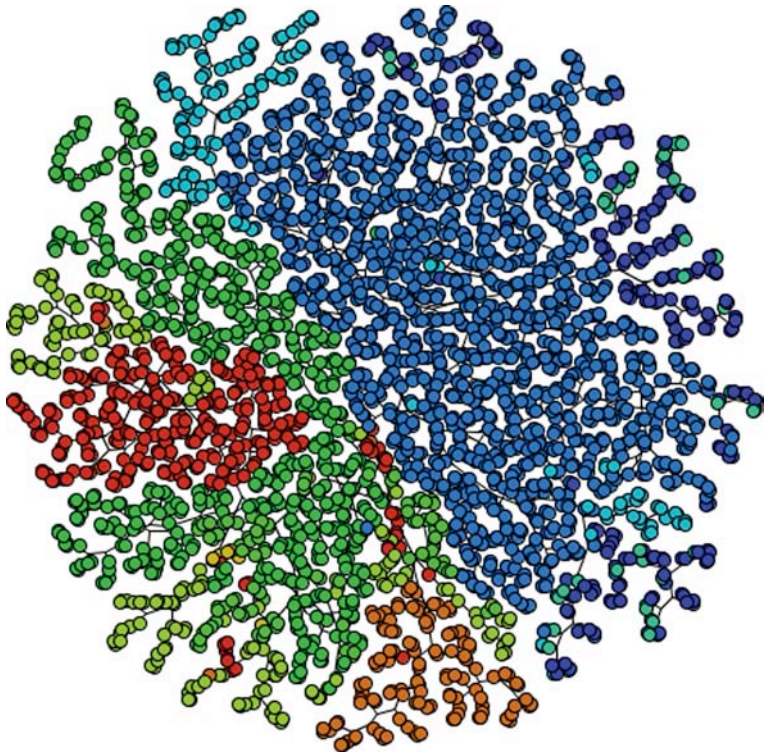
- **NJ**



- **LSP**

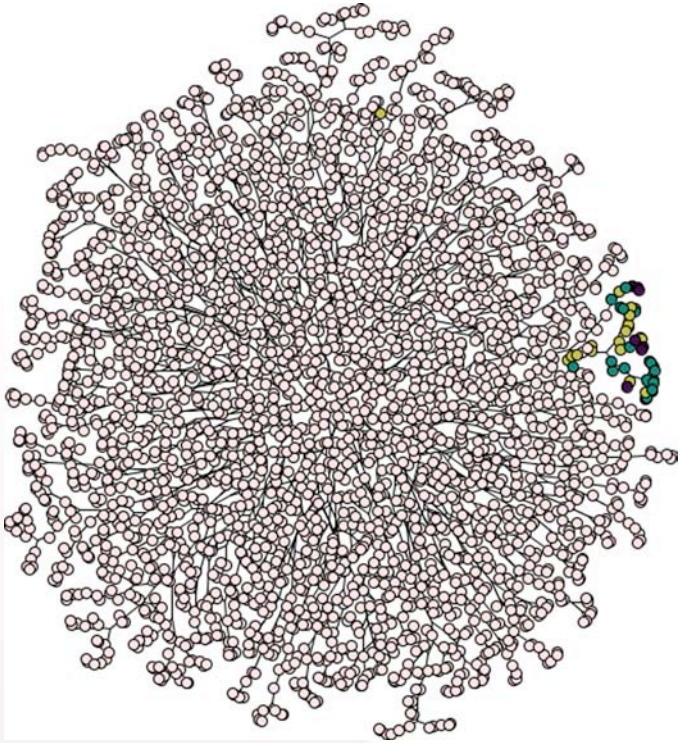
Results

- All scientific data set together using NCD similarity [8]

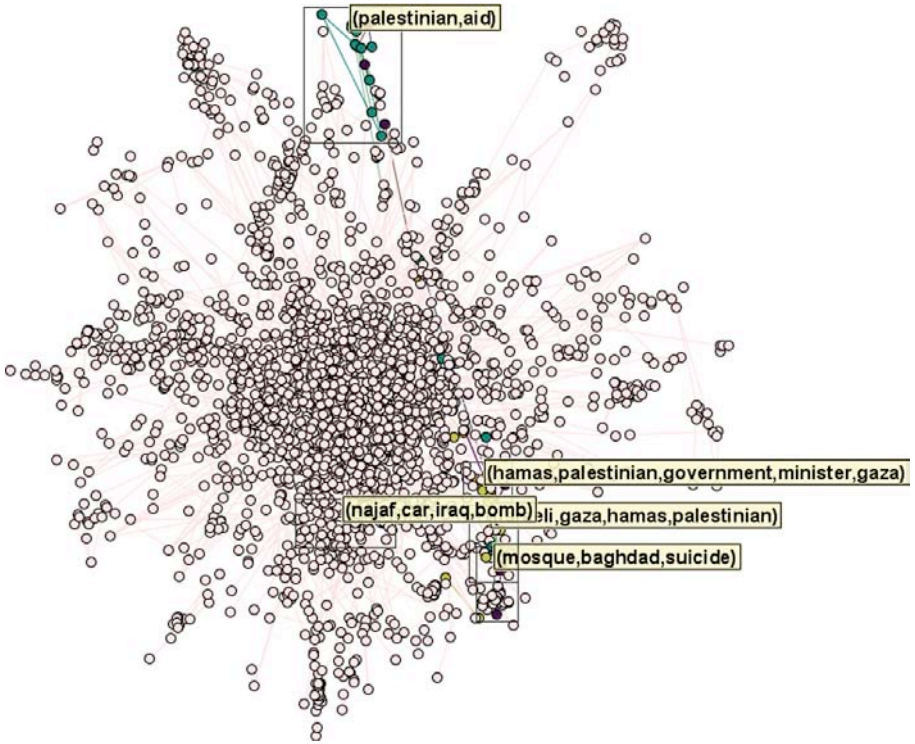


Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - NEWS, 2,684 files



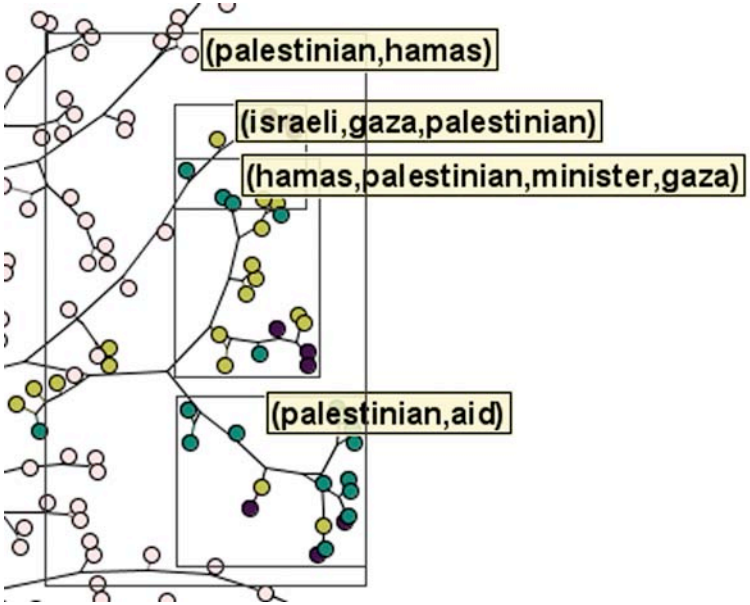
- **NJ**



- **LSP**

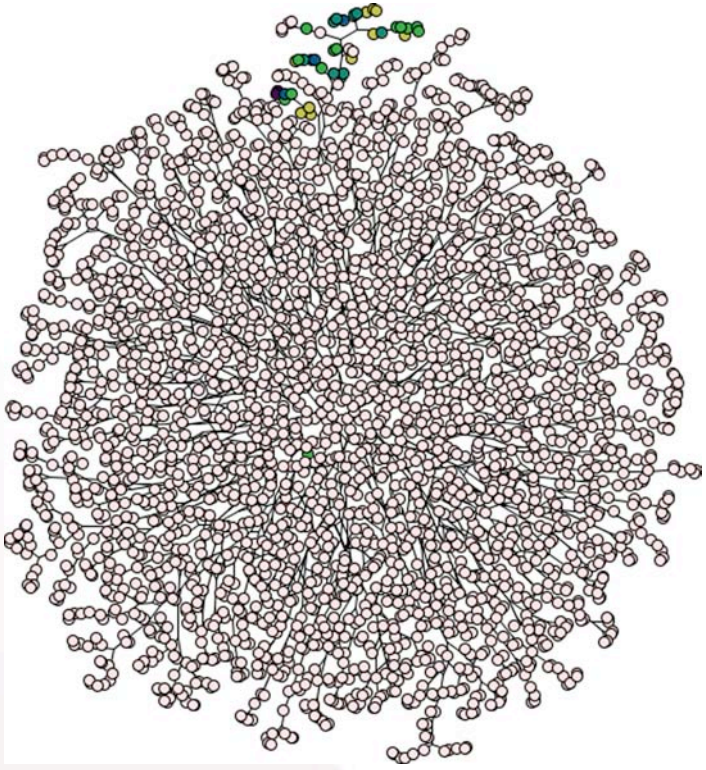
Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - Corpus NEWS, 2,684 files

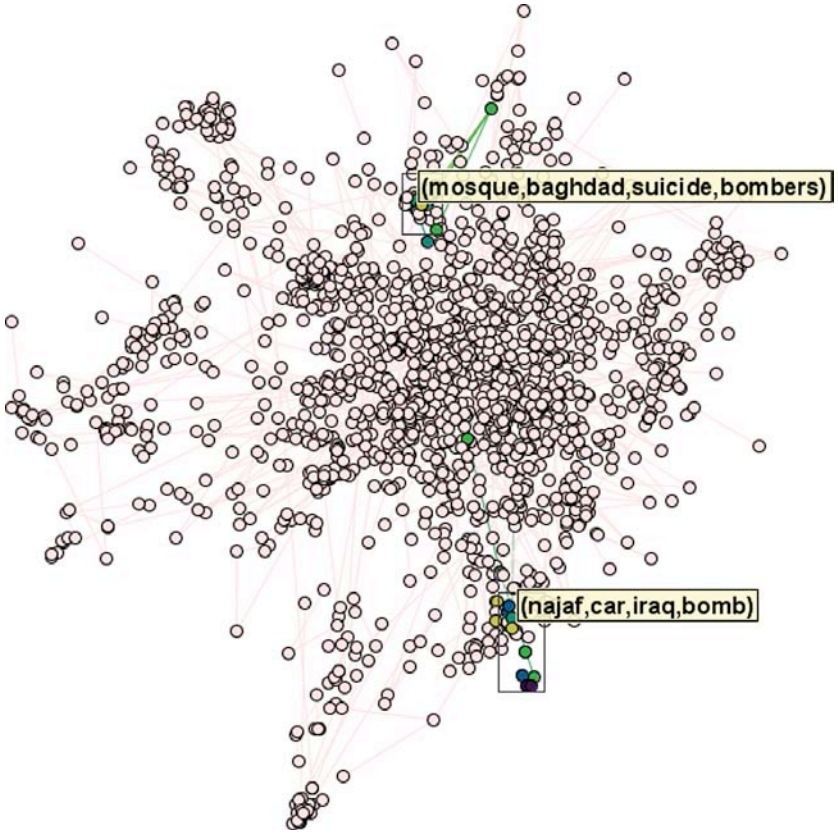


Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - Corpus NEWS, 2,684 files



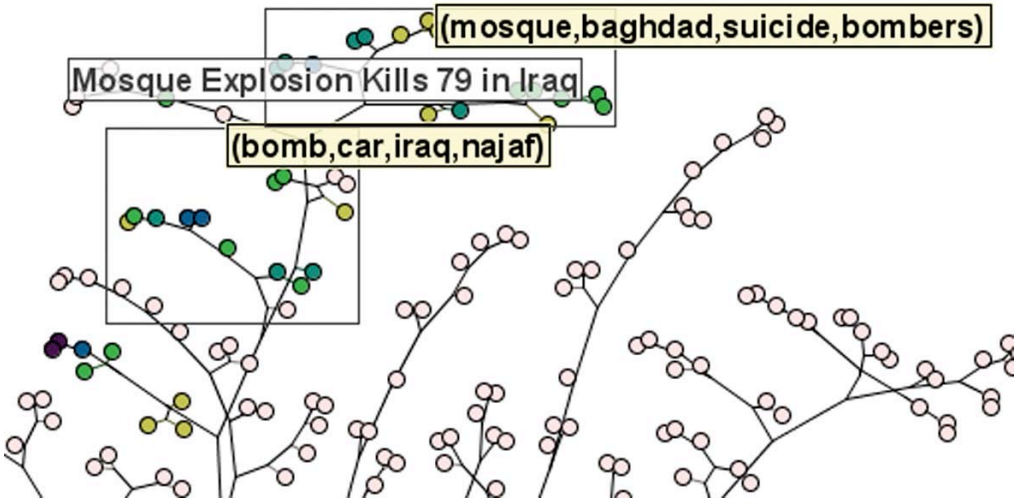
- **NJ**



- **LSP**

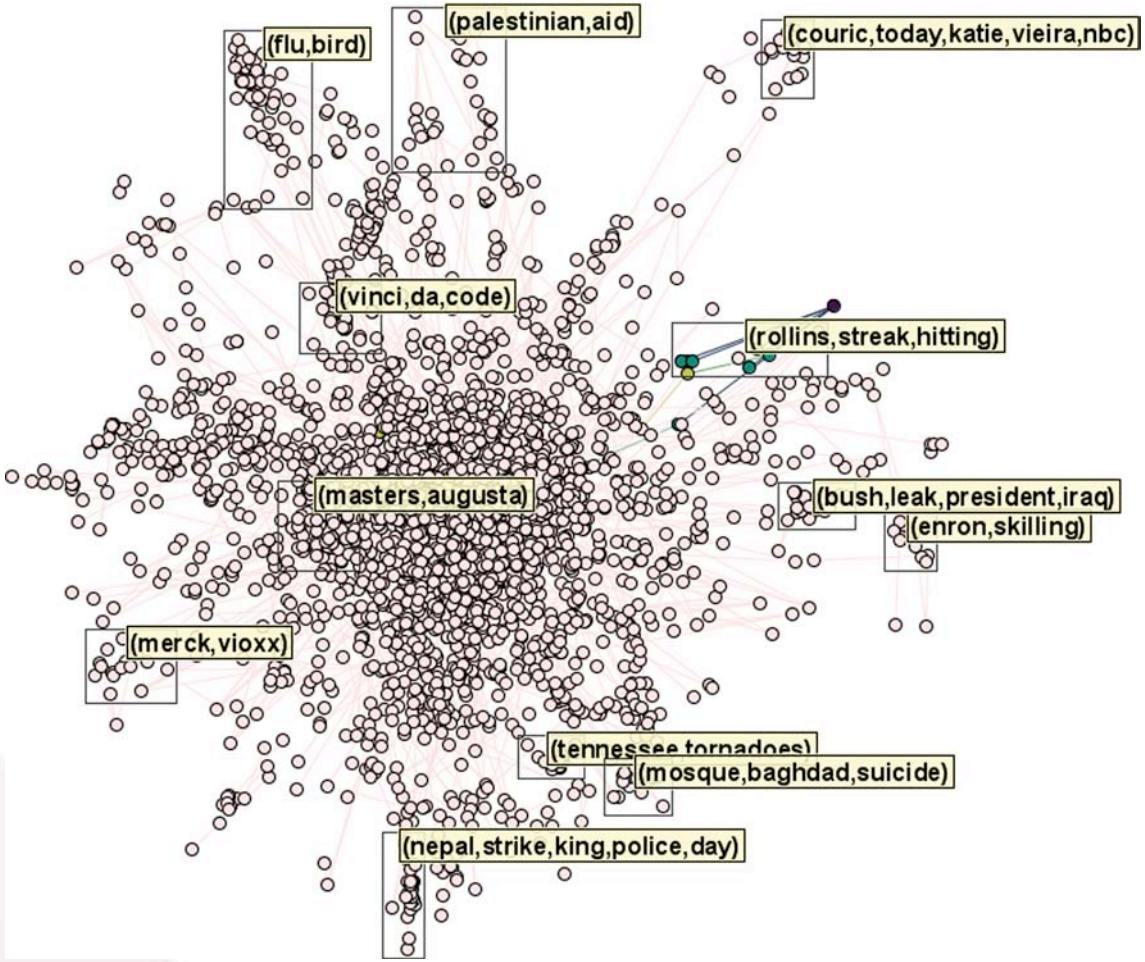
Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - Corpus NEWS, 2,684 files



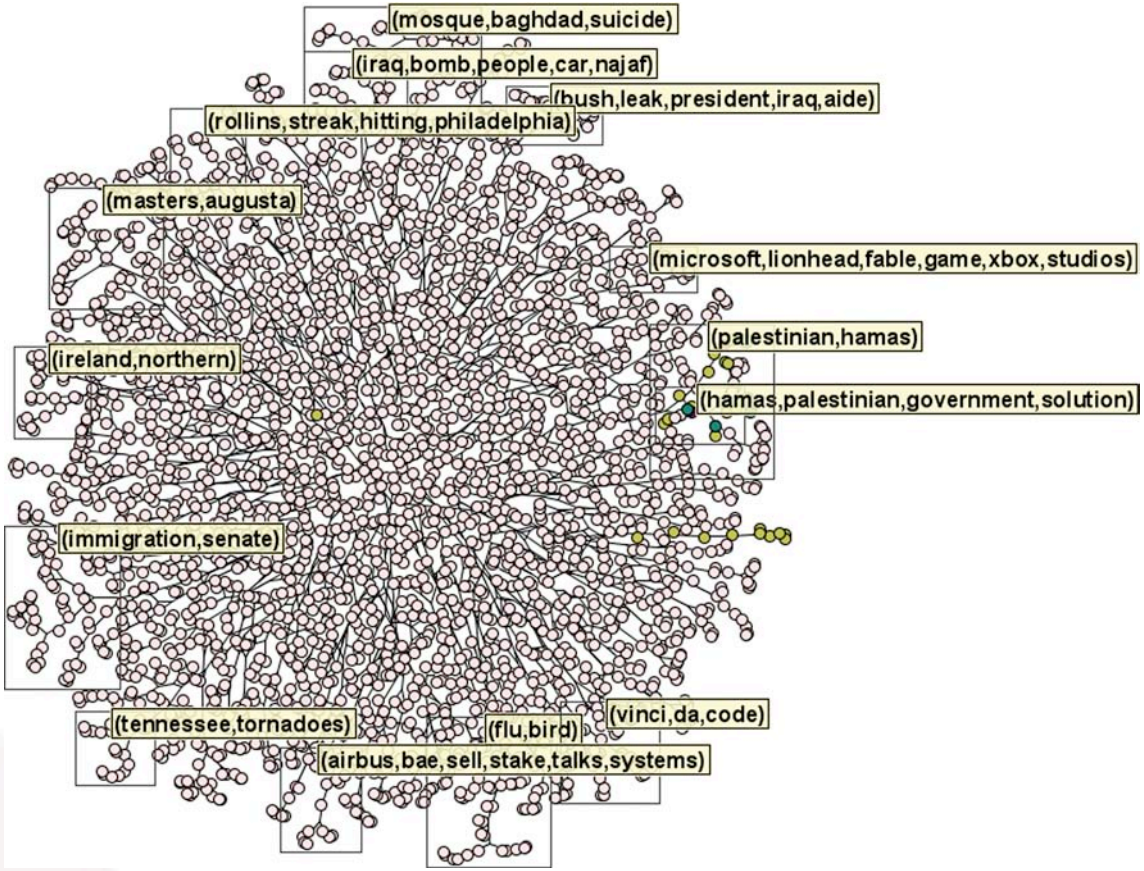
Results

- Grouping by topic (LSP)



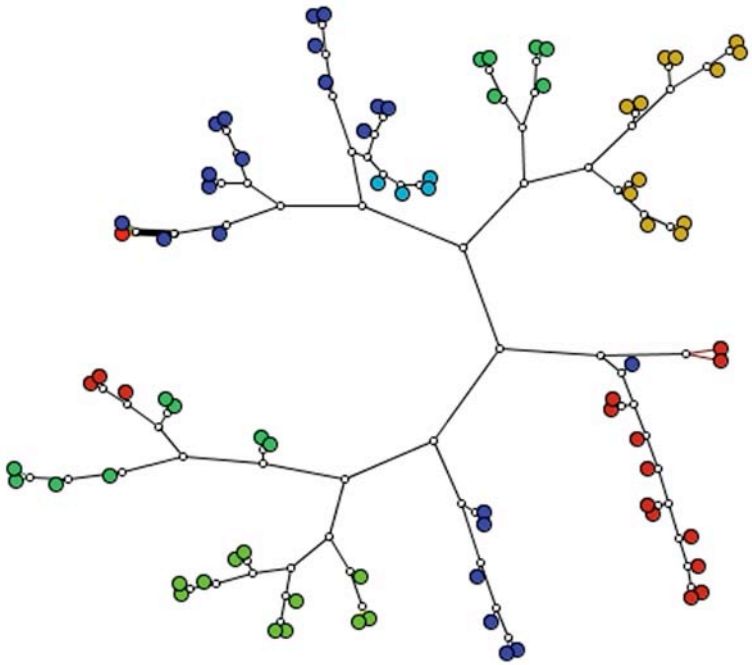
Results

- Grouping by topic (NJ)



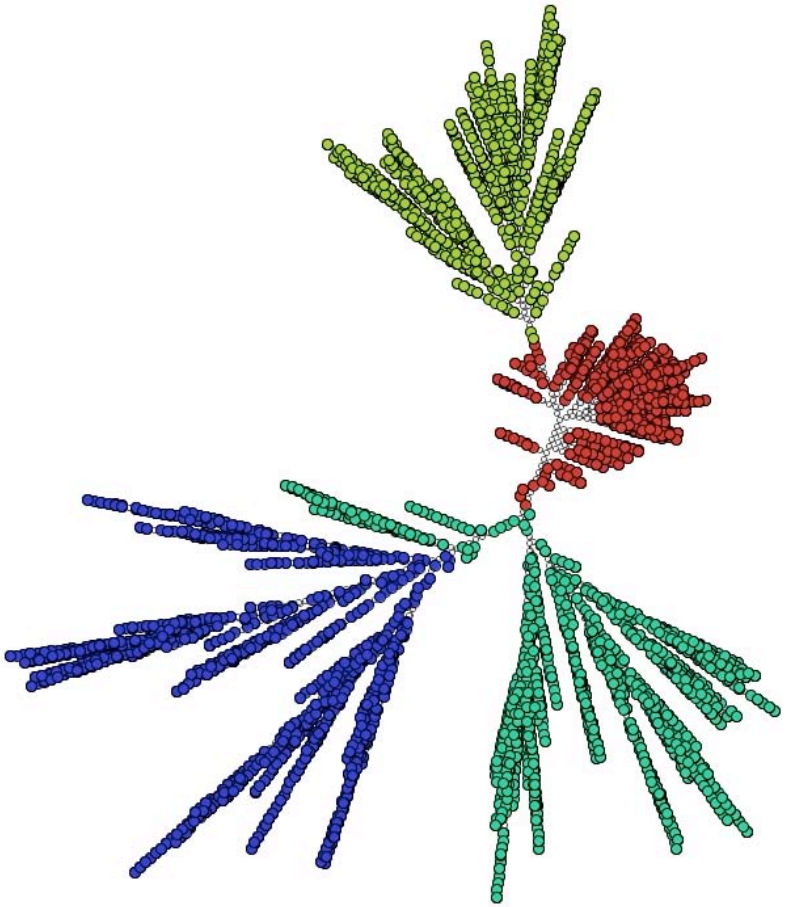
Results

- **Stream-flow in hydroelectric plants of Paraná River (Brazil)**
  - Color is sub-basin of the river



Results

- **Quadrupeds mammals data set**
  - 10,000 data instances



Results

- Time in seconds to create maps in a 3.2 GHz Pentium 4

Data Set	NJ	Layout	Total
CBR+ILP+IR+SON	4,55	0,52	5,17
KDVis	66,20	1,26	67,46
INFOVIS04	1,83	0,45	2,28
ALL	454,66	2,17	456,83
MESSAGES	0,35	0,31	0,66
NEWS	359,63	1,70	361,33

Concluding remarks

- NJ
  - Reflects content relationship visually
  - Constructs a hierarchy
- Interpretation of display
  - Makes good use of the visual space
  - Complementary of the projections
- Same distance matrix always generates the same tree
  - Helps evaluating the similarity measurement

Concluding Remarks

- Further work
  - Construct a set of tools for proper exploration of phylogenetic trees
  - Reduce processing time
  - Hybrid and hierarchical approaches
  - Test with other phylogenetic reconstruction methods

- Brazilian financial agencies
  - FAPESP
  - CNPq

References

[1] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3/4):166–181, 2002.

[2] C. Bachmaier, U. Brandes, and B. Schlieper. Drawing phylogenetic trees. In X. Deng and D. Du, editors, *Proc. Intl. Symp. on Alg. And Comp. , ISAAC 2005*, volume 3827, pages 1110–1121, 2005.

[3] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, second edition, 2000.

[4] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[5] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2 edition,2002.

[6] F. V. Paulovich and R. Minghim. Text map explorer: a tool to create and explore document maps. In *IV '06: Proc. of the conf. on Information Visualization*, pages 245–251, Washington, DC, USA, 2006. IEEE Computer Society Press.

References

[7] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.

[8] G. Telles, R. Minghim, and F. Paulovich. Normalized compression distances for visual analysis of document collections. *Computer & Graphics, Special Issue on Visual Analytics*, 2007.

[9] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Readings in information visualization: using vision to think*, pages 442–450, San Francisco, CA - USA, 1995. Morgan Kaufmann Publishers Inc.

[10] M. Salemi, A. Vandamme. The phylogenetic handbook. A practical approach to DNA and protein phylogeny. Cambridge University Press, 2003.

[11] F. V. Paulovich and L. G. Nonato and R. Minghim and Haim Levkowitz. Least Square Projections; A Fast High-precision multidimensional projection technique and its application to document mapping. IEEE TVCG (to appear), 2008.