

Interactive Visualization of a News Clips Network: *A Journalistic Research and Knowledge Discovery Tool*

José Devezas and Álvaro Figueira

CRACS/INESC TEC, Faculdade de Ciências, Universidade do Porto
Rua do Campo Alegre, 1021/1055, 4169-007 Porto, Portugal

`jld@dcc.fc.up.pt, arf@dcc.fc.up.pt`

4th International Conference on Knowledge Discovery and Information Retrieval

October 5, 2012

Introduction

- Breadcrumbs is a social network based on the relations established by collections of text fragments taken from online news.
- It can be used to collect and store fragments of text, called news clips, from online sources, in a Personal Digital Library.
- Text fragments are then semantically organized, based on several latent features found in the text, tags and comments assigned by the users.

Introduction

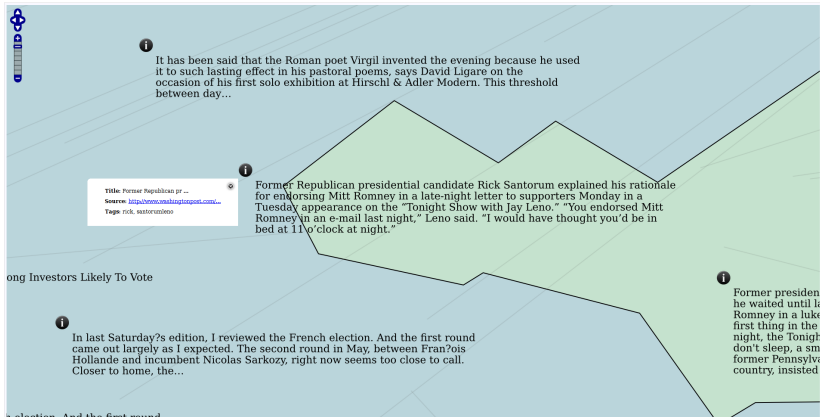
- We present two interactive visualization tools based on a multidimensional network of news clips from the Breadcrumbs system:
 1. A multiresolution visualization created with *gvmap* (Gansner et al., 2010), a GraphViz tool to generate static illustrations of graphs as maps.
 2. A web-based force-directed visualization developed using the data-driven approach of *d3.js* (Bostock et al., 2011).

News Clips Network

- ▶ We used a small network of news clips with:
 - ▶ 94 nodes representing news clips
 - ▶ 166 edges representing entity coreferences¹ in three dimensions:
 - ▶ 17 people (*who* dimension)
 - ▶ 106 places (*where* dimension)
 - ▶ 43 dates (*when* dimension)
- ▶ The following attributes were available for each node:
 - ▶ Clip ID
 - ▶ Creation Date
 - ▶ Source URL
 - ▶ Text Fragment
 - ▶ Community Membership ID
- ▶ The following attributes were available for each edge:
 - ▶ Number of Co-occurrences
 - ▶ Dimension Description (*who*, *where* or *when*)
 - ▶ DBpedia entity class (e.g. *dbpedia-owl:Scientist*)
 - ▶ DBpedia entity URI

¹Entity references were discovered independently from the idiom and resolved to their corresponding URI. This way connections are also established between news clips written in different languages (IAENG IJCS, Devezas, 2012).

Mapping the Relationships



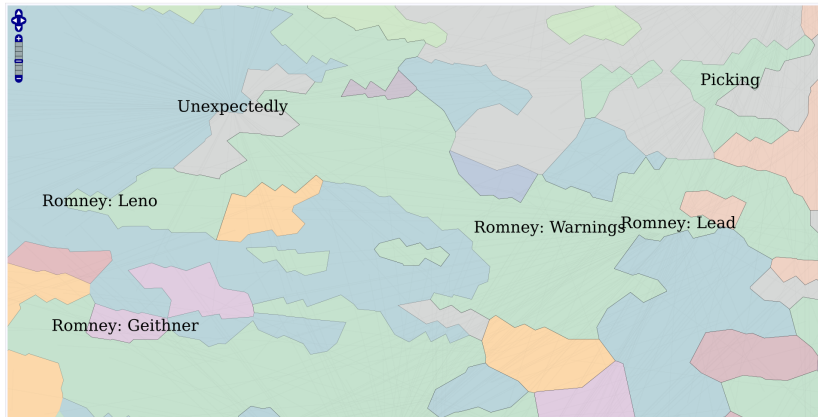
Mapping the Relationships

- ▶ Each zoom level was created from a single precomputed GraphViz file for the maximum resolution (illustrated in the previous figure).
 - ▶ The original input file described the network and defined news clips text as node labels.
 - ▶ Node positioning and spacing was done according to this initial file using *sfdp* and *gvmap*.
 - ▶ The resulting GraphViz file was then used to create the highest resolution zoom level PNG and as base for the lower levels.

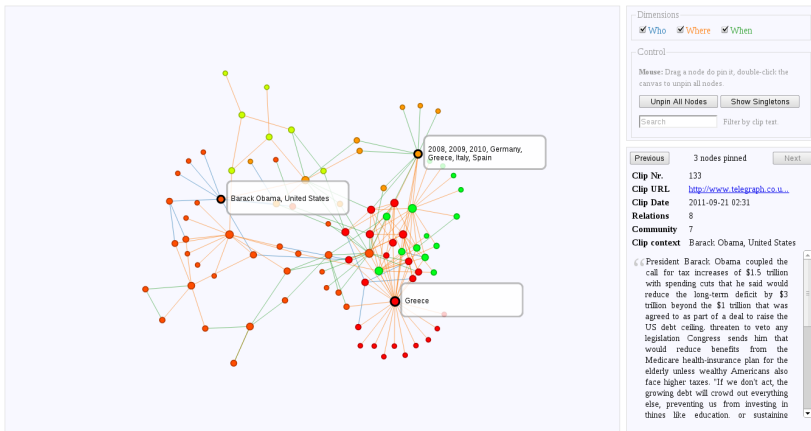
Mapping the Relationships

- ▶ For lower levels, clip text was replaced with a double label (illustrated next).
- ▶ To keep the initially computed map structure across zoom levels, only the node label was replaced in the original GraphViz file.
- ▶ For the n most central nodes, according to PageRank, where n depends on the level, we retrieved the word with highest TF-IDF for the concatenated text of the node's community, as well as for the node's text.
- ▶ This resulted in a double label in the format "*community label: node label*", where *community label* illustrates the global topic, while *node label* illustrates the local topic of the node.

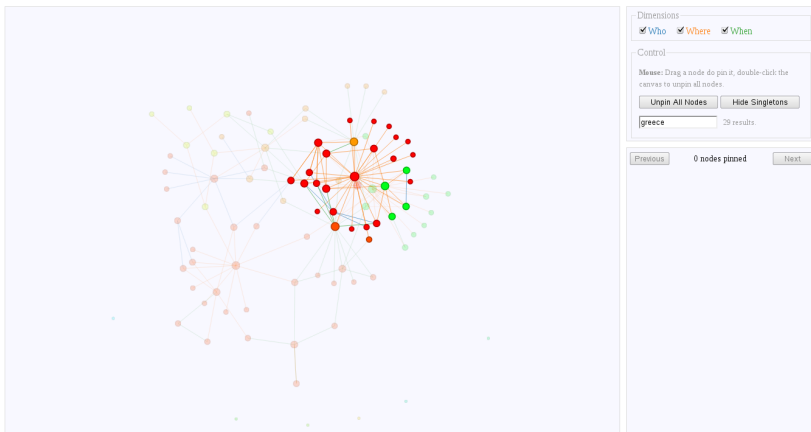
Mapping the Relationships



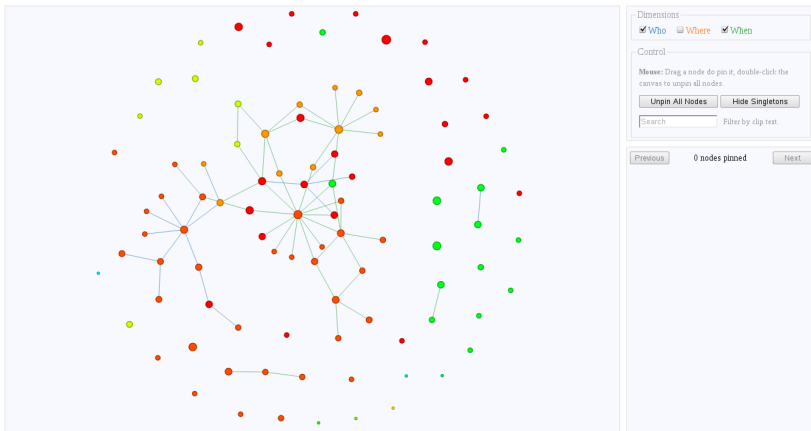
Multidimensional Network: Node Pinning



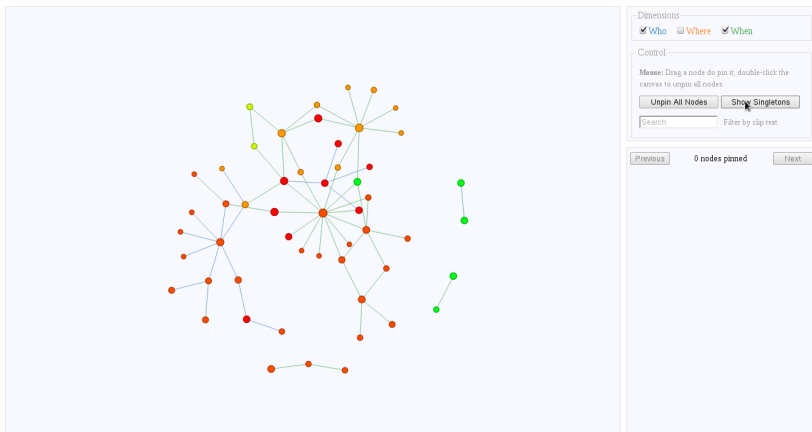
Multidimensional Network: Node Filtering



Multidimensional Network: Dimension Filtering



Multidimensional Network: Toggle Singleton Nodes Visibility



Journalistic Research

- ▶ The goal: to provide visual tools that support the journalistic research process.
- ▶ We provide two complementary solutions:

1. A network map for global topic exploration.

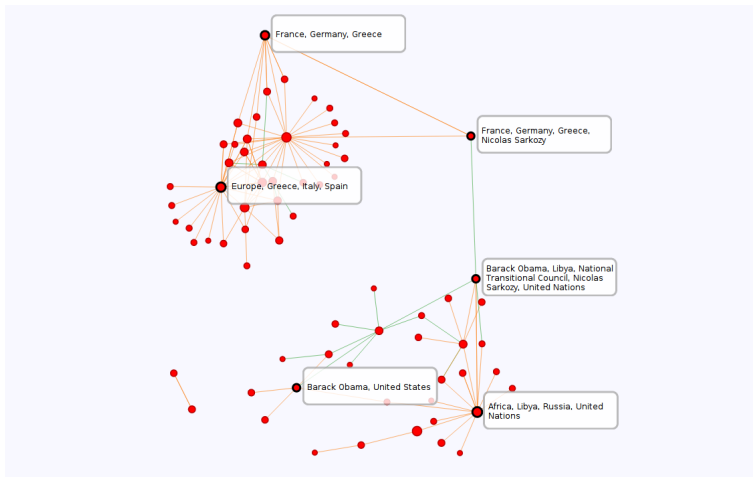
- ▶ Maps are familiar to people and provide a much more intuitive way to explore networks with identified community structure.
- ▶ The semantic aggregation of news clips provided by the colored communities should help identify the different groups of related topics, as well as the boundaries where topics transition to new subjects.
- ▶ Each “country” in the map contains a group of “semantically close” news clips.

Journalistic Research

2. A web-based multidimensional network.

- ▶ Journalists can explore the connections between news clips based on three of the Five-Ws: *who*, *where* and *when*.
- ▶ This enables, for instance, the discovery of important people that serve as connectors between different topics (e.g. Nicolas Sarkozy is a bridge between Europe and international affairs, being connected to the United Nations, Libya and Barack Obama — illustrated next).
- ▶ The journalist can use this to discover and question connections (e.g. Why is Nicolas Sarkozy being referenced alongside such distinct entities? What role does this entity play in the whole?).

Journalistic Research



Conclusions

- ▶ The multiresolution map visualization was effective in producing a clear illustration of the network's nodes and clusters.
- ▶ However it didn't provide by itself a very rich interaction to the user apart from a semantic zooming behavior and quick access to news clips metadata.
- ▶ GraphViz only provides the means to generate a single static image for the network map, thus making it hard to create interactive and dynamic visualizations.

Conclusions

- ▶ The visualization of the multidimensional network of news clips, developed using a data-driven approach, enabled the user to organize and filter nodes, as well as to visually toggle any of the available edge dimensions.
- ▶ This allowed the users to interactively explore several aspects of the data that would otherwise be difficult to interpret, resulting in a tool that can be helpful in journalist research.

Future Work

- ▶ Improve on the existing network map visualization, specially in regards to the method of community and news clip topic discovery, when computing the pair of node labels.
- ▶ Evaluate the developed visualization tools based on human input, assessing user experience and usability, with a focus on the journalistic community.