Creating Mindmaps of Documents

Using an Example of a News Surveillance System

Oskar Gross
Hannu Toivonen
Teemu Hynonen
Esther Galbrun

February 6, 2011

Outline

- Motivation
- Bisociation Network
- Tpf-Idf-Tpu Measure
- News Surveillance System
- Bisociations for Computational Creativity

Motivation

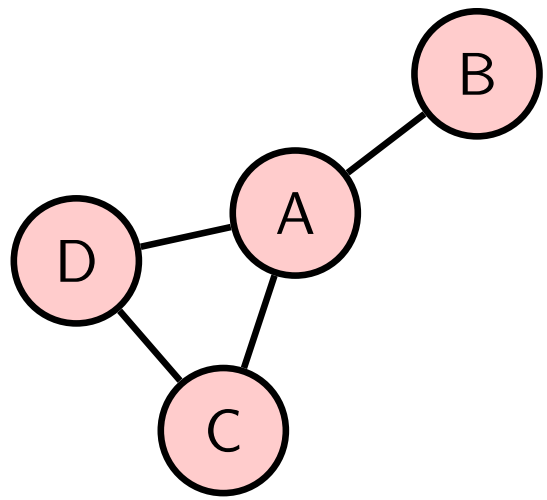
- Epic information overload
- Finding connections between concepts
- Discovering novel (hopefully interesting) connections

Bisociation Networks

- Networks constructed of item (in our case term) pairs
- For an example consider the following set of item pairs:

$$P = \{(A, B), (A, C), (C, D), (D, A)\}$$

- Now treating items as nodes and drawing an undirected connection between each pair gives us a graph



Text to Bisociation Network: Step 1 - Preprocessing

- Our goal is to apply this method on everyday texts
- Reasonable preprocessing is needed
  - Wonderful Python package NLTK
  - HTML $\rightarrow$ plain text
  - Named Entity Recognition
  - Removing Stopwords
  - Stemming

Text to Bisociation Network: Step 2 - Creating Pairs

- Tokenize document into sentences
- Sort words in sentences
- Remove duplicates
- Create Pairs
- Example:
  - Consider the following text
    Thank you for the dinner and a very pleasant evening. Have your car take me to the airport. Mr Corleone is a man who insists on hearing bad news at once.
  - Which is after preprocessing
    dinner even pleasant thank veri . airport bad car insist take . hear mr_corleon man new onc .

Step 3 - Calculate Measure (1)

- Term pair frequency (tpf)

$$tpf_{sen}(\{t, u\}, d) = \frac{|\{s \in d | \{t, u\} \subset s\}|}{|\{s \in d\}|},$$

where s is a sentence, d is a document.

- Inverse document frequency (idf)

$$idf_{doc}(t, u) = \log \frac{|C|}{|\{d \in C | \{t, u\} \subset d\}|},$$

where C is document collection, d is a document, (t, u) is a term pair.

Step 3 - Calculate Measure (2)

- Term pair uncorrelation (tpu)

$$tpu_{sen}(\{t,u\},d) = \min_{v \in \{t,u\}} \left(2 - \frac{|\{d \in C | \exists s \in d \ s.t. \ \{t,u\} \subset s\}|}{|\{v \in d\}|}\right)$$

- Finally getting the tpf-idf-tpu measure

$$M = tpf_{sen} \cdot idf_{doc} \cdot tpu_{sen}$$

Applying to News Stories

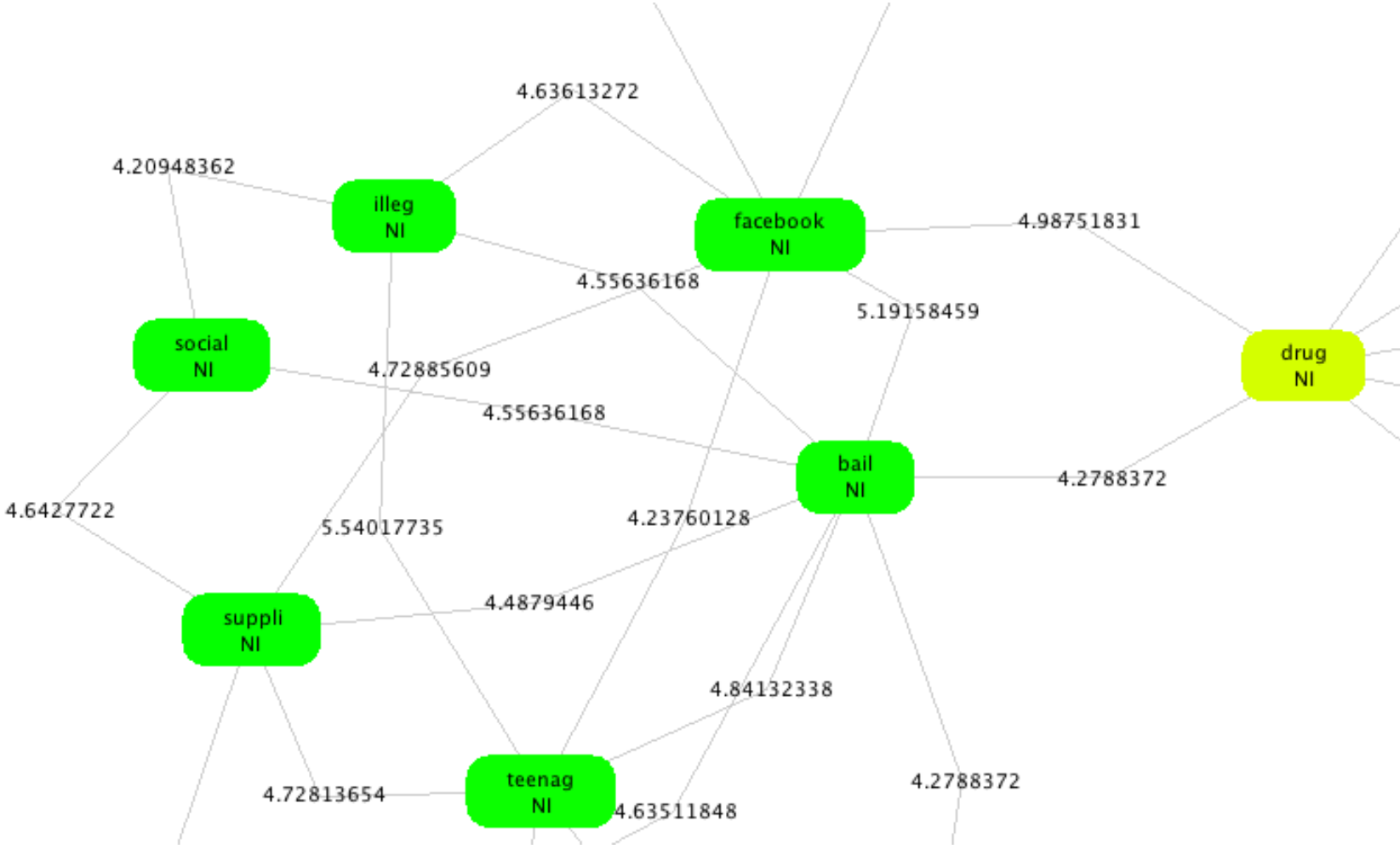
- Currently crawling 7 news sources
- The corpus size is ≈ 65000 with $\approx 47 \cdot 10^6$ term pairs
- Incremental implementation

Goals for a News Surveillance System

- What is really new in a news story?
- Create a summary of a news story
- Decide in a glance whether the news story provides me anything
- Find related news stories

What is new?

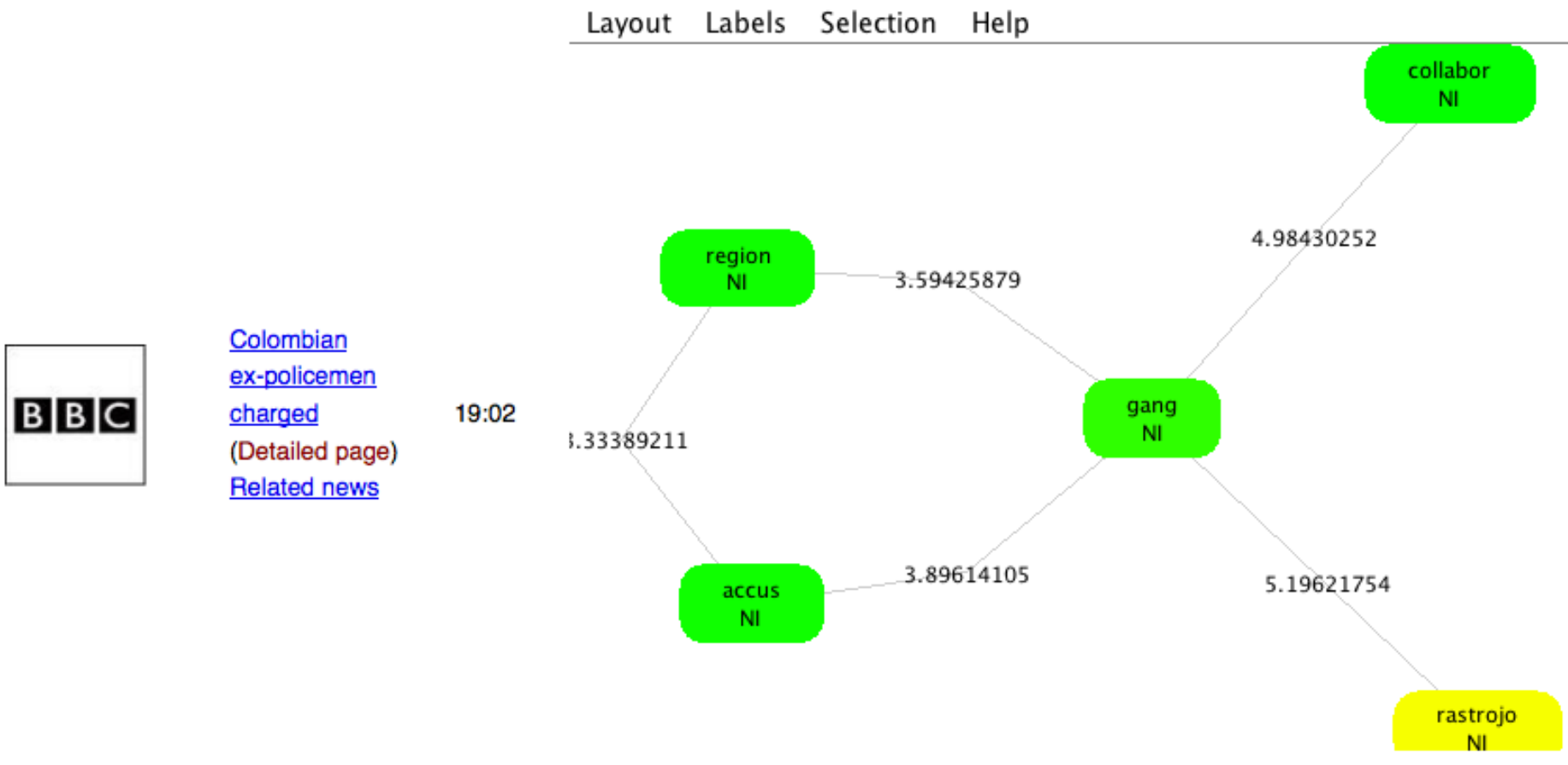
- Sample from a news story which was published yesterday



Summary Generation

- For the sake of clarity, the summary is copy-pasted
- Generated by using the highest scoring term pairs and taking out the sentences from news story
  Northamptonshire Police seized computer equipment, drugs paraphernalia and mobile phones during the arrest of the 17-year-old from Corby. A teenager has been released on bail after being questioned by police about the supply of illegal drugs via the Facebook social media website.
- Randomly generated summary
  Police said a Facebook page, which had more than 200 friends, was shut down. Officers said they would be taking part in activities in schools to promote internet safety.

Glance on a News Story



Related news story published on February 6

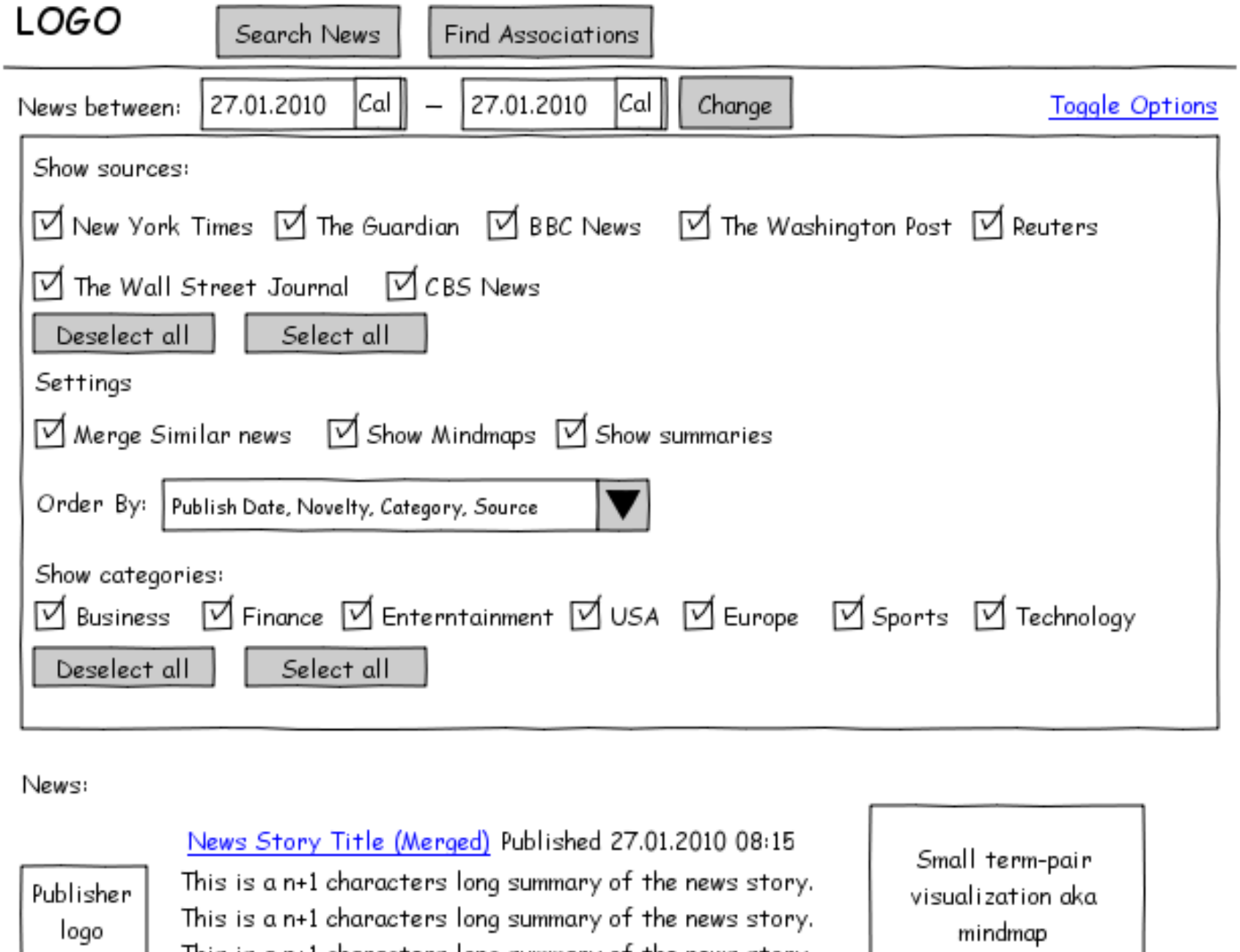
- Story headline "Shake-up in Egyptian ruling party"

Mubarak moves to restart economy (Detailed page)	Sat, 05 Feb, 2011 09:02	229
Cairo protest thousands hold firm (Detailed page)	Wed, 02 Feb, 2011 23:02	14
Egypt braces for 'day of departure' rallies (Detailed page)	Fri, 04 Feb, 2011 07:02	13
Protesters reject Mubarak speech (Detailed page)	Wed, 02 Feb, 2011 05:02	12

Future Work

- Create intuitive and functional GUI
- Merging news stories
- We are still looking for a method for validating if any of this makes any sense
- Something like on the next slide

Usable News Surveillance System



Computational Creativity & Novelty

- One way for creating background associations of a domain
- Considering two backgrounds graphs from different domains
  - Find an interesting association
  - Translate through high abstraction to another
  - Propose new "creative" connection in the other domain
- The background graph can also be used for novelty detection

Background Generation

- Extract keywords with $tf - idf$ algorithm
- Extract term pairs using log likelihood or $tpf - idf$ measure
- Take n top keywords and add them as nodes to graph G
- Take m term pairs and add them to the graph G
- If we have many components in G
  - Connect components using Wordnet Synsets or extracted term pairs

The end

Questions?

It's amazing that the amount of news that happens in the world every day always just exactly fits the newspaper.

Jerry Seinfeld