Terms in Time and Times in Context: A Graph-based Term-Time Ranking Model

Andreas Spitz, Jannik Strötgen, Thomas Bögel and Michael Gertz

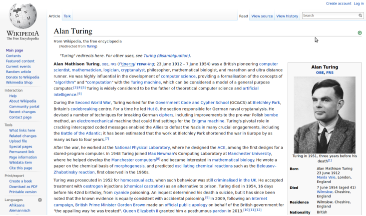
Heidelberg University
Institute of Computer Science
Database Systems Research Group
http://dbs.ifi.uni-heidelberg.de

spitz@informatik.uni-heidelberg.de

5th Temporal Web Analytics Workshop
Florence, May 18, 2015

What happened on June 15, 1215?

A simple question.
How simple is the answer?



With structured data:
quite simple

Based on unstructured text data:
much more challenging

Data Set and Approach

A corpus of all English Wikipedia articles:

- Only text is considered, no info-boxes
- $3,079,620$ documents with time expressions

Problem statement, given such a corpus:

- Extract and normalize temporal expressions (dates)
- Find key terms that best summarize a given date

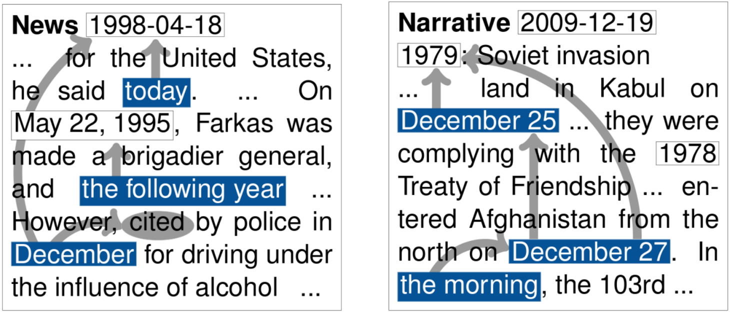
Outline

Outline of the approach:

- Represent date-term co-occurrences efficiently
  - Extract and normalize temporal expressions (dates)
  - Extract content words that co-occur with dates
  - Generate an efficient data structure
- Based on this representation
  - Identify relevant terms for any given date
  - Identify similar dates for any given date
- Example applications

Extraction of Temporal Expressions

- Normalization, e.g., May 18, 2015 → 2015-05-18
- Handling relative temporal expressions, e.g., *in May*
- Considering the document type



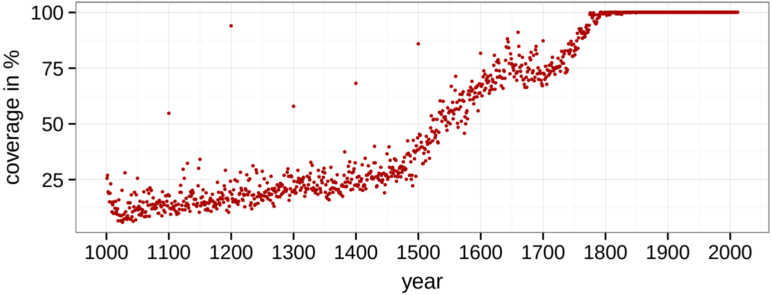
Source: Strötgen, Gertz *Multilingual and Cross-domain Temporal Tagging* (2013)

Coverage of Dates

We use a combination of dates of three granularities:

- YYYY-MM-DD (day)
- YYYY-MM (month)
- YYYY (year)

Percentage of dates that are included in the data per year



Extraction of Terms and Representation

For all sentences s in any Wikipedia document:

The Demolition of the Berlin Wall officially began on 13 June 1990.

Extraction of Terms and Representation

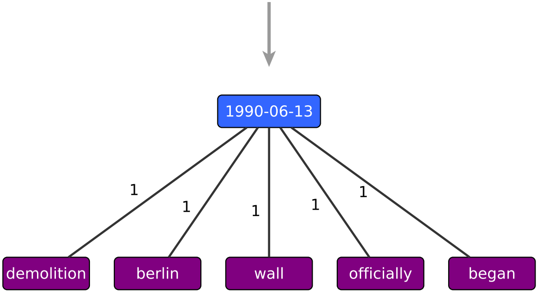
Identify/normalize dates and remove stop words

The $\boxed{\text{Demolition}}$ of the $\boxed{\text{Berlin}}$ $\boxed{\text{Wall}}$ $\boxed{\text{officially}}$ $\boxed{\text{began}}$ on $\boxed{\text{13 June 1990}}$.

Extraction of Terms and Representation

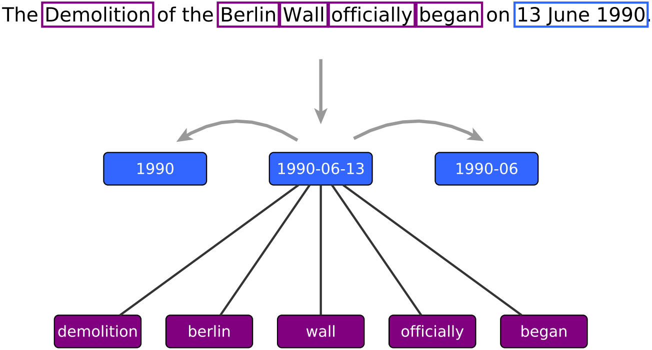
Create a bipartite graph $G_s = (T_s \cup D_s, E_s)$ with weights ω_s

The $\boxed{\text{Demolition}}$ of the $\boxed{\text{Berlin}}$ $\boxed{\text{Wall}}$ $\boxed{\text{officially}}$ $\boxed{\text{began}}$ on $\boxed{\text{13 June 1990}}$.



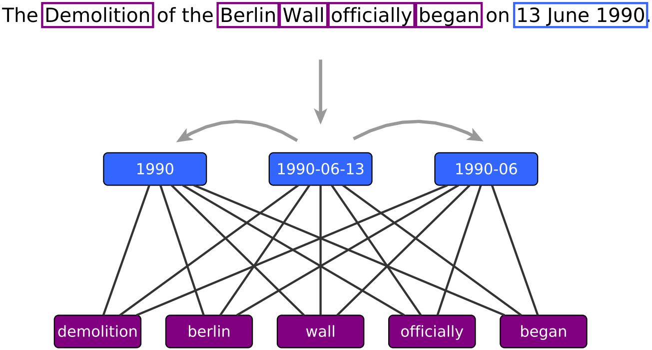
Extraction of Terms and Representation

Satisfy the inclusion condition for dates

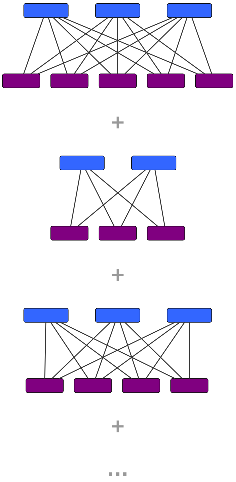


Extraction of Terms and Representation

Satisfy the inclusion condition for dates



Graph aggregation



Aggregate the sentence-graphs G_s:

- $T := \bigcup T_s$
- $D := \bigcup D_s$
- $E := \bigcup E_s$
- $\omega(e) := \sum \omega_s(e)$

We obtain $G = (T \cup D, E, \omega)$ with:

- $|T| = 3,748,730$ terms
- $|D| = 210,375$ dates
- $|E| = 110,639,525$ edges

Formalising the Question

What happened on June 15, 1215?

↓

Which terms in the graph co-occur in a significant manner with the date 1215-06-15?

Ranking

We need a ranking-function from dates D to a list of terms T

- $r : D \rightarrow \mathbb{R}^{|T|}$
- $r(d) :=$ ranking of terms $t \in T$ by their significance for d

Ranking

We need a ranking-function from dates D to a list of terms T

- $r : D \to \mathbb{R}^{|T|}$
- $r(d) :=$ ranking of terms $t \in T$ by their significance for d

Idea: adapt *tf-idf* to the bipartite graph

$$tf\text{-}idf := tf \cdot \log \frac{1}{df}$$

- *tf*: frequency of term in document
- *df*: fraction of documents that contain the term

Adapting tf-idf

How does this relate to the graph?

- Identify dates with documents, i.e., dates *contain* terms
- Term frequency given by edge weights:
  $tf(d,t) \approx \omega(d,t)$
- Inverse document frequency given by number of neighbours:
  $idf(t) \approx \frac{|D|}{deg(t)}$

$$tf\text{-}idf := tf \cdot \log \frac{1}{df} \qquad \Rightarrow \qquad tf\text{-}idf(d,t) := \omega(d,t) \log \frac{|D|}{deg(t)}$$

June 15, 1215

Query: "1215-06-15"

	tf–idf	ω	*deg(t)*
carta	79.7	14	709
magna	71.2	14	1298
barons	46.9	10	1928
runnymede	40.5	6	247
king	20.4	12	38400
oaths	17.1	3	714
king's	15.1	5	10200
repudiation	13.6	2	231
fealty	12.4	2	424
john	11.8	11	71893

June 15, 1215

Query: "1215-06-15"

	tf–idf	ω	*deg(t)*
carta	79.7	14	709
magna	71.2	14	1298
barons	46.9	10	1928
runnymede	40.5	6	247
king	20.4	12	38400
oaths	17.1	3	714
king’s	15.1	5	10200
repudiation	13.6	2	231
fealty	12.4	2	424
john	11.8	11	71893

On June 15, 1215 at Runnymede, King John of England
and a council of rebellious barons agreed to the Magna Carta.

A Ranking for Dates

Ranking dates by term works analogously:

$$tf\text{-}idf(t,d) := \omega(t,d) \log \frac{|T|}{deg(d)}$$

Query: "Tsunami"

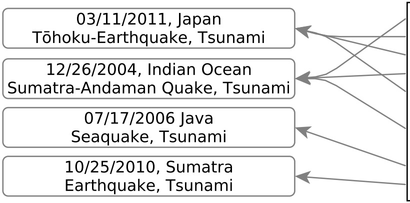
	tf–idf	ω	*deg(t)*
2004	3097.2	1374	393475
2011	2753.9	1313	460264
2011-03	1878.5	464	65407
2004-12-26	1658.0	238	3536
2011-03-11	1474.2	226	5508
2005	1030.6	476	430107
2004-12	734.8	162	40186
2005-01	465.5	102	39062
2006	301.7	147	481555
2010	295.2	148	510254

A Ranking for Dates

Ranking dates by term works analogously:

$$tf\text{-}idf(t,d) := \omega(t,d) \log \frac{|T|}{deg(d)}$$

Query: "Tsunami"



	tf–idf	*ω*	*deg(t)*
2004	3097.2	1374	393475
2011	2753.9	1313	460264
2011-03	1878.5	464	65407
2004-12-26	1658.0	238	3536
2011-03-11	1474.2	226	5508
2005	1030.6	476	430107
2004-12	734.8	162	40186
2005-01	465.5	102	39062
2006	301.7	147	481555
2010	295.2	148	510254

Ranking Nodes by Similarity Within a Set

Can we...

- ... create a ranking for dates by dates?
- ... or for terms by terms?

Ranking Nodes by Similarity Within a Set

Can we...

- ... create a ranking for dates by dates?
- ... or for terms by terms?

Formally this is a *one-mode projection* of the bipartite graph:

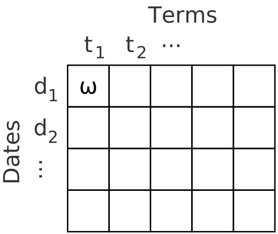
- Reduce graph to a single set of nodes T or D
- Connect nodes that share neighbours in the bipartite graph
- This results in a very dense graph

$\Rightarrow$ How can we identify relevant edges in the projection?

Cosine Similarity of Adjacency Vectors

In a lesson from *Collaborative Filtering*:
use a cosine similarity of adjacency vectors

$$cos(t_a, t_b) := \frac{\sum t_{a_i} \cdot t_{b_i}}{\sqrt{\sum t_{a_i}^2} \cdot \sqrt{\sum t_{b_i}^2}}$$



Evaluation

Ground truth: *U.S. Election Days (1848 - 2013)*

- Recurs annually
- Always on Tuesday after the first Monday in November (Nov 2 - Nov 8)
- Every four years: presidential election

Evaluation

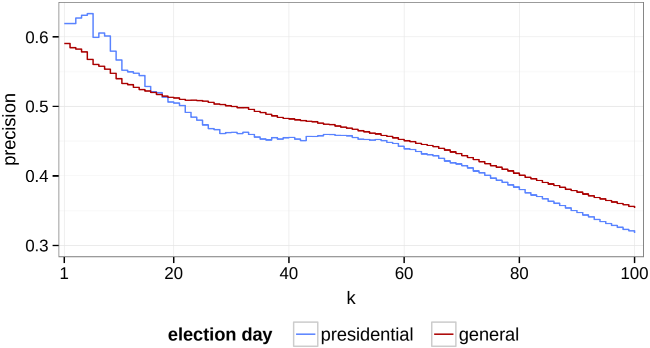
Ground truth: *U.S. Election Days (1848 - 2013)*

- Recurs annually
- Always on Tuesday after the first Monday in November (Nov 2 - Nov 8)
- Every four years: presidential election

Expectation:

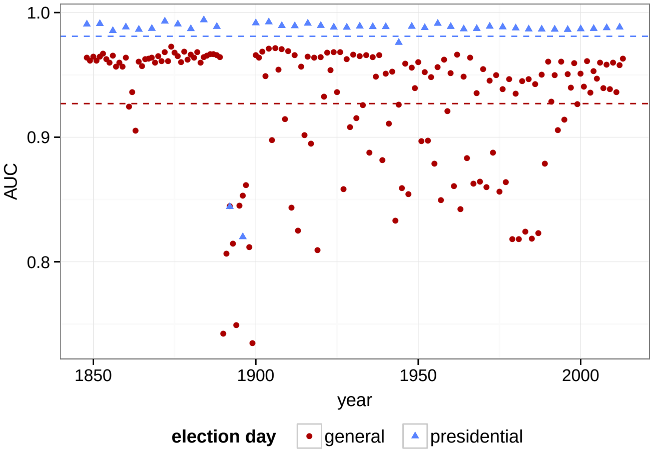
- For a given election day, election days in other years are ranked highly
- For presidential election days, other presidential election days are ranked highly

Precision at k



$$prec_k := \frac{|\text{Election days in top } k \text{ ranks}|}{k}$$

Area Under the ROC Curve



Practical Application: Hot Spots & Key Players

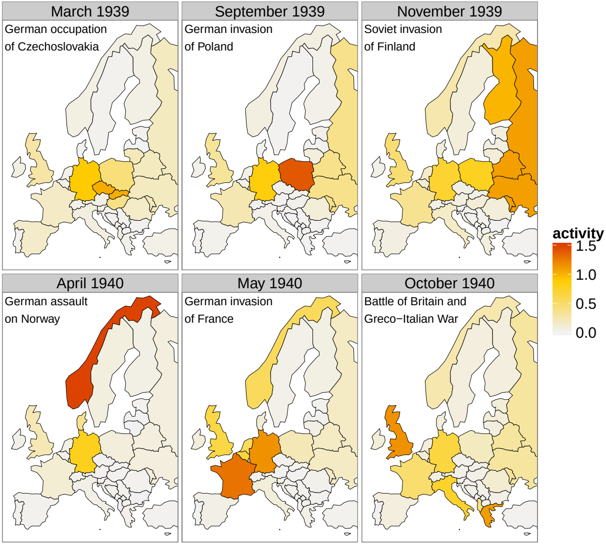
Here: approximation of countries' activity during given months

For each European country c,

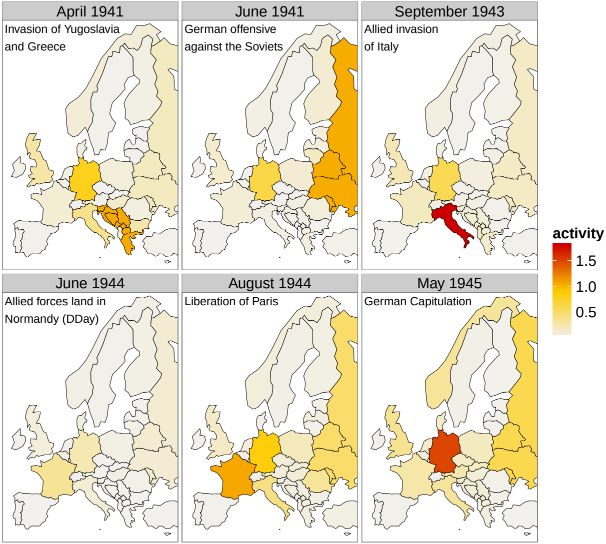
- define its name, e.g. $t_n(c) =$ *italy*,
- define the countries adjectival form, e.g. $t_a(c) =$ *italian*,
- compute individual *tf-idf* scores for terms and combine.

$$act(c,d) := \frac{tf\text{-}idf(d, t_n(c)) + tf\text{-}idf(d, t_a(c))}{\max[tf\text{-}idf(d, \cdot)]}$$

Activity by Country During World War II



Activity by Country During World War II (2)



Summary

Approach:

- Extract dates and terms from unstructured text
- Construct a bipartite date-term graph
- Allows ranking dates / terms according to co-occurrences

Benefits:

- Simple measures already yield good results
- Efficient: 4GB Memory and real-time queries
- Flexibility of ranking methods

Ongoing Work

Query: "2016-05"

	ω
Multi-partite graphs: Dates Persons Locations	1
Terms as n-Grams	2
Ranking-Functions	3

Thank you!

Questions?