

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

Alan Turing

Alan Mathison Turing, FRS (23 June 1912 – 7 June 1954) was a British pioneer in computer science, mathematical logic, cryptanalysis, philosophical, mathematical biology, and mathematics and its education. He was highly influential in the development of computer science, providing a formalisation of the concepts of "algorithm" and "computation" with the Turing machine, which can be considered a model of a general-purpose computer. Turing is widely considered to be the father of theoretical computer science and artificial intelligence.

During the Second World War, Turing worked for the Government Code and Cipher School (GC&CS) at Bletchley Park, Britain's communications centre. For a time he helped to develop techniques for German naval cryptanalysts, he devised a number of techniques for breaking German rotor machines, including improvements to the pre-war Polish bombe method, an electromechanical machine that could find settings for the Enigma machine. Turing is credited with cracking intercepted coded messages that enabled the Allies to defeat the Axis in many crucial engagements, including the Battle of the Atlantic. It has been estimated that the work of Bletchley Park shortened the war in Europe by as much as two to four years.^[c]

After the war, he worked at the National Physical Laboratory, where he designed the ACE, among the first designs for a stored-program computer. In 1948 Turing joined Max Newman's Computing Laboratory at Manchester University, where he helped develop the Manchester computer *ENIAC* and became interested in mathematical biology. He wrote a paper on the chemical basis of morphogenesis, and predicted oscillating chemical reactions such as the Belousov-Zhabotinsky reaction, first observed in the 1950s.

Turing was prosecuted in 1952 for homosexual acts, when such behaviour was still criminalised in the UK. He accepted treatment with hormone injections to chemical castration as an alternative to prison. Turing died in 1954, 42 days before his 43rd birthday. There is a public inquiry. An inquest determined his death a suicide, but it has since been noted that the inquest evidence is in many respects contradictory.^[d] In 2009, following an internet campaign, British Prime Minister Gordon Brown made an official public apology on behalf of the British government for "the appalling way he was treated". Queen Elizabeth II granted him a posthumous pardon in 2013.^[e]^[f]^[g]

Born	Alan Mathison Turing 23 June 1912 Woolwich, London, England
Died	7 June 1954 (aged 42) Milton Keynes, England
Residence	Milton Keynes, England
Nationality	British

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

News 1998-04-18
 ... for the United States,
 he said **today**. ... On
May 22, 1995, Farkas was
 made a brigadier general,
 and **the following year** ...
 However, cited by police in
December for driving under
 the influence of alcohol ...

Narrative 2009-12-19
1979: Soviet invasion
 ... land in Kabul on
December 25 ... they were
 complying with the **1978**
 Treaty of Friendship ... en-
 tered Afghanistan from the
 north on **December 27**. In
the morning, the 103rd ...

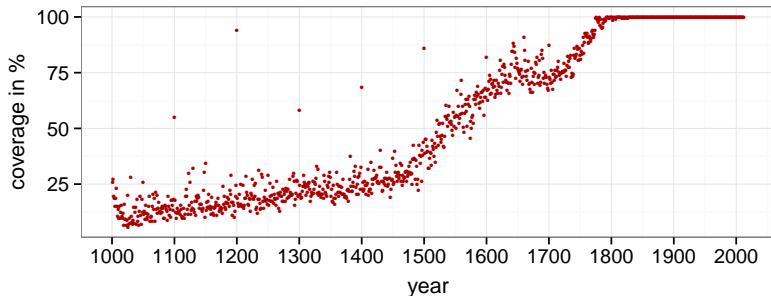
Source: Strötgen, Gertz *Multilingual and Cross-domain Temporal Tagging* (2013)

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

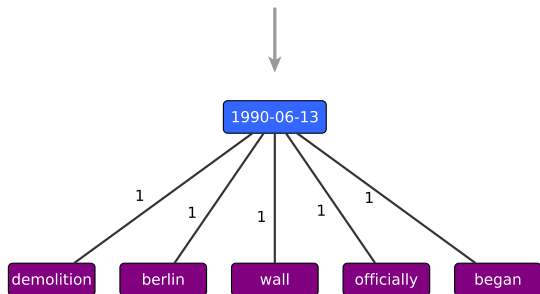
Identify/normalize dates and remove stop words

The **Demolition** of the **Berlin Wall** **officially began** on **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

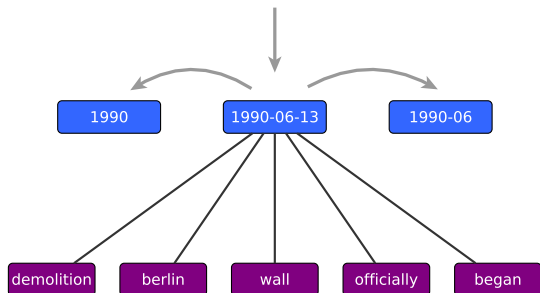
The **Demolition** of the **Berlin Wall** **officially** **began** on **13 June 1990**.



Extraction of Terms and Representation

Satisfy the inclusion condition for dates

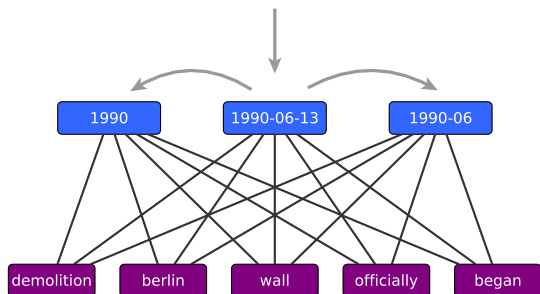
The **Demolition** of the **Berlin Wall** **officially** **began** on **13 June 1990**.



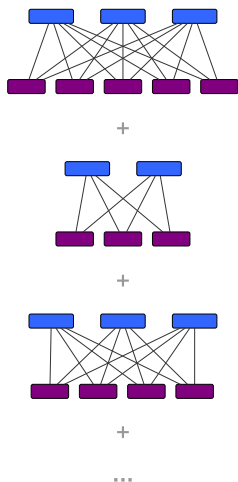
Extraction of Terms and Representation

Satisfy the inclusion condition for dates

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



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

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

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

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

- *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:
 $idf(t) \approx \frac{|D|}{deg(t)}$

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

June 15, 1215

Query: "1215-06-15"

	<i>tf-idf</i>	ω	<i>deg(t)</i>
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"

	<i>tf-idf</i>	ω	<i>deg(t)</i>
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:

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

Query: "Tsunami"

	<i>tf-idf</i>	ω	<i>deg(t)</i>
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:

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

Query: "Tsunami"

	<i>tf-idf</i>	ω	<i>deg(t)</i>	
03/11/2011, Japan Tōhoku-Earthquake, Tsunami	2004	3097.2	1374	393475
	2011	2753.9	1313	460264
	2011-03	1878.5	464	65407
12/26/2004, Indian Ocean Sumatra-Andaman Quake, Tsunami	2004-12-26	1658.0	238	3536
	2011-03-11	1474.2	226	5508
07/17/2006 Java Seaquake, Tsunami	2005	1030.6	476	430107
	2004-12	734.8	162	40186
	2005-01	465.5	102	39062
10/25/2010, Sumatra Earthquake, Tsunami	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

⇒ 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

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

		Terms			
		t_1	t_2	...	
Dates	d_1	ω			
	d_2				
	\vdots				

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

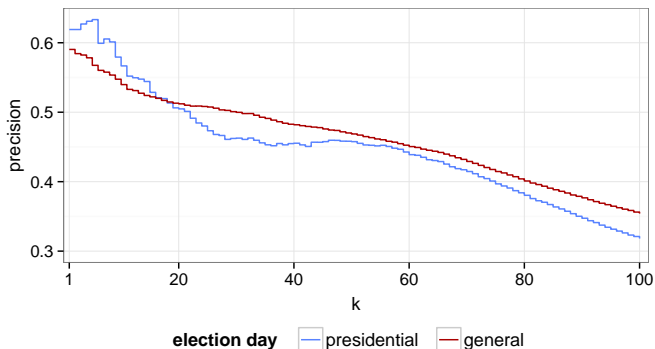
Ground truth: *U.S. Election Days (1848 - 2013)*

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

Expectation:

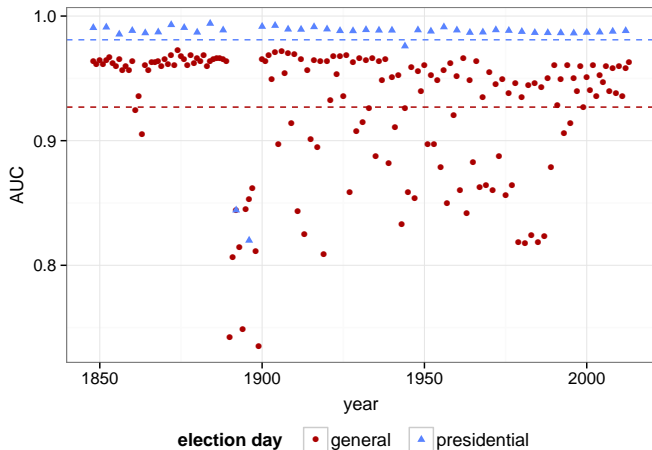
- For a given election day, election days in other years are ranked highly
- For presidential election days, other presidential election days are ranked highly

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

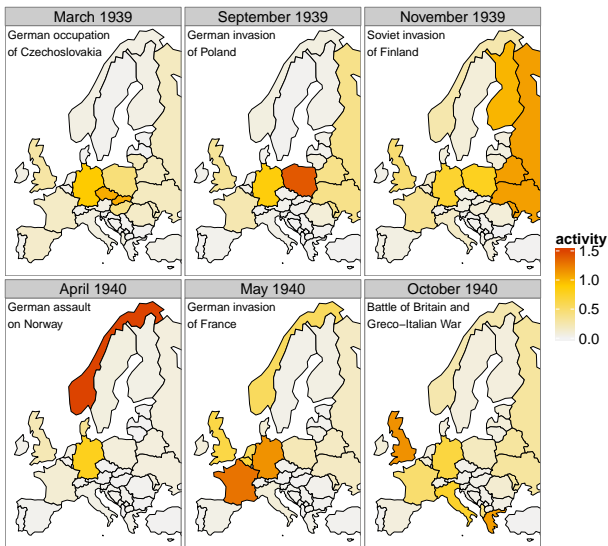
Here: approximation of countries' activity during given months

For each European country c ,

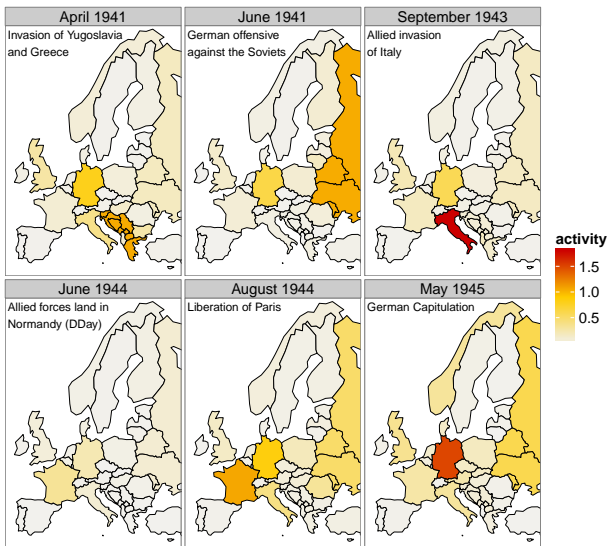
- define its name, e.g. $t_n(c) = \textit{italy}$,
- define the countries adjectival form, e.g. $t_a(c) = \textit{italian}$,
- compute individual *tf-idf* scores for terms and combine.

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

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"

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

Thank you!

Questions?