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?
Motivation

Co-occurrence Graphs

Term-Ranking

Projection

Application

Summary

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 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 Demolition of the Berlin Wall officially began on 13 June 1990.
Create a bipartite graph $G_s = (T_s \cup D_s, E_s)$ with weights $\omega_s$.

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

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

\[
\text{tf-idf} := tf \cdot \log \frac{1}{df} \quad \Rightarrow \quad \text{tf-idf}(d, t) := \omega(d, t) \log \frac{|D|}{\deg(t)}
\]
## Query: "1215-06-15"

<table>
<thead>
<tr>
<th>Term</th>
<th>tf-idf</th>
<th>$\omega$</th>
<th>$deg(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>carta</td>
<td>79.7</td>
<td>14</td>
<td>709</td>
</tr>
<tr>
<td>magna</td>
<td>71.2</td>
<td>14</td>
<td>1298</td>
</tr>
<tr>
<td>barons</td>
<td>46.9</td>
<td>10</td>
<td>1928</td>
</tr>
<tr>
<td>runnymede</td>
<td>40.5</td>
<td>6</td>
<td>247</td>
</tr>
<tr>
<td>king</td>
<td>20.4</td>
<td>12</td>
<td>38400</td>
</tr>
<tr>
<td>oaths</td>
<td>17.1</td>
<td>3</td>
<td>714</td>
</tr>
<tr>
<td>king’s</td>
<td>15.1</td>
<td>5</td>
<td>10200</td>
</tr>
<tr>
<td>repudiation</td>
<td>13.6</td>
<td>2</td>
<td>231</td>
</tr>
<tr>
<td>fealty</td>
<td>12.4</td>
<td>2</td>
<td>424</td>
</tr>
<tr>
<td>john</td>
<td>11.8</td>
<td>11</td>
<td>71893</td>
</tr>
</tbody>
</table>
**Query: "1215-06-15"**

<table>
<thead>
<tr>
<th>Term</th>
<th>$tf-idf$</th>
<th>$\omega$</th>
<th>$deg(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>carta</td>
<td>79.7</td>
<td>14</td>
<td>709</td>
</tr>
<tr>
<td>magna</td>
<td>71.2</td>
<td>14</td>
<td>1298</td>
</tr>
<tr>
<td>barons</td>
<td>46.9</td>
<td>10</td>
<td>1928</td>
</tr>
<tr>
<td>runnymede</td>
<td>40.5</td>
<td>6</td>
<td>247</td>
</tr>
<tr>
<td>king</td>
<td>20.4</td>
<td>12</td>
<td>38400</td>
</tr>
<tr>
<td>oaths</td>
<td>17.1</td>
<td>3</td>
<td>714</td>
</tr>
<tr>
<td>king’s</td>
<td>15.1</td>
<td>5</td>
<td>10200</td>
</tr>
<tr>
<td>repudiation</td>
<td>13.6</td>
<td>2</td>
<td>231</td>
</tr>
<tr>
<td>fealty</td>
<td>12.4</td>
<td>2</td>
<td>424</td>
</tr>
<tr>
<td>john</td>
<td>11.8</td>
<td>11</td>
<td>71893</td>
</tr>
</tbody>
</table>

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:

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

<table>
<thead>
<tr>
<th>Year</th>
<th>$tf-idf$</th>
<th>$\omega$</th>
<th>$deg(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>3097.2</td>
<td>1374</td>
<td>393475</td>
</tr>
<tr>
<td>2011</td>
<td>2753.9</td>
<td>1313</td>
<td>460264</td>
</tr>
<tr>
<td>2011-03</td>
<td>1878.5</td>
<td>464</td>
<td>65407</td>
</tr>
<tr>
<td>2004-12-26</td>
<td>1658.0</td>
<td>238</td>
<td>3536</td>
</tr>
<tr>
<td>2011-03-11</td>
<td>1474.2</td>
<td>226</td>
<td>5508</td>
</tr>
<tr>
<td>2005</td>
<td>1030.6</td>
<td>476</td>
<td>430107</td>
</tr>
<tr>
<td>2004-12</td>
<td>734.8</td>
<td>162</td>
<td>40186</td>
</tr>
<tr>
<td>2005-01</td>
<td>465.5</td>
<td>102</td>
<td>39062</td>
</tr>
<tr>
<td>2006</td>
<td>301.7</td>
<td>147</td>
<td>481555</td>
</tr>
<tr>
<td>2010</td>
<td>295.2</td>
<td>148</td>
<td>510254</td>
</tr>
</tbody>
</table>
A Ranking for Dates

Ranking dates by term works analogously:

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

**Query: "Tsunami"**

<table>
<thead>
<tr>
<th>Date</th>
<th>tf-idf</th>
<th>(\omega)</th>
<th>deg(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>3097.2</td>
<td>1374</td>
<td>393475</td>
</tr>
<tr>
<td>2011</td>
<td>2753.9</td>
<td>1313</td>
<td>460264</td>
</tr>
<tr>
<td>2011-03</td>
<td>1878.5</td>
<td>464</td>
<td>65407</td>
</tr>
<tr>
<td>2004-12-26</td>
<td>1658.0</td>
<td>238</td>
<td>3536</td>
</tr>
<tr>
<td>2011-03-11</td>
<td>1474.2</td>
<td>226</td>
<td>5508</td>
</tr>
<tr>
<td>2005</td>
<td>1030.6</td>
<td>476</td>
<td>430107</td>
</tr>
<tr>
<td>2004-12</td>
<td>734.8</td>
<td>162</td>
<td>40186</td>
</tr>
<tr>
<td>2005-01</td>
<td>465.5</td>
<td>102</td>
<td>39062</td>
</tr>
<tr>
<td>2006</td>
<td>301.7</td>
<td>147</td>
<td>481555</td>
</tr>
<tr>
<td>2010</td>
<td>295.2</td>
<td>148</td>
<td>510254</td>
</tr>
</tbody>
</table>

**Dates and Events:**
- 03/11/2011, Japan Tōhoku-Earthquake, Tsunami
- 12/26/2004, Indian Ocean Sumatra-Andaman Quake, Tsunami
- 07/17/2006 Java Seaquake, Tsunami
- 10/25/2010, Sumatra Earthquake, Tsunami
Can we...

- ... create a ranking for dates by dates?
- ... or for terms by terms?
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?
In a lesson from *Collaborative Filtering*: use a cosine similarity of adjacency vectors

$$cos(t_a, t_b) := \frac{\sum t_{ai} \cdot t_{bi}}{\sqrt{\sum t_{ai}^2} \cdot \sqrt{\sum t_{bi}^2}}$$
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 $k$

\[ \text{prec}_k := \frac{|\text{Election days in top } k \text{ ranks}|}{k} \]
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.

$$act(c, d) := \frac{tf-idf(d, t_n(c)) + tf-idf(d, t_a(c))}{\max[tf-idf(d, \cdot)]}$$
Activity by Country During World War II

March 1939: German occupation of Czechoslovakia
September 1939: German invasion of Poland
November 1939: Soviet invasion of Finland

April 1940: German assault on Norway
May 1940: German invasion of France
October 1940: Battle of Britain and Greco–Italian War

Andreas Spitz
Activity by Country During World War II (2)

- April 1941: Invasion of Yugoslavia and Greece
- June 1941: German offensive against the Soviets
- September 1943: Allied invasion of Italy
- June 1944: Allied forces land in Normandy (D-Day)
- August 1944: Liberation of Paris
- May 1945: German Capitulation
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
**Query: "2016-05"**

<table>
<thead>
<tr>
<th>Multi-partite graphs:</th>
<th>$\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dates</td>
<td>1</td>
</tr>
<tr>
<td>Persons</td>
<td></td>
</tr>
<tr>
<td>Locations</td>
<td></td>
</tr>
<tr>
<td>Terms as n-Grams</td>
<td>2</td>
</tr>
<tr>
<td>Ranking-Functions</td>
<td>3</td>
</tr>
</tbody>
</table>
Thank you!

Questions?