Identifying Events from Co-Occurrences and Context across Large Document Collections

Andreas Spitz
Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

**What describes an event?**

Textual representations and mentions of events have a spatial and temporal component and involve a set of actors.

**First Approach: Temporal Profiles**

With the English Wikipedia as corpus:
- Extract terms (content words)
- Extract dates at granularity levels day, month and year

Concept: Use term-time co-occurrences to construct temporal term profiles and contextual date profiles to extract information from the text [1].

**Term-Time Co-occurrences**

Create bipartite sentence-graphs:

- The Demolition of the Berlin Wall officially began on [23 June 1989].

Combine into one graph representation $G = (T \cup D, E, \omega)$ with $\omega : E \rightarrow \mathbb{N}$:
- $|T| = 3,748,730$ terms
- $|D| = 210,375$ dates
- $|E| = 110,639,525$ edges

**Evaluation**

**Ranking for U.S. Election Days**
- Annually, varies between Nov 2 - Nov 8
- Presidential election every 4 years
- Idea: similar dates to Election Days are also Election Days (in different years)

**Outlook: The LOAD Model**

**Extension to a Multi-Party Graph**
- Include spatial information
- Include persons
- Include organizations
- Include relationships between terms

**Identifying Events:**
- Based on incomplete information
- Without need for disambiguation

**Significance- & Similarity-Measures**

**Information Retrieval**

Leverage the graph representation to identify significant co-occurrences of dates and terms in the corpus and structural equivalence. Therefore, introduce ranking functions between sets of nodes:

$$r_{XY} : X \rightarrow \mathbb{R}^{|Y|}$$

where $X, Y \in \{D, T\}$.

**Homogeneous ranking ($X = Y$):**
- Similarity within the same node set
- No direct links available
- Average similarity between links sets
- Use for example a cosine similarity of adjacency vectors:

$$cos(t_w, t_b) := \frac{\sum t_w \cdot t_b}{\sqrt{\sum t_w^2 \cdot \sum t_b^2}}$$

**Examples of rankings:**

<table>
<thead>
<tr>
<th>Query: &quot;2319-09-15&quot;</th>
<th>Query: &quot;1945&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>june, 15</td>
<td>june 15</td>
</tr>
<tr>
<td>june, 1945</td>
<td>june 1945</td>
</tr>
<tr>
<td>june 15, 1945</td>
<td>june 15, 1945</td>
</tr>
</tbody>
</table>

**Outlook: Weight decay by distance**

Beyond a bag of words:
- Consider co-occurrences in the entire document, not just in sentences
- Use a weight function for edges that decays with distance
- Aggregate individual co-occurrences
- When applied to the construction of a social network from person mentions on Wikipedia: results in relationships with natural age profiles [2].

**References**


This work was presented at the 3rd Heidelberg Laureate Forum, 23.08 – 28.08, 2015, Heidelberg, Germany.

Contact Information:
Andreas Spitz
spitz@informatik.uni-heidelberg.de
http://dbs.ifi.uni-heidelberg.de/