Temporal Information Access (Temporalia-2)

NTCIR-12 Core Task
http://ntcirtemporalia.github.io

Hideo Joho (University of Tsukuba)
Adam Jatowt (Kyoto University)
Roi Blanco (Yahoo! Research)
Haitao Yu (University of Tsukuba)
Shuhei Yamamoto (University of Tsukuba)
Motivation: Temporal IR

- 1.5% of queries are **explicit temporal queries**
  - *Germany 1920s, Olympics 1956*

- Many queries are **implicitly temporal**, e.g.:
  - *einstein life, fashion trend, tokyo weather, olympics*

- Temporal queries **require special treatment**
  - By matching query temporality with document temporality

---

Arikan, I., Bedathur, S.J. and Berberich, K. Time Will Tell: Leveraging Temporal Expressions in IR. WSDM2009
Berberich et al. Language Modeling Approach for Temporal Information Needs. ECIR 2010
Kanhabua, N. and Norvag,K., Determining Time of Queries for Reranking Search Results. ECDL2010
Nunes, S., Ribeiro, C., and David, G. Use of Temporal Expressions in Web Search. ECIR2008
Motivation: Importance of Time in IR

• Example results that do not match query temporal intent:
  • *tokyo weather* → doc with yesterday’s results (?)
  • *economic forecast japan* → doc with past forecasts (?)
  • *previous tokyo olympics* → doc about Olympics 2020 (?)
(PAST) Temporalia-1: English Lang.

1. **Temporal Query Intent Classification (TQIC)**
   - Classify query to temporal classes

2. **Temporal Information Retrieval (TIR)**
   - Rank documents for temporal queries

<table>
<thead>
<tr>
<th>Team ID</th>
<th>Institution</th>
<th>Country</th>
<th>TQIC</th>
<th>TIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HITSZ</td>
<td>Graduate School of Harbin Institute of Technology at Shenzhen</td>
<td>P.R.C.</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>HULTECH</td>
<td>University of Caen</td>
<td>France</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>TUTA1</td>
<td>University of Tokushima</td>
<td>Japan</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Andd7</td>
<td>Dhirubhai Ambani Institute of Information and Communication Technology</td>
<td>India</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>MPII</td>
<td>Max Planck Institute for Informatics</td>
<td>Germany</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>UniMAN</td>
<td>University of Manchester</td>
<td>UK</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>BRKLY</td>
<td>U.C. Berkeley</td>
<td>USA</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>OSKAT</td>
<td>Osaka Kyoiku University</td>
<td>Japan</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>ORG</td>
<td>Temporalia Organizers</td>
<td>Japan, Spain</td>
<td>☑</td>
<td></td>
</tr>
</tbody>
</table>

1. **Temporal Intent Disambiguation (TID)**
   - For a given query, determine probability distribution over temporal intent classes
     - **Input:** query, query timestamp
     - **Output:** temporal intent class distribution

2. **Temporally Diversified Retrieval (TIR)**
   - For a given topic, retrieve documents relevant to each of temporal intent classes
     - **Input:** set of topic descriptions, timestamp and fixed temporal intent class distribution
     - **Output:** diversified ranking of documents
Temporal Classes of Queries/Topics

- **Recency**: On recent things/events
- **Past**: On past things/events
- **Future**: On future things/events
- **Atemporal**: No temporal intent or navigational
English Document Collection

- LivingKnowledge Corpus [http://livingknowledge.europarchive.org/]
  - Crawl time: April 2011 and March 2013
  - Source: news and blog feeds
  - Size: 20GB uncompressed (>5GB zipped)
    - 3.8 million documents from 1500 different blogs and news sources
    - Text only

- Provided annotations: Temporal Expressions, Named Entities and Sentence Boundaries
  - Participants can choose to use annotations or not
  - Participants can use external collections such as Wikipedia
Chinese Document Collection

• **Dry Run**: SogouCA, a collection of news web pages
  • **Time span**: June and July in 2012
  • **Source**: several popular news websites in China
  • **Size**: the total uncompressed size is 711MB

• **Formal Run**: Distilled subset of the web page corpus SogouT 2012
  • **Crawl time**: mid 2012
  • **Size**: subset of 135.4 million Web pages from 5.3 million Chinese Web sites

• **Provided annotations**:
  • Similar as in the English collection
# Draft Schedule

## Important Dates (Tentative)

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb 28, 2015</td>
<td>NTCIR-12 Kick-off Event</td>
</tr>
<tr>
<td>Jun 2015</td>
<td>Final task guideline release. Task Registration due. Document Collection release</td>
</tr>
<tr>
<td>Jun 2015</td>
<td>Dry Run: Test topics release</td>
</tr>
<tr>
<td>Aug 2015</td>
<td>Dry Run: Run submission due</td>
</tr>
<tr>
<td>Sep 2015</td>
<td>Dry Run: Evaluation results returned</td>
</tr>
<tr>
<td>Oct 2015</td>
<td>Formal Run: Test topics release</td>
</tr>
<tr>
<td>Dec 2015</td>
<td>Formal Run: Run submission due</td>
</tr>
<tr>
<td>Feb 2016</td>
<td>Formal Run: Evaluation results returned</td>
</tr>
<tr>
<td>Mar 2016</td>
<td>Paper submission due</td>
</tr>
<tr>
<td>May 2016</td>
<td>Camera-ready submission due</td>
</tr>
<tr>
<td>Jun 2016</td>
<td>NTCIR-12 Conference</td>
</tr>
</tbody>
</table>
Thank you!

Website: http://ntcirtemporalia.github.io
Facebook: https://www.facebook.com/ntcirtemporalia
Twitter: https://twitter.com/ntcirtemporalia
Email: tc4fia@googlegroups.com