Exploring semantically-related concepts from Wikipedia: the case of SeRE

Daniel Hienert, Dennis Wegener and Siegfried Schomisch
GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany

International UDC Seminar 2013,
25th October 2013
The Hague, Netherlands
1. Introduction
**Brief overview**

- **Visual Search Engines** like Kartoo, Grooker or MapStan for the presentation of search engine results
  
  Börner & Chen, 2002:
  - Visual interfaces for searching & browsing, showing semantic links -> support exploration
  - Get an overview of the entire document collection (Clustering, Categories)
  - Visualization of user interaction data

- **Visualization of relationships between concepts**: Relfinder, Eyeploker, gFacet, Oobian Insight -> **Concept Explorers**

  - To get an overview of the area and to make comparisons of groups and concepts inside the topic (Eppler & Stoyko, 2009)
  - Showing relationships between concepts -> Browsing between concepts
  - Results can be classified
  - Concept facets can be used for filtering
  - Using different visualization techniques like network graphs, maps, circular design, hierarchical text filtering
Goal

• Create an interactive user interface, that let user search for arbitrary *concepts* in any language
• Related concepts are then computed on the basis of knowledge bases like Wikipedia and DBpedia
• They are shown with thumbnails sorted by semantic relatedness and text snippets describing the *relationship*
2. Computing semantically-related concepts
Steps to compute semantically-related concepts

The user enters a keyword in the search form
Steps to compute semantically-related concepts

*Step 1*: Query the Wikipedia API for an article page with a matching concept
Steps to compute semantically-related concepts

Step 2: Query in/outlinks from Wikipedia and broader/narrower terms, categories from DBpedia
Steps to compute semantically-related concepts

**Step 3:**

- For each concept the semantic relatedness (SR) is computed.
- We use the Normalized Google Distance formula, but take Wikipedia full text search hits, instead of search engine results.
- This approach achieves a Spearman correlation up to 0.729 for human judged datasets and P(20) up to 0.934 for semantic relation datasets within the sim-eval framework.

\[
SR = \frac{\log_{10}(\max(A, B)) - \log_{10}(A \cup B)}{\log_{10}(W) - \log_{10}(\min(A, B))}
\]

*Step 1: Find Matching Wikipedia Article*

*Step 2: Query in/outlinks, related terms*

*Step 3: Compute Semantic Relatedness*

*Step 4: Additional Information*

*List of related concepts*
Steps to compute semantically-related concepts

**Step 4:**
- Query category information, thumbnail and text snippets describing the relation to the search term
- Computing most common category

All these processing steps are computed live, in a parallel manner, with several hundred queries in parallel

-> this allows the implementation in an interactive system
3. User Interface
The German Chancellor Angela Merkel and her connection to Helmut Kohl

www.vizgr.org/sere
4. User Study
User Study

Method: Task-based user test with 9 scientists of computer science. Tasks were first conducted with Google, then with SeRE

Task & Questions:
1. Find five persons who played a major role in the political career of Angela Merkel.
2. Find information about possible relations of Angela Merkel and Jean-Claude Juncker.
3. Cite the five most important banks in the context of the current euro crisis.
## Results

*Table 1: Found answers for Task 1 to 3, A= absolute answers, C=confidence scores (1=very unsure to 5=very sure)*

<table>
<thead>
<tr>
<th>Task</th>
<th>Google</th>
<th>A</th>
<th>C</th>
<th>SeRE</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1: Five important persons that played a major role in the political career of Merkel</strong></td>
<td>1. Helmut Kohl</td>
<td>7</td>
<td>4.57</td>
<td>Christian Wulff</td>
<td>6</td>
<td>3.16</td>
</tr>
<tr>
<td>2. Wolfgang Schäuble</td>
<td>7</td>
<td>4.28</td>
<td>Helmut Kohl (1.)</td>
<td>3</td>
<td>3.33</td>
<td></td>
</tr>
<tr>
<td>3. Lothar de Maizière</td>
<td>5</td>
<td>3.4</td>
<td>Franz Müntefering</td>
<td>3</td>
<td>3.33</td>
<td></td>
</tr>
<tr>
<td>4. Gerhard Schröder</td>
<td>2</td>
<td>4</td>
<td>Nicolas Sarkozy</td>
<td>2</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>5. Edmund Stoiber</td>
<td>2</td>
<td>2</td>
<td>Gerhard Schröder (4.)</td>
<td>2</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td><strong>2: Relations between Merkel and Juncker</strong></td>
<td>Topics referring to euro crisis</td>
<td>5</td>
<td>4.2</td>
<td>Karlspreis</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>Juncker supported Merkel, e.g. in elections</td>
<td>6</td>
<td>4.6</td>
<td>Frankfurter Runde</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Party affiliation</td>
<td>1</td>
<td>4</td>
<td>Christine Lagarde</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hermann van Rompuy</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>José Manuel Barroso</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>3: Five important banks in the euro crisis</strong></td>
<td>1 EZB</td>
<td>5</td>
<td>4.2</td>
<td>EZB (1.)</td>
<td>8</td>
<td>3.9</td>
</tr>
<tr>
<td>2. Lehmann Brothers</td>
<td>3</td>
<td>4.6</td>
<td>Deutsche Bundesbank (4.)</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3. Commerzbank</td>
<td>3</td>
<td>4.3</td>
<td>Lehmann Brothers (2.)</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4. Deutsche Bank</td>
<td>3</td>
<td>4</td>
<td>Banco de Portugal</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5. Goldmann Sachs</td>
<td>2</td>
<td>4</td>
<td>Bank of England</td>
<td>3</td>
<td>2.6</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Task</th>
<th>Google (average, standard deviation)</th>
<th>SeRE (absolute, standard deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Important persons – Merkel (absolute, average)</td>
<td>(40, 4.44)</td>
<td>(39, 4.33)</td>
</tr>
<tr>
<td>Confidence</td>
<td>sure (4.05, 0.93)</td>
<td>normal (3.18, 1.18)</td>
</tr>
<tr>
<td>Difficulty</td>
<td>normal (0.44, 0.73)</td>
<td>normal (-0.44, 1.24)</td>
</tr>
<tr>
<td>2: Relations between Merkel – Juncker (absolute, average)</td>
<td>(25, 2.77)</td>
<td>(18, 2)</td>
</tr>
<tr>
<td>Confidence</td>
<td>sure (4.20, 0.96)</td>
<td>normal (3.44, 1.15)</td>
</tr>
<tr>
<td>Difficulty</td>
<td>normal (0.33, 0.87)</td>
<td>normal (0.00, 1.00)</td>
</tr>
<tr>
<td>3: Important banks in the euro crisis (absolute, average)</td>
<td>(37, 4.11)</td>
<td>(35, 3.88)</td>
</tr>
<tr>
<td>Confidence</td>
<td>normal (3.89, 0.94)</td>
<td>normal (3.46, 1.40)</td>
</tr>
<tr>
<td>Difficulty</td>
<td>normal (-0.67, 0.87)</td>
<td>normal (-0.44, 1.13)</td>
</tr>
<tr>
<td>Final evaluation</td>
<td>normal (-0.33, 1.00)</td>
<td>normal (-0.22, 0.97)</td>
</tr>
<tr>
<td>Sorting of search results by semantic relatedness</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Google
- Broad data basis and different data sources
- One can use search terms in combinations
- Text information presented at a glance
- Snippets could be seen immediately, more extensive information
- No concrete concepts only websites
- A lot of redundancy
- Results could not be filtered according to special categories
- Difficult to search for related entities

SeRE
- No redundancy
- Good presentation of results
- Sorting by semantic relatedness
- Snippets helpful
- Easier to search for related entities
- Only Wikipedia as a search basis
- Snippets too short
- No combination of search terms

Main challenge for concept explorers: Meaningful natural languages relationships between concepts!
Thank you!

Daniel Hienert
GESIS – Leibniz-Institute for the Social Sciences
Unter Sachenhausen 6-8
50667 Cologne
Germany

daniel.hienert@gesis.org

http://www.gesis.org
http://vizgr.org/sere