Connecting Knowledge for A New Kind of Search

Andias Wira Alam
Email: andias.wira-alam@gesis.org
Motivation and Background

Web Search

yesterday

WWW

user

formulating queries

retrieved documents

information needs

WWW

Knowledge Base
(e.g. Wikipedia)

user

enriching information needs

retrieved documents

matching entities

formulating queries

today
Example on Query Answering

Who is the 44th President of the USA?

Retrieved Documents:

- President Barack Obama | The White House
  www.whitehouse.gov | The Administration
  Barack H. Obama is the 44th President of the United States. His story is the American story — values from the heartland, a middle-class upbringing in a strong...

- The answer lies in the snippet (from one document)
- What if the answer lies in more documents?
Query Results (in general)

Bonn Berlin

Retrieved Snippet:

- Berlin
- Bonn
- Germany
Google Knowledge Graph

Is there a “Why” Button?
“Lessons learned” from Google Knowledge Graph

What does Thomas Edison have to do with Einstein?

User needs to explore by itself and there is no initial headwords to start with!
Challenging Tasks

- connecting two points of knowledge (entities)
- *describing the “connections”*
  - explaining the relationship
  - giving meaningful, human-readable information
  - unexpectedness, usefulness (serendipity)
- ...

...
Making sense of connections?

- Some terms / entities are more connected than others
  - “Crime” and “victim” are close
  - “Crime” and “love” not so much
- For text analysis we need to know how much they are related (on a scale from 0..1 preferably)
- For humans, we are also interested in what sense are they related (in text preferably)
“Semantic Relatedness”: 
- connecting two entities: 
  - A and B 
  - Result: xy% similar 

“Semantic Relatedness ++” 
- connecting two entities: 
  - A and B 
  - Result: xy% similar 
  - describing the “relatedness” 
    - (some explaining texts)
Example

- Relationship between two entities:
  - **Radio** and **Television**
  - Result: 70% related

- Relationship (+ explanation) between two entities:
  - **Radio** and **Television**
  - Result:
    - 70% related
    - *Television sends the picture as AM and the sound as AM or FM, with the sound carrier a fixed frequency (4.5 MHz in the NTSC system) away from the video carrier.*
Connecting the Dots: Chain of News Articles

Earthquake in Japan → Nuclear disaster → Fear of nuclear reactor

Demonstration against nuclear power plants → Election in Baden Württemberg, Germany → Green party governs

Shahaf et al., KDD 2010
Knowledge representation and knowledge engineering are central to AI research. Many of the problems machines face are due to difficulties in understanding the world and natural language. These difficulties have led to a broad range of research and development in areas such as the Semantic Web, which involves developing methods and technologies to allow machines to understand the meaning of information.

Development of the Semantic Web has included the development of XML-based knowledge representation languages and standards, including RDF, RDF Schema, Topic Maps, DARPA Agent Markup Language (DAML), Ontology Inference Layer (OIL), and Web Ontology Language (OWL).

Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies, and is endorsed by the World Wide Web Consortium. This family of languages is based on ....

Ontologies are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it.

(Wira-Alam, Zapilko, Mayr - 2010)
Important References

- Connecting the Dots, Metro Maps, Large Scale Maps of Information (Shahaf et al., 2010 – 2013)
- Serendipitous Entity Search (Bordino et al., 2013)
- Improving Search Result Summaries (Ageev et al., 2013)
# Wikipedia Datasets

<table>
<thead>
<tr>
<th>Wikipedia (EN)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Page Articles</td>
<td>10,276,554</td>
</tr>
<tr>
<td>Page-to-Page Links</td>
<td>315,255,359</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wikipedia (DE)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Page Articles</td>
<td>2,673,528</td>
</tr>
<tr>
<td>Page-to-Page Links</td>
<td>82,515,479</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wikischools for Filtering (School related Wikipedia pages)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Short List</td>
<td>1,875</td>
</tr>
<tr>
<td>Complete List</td>
<td>8,944</td>
</tr>
</tbody>
</table>
Start : Albert Einstein
End : Thomas Edison
Distance : max 3 hops
start to end: How to reach the end?

Choosing a highest reachability score. This score describes a “probability” of reaching a node end given a node start.
Similarity Score between two nodes A and B:

$$\text{sim}(A, B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$  \hspace{1cm} (1)

Reachability between two nodes A and B (in a path):

$$r_{ij} = \prod_{i}^{j-1} \text{sim}(\text{term}_i, \text{term}_{i+1})$$  \hspace{1cm} (2)
How to rank the excerpts:

- Based on position on the text, e.g. abstracts have “best score”

- Based on normalized term frequency
  - Excerpt: Berlin is the capital of Germany
  - Stopwords: is, the, of
  - Important words: Berlin, capital, Germany

\[
score = \frac{freq_{Berlin} + freq_{capital} + freq_{Germany}}{\# \text{ words} \ \# \text{ stopwords}}
\]
Short Demo

http://tinyurl.com/knowescape-demo-v3
Some Points....

• help users gain a better understanding about connections between point of knowledge

• while preliminary, suggests a promising direction for further investigation

• challenge for us to improve the current results
Thank You for Your Attention!

I am looking forward to your remarks, questions, critics, feedback.