Similarity-Based Navigation in Visualized Collections of Historical Documents

Yevgeni Berzak, Michal Richter, Carsten Ehrler

Saarland University
Introduction

• Historians work with textual sources.
• Growing amount of digitized historical documents.
• However, lack of tools to work with electronic collections.
Introduction

Goal:

• Enrich exploration of historical collections beyond keyword search.

• Support historical research by revealing links, structure and knowledge in the collection.
Introduction

Approach: Similarity Visualization
• Extract historically relevant information.
• Measure inter-document similarities.
• Visualize collection according to similarities.
• Integrate visualization within an IR system.
Introduction
Information Extraction

• Question: Which information is important for the historical domain?
Information Extraction

• Question: Which information is important for the historical domain?
• How about:

Persons
Information Extraction

• Question: Which information is important for the historical domain?
• How about:

Persons

Locations
Information Extraction

• Question: Which information is important for the historical domain?

• How about:

Persons  Locations  Organizations
Information Extraction

• Rings a bell?
• Rings a bell?
• Stanford Named Entity Recognizer.

“We have been hoping for a long time that Comrade Brezhnev would visit Cuba. The relations between the CPSU and the communist party of Cuba, relations between our governments and peoples are developing as well as possible.”
Information Extraction

• Rings a bell?
• Stanford Named Entity Recognizer.

“We have been hoping for a long time that Comrade Brezhnev would visit Cuba. The relations between the CPSU and the communist party of Cuba, relations between our governments and peoples are developing as well as possible.”
Information Extraction

• Represent documents in vector space indexed by the extracted named entities and general vocabulary.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Count</th>
<th>Location</th>
<th>Count</th>
<th>Location</th>
<th>Count</th>
<th>Count</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castro</td>
<td>1</td>
<td>Cuba</td>
<td>1</td>
<td>CPSU</td>
<td>2</td>
<td>declare</td>
<td>6</td>
</tr>
<tr>
<td>Breznev</td>
<td>1</td>
<td>Moscow</td>
<td>1</td>
<td>Health Ministry</td>
<td>3</td>
<td>visit</td>
<td>4</td>
</tr>
<tr>
<td>Batista</td>
<td>0</td>
<td>Volgograd</td>
<td>1</td>
<td>Columbian Army</td>
<td>1</td>
<td>relations</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Information Extraction

- Problem: Name variations (and referring expressions) of entities.

Fidel Castro
Dr. Fidel Castro
Maj. Fidel Castro
Castro

Cuban Communist Party
Communist Party of Cuba

...
Information Extraction

- **Solution: Aliasing**
- **Measure inter-entity string similarity using string kernels.**

![NE Similarity matrix]

- **NE Similarity matrix**

<table>
<thead>
<tr>
<th></th>
<th>Castro</th>
<th>Dr. Castro</th>
<th>Batista</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castro</td>
<td>1</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>Dr. Castro</td>
<td>0.8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Batista</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Information Extraction

- Solution: Aliasing
- Measure inter-entity string similarity using string kernels.
- Boost document (and query) vectors with aliases.

<table>
<thead>
<tr>
<th></th>
<th>Castro</th>
<th>Dr. Castro</th>
<th>Batista</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castro</td>
<td>1</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>Dr. Castro</td>
<td>0.8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Batista</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

NE Similarity matrix

Document Vector
Information Extraction

- Solution: Aliasing
- Measure inter-entity string similarity using string kernels.
- Boost document (and query) vectors with aliases.

\[
\begin{bmatrix}
  \text{Castro} & \text{Dr. Castro} & \text{Batista} \\
  1 & 0.8 & 0 \\
  0.8 & 1 & 0 \\
  0 & 0 & 1
\end{bmatrix}
\]

NE Similarity matrix

\[
\begin{bmatrix}
  \text{Castro} & \text{Dr. Castro} & \text{Batista} \\
  1 & 0.8 & 0 \\
  0.8 & 1 & 0 \\
  0 & 0 & 1
\end{bmatrix}
\]

Document Vector

Aliased Document Vector
Inter-Document Similarity

• Use Cosine measure to determine similarity between each pair of documents.

• Similarity is measured according to each type of Named Entities and general vocabulary.

Assumption: similar documents are related in an interesting way.
Visualization

Similarity Graph
Visualization

Similarity Graph
• Documents = Nodes
Visualization

Similarity Graph

• Documents = Nodes
• Similarities = Edges
Visualization

Layouting

• Flexible positioning
• Force Directed Placement (FDP)
• Graph Clustering (Chinese Whispers)
Visualization

Major Benefits

• Recommender: links between documents
• Structure (e.g. topics, events, etc..)
Visualization

Major Benefits

• Recommender: links between documents
• Structure (e.g. topics, events, etc..)
• Visual IR
The History Explorer

- Open Text, Distance Filter, Mark Neighbors

---

Santiago Municipality Speech

Santiago Chile PRENSA LATINA in Spanish to PRENSA LATINA Havana 2256 GMT 26 Nov 71 C--FOR OFFICIAL USE ONLY

[Text] Santiago, Chile, 26 Nov--Herewith is the text of the speech given by Cuban Prime Minister Fidel Castro in a municipality in this city yesterday, 25 November 1971:

Once upon a time there was the miracle of the fishes and the bread--a miracle which we revolutionaries way we want to repeat, but this time there has also been the miracle of time. At this time yesterday we were in Rancagua dedicating a labor union hall, then there was the rally, then we had to go to Coya, and later we had to tour Colchagua Province. We had to visit an agrarian reform center--there are various designations and I do not want to make a mistake--(laughter) and then we had to tour the streets of the provincial capital. Later we had to attend a ceremony with the peasants in Santa Cruz.

By yesterday the program planners were saying: "At what time will the municipality ceremony be?" I believe that they had decided to propose a change of date to the mayor and the councilmen.
The History Explorer

- Query: “health”
The History Explorer

• Query: “health”
The History Explorer

• Query: “Giron Kennedy”
The History Explorer

• Query: “Giron Kennedy”
Recap

• Extract domain-relevant information.
• Visualize collections according to similarities with regard to this information.
• Visual Information Retrieval:
  – Links between documents
  – Interpretable structure
  – Interactivity, customization

→ Knowledge Discovery
Outlook

• Co-reference resolution.
• New databases.
• Other domains (legal, medical, etc).
• Other methods for feature extraction.
• Expert User Evaluation.
Thank You