gesis
Leibniz-Institut für Sozialwissenschaften

CLASSIFICATION & Visualization
International UDC Seminar 2013

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, Eyeplorer, 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

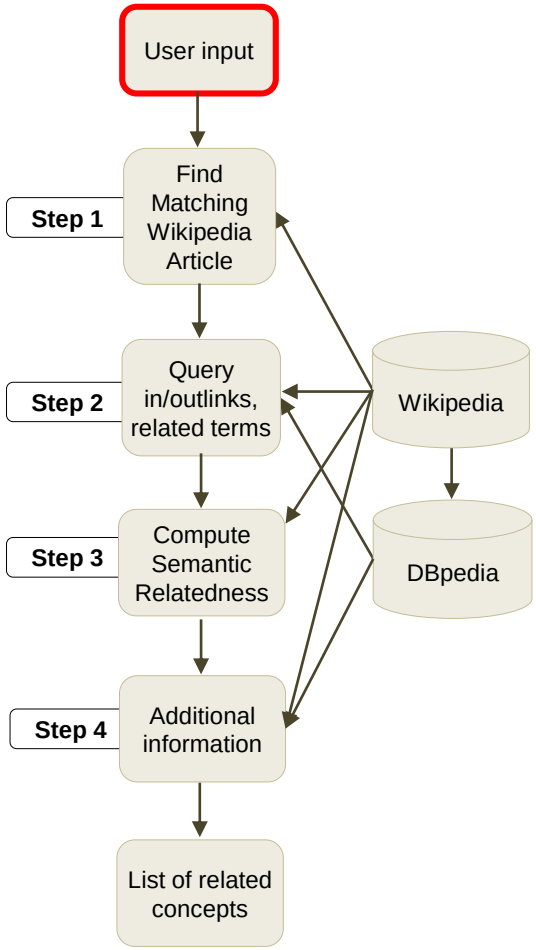
Goal

- 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

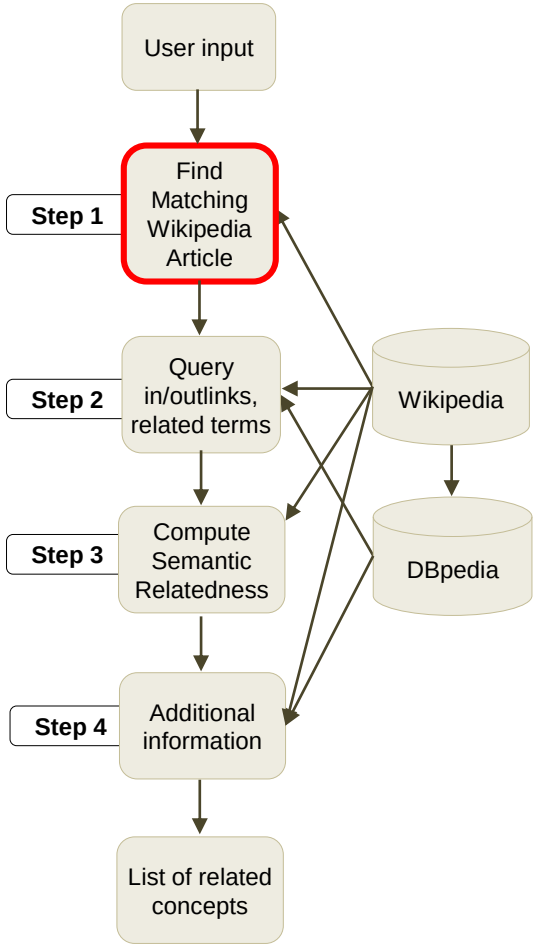
Steps to compute semantically-related concepts

The user enters a keyword in the search form



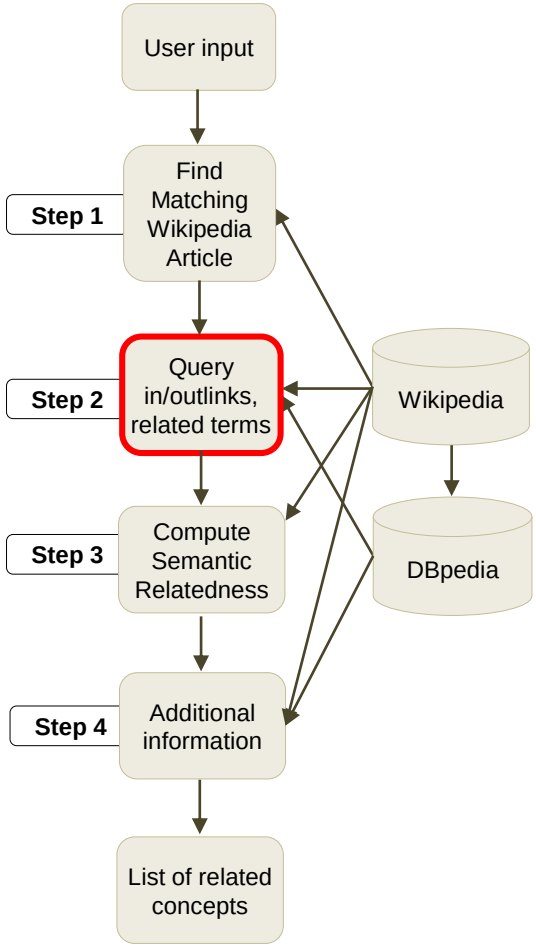
Steps to compute semantically-related concepts

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



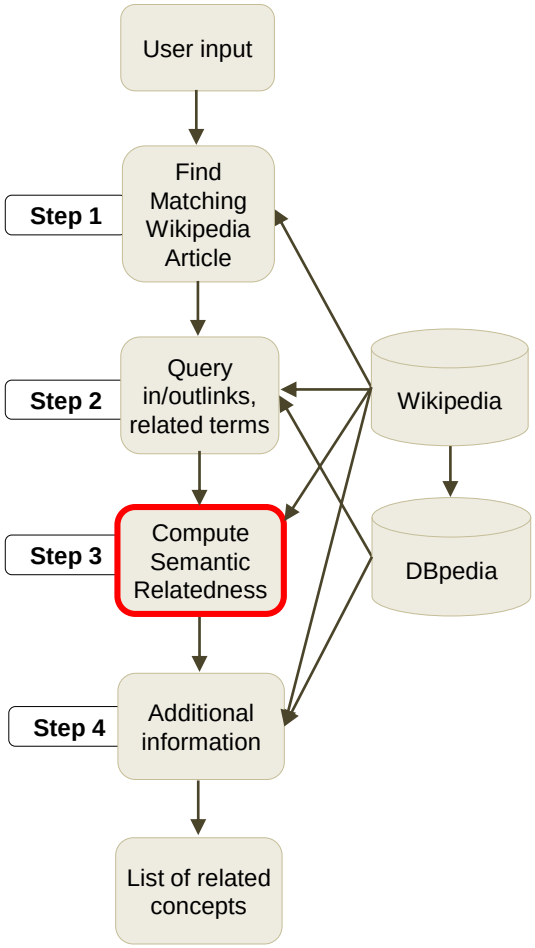
Steps to compute semantically-related concepts

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

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

A: Number of full text search hits in Wikipedia for concept one
B: Number of full text search hits in Wikipedia for concept two
$A \cup B$: Number of full text search hits in Wikipedia for concept one AND concept two.
W: Number of articles in Wikipedia.



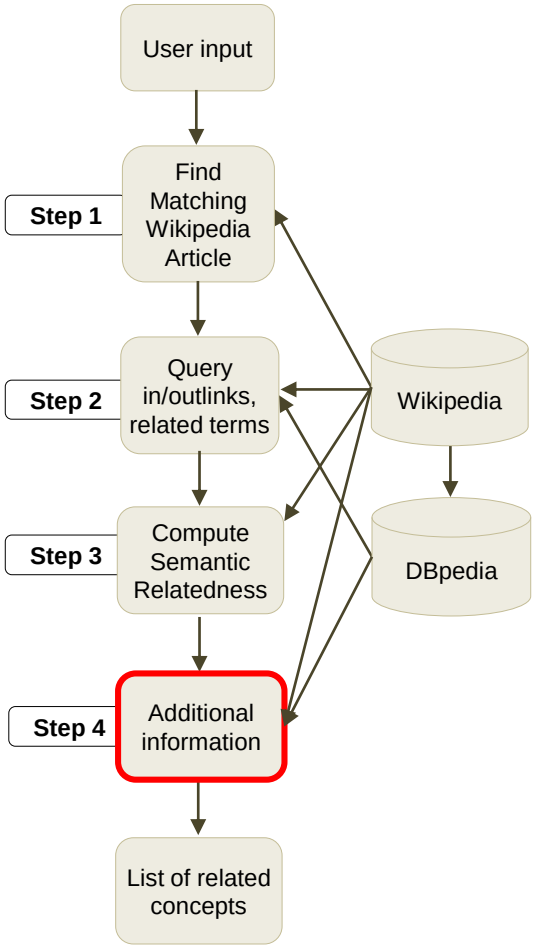
Steps to compute semantically-related concepts

Step 4:

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

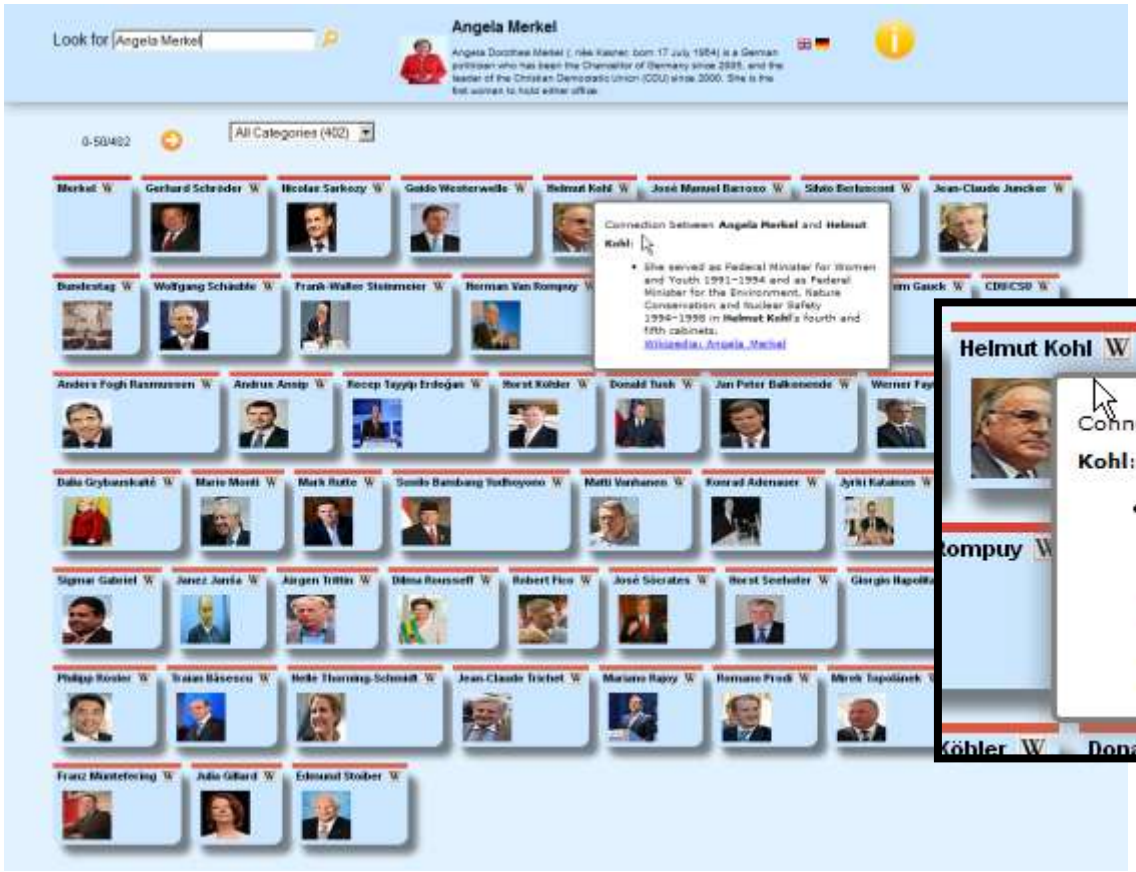
All these processing steps are computed live, in a parallel manner, with several hundred queries in parallel

-> 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)

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

Results

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

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