SIEMENS

Corporate Technology

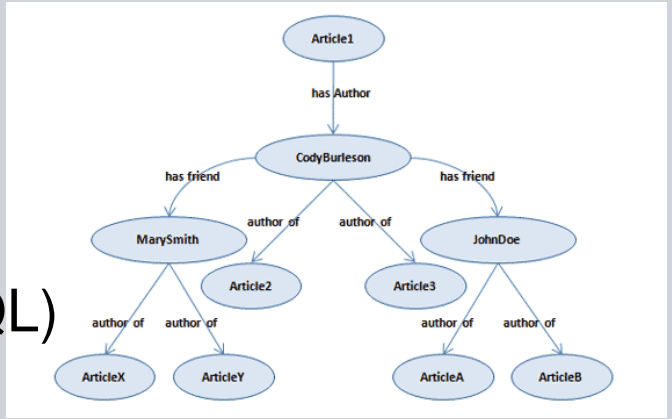
Combined Structured and Keyword-Based Search in Textually Enriched Entity-Relationship Graphs

Davide Magatti – DISCo – Università degli Studi di Milano Bicocca
Florian Steinke – Siemens AG Corporate Research
Markus Bundschus – Siemens AG Corporate Research
Volker Tresp – Siemens AG Corporate Research

SIEMENS

Background

Two mainly separated worlds regarding Information Representation Storage and Retrieval



- ***Semantic World***
  - Entity-Relationship Graphs (e.g. YAGO)
  - RDF Triple Stores
  - *Complex* but *precise* query languages (SPARQL)

- ***Traditional IR World***
  - Huge text collections
  - Construction of inverted indexes
  - *Simple* but *imprecise* keyword based search

SIEMENS

Motivation

- **Many repositories are a mixture of textual data and structured information**
  - Textually enriched Entity-Relationship Graphs
  - e.g. Combination of YAGO+ Wikipedia sources

- **<u>Idea:</u> Combine Structured Search and Keyword-Based Search:**
  - Exploit the strengths of keyword search on unstructured text (i.e. simplicity)
  - Propagate the relevance obtained by traditional keyword search to the structured graph (**Step I**)
  - Use the power and preciseness of structured queries -SPARQL- to narrow down ("filter") the results according to user needs (**Step II**)

→ ***HYBRID-SEARCH ENGINE***

SIEMENS

Proposed Hybrid Search Engine – Step I

- Let *G=(V,E)* be a Entity-Relationship graph and *D* be a set of linked documents
  - Build an index for each node $v_k \in V$ which is associated with a document $d_k \in D$
  - At query time, use the keyword-based ranking to assign relevance scores to nodes in the graph
  - Based on the idea of spreading activation [Crestani 1997], propagate the scores through the graph such that the score r_i of a node $v_i \in V$ is:

$$r_i = l_i + \lambda \sum_{j:(j,i)\in E} \frac{r_j}{d_j}$$

  - With l_i the Lucene relevance score, λ the spreading factor and d_j the (out-)degree of node *j*

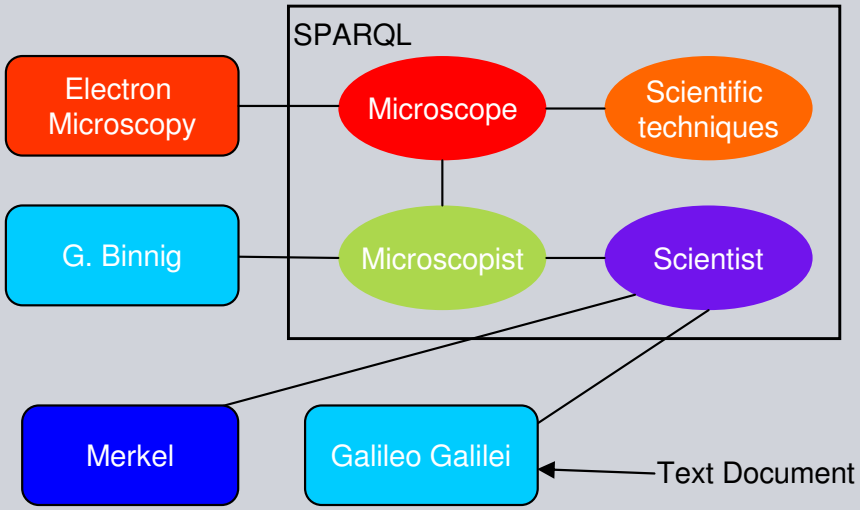
SIEMENS

Hybrid Search Engine – Step I+II

- Computing the spread activation is equivalent to compute the solution for the sparse linear system:

$$(I - \lambda G) \cdot \underline{r} = \underline{l}$$

- Then a *SPARQL–SELECT* query is performed to filter only relevant entities from the ER- graph (**Step II**)



r	Category	score
0	wikicategory wordnet microscope	22079.62
1	wikicategory Scientific techniques	11873.37
2	wordnet disease	7419.3
3	wikicategory wordnet optics	6621.56
4	wordnet person	4254.36

SIEMENS

Experimental Setup

Data

- ER Graph: YAGO
- Text: Wikipedia pages referenced by YAGO predicate "`describes`"

Size:

- 2,582,063 Nodes *(V)*
- 8,647,952 Edges *(E)*
- 2,082,598 Wikipedia pages *(D)*

Storage:

- Apache Lucene
- Ontotext Swift-OWLIM

Computations:

- Java Matrix Toolkit (GMRES)

SIEMENS

Experiments - 1

Context – Aware entity search

- Retrieve facts about entities where the context is specified via keywords

r	Company	Employees	Revenue	Score
1	General Electric	327000	$ 172.738 M.	9083.89
2	GE Healthcare			8419.13
3	Philips	125500	26.976 M.	8404.8
4	Siemens AG	430000	$ 110,820 M.	8092.45
5	Neusoft Group	12000	$ 355 M.	4640.18
6	SRI International			4299.93
7	Agfa-Gevaert	13565	3.300 M.	3759.03
8	Foster-Miller			3055.97
9	Ellex Medical Las.			3011.97
10	Turtle Beach Syst.			2958.37

Give me a list of Companies with # of Employees and Revenue related to keyword "`ultrasound`"

Experiments - 2

Context – Aware entity search

r	Physicist	Advisor	Score r_i
0	Robert Oppenheimer	Max Born	86579.94
1	David Bohm	Robert Oppenheimer	77108.46
2	Willis Lamb	Robert Oppenheimer	62835.95
3	Philip Morrison	Robert Oppenheimer	61497.02
4	Richard Feynman	John Archibald Wheeler	54672.82
5	George Zweig	Richard Feynman	52808.75
6	Chen Ning Yang	Edward Teller	47098.77
7	Edward Teller	Werner Heisenberg	46347.56
8	Lincoln Wolfenstein	Edward Teller	45903.60
9	John von Neumann	Leopold Fejr	34172.28
10	Emilio G. Segre'	Enrico Fermi	31561.64
10	Enrico Fermi	Luigi Puccianti	31561.64

Give me a list of physicist and their advisor who worked on "`quantum mechanics`" and have some relations to "`Los Alamos`"

Experiments - 3

Context-Aware category search

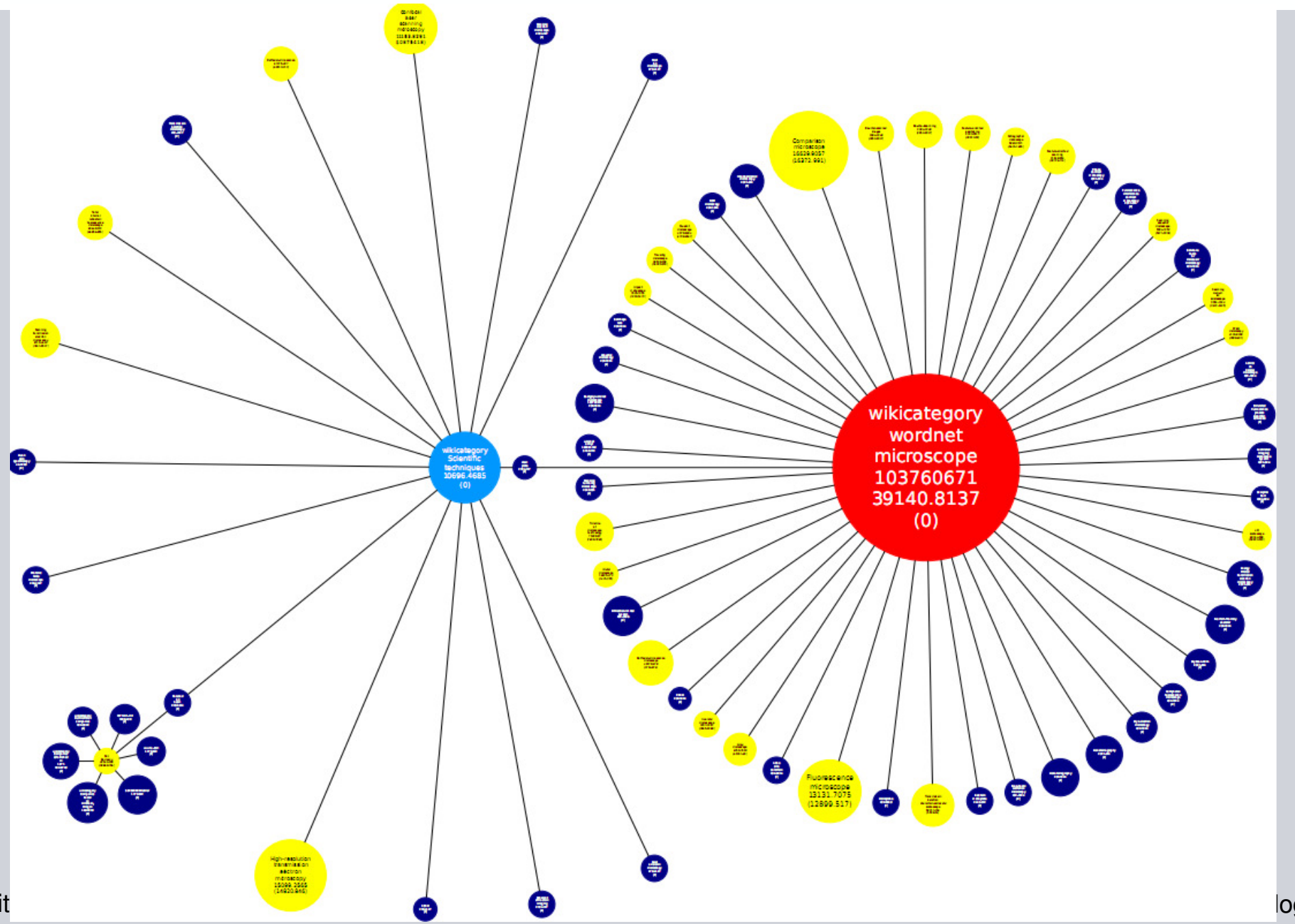
- Retrieve general concept not connected to text articles, through spread activation of the keywords

r	Category	score
0	wikicategory wordnet microscope	1234643.5
1	wikicategory Scientific techniques	1112785.29
2	wordnet disease	962138.78
3	wordnet person	632517.98
4	wikicategory wordnet optics	521473.51
5	wikicategory Plant pathogens and diseases	486216.12
6	wikicategory wordnet measuring instruments	322737.59
7	wikicategory X-rays	281627.11
8	wordnet anatomy	280101.66
9	wikicategory wordnet igneous rock	230442.75

Give me categories related to "microscope"

SIEMENS

Example of Spreading Activation



SIEMENS

Conclusions and Future works

- **Conclusions**
  - Novel method to combine flexible keyword based search with expressive structured queries
  - Hybrid search can avoid the complexity of structured queries by offering a faceted-style user interface for the SPARQL query and the familiar keyword-search
  - The system can retrieve combination of facts related an entity (e.g. company, revenue, employees) and not only list of entities.

- **Future Works**
  - Quantitative ranking evaluation (NDCG) for TREC data (ongoing)
  - Experiments with SPARQL reasoning (ongoing)

SIEMENS

Bibliography

- Crestani, F. (1997). **Application of spreading activation techniques in information retrieval.** Artificial Intelligence Review, 11, 453–482.ss.
- Kasneci, G., Suchanek, F., Ifrim, G., Ramanath, M., & Weikum, G. (2008). **Naga: Searching and ranking knowledge**. Proc. of ICDE (pp. 1285–1288).
- Mangold, C. (2007). **A survey and classification of semantic search approaches.** International Journal of Metadata, Semantics and Ontologies, 2, 23–34.
- Rocha, C., Schwabe, D., & Aragao, M. (2004). **A hybrid approach for searching in the semantic web**. Proceedings of the 13th international conference on World Wide Web (pp.374–383).
- Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). **Yago: A Core of Semantic Knowledge.** 16th international World Wide Web conference (WWW 2007). New York, NY, USA: ACM Press.
- Tran, T., Cimiano, P.,Rudolph, S., & Studer, R. (2007). **Ontology-based interpretation of keywords for semantic search.** Lecture Notes in Computer Science, 4825, 523.
- W3C **SPARQL Query Language for RDF**. http://www.w3.org/TR/rdf-sparql-query/.
- Ontotext (2009). http://www.linkedlifedata.com/ **Linked life data**.

