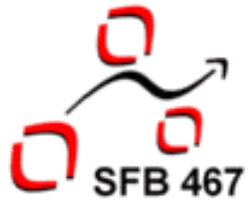


Intranet Search

Exploiting Databases for Document Retrieval

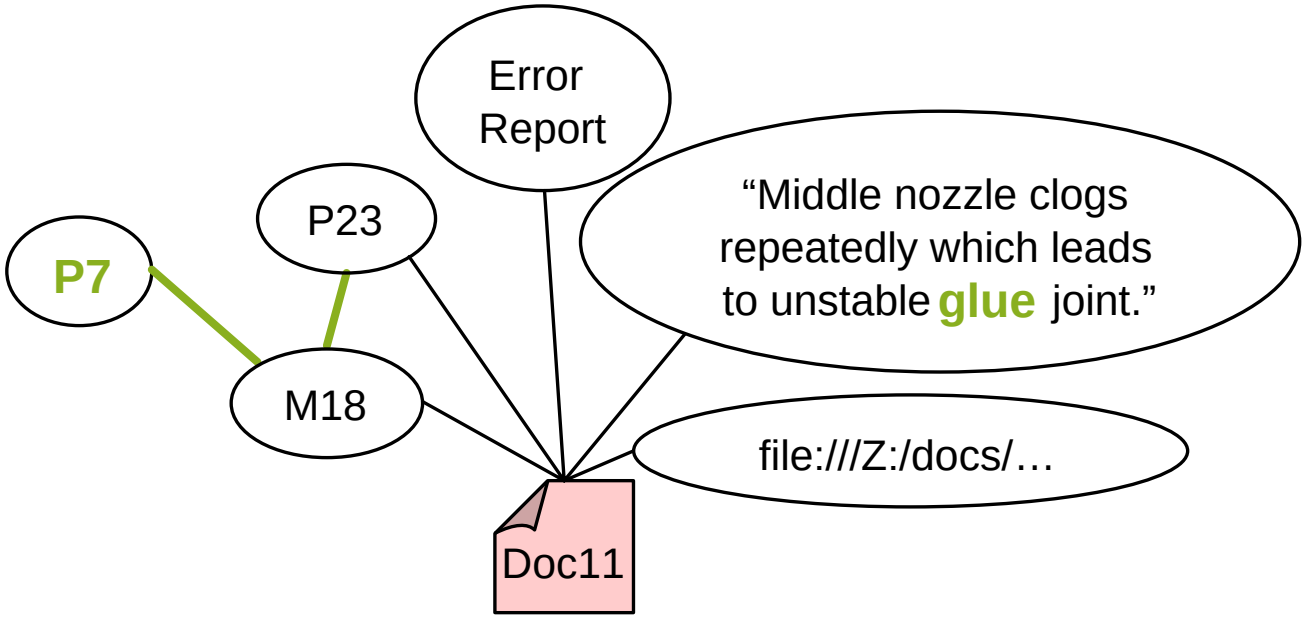
Christoph Mangold

Universität Stuttgart

The Big Picture: Assume….

… there is a glueing problem with product P7

- Has this happened before?
- Is there any document about the problem?
- Search for: **`P7 glue`**



→ **Aim: Rank Doc11 as highly relevant**

Overview

- The ContextGraph (1)
- Ranking (4)
- Computing the context (1)
- Implementation & Performance (2)
- Related Work (3)
- Future Work & Summary (2)

ContextGraph & Semantic Distance

ErrorReport

DocID	URL	Abstract	ProductID (FK)	MachineID (FK)	…
Doc11	file:// …	Middle noz…	P23	M18	…

Product

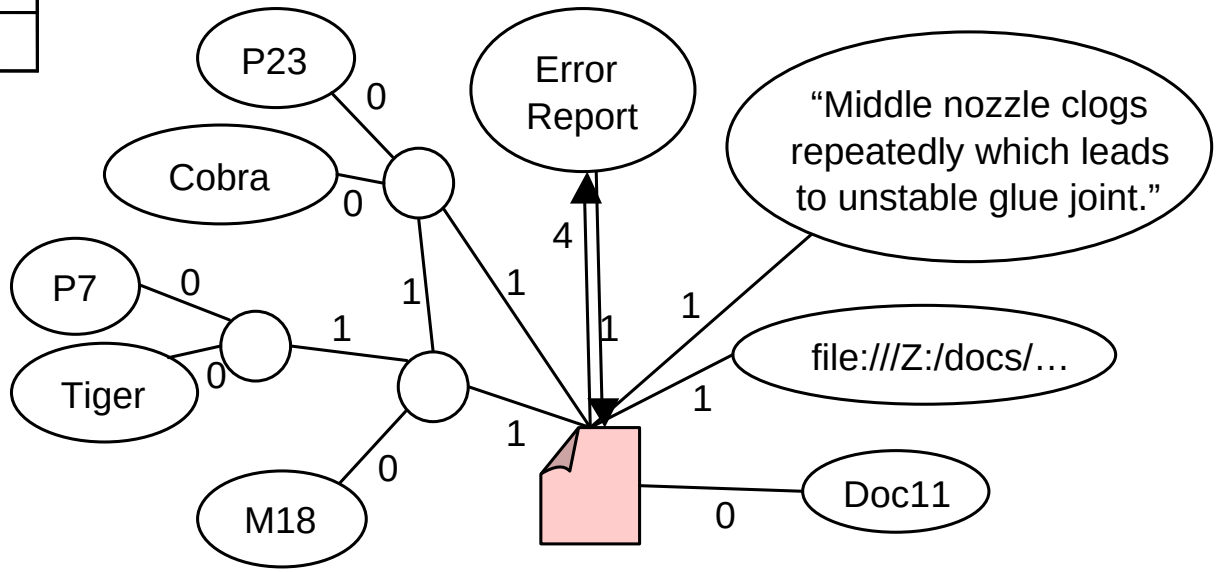
ProductID	Name	…
P7	Tiger	
P23	Cobra	

Machine

MachineID	Location	Type	…
M18	…	…	…

Production

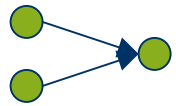
ProductID (FK)	MachineID (FK)
P7	M18
P23	M18



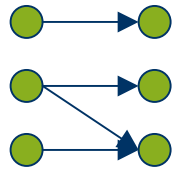
Ranking

Idea: Transfer well-proven ranking measures to the context-based scenario

- “What‘s Related“:
  Exploit the web structure



- Query independent:
  Google’s PageRank / ObjectRank



- Query specific:
  Vector space model & tf.idf

coming up next …

Ranking: Vector Space Model & tf.idf

- Documents and queries are vectors in a $|T|$-dimensional vector-space where T is the set of all terms.
  - Similar vectors denote similar documents/queries

Term	d_1	d_2	d_3		q
clog	0.3	0.3	0		0
glue	0.6	0	0	…	0.5
…					
nozzle	0.2	0.1	0.1		0.2

- Vector entries are calculated by means of tf•idf
  - **tf** (term frequency):
    How **often** does the term appear in the document/query
  - **idf** (inverse document frequency):
    How **rare** is the term in the document collection

Ranking: tf ➔ ctf

Consider the text only

tf: How often does the term appear in a document?

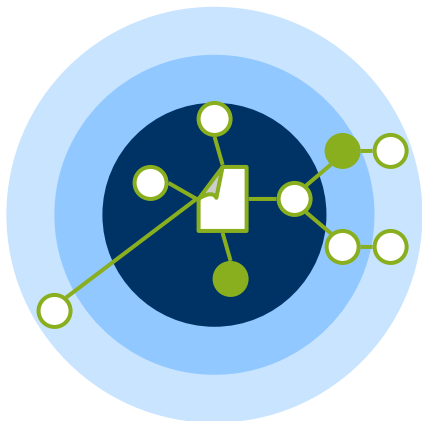
$$\mathrm{tf}(t,d) = \frac{\mathrm{freq}(t,d)}{\max_{\ell \in d}(\mathrm{freq}(\ell,d))}$$

There is a glueing problem on M18 …
When the glue gets …
….

Consider context and semantic distances

ctf: How often does the term appear in the context of a document?

$$\mathrm{ctf}(t,d) = \frac{\sum_{k=1}^{|H^t|} \mathrm{sim}(d, h_k^t)}{\max_{\ell \in \mathrm{Context}(d)} \sum_{k=1}^{|H^\ell|} \mathrm{sim}(d, h_k^\ell)}$$



Ranking: idf ➔ icf

Consider the text only

idf: How rare is the term in the document collection?

$$\mathrm{idf}(t) = \frac{1}{|\{\delta \in D \mid t \in \delta\}|}$$

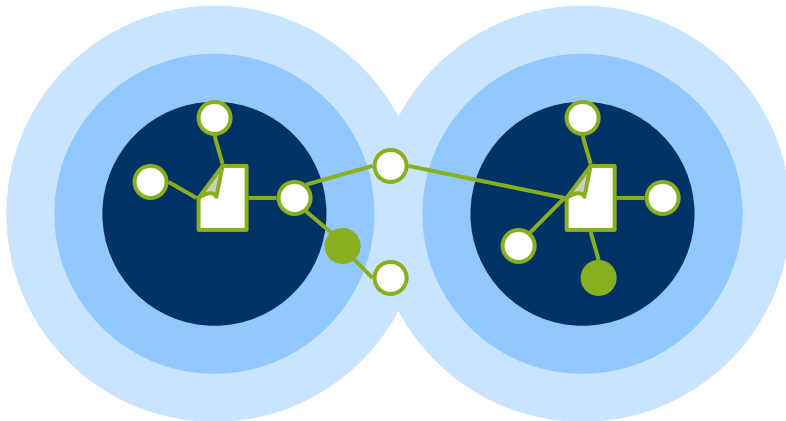
Consider context and semantic distances

icf: How rare is the term in the contexts of all documents?

$$\mathrm{icf}(d,t) = \frac{1}{|\{\delta \in D \mid \exists n \in V_t : \mathrm{sim}(\delta, n) \geq \mathrm{sim}(d, n)\}|}$$

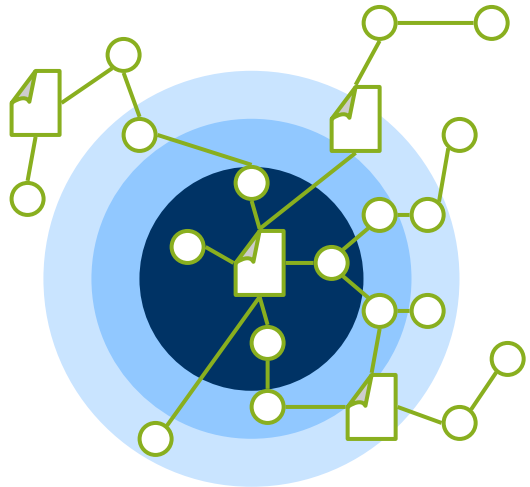
Term	d_1	d_2	d_3	
clog	2	2	0	$\mathrm{idf}(\mathrm{clog}) = 1/4$
glue	1	0	0	$\mathrm{idf}(\mathrm{glue}) = 1/1$
…				
nozzle	2	1	1	$\mathrm{idf}(\mathrm{nozzle}) = 1/5$

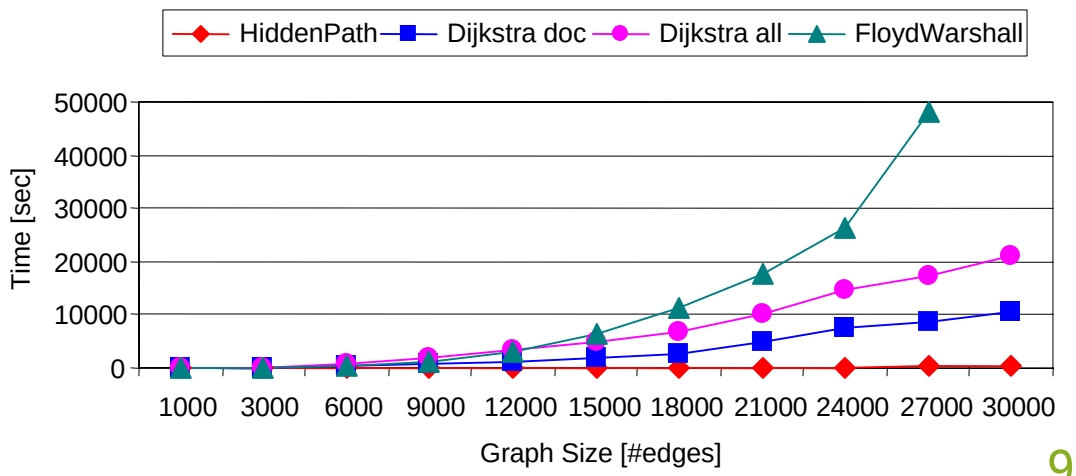
$\rightarrow$



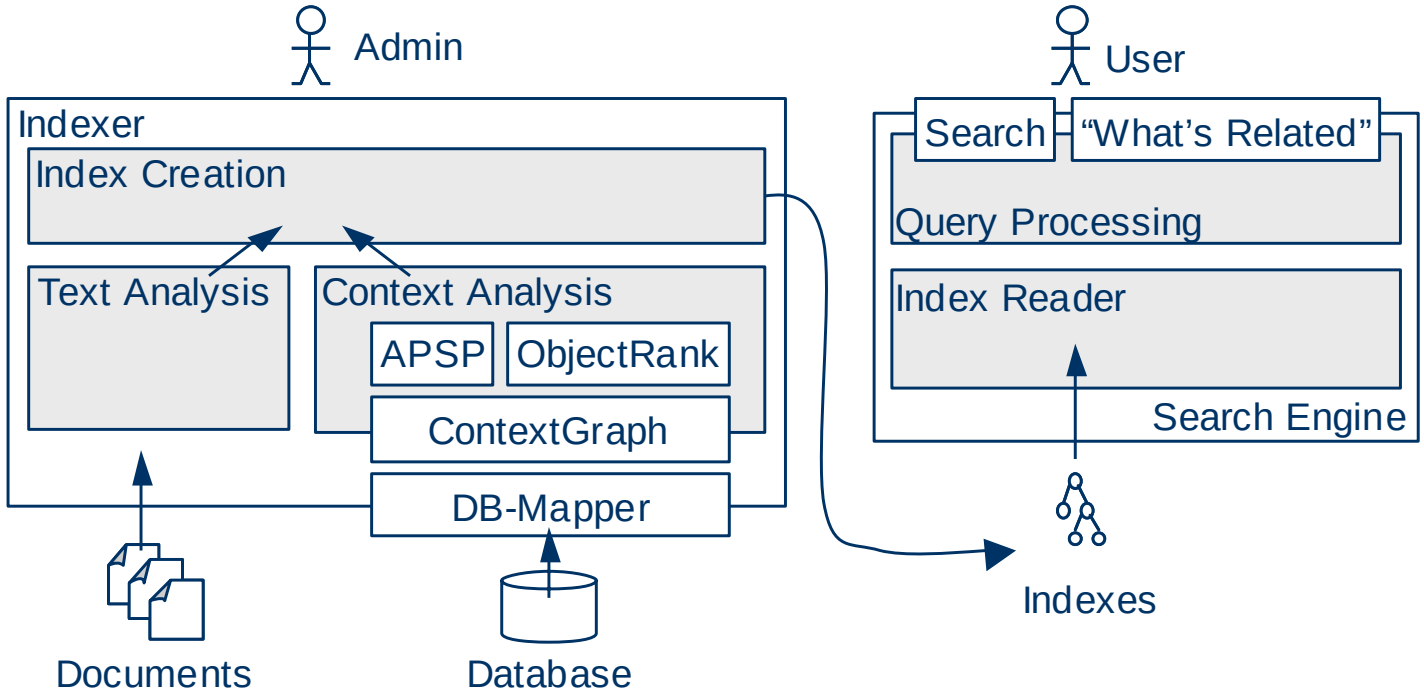
Computing the Context

- All Pairs Shortest Path (APSP)
- Optimizations:
  - Neighborhood only
  - Documents only
- Implementation
  - FloydWarshall
  - Neighborhood Dijkstra
  - Document Neighborhood Dijkstra
  - Neighborhood HiddenPath [Karger ‘93]



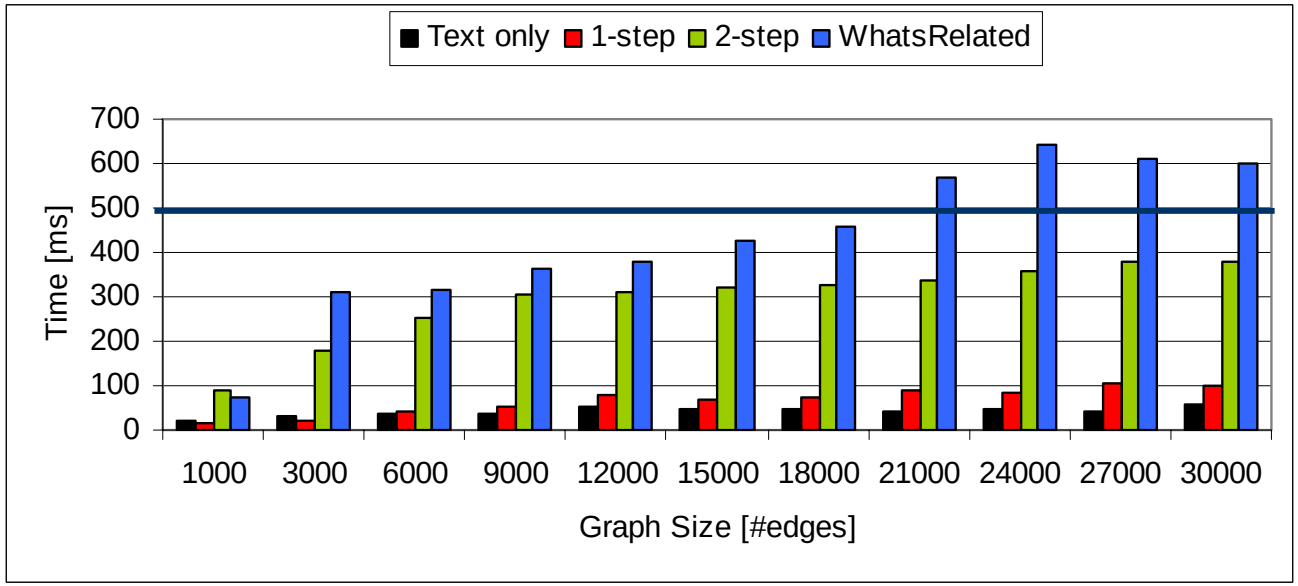


Implementation: Architecture & Technologies



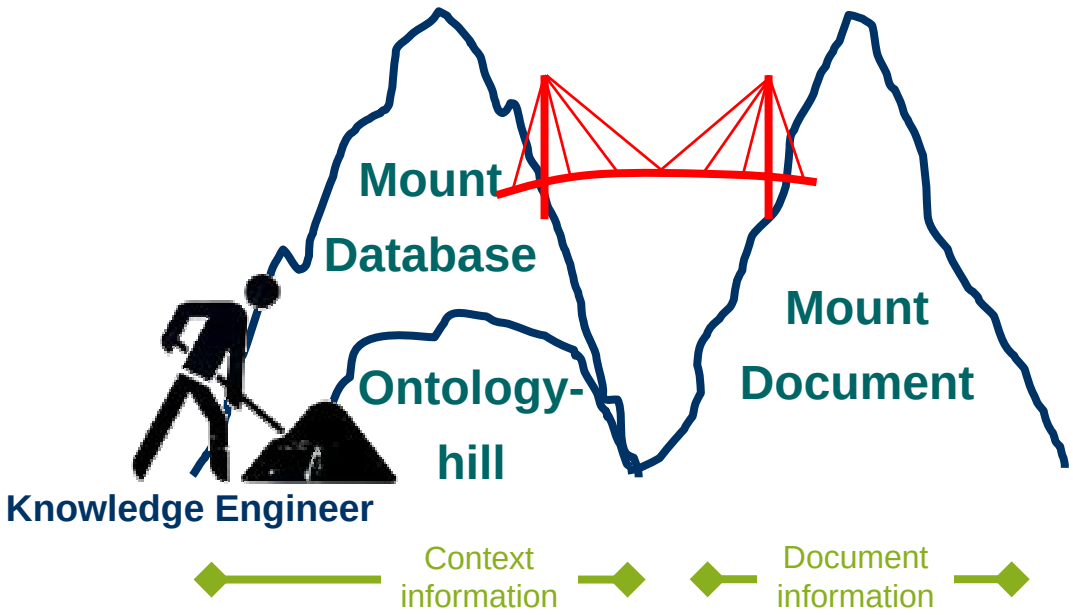
- Java
- Lucene (Apache’s search engine)
- D2RQ (DB-ontology mapping tool, FU Berlin)
- Jena (hp’s semantic web framework)
- OWL / RDF (W3C’s ontology description language)

Performance: Query Time



Related Work: Semantic Search

- Surveyed 21 approaches
- Semantic Web
- Contextual knowledge is modeled in (handcrafted) ontologies
- User interaction
- Different ontology structures require / enable a large variety of search engines



Related Work: Keyword Search in Databases

- [Goldman, VLDB‘98] Lorel DB
  - FIND ... NEAR
  - Shortest Path
- [Bhalotia, ICDE‘02] BANKS
  - Relational DB as a graph
  - Search for subgraphs
- [S. Agrawal, ICDE’02] DBXplorer,
  [Hristidis, VLDB’02, VLDB’03] DISCOVER
  - Join tables to retrieve tuples that contain all search terms

Related Work: Combining Structured & Unstructured Data

Using SQL queries

- [Dessloch, VLDB'97]
- [Goldman, SIGMOD'00] WSQ
  - Unstructered data as virtual tables
  - Computes e.g. number of appearances of search terms

Using OLAP techniques

- [Cody, IBM Sys. Journal **41**(4), 2002] BIKM
  - Information Extraction
  - Data Warehouse

Future Work

- Assess semantic correctness
- Integration of ontologies / semantic search
- External memory shortest path algorithm

Summary

- Exploit DB-Information to support Document Retrieval
- ContextGraph
- Semantic-distance based ranking à la tf.idf
- Architecture incorporates text- and context-search
- Performance evaluations promise little overheads only
- Related work: Semantic Search & DB Keyword Search