Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

The Eight International Conference on Advances in Databases, Knowledge, and Data Applications

June 26 - 30, 2016 - Lisbon, Portugal

Semantic Suggestions in Information Retrieval

Andreas Schmidt

**Department of Informatics and
Business Information Systems
University of Applied Sciences Karlsruhe
Germany**

**Institute for Applied Computer Sciences
Karlsruhe Institute of Technologie
Germany**

Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Outlook

- Introduction
- Query Principles
- Implementation Aspects
- Summary & Outlook

STICS [1]

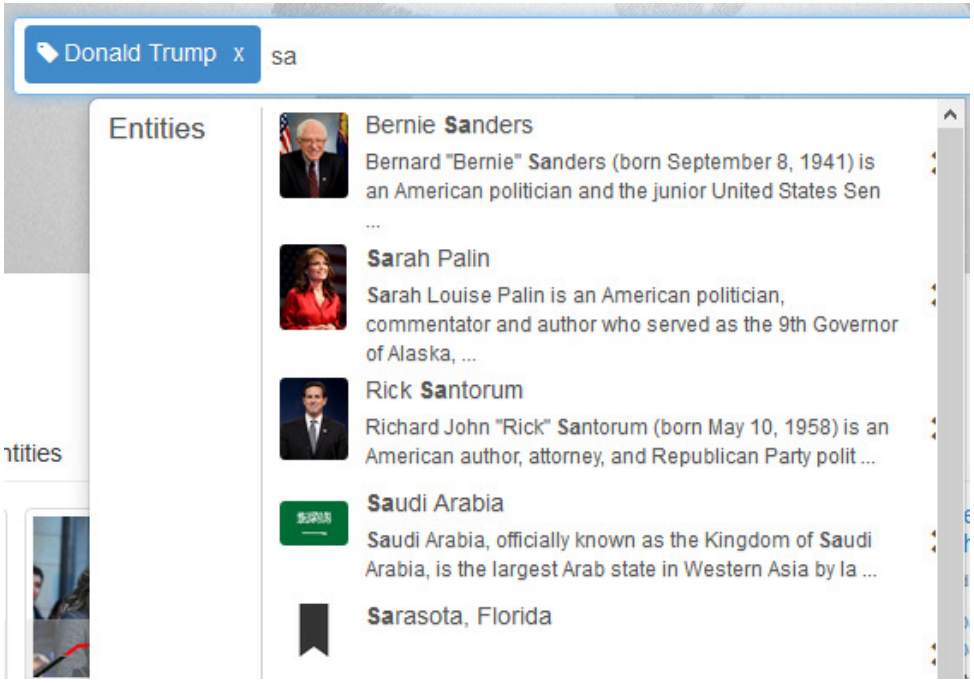
- Semantic Search engine for entities and categories, developed at MPI for Informatics
- Every document in the search space is preprocessed with
  - Named Entity Recognition (NER)
  - Named Entity Disambiguation (NED)
- Categories and entities based on YAGO knowledge base
- Auto-completion feature for a given prefix, based on global relevance of a entity, category (ranking)
- This often leads to empty resultsets

Problem Statement

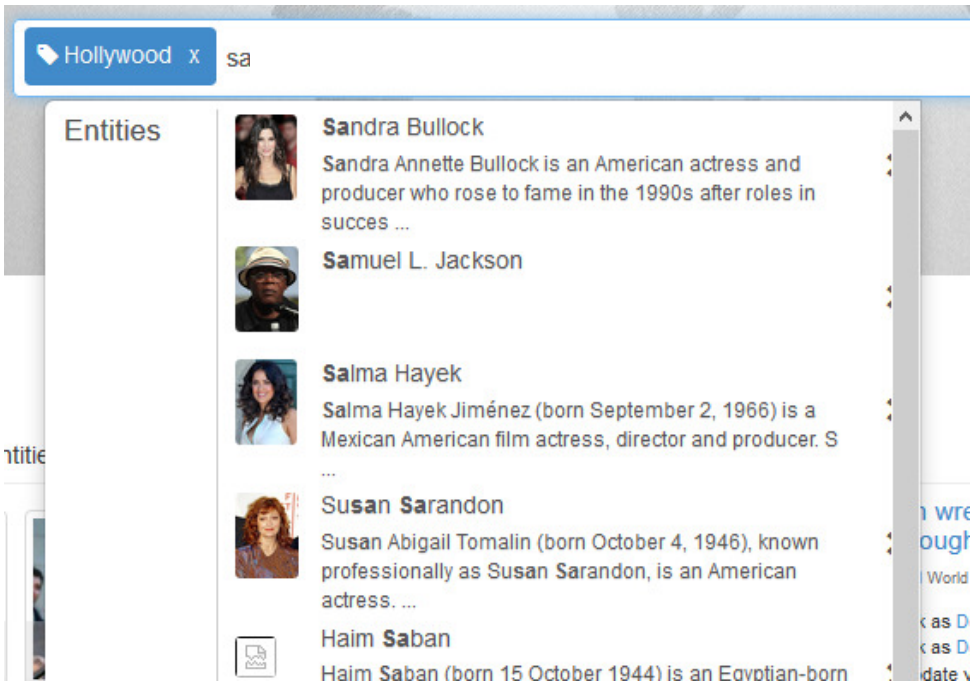
- Given:
  - News article collection (~4 million news articles)
  - 600.000 different mentioned entities
  - 60 million occurences of entities in collection
- Build a query interface, so that ...
  - Given a number of previously chosen entities and one or more prefixes, suggest related entities, so that the result set is not empty
  - Rank the suggestions based on relevance
  - The suggestions must be calculated fast

Example

- Entity Donald Trump
- Prefix: ’sa’



- Entity Hollywood
- Prefix: ’sa’



Demo: https://stics.mpi-inf.mpg.de/

Query Principle

If no entity is already given:

- Extend the prefix with matching entities, ranked by their global relevance

One or more entities given:

- Select all news articles, that contain the already given entities
- Further restrict this set, by deleting all news articles not containing further entities with the given prefixes
- Extract from this set of news articles all entities with the given prefix(es)
- Rank these entities according to some relevance criteria

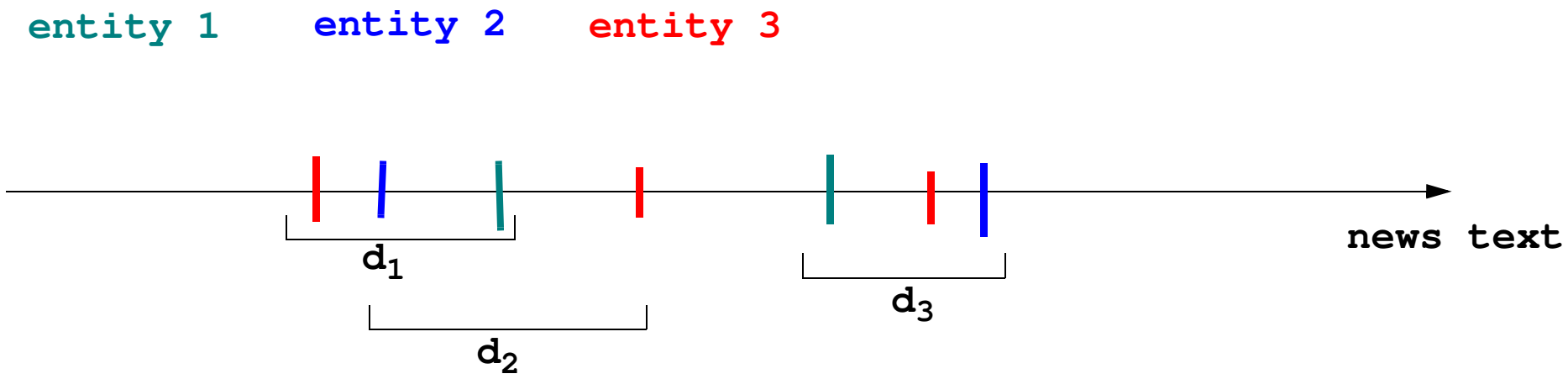
Ranking of Entities

- Based on some information extracted from YAGO knowledge (Wiki links)
  - Milne Witten [5]
  - Kore [6]
  - ...
- Based on „dynamic document frequency“ (in how many news documents of the resulting news document does the entity appear)
- Based on co-occurence of entities in an interval of words inside documents

static

corpus adaptive

Calculation of relatedness based on co-occurence in news



- Search for tuples, triples, quadruples of entities in the news text
- Entities in tuples must be inside an interval of length d_{max} (a priory fixed)

  => $d_{max} = max(d_1 , d_2 , d_3)$
- relatedness $(e_1, e_2, e_3) = \log_2(1/d_1) + \log_2(1/d_2) + \log_2(1/d_3) + ...$

Karlsruhe Institute of Technology

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

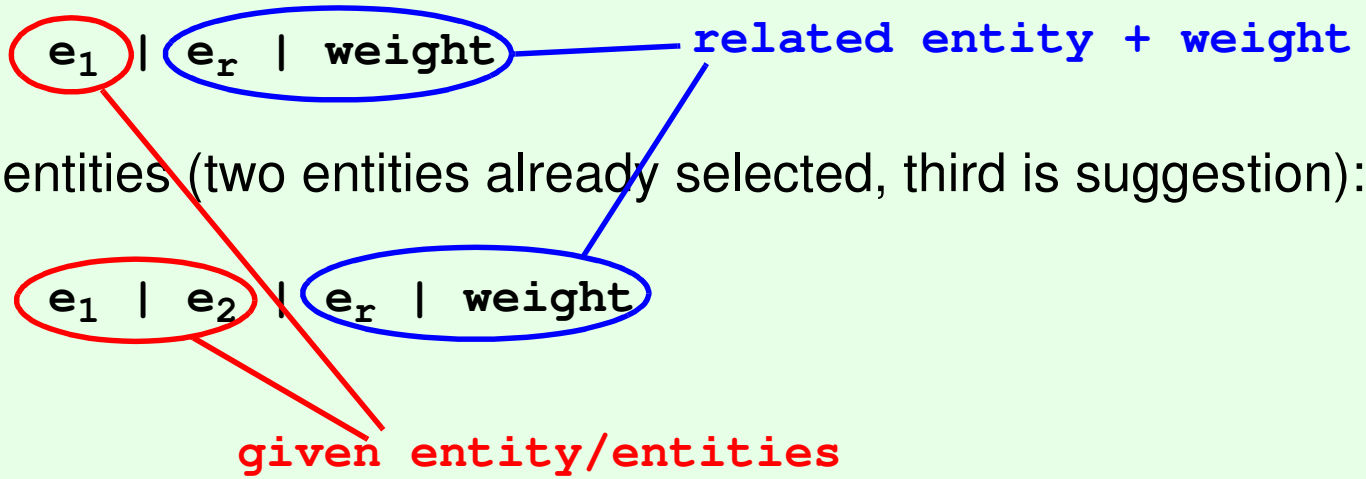
Calculation of relatedness based on co-occurence in news

- Because of strong time constrainst ...
  - Precalculate „relatedness“ for all tuples, triples, quadruples, ... of entities based on document collection (n >1, n < 7)
    (e_1, ..., e_{n-1}) -> (e_n, $rel_{1, ... , n-1}$)

- Some numbers (based on 3.582.098 news articles)
  - 5.594.390 cooccurence tuples (max dist.: 30)
  - 5.022.237 coocurence triples (max. dist. 42)
  - 2.814.076 coocurrence quadruples (max. dist: 51)
  - 2.336.808 cooccurence quintuples (max. dist.: 60)
  - 1.454.580 cooccurence 6-tuples (max. dist.: 67)

Implementation based on relational DB

- Precalculation of tuples/triples of related entities together with a weight
- 2 entities (one entity already selected, second is suggestion):



- 3 entities (two entities already selected, third is suggestion):
- ...
- Entity Frequency (if no entitity is given so far)

```
e | frequency
```

Prefix Handling

- Every entity has a short description of avg : 2.5 words

entity_id	entity_value	human_readable_name
20195899	Boston_University_Bridge	Boston University Bridge
12905648	Boston_University_College_of_Communication	Boston University College of Communication
21356536	Boston_University_School_of_Law	Boston University School of Law
21181981	Boston_University_School_of_Management	Boston University School of Management
12738583	Boston_University_School_of_Medicine	Boston University School of Medicine
20146803	Boston_University_School_of_Public_Health	Boston University School of Public Health
11722953	Boston_University_School_of_Social_Work	Boston University School of Social Work
20534109	Boston_University_School_of_Theology	Boston University School of Theology
7782976	Boston_University_Tanglewood_Institute	Boston University Tanglewood Institute
20942026	Boston_University_Terriers	Boston University Terriers
19475745	Boston_University_Terriers_men\u0027s_ice_h...	Boston University Terriers men's ice hockey

- Prefix match every start of a word
  i.e prefix('bos','hea') -> (Boston_Public_Health_Commission,
  Boston_University_School_of_Public_Health)

- Table prefix_entity (~10 million entries)

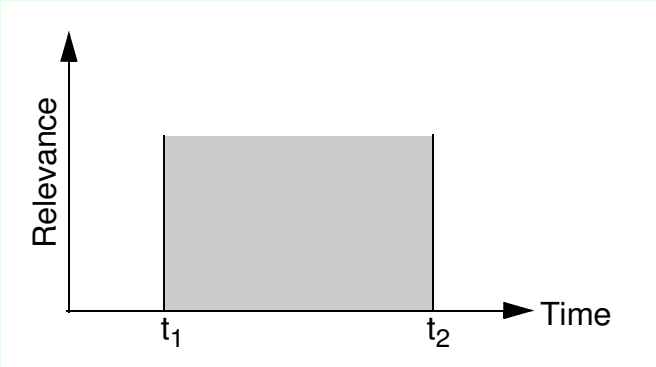
	id	entity	word_match
	bos	1112035	0
	bos	1885758	1
	bos	2303237	0
	bos	2417071	0
	bos	2449991	0
	bos	2801432	0
	bos	3254453	0
	bos	3702676	1
	bos	3965776	0
	bos	4476312	0

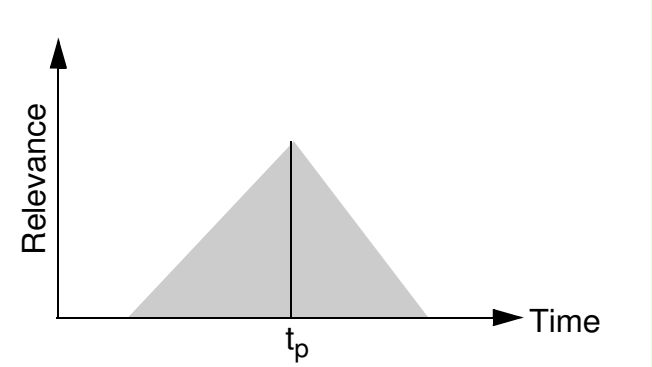
Additional ranking factors:

- Full word match
  (prefix ’frank’ in „Frank Walter Steinmeier“ vs. „Frankfurt am Main“
- Overlap of prefix with words in column ’human_readable_name’
  (prefix ’us’ in „USA“ vs. „Usenet“
- Number of words in ’human_readable_name’ column

Extensions I

- Time travel queries
  - Queries restricted to an interval of time
  - Time point queries





  - Restrict suggestions on news documents inside a given time interval (or point)
- Approach:
  - Split precalculated data into „slices“ of one month length
  - Calculation of bigger time intervals based on aggregation over month slices

Extensions II

- Beside entities, also categories can be used as query input.
- Integration of categories (also from wikipedia)
- Categories form a taxonomy
- Quantative aspects
  - ~250.000 categories
  - avg(6.3) categories/entity

Semantic of Categories in Queries

- Input:
  - Entities e_1, ..., e_n
  - Categories c_1, ..., c_m
  - Prefix p (can be an entity or a category)
- Output (Suggestion):
  - **Entities** with prefix p which can be found in news articles which contain
    (1) each given entity (e_1, ..., e_n) and
    (2) at least one entity of each given category (c_1, ..., c_m)
  - **Categories** with prefix p from entities which can be found in news articles which contain
    (1) each given entity (e_1, ..., e_n) and
    (2) at least one entity of each given category (c_1, ..., c_m)

Summary

- Auto-completion system for query input
- Document set with preclassified entities (Disambiguation via AIDA)
- Input can be entities, categories and prefixes
- Entities and categories are from YAGO (base Wikipedia)
- High time constraints (< 0.1 sec.)
- Precalculation of „relatedness“ based on document corpus
- Actual implementation based on relational database

Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Outlook

- Develop sophisticated data-structure to also handle higher volumes of data
- Integration of „normal“ words (not only entities and categories)
- Incremental updates of precalculated data-structure
- Integration of corpus independent knowledge

Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

References

[1] J. Hoffart, D. Milchevski, and G. Weikum. STICS: Searching with Strings, Things, and Cats. Demo at SIGIR 2014, Gold Coast, Australia, 2014.

[2] M. Amir Yosef, J. Hoffart, et. al. AIDA - An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. Proceedings of the 37th International Conference on Very Large Databases, VLDB 2011, Seattle, WA, 2011

[3] Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum. YAGO - A Core of Semantic Knowledge. 16th international World Wide Web conference

[4] Tim van de Cruys; Two Multivariate Generalizations of Pointwise Mutual Information; DiSCo 2011, Oregon, pp 16-20

[5] D. Milne, H. Witten. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links, 2008

[6] J. Hoffart et al. KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. CIKM'12, Maui/USA, 2012