Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

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

May 21 - 26, 2017 - Barcelona, Spain

Exploring the Wikipedia-Graph

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

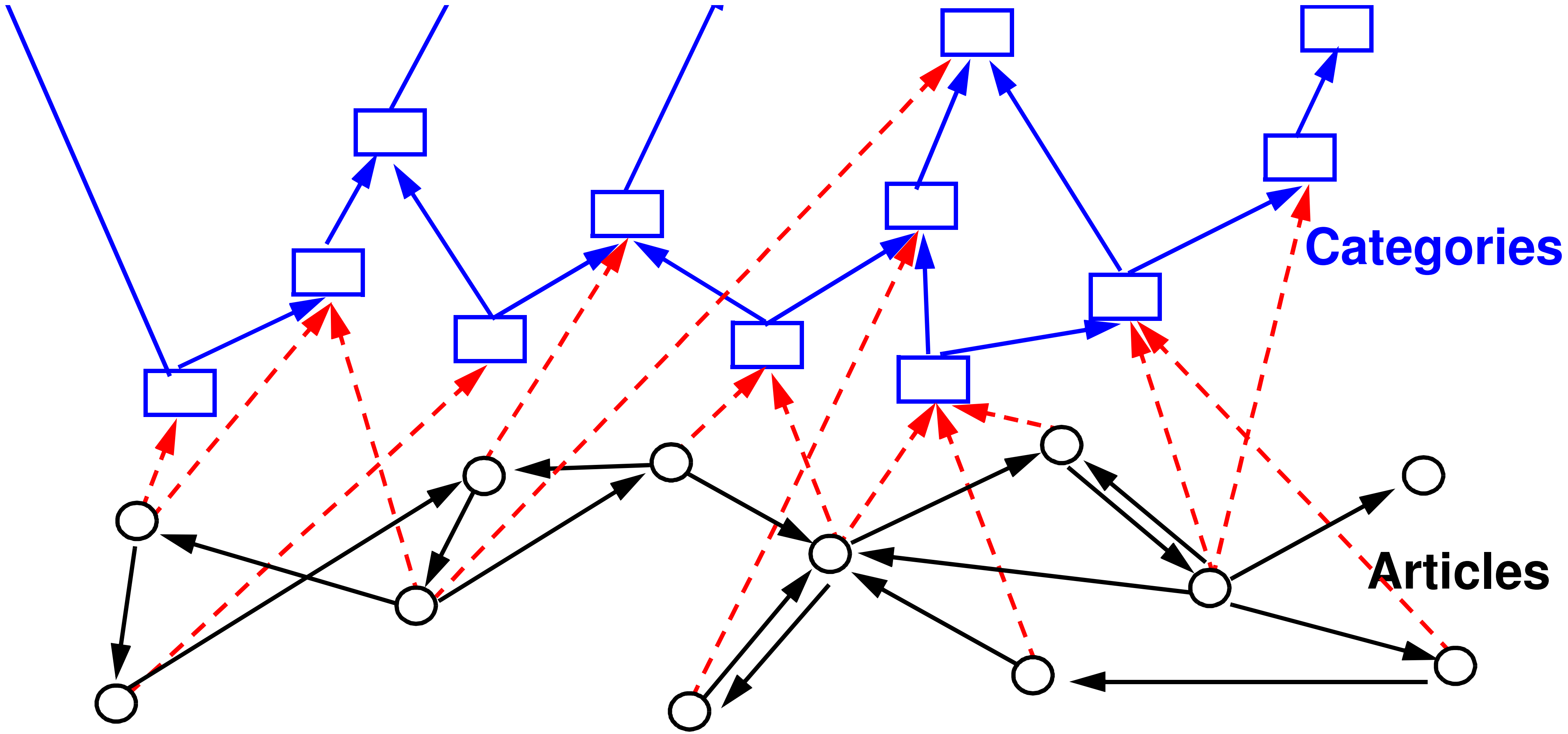
Outlook

- Introduction
- Relatedness Measures
- Concepts
- Examples
- Implementation Aspects
- Summary & Outlook

Wikipedia

- Over 4 million individual articles (english version)
- Wikipedia articles can link to each other
- Each Wikipedia article describes a concrete concept in the real world (Entity)
- Wikipedia categories to classify each article in one or more classes
- Categories form a hierarchy
- Automatic generated pages which list all articles of one category (links)

Wikipedia Structure



Semantic Relatedness between Entities

- Jaccard Koefficient
- Cosine measure in n-dimensional space
- Milne-Witten

Jaccard Koefficient

- Based on the quotient of the cardinality of the intersection and union of two sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

- Example for the calculation of two wikipedia articles:
  - Extract all words from an article
  - Stopword elimination
  - Stem the words and build a set from it
  - Calculation of similarity between two articles based on the cardinality of the intersection and union of two word-sets.

Cosine-Measure

- Each document is represented as an vector
- Vector space defined by language (each word represent a dimension)
- Similarity between two vectors, based on the cosine of the angle between the vectors

$$\cos(\theta) = \frac{a \cdot b}{\|a\| \|b\|} = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i)^2}}$$

- Often combined with tf*idf, to capture different importance of words

tf : Number of times a word t apperars inside a document

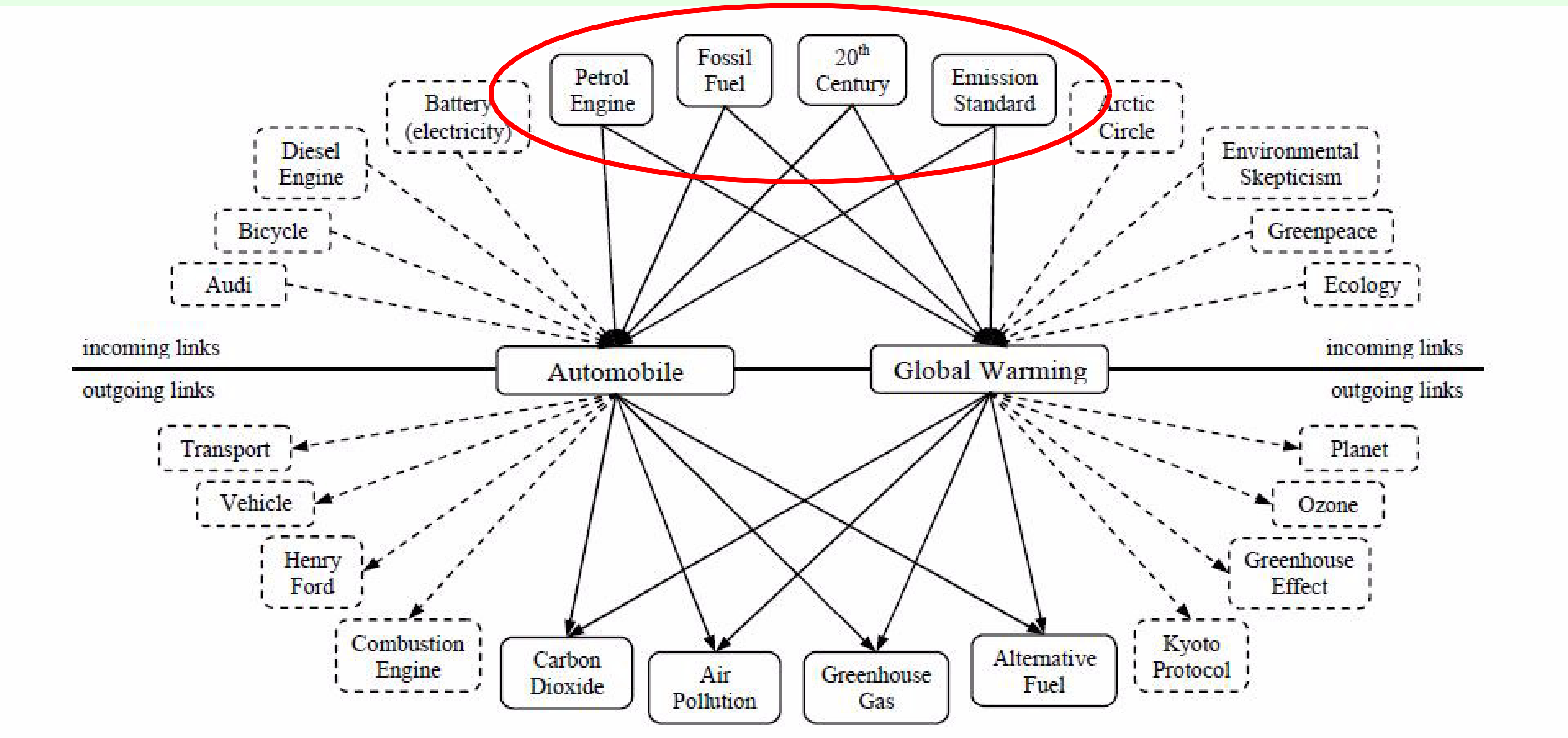
idf_t : $\log(N_d/f_t)$

N_d : Number of documents in the collection

f_t : Number of documents in the collection with term t

Milne Witten [1]

- Use of hyperlink structure in wikipedia to measure semantic relatedness
- Example (from [1]):



Milne Witten [1]

- Measure is based on the weight of a link between articles `s` and `t`
- Measure (normalized google distance):

$$sr(a,b)=\frac{\log(\max(|A|,|B|))-\boxed{\log(|A\cap B|)}}{\log(|W|)-\log(\min(|A|,|B|))}$$

A: Artiles that link to page a
B: Articles that link to page b
W: The set of all wikipedia articles

```
s(a,b) = 0: higly related
         1: not related
```

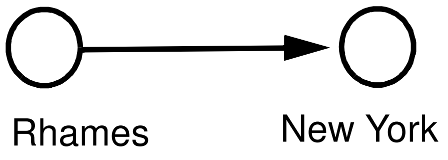
Karlsruhe Institute of Technology

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

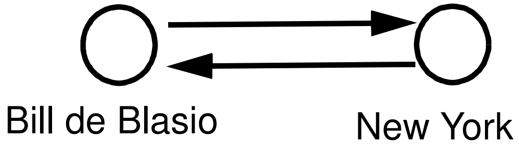
Our Concept

- Uncover hidden relationships between two Entities in Wikipedia
- Relatedness is based on linking structure between article pages
- Examples:

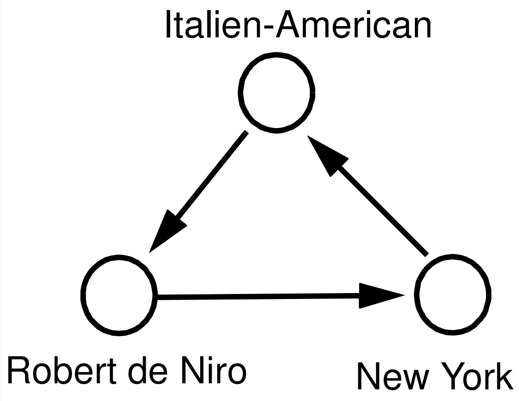
Unidirectional Link



Bidirectional Link

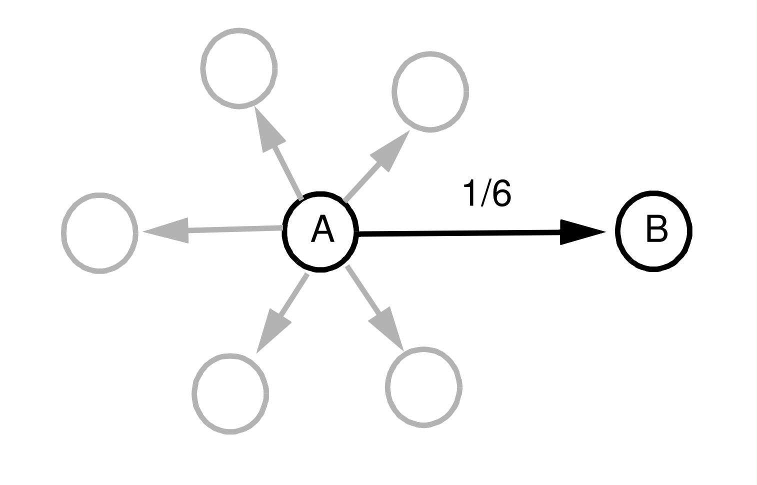


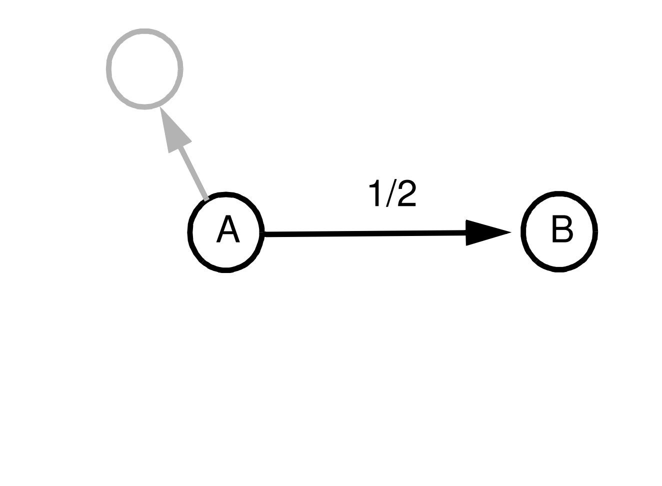
Indirect Backlink



Relevance of a Link

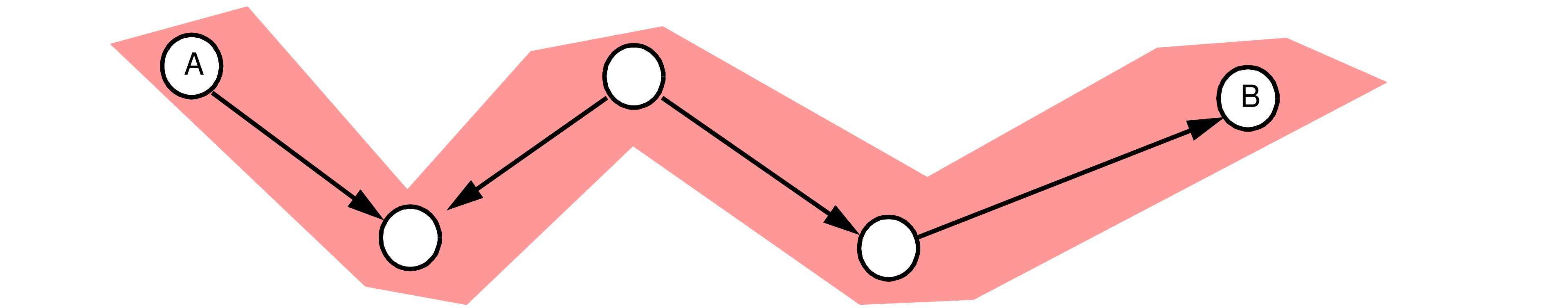
- Relevance of a link is based on the number of further outgoing links



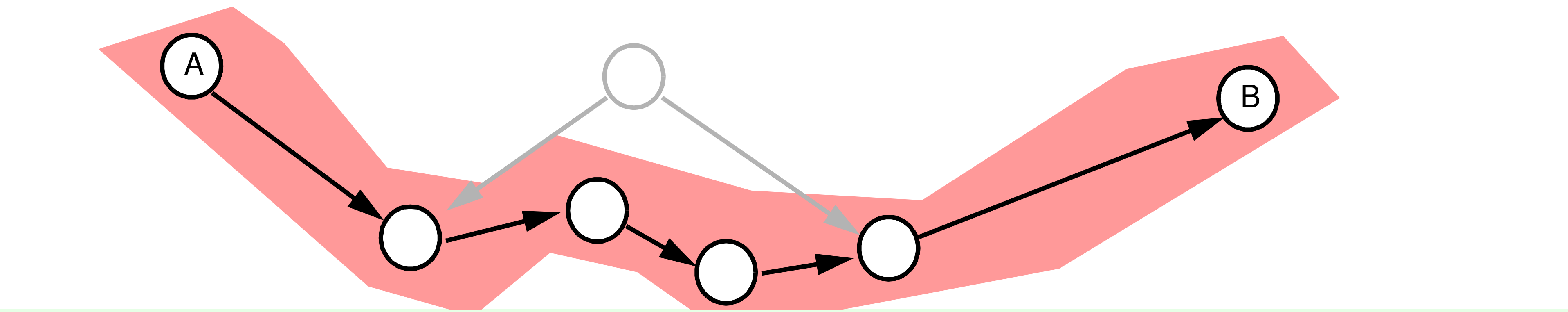


Path Types

- Arbitrary links

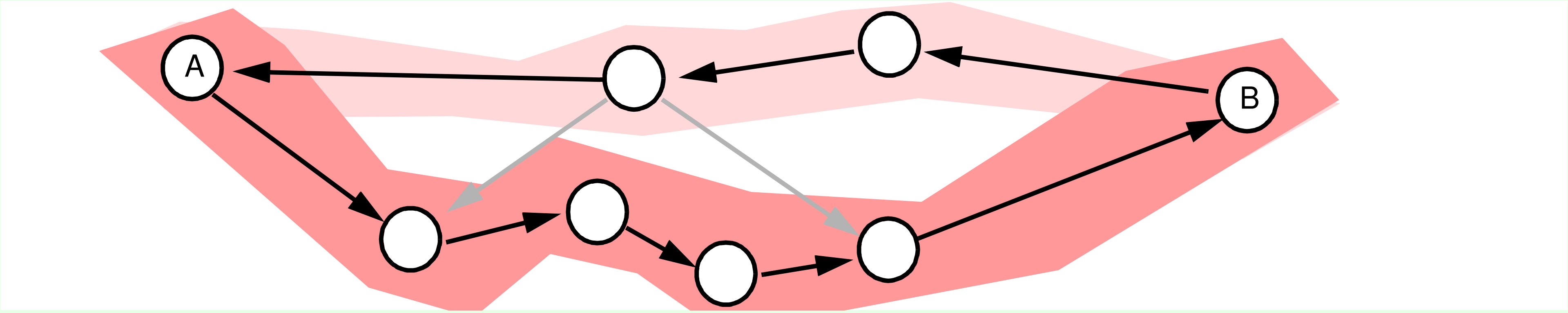


- Directed links

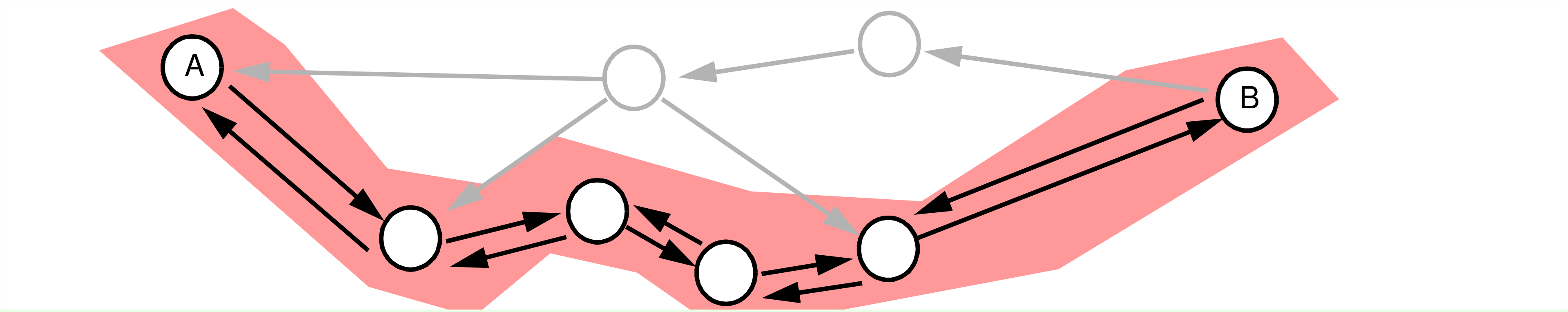


Path Types

- Strong Component Links

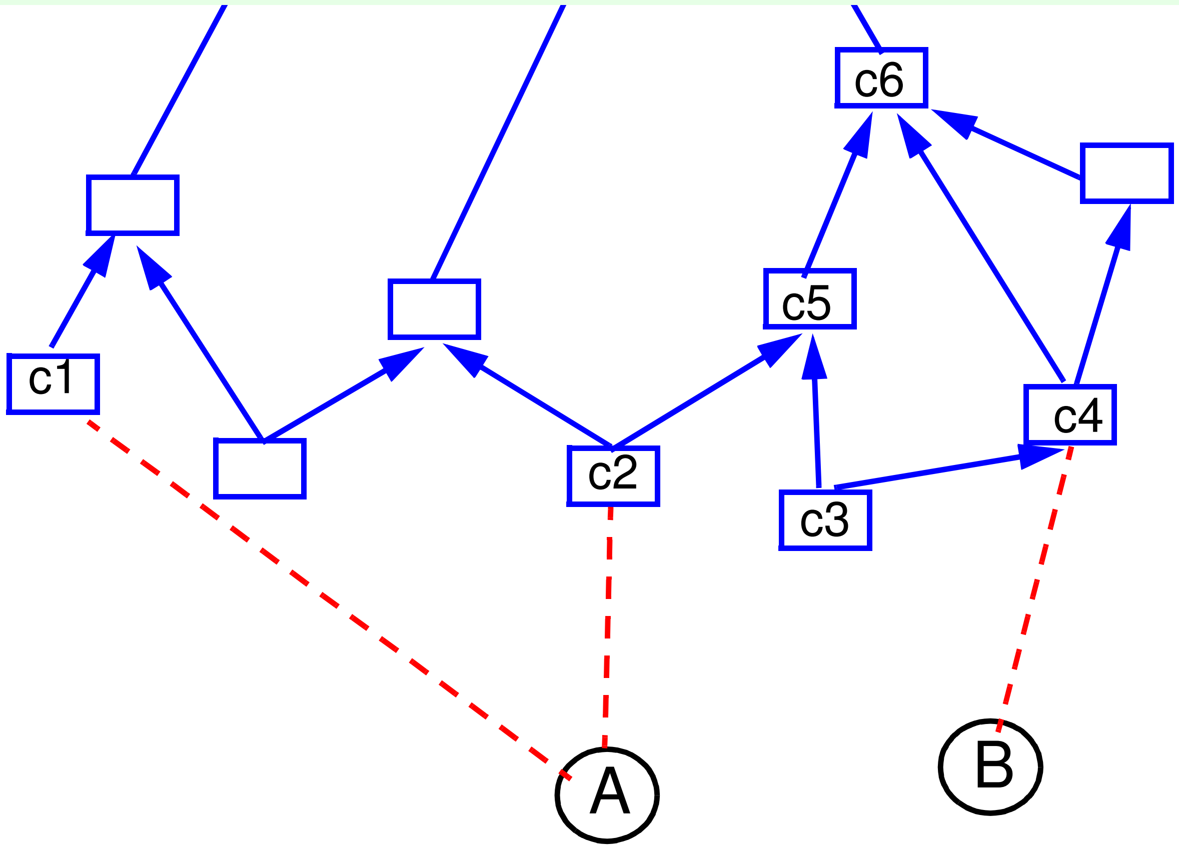


- Sequence of Bidirectional Links



Relevance of a Link

- Nearest Common Super-category

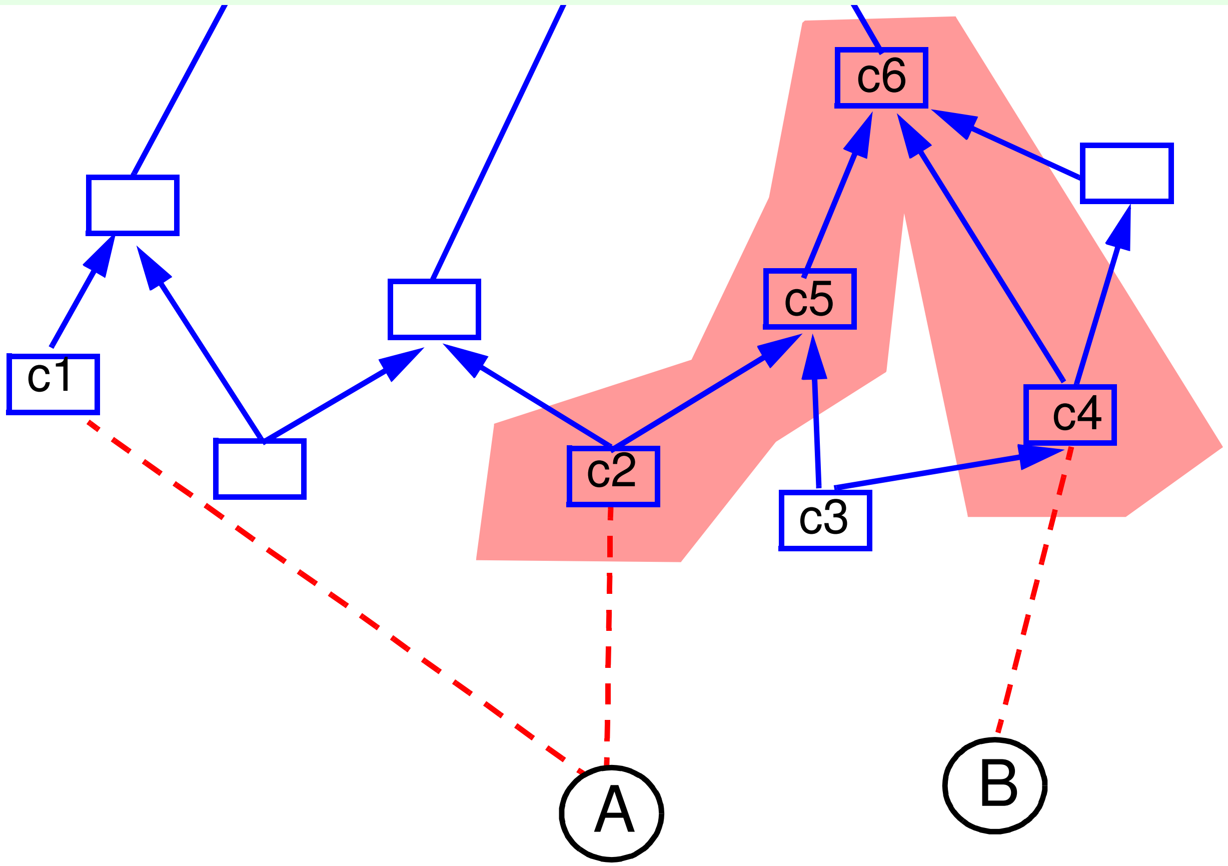


Categories

Articles

Relevance of a Link

- Nearest Common Super-category

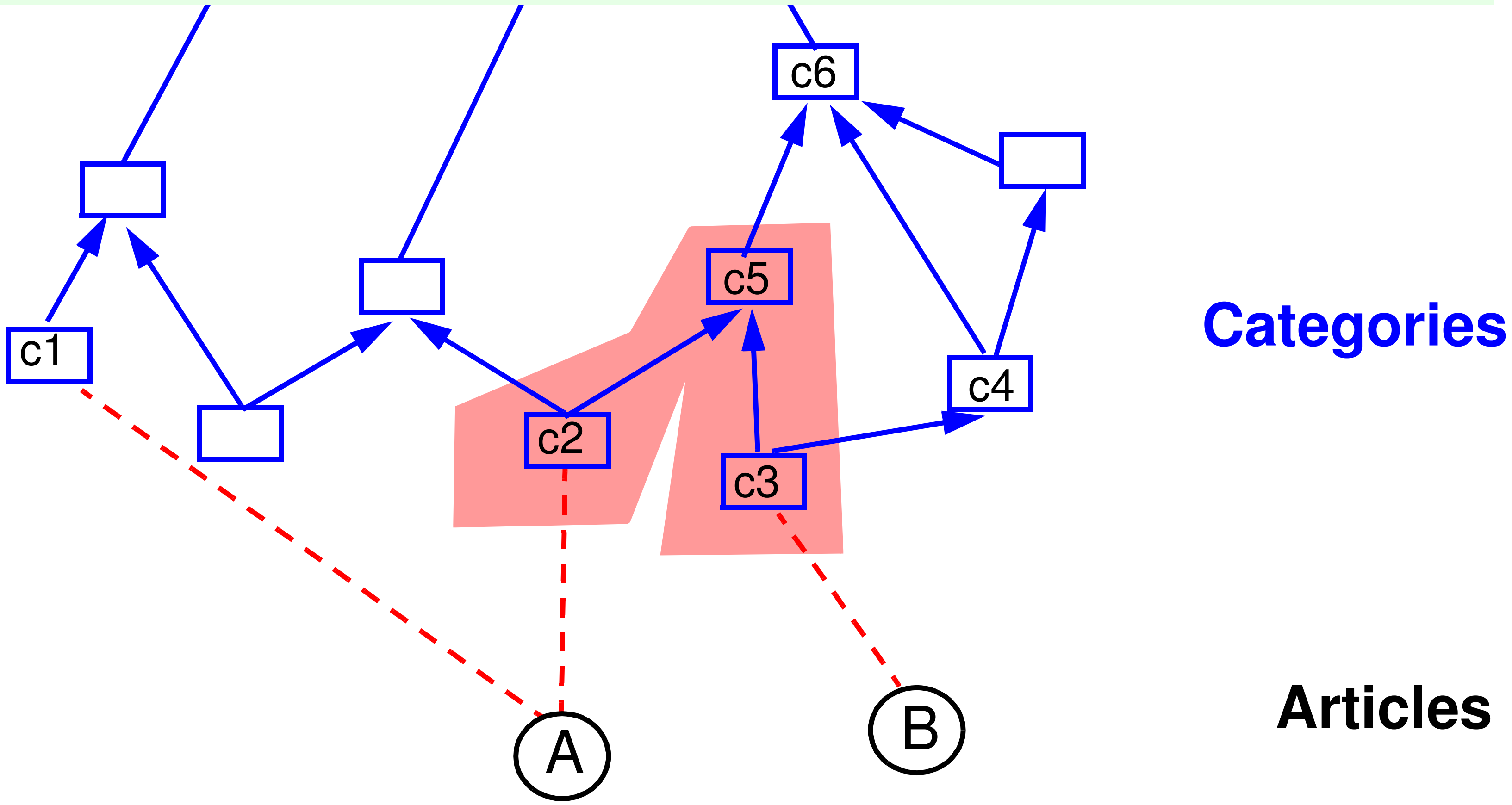


Categories

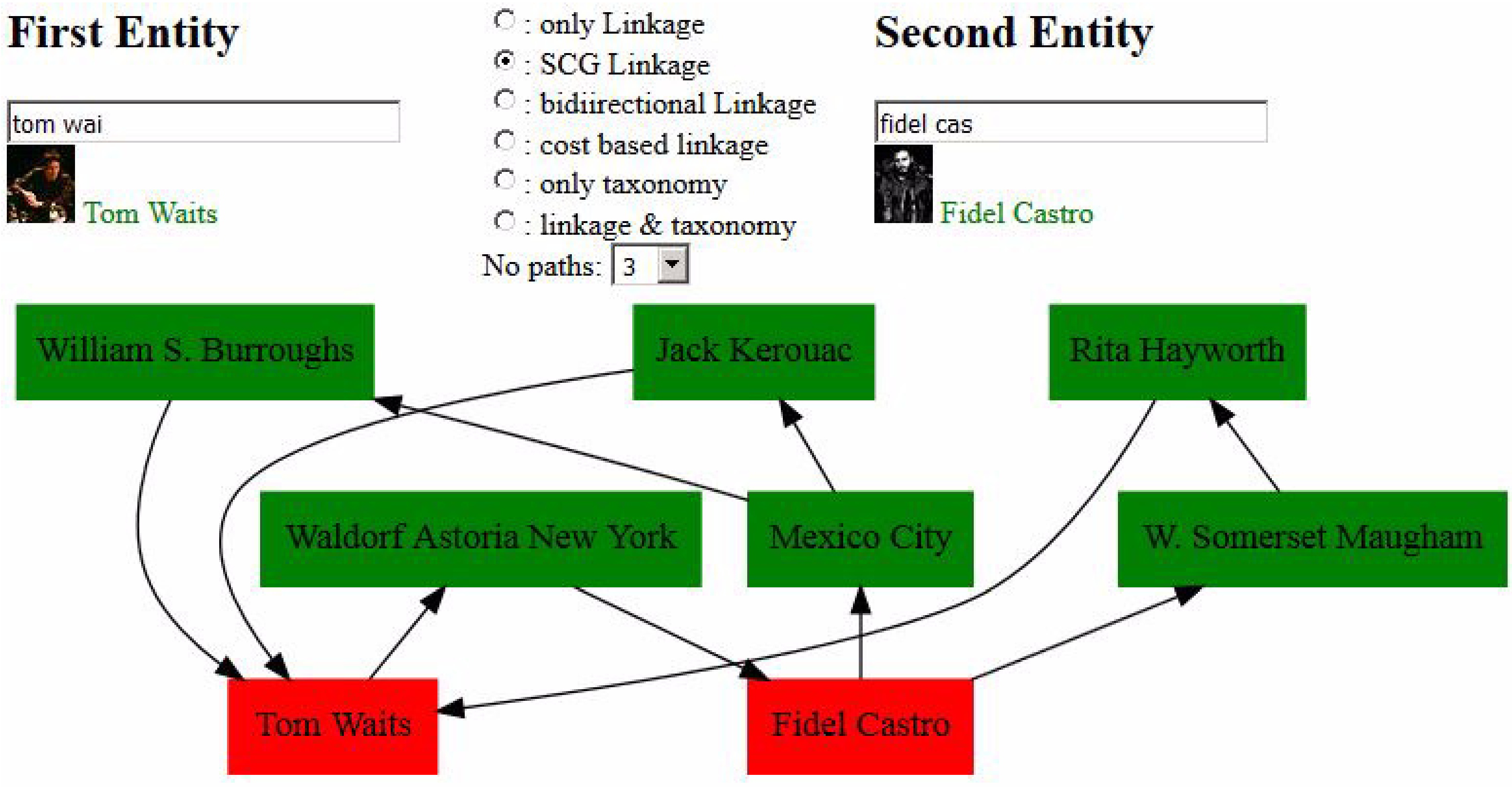
Articles

Relevance of a Link

- Nearest Common Super-category



Examples



First Entity

tom wai

Tom Waits

○ : only Linkage
● : SCG Linkage
○ : bidiirectional Linkage
○ : cost based linkage
○ : only taxonomy
○ : linkage & taxonomy

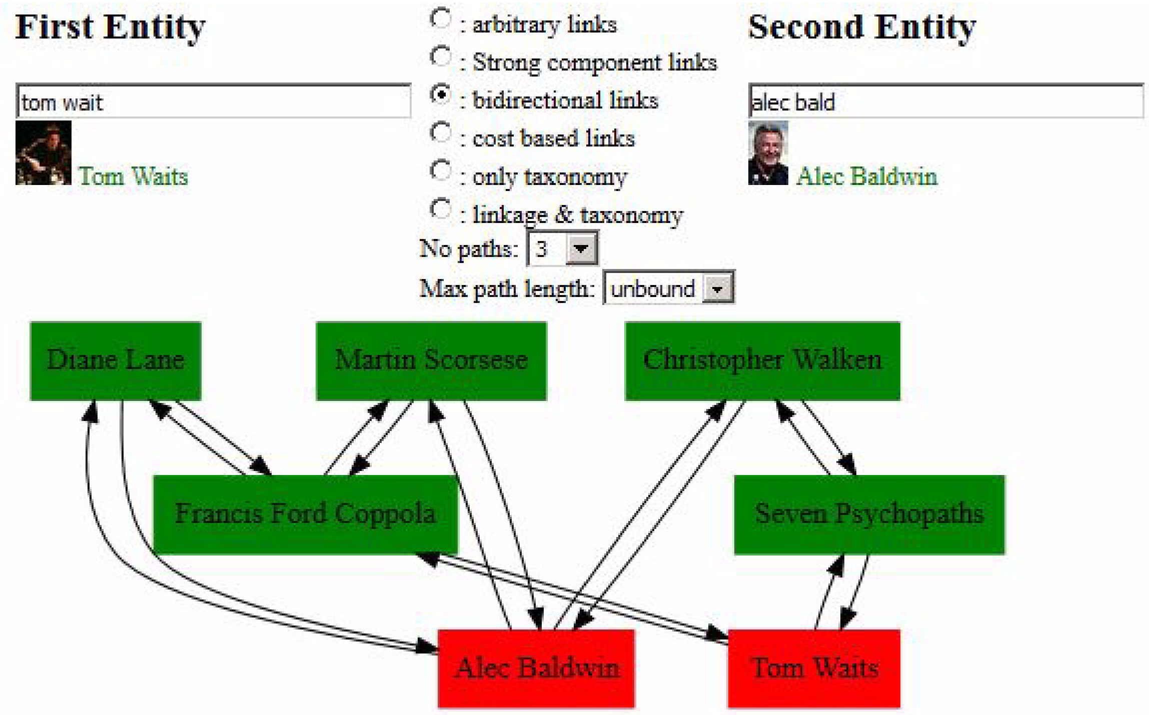
No paths: 3

Second Entity

fidel cas

Fidel Castro

Examples



First Entity

tom wait

Tom Waits

○ : arbitrary links
○ : Strong component links
● : bidirectional links
○ : cost based links
○ : only taxonomy
○ : linkage & taxonomy

No paths: 3

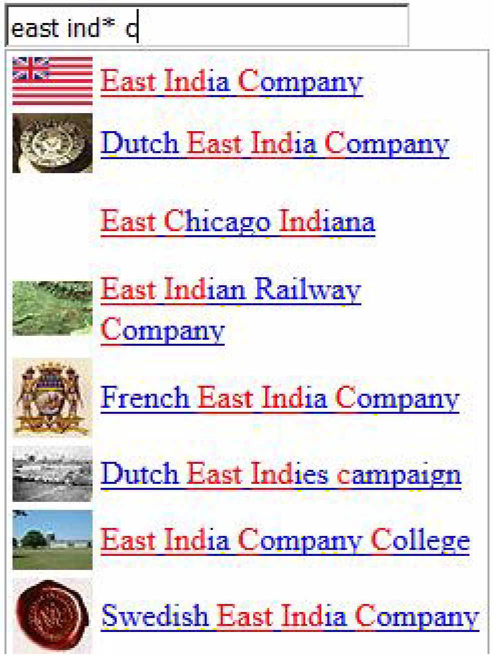
Max path length: unbound

Second Entity

alec bald

Alec Baldwin

Entity Identification



east ind* c

- East India Company
- Dutch East India Company
- East Chicago Indiana
- East Indian Railway Company
- French East India Company
- Dutch East Indies campaign
- East India Company College
- Swedish East India Company

- Mixed Word, Prefix Search
- Last word always considered as a a prefix
- Previous words need a asterisk at the end to be considered as prefix
- Ranking based on
  - global relevance
  - Coverage of words/prefixes
  - Lucene ranking

Quantitative Aspects

- Data ground: English wikipedia (YAGO) [2,3]
- ~4,340,000 entities
- ~83,000,000 links
- Time behaviour: Path of length 12 returned within 1 second.

Implementation Aspects

- Neo4j Database
- Native Java-Api, Traversal API
- Implemented as Unmanaged Server Extension
- Full text search for entity identification based on Lucene index
- Web-based frontend
- Vizualisation using Graphviz [4]

Summary

- Tool for uncovering and vizualisation of relationships between Wikipedia entities
- Using link-structure and classification hierarchy for the calculation of relationships
- Easy selection of entities based on autocompletion mechanism
- Support for different link charateristics
- Graphical vizualisation of link path/classification tree between entities

Literature

- [1] Ian Witten and David Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links." Paper presented at the meeting of the Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA, 2008.
- [2] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA
- [3] YAGO Download, https://www.mpi-inf.mpg.de/de/departments/databases-and-information-systems/research/yagonaga/yago/downloads/, last accessed 11,2,2017
- [4] Graphviz – Graph Visualisation Software. http://www.graphviz.org/, last accessed: 11.2.2017