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)
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

\[
\begin{align*}
\text{tf} & : \text{Number of times a word t apperars inside a document} \\
\text{idf}_t & : \log (\frac{N_d}{f_t}) \\
N_d & : \text{Number of documents in the collection} \\
f_t & : \text{Number of documents in the collection with term t}
\end{align*}
\]
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 \( t \)

\[
sr(a,b) = \frac{\log\left(\max(|A|,|B|)\right) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}
\]

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

\( s(a,b) = 0 \): highly related
\( 1 \): not related
Our Concept

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

Unidirectional Link

Ving Rhames → New York

Bidirectional Link

Bill de Blasio ↔ New York

Indirect Backlink

Italien-American

Robert de Niro → New York
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

Diagram:

- Categories:
  - c1
  - c2
  - c3
  - c4
  - c5
  - c6

- Articles:
  - A
  - B
Relevance of a Link

- Nearest Common Super-category
Relevance of a Link

- Nearest Common Super-category

![Diagram showing categories and links]

Categories

Articles
Examples

First Entity

William S. Burroughs
Waldorf Astoria New York
Tom Waits

Second Entity

Jack Kerouac
Mexico City
Fidel Castro

Rita Hayworth
W. Somerset Maugham

Examples
Examples

First Entity

- Tom Waits
- Francis Ford Coppola
- Diane Lane

Second Entity

- Alec Baldwin
- Tom Waits
- Christopher Walken
- Seven Psychopaths
- Martin Scorsese

Options:
- arbitrary links
- strong component links
- bidirectional links
- cost based links
- only taxonomy
- linkage & taxonomy

No paths: 3
Max path length: unbound
Entity Identification

- Mixed Word, Prefix Search
- Last word always considered as a prefix
- Previous words need an 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
- Visualization using Graphviz [4]
Summary

- Tool for uncovering and visualization 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 characteristics
- Graphical visualization of link path/classification tree between entities
• [1] Ian Witten and David Milne, "An effective, low-cost measure of semantic relat-
edness obtained from Wikipedia links." Paper presented at the meeting of the Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolv-

ence on World Wide Web (WWW '07). ACM, New York, NY, USA

information-systems/research/yagonaga/yago/downloads/, last accessed 11,2,2017