Extraction and Applications of Implicit Networks from Unstructured Text

Andreas Spitz

Heidelberg University, Institute of Computer Science
Database Systems Research Group

spitz@informatik.uni-heidelberg.de

Max Planck Institute for Informatics
Saarbrücken, September 14, 2016

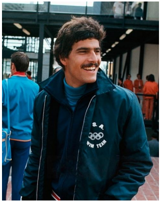
The following is (in part) joint work with:

Johanna Geiß

Michael Gertz

Jannik Strötgen


S. A.
WIM TEAM

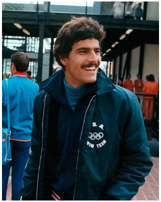


Mark Spitz

From Wikipedia, the free encyclopedia

Mark Andrew Spitz (born February 10, 1950) is an American former competition swimmer, nine-time Olympic champion, and former world record-holder in seven events. He won seven gold medals at the 1972 Summer Olympics in Munich, an achievement surpassed only by Michael Phelps, who won eight golds at the 2008 Summer Olympics in Beijing. Spitz set new world records in all seven events in which he competed in 1972, an achievement that still stands. Since the year 1900, no other swimmer has gained so great a percentage of all the medals awarded for Olympic events held in a single Games.

WIKIPEDIA
The Free Encyclopedia



Mark Spitz

From Wikipedia, the free encyclopedia

Mark Andrew Spitz (born February 10, 1950) is an American former competition swimmer, nine-time Olympic champion, and former world record-holder in seven events. He won seven gold medals at the 1972 Summer Olympics in Munich, an achievement surpassed only by Michael Phelps, who won eight golds at the 2008 Summer Olympics in Beijing. Spitz set new world records in all seven events in which he competed in 1972, an achievement that still stands. Since the year 1900, no other swimmer has gained so great a percentage of all the medals awarded for Olympic events held in a single Games.

WIKIPEDIA
The Free Encyclopedia

Olympia Schwimmhalle

From Wikipedia, the free encyclopedia

The **Olympia Schwimmhalle** is an aquatics centre located in the Olympiapark in Munich, Germany. It hosted the swimming, diving, water polo, and the swimming part of the modern pentathlon events at the 1972 Summer Olympics. At the 1972 Olympics, the stadium had a 9000-seat capacity which was reduced to 1,500 soon after. During the 1972 Olympics, the Olympic Records in all 29 Olympic swimming events were broken as well as the World Records in 20 events.[*citation needed*]

The Schwimmhalle is unique for its roof construction which is a lightweight stressed-skin structure. This curved structure bears loads through tension only, not compression. The double curvature in the roof design is what provides support which is further stabilized through pretensioned guy wires.

The Olympia Schwimmhalle is where swimmer Mark Spitz broke the record for most individual gold medals won in a single Olympics with seven gold medals. This record was not surpassed until fellow swimmer Michael Phelps won eight gold medals at the 2008 Summer Olympics in Beijing.

1972 Summer Olympics

From Wikipedia, the free encyclopedia

The **1972 Summer Olympics** (German: *Olympische Sommerspiele 1972*), officially known as the **Games of the XX Olympiad**, was an international multi-sport event held in Munich, West Germany, from August 26 to September 11, 1972.

Steve Genter

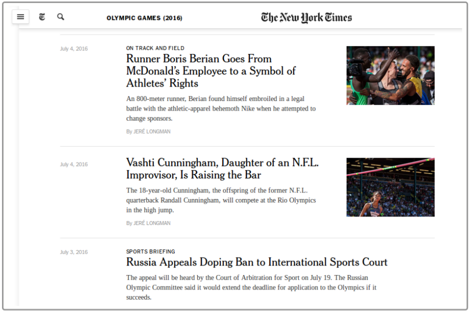
From Wikipedia, the free encyclopedia

Robert Steven Genter (born January 4, 1951) is an American former competition swimmer and three-time Olympic medalist. He was freestyle specialist who earned a gold medal as a member of the winning U.S. team in the 4×200-meter freestyle relay at the 1972 Summer Olympics in Munich, Germany. He also won silver medals in the 200-meter and 400-meter freestyle events.

Swimming at the 1972 Summer Olympics

From Wikipedia, the free encyclopedia

The 1972 Summer Olympics were held in Munich, West Germany, 29 events in **swimming** were contested. There was a total of 532 participants from 52 countries competing.



OLYMPIC GAMES (2016) The New York Times

July 4, 2016

ON TRACK AND FIELD

Runner Boris Berian Goes From McDonald's Employee to a Symbol of Athletes' Rights

An 800-meter runner, Berian found himself embroiled in a legal battle with the athletic-apparel behemoth Nike when he attempted to change sponsors.

By JERÉ LONGMAN

July 4, 2016

Vashti Cunningham, Daughter of an N.F.L. Improvisor, Is Raising the Bar

The 18-year-old Cunningham, the offspring of the former N.F.L. quarterback Randall Cunningham, will compete at the Rio Olympics in the high jump.

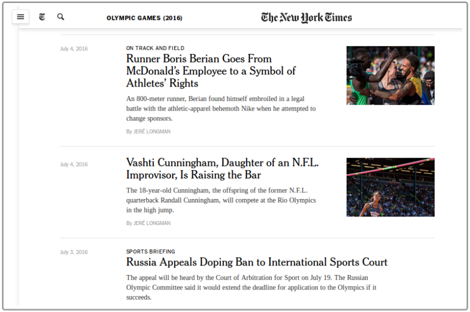
By JERÉ LONGMAN

July 3, 2016

SPORTS BRIEFING

Russia Appeals Doping Ban to International Sports Court

The appeal will be heard by the Court of Arbitration for Sport on July 19. The Russian Olympic Committee said it would extend the deadline for application to the Olympics if it succeeds.


OLYMPIC GAMES (2016) The New York Times

July 4, 2016
ON TRACK AND FIELD
Runner Boris Berian Goes From McDonald's Employee to a Symbol of Athletes' Rights
An 800-meter runner, Berian found himself embroiled in a legal battle with the athletic-apparel behemoth Nike when he attempted to change sponsors.
By JERÉ LONGMAN

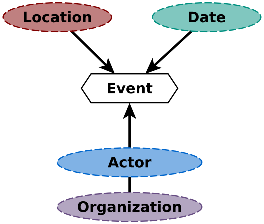
July 4, 2016
Vashti Cunningham, Daughter of an N.F.L. Improvisor, Is Raising the Bar
The 18-year-old Cunningham, the offspring of the former N.F.L. quarterback Randall Cunningham, will compete at the Rio Olympics in the high jump.
By JERÉ LONGMAN

July 3, 2016
SPORTS BRIEFING
Russia Appeals Doping Ban to International Sports Court
The appeal will be heard by the Court of Arbitration for Sport on July 19. The Russian Olympic Committee said it would extend the deadline for application to the Olympics if it succeeds.

Motivation

Definition: Event

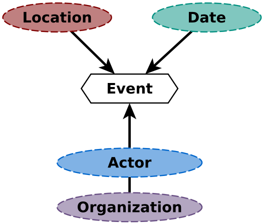
"Something that happens at a given **place** *and* **time** *between a* **group** *of* **actors***."* [CSG+02]



Motivation

Definition: Event

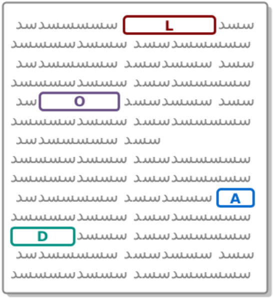
"Something that happens at a given **place** *and* **time** *between a* **group** *of* **actors***."* [CSG+02]



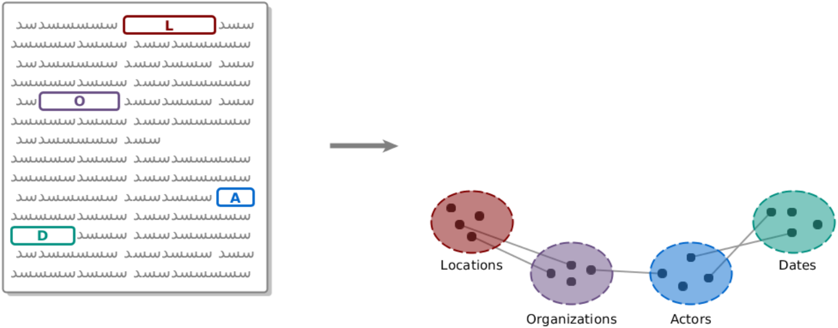
For large document collections,
how can we...

- obtain events from unstructured text?
- identify connections across documents?
- support ad-hoc event search?

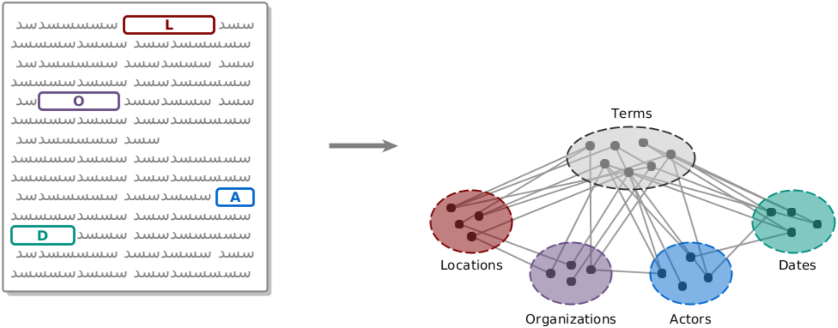
Graph Extraction from Unstructured Text



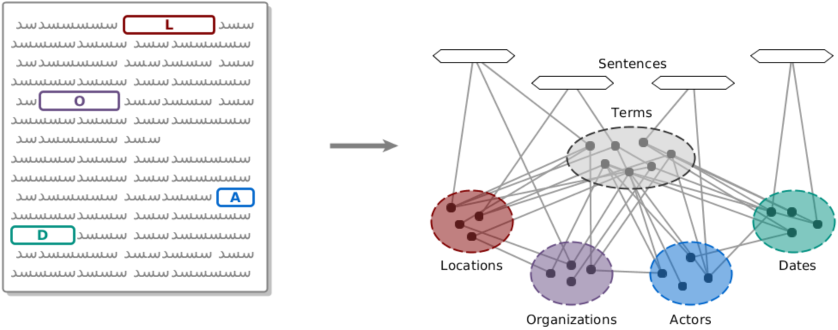
Graph Extraction from Unstructured Text



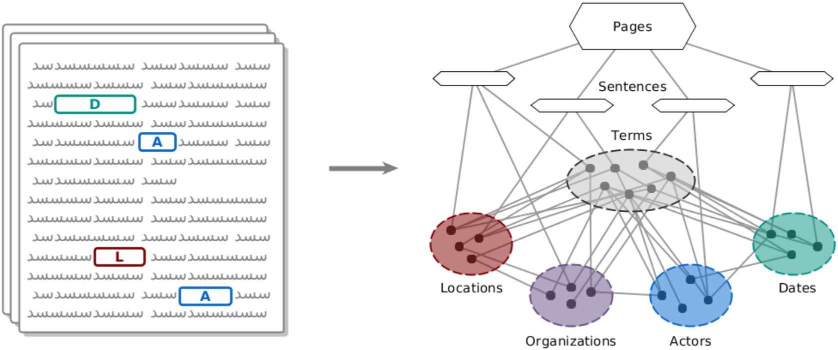
Graph Extraction from Unstructured Text



Graph Extraction from Unstructured Text

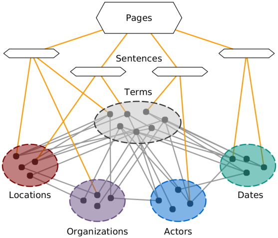


Graph Extraction from Unstructured Text



[SG16]

Edge Weight Generation

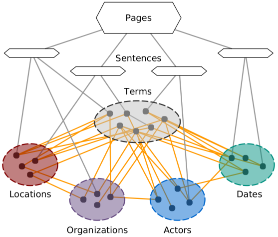


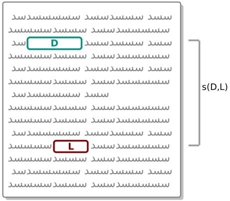
For edges (x, y) for which y is a page or sentence, count only (co-) occurrences:

$$\omega(x,y) = \begin{cases} 1 & \text{if } y \text{ contains } x \\ 0 & \text{otherwise} \end{cases}$$

[SG16]

Edge Weight Generation





For edges (x, y) for which y is a page or sentence, count only (co-) occurrences:

$$\omega(x,y) = \begin{cases} 1 & \text{if } y \text{ contains } x \\ 0 & \text{otherwise} \end{cases}$$

For edges (x, y) between entity types and terms, aggregate co-occurrence instances I: sum over similarities derived from sentence distances s.

$$\omega(x,y) := \sum_{i \in I} \exp(-s(x,y,i))$$

[SG16]

LOADing Wikipedia

For the entire English Wikipedia
($\sim$ 4.5M articles with annotations):

- use only **unstructured** text.
- exclude pages of lists.
- exclude info boxes.
- exclude references.

Extract named entities with:

- Stanford NER for locations, organizations and actors [FGM05]
- Heideltime for dates [SG13]

Wikipedia LOAD Graph

edges	LOC	ORG	ACT	DAT	TER	SEN	PAG
LOC	0						
ORG	91	0					
ACT	276	106	0				
DAT	83	46	128	0			
TER	183	94	317	57	0		
SEN	71	21	84	38	412	0	
PAG	0	0	0	0	0	54	0
nodes	2.7	3.4	7.1	0.2	4.9	53.5	4.5

Number of edges and nodes (in millions) of the LOAD graph of the English Wikipedia. $\sim$ 2B edges and $\sim$ 76M nodes in total.

Single Entity Queries

How can we rank nodes in one set Y by their neighbours in set X?
Adapt *tf-idf* scores to the graph [RV13]:

- Term frequency:
  edge weights
  $tf(x, y) \approx \omega(x, y)$
- Inverse document frequency:
  number of neighbours
  $idf(x) \approx \frac{|Y|}{deg_Y(x)}$

$$r(x, y) \approx \omega(x, y) \log \frac{|Y|}{deg_Y(x)}$$

Single Entity Queries

How can we rank nodes in one set Y by their neighbours in set X?
Adapt *tf-idf* scores to the graph [RV13]:

- Term frequency:
  edge weights
  $tf(x, y) \approx \omega(x, y)$
- Inverse document frequency:
  number of neighbours
  $idf(x) \approx \frac{|Y|}{deg_Y(x)}$

$$r(x, y) \approx \omega(x, y) \log \frac{|Y|}{deg_Y(x)}$$

$\langle LOC : (ACT, \text{Mark Spitz})\rangle$

location	score
munich	1.00000
us	0.70651
states	0.49010
united states	0.46918

Query: $\langle Y : (X, \text{value})\rangle$

Multi-Entity Queries

How can we rank nodes in Y by neighbours in multiple sets X^n?
Combine individual set scores:

$$r(\vec{x}, y) := \frac{1}{n}\eta(\vec{x}, y)\sum_{i=1}^{n} r(x_i, y)$$

Multi-Entity Queries

How can we rank nodes in Y by neighbours in multiple sets X^n?
Combine individual set scores:

$$r(\vec{x}, y) := \frac{1}{n} \eta(\vec{x}, y) \sum_{i=1}^{n} r(x_i, y)$$

Ensure triangular cohesion when combining results:

$$\eta(\vec{x}, y) := \begin{cases} 1 & \text{if } \sum_{i=1}^{n} \sum_{j>i}^{n} M_{yx_i} M_{yx_j} > 1 \\ 0 & \text{otherwise} \end{cases}$$

Where M is the adjacency matrix of the graph.

Summarization: Sentence Queries

How can sentences in S be used to describe combinations of entities in X^n?

Find a sentence that contains them:

$$r(\vec{x}, s) := \sum_{i=1}^{n} M_{sx_i}$$

Summarization: Sentence Queries

How can sentences in S be used to describe combinations of entities in X^n?

Find a sentence that contains them:

$$r(\vec{x}, s) := \sum_{i=1}^{n} M_{sx_i}$$

$\langle SEN : (ACT,$ Mark Spitz$)\rangle$

Mark Spitz of the United States had a spectacular run, lining up for seven events, winning seven Olympic titles and setting seven world records.

Entity Linking: Document Queries

Since we created the LOAD graph from Wikipedia, can we link entities in X^n to pages P?

Use sentences to find the page that contains them most frequently:

$$r(\vec{x}, p) := \sum_{s \in S} \sum_{i=1}^{n} M_{sx_i} M_{sp}$$

Entity Linking: Document Queries

Since we created the LOAD graph from Wikipedia, can we link entities in X^n to pages P?

Use sentences to find the page that contains them most frequently:

$$r(\vec{x}, p) := \sum_{s \in S} \sum_{i=1}^{n} M_{sx_i} M_{sp}$$

$\langle PAG : (ACT, \text{Mark Spitz}) \rangle$

Wiki page ID 66265: Mark Spitz

Sentence and Document Queries

$\langle SEN : (ACT, \text{Mark Spitz})\rangle$

Mark Spitz of the United States had a spectacular run, lining up for seven events, winning seven Olympic titles and setting seven world records.

Sentence and Document Queries

$\langle SEN : (ACT,$ Mark Spitz$)\rangle$

Mark Spitz of the United States had a spectacular run, lining up for seven events, winning seven Olympic titles and setting seven world records.

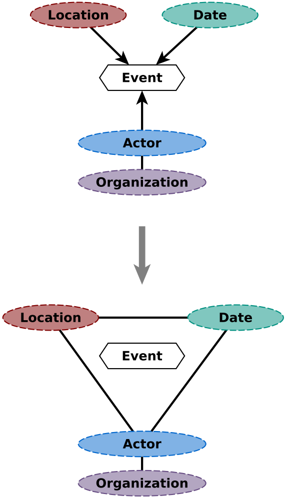
$\langle PAG : (ACT,$ Mark Spitz$)\rangle$

Wiki page ID 66265: Mark Spitz

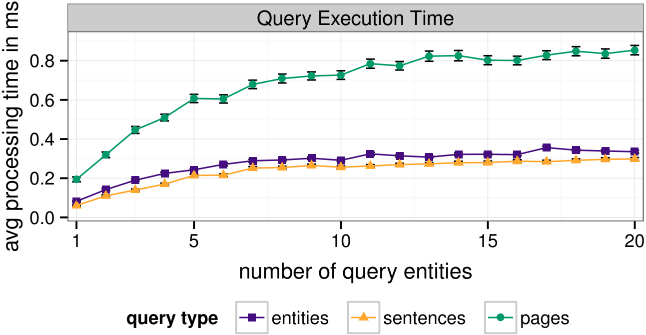
Event Extraction and Completion

Intuition:

- Events correspond to triangular structures in the network
- Participating entities can be used to complete events



Query Answering Speed

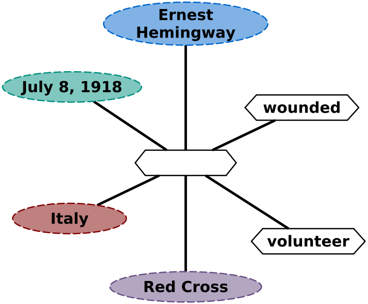


Asymptotic complexity of entity queries: $\mathcal{O}(deg_X(y)\ deg_Y(x))$

Historic Event Evaluation Data

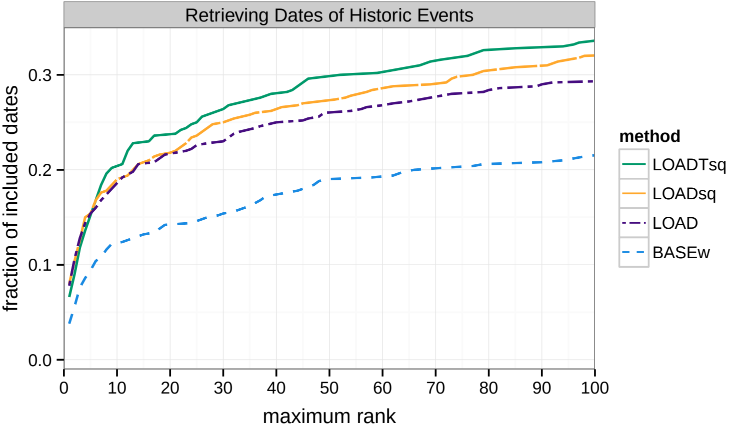
Evaluation data set from a "This Day in History" website

- old enough to not contain Wikipedia data
- exactly one date per sentence
- 500 hand-annotated historic events
- example: Ernest Hemingway, Red Cross volunteer, wounded in Italy on 1918-07-08.



[SG16]

Evaluation on Historic Event Data



LOAD Network: Summary

The Good:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and parallelizable

LOAD Network: Summary

The Good:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and parallelizable

The Bad (i.e. ongoing work):

- no streaming data support (yet)
- entity triangles $\neq$ events: requires filtering

LOAD Network: Summary

The Good:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and parallelizable

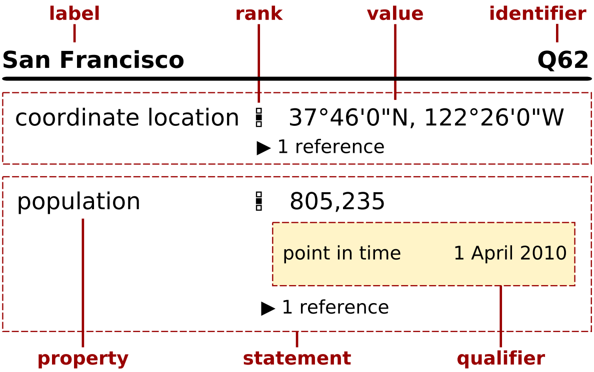
The Bad (i.e. ongoing work):

- no streaming data support (yet)
- entity triangles $\neq$ events: requires filtering

The Ugly:

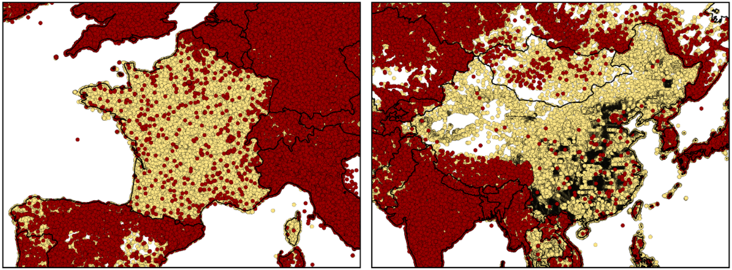
- strong dependence on quality of NER

Adding Knowledge Base Support: Wikidata



Named Entity Extraction in Wikipedia & Wikidata

Wikidata Challenges: Location, Location, Location



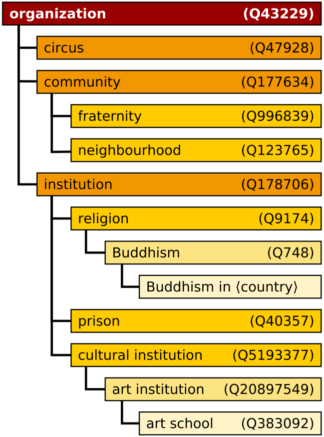
Coverage comparison of populated places in GeoNames (yellow) and human settlements in Wikidata (red).

[SDR+16]

Wikidata Challenges: Organizational Issues

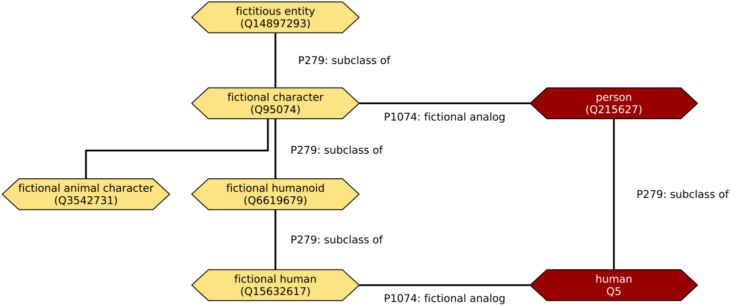
Subclasses of organization (Q43229)

- overlap with locations
  (company headquearters)
- overlap with persons
  (small architecture and law firms)
- form a complicated hierarchy
  that is difficult to clean



[SDR+16]

Wikidata Challenges: Actors Acting Up

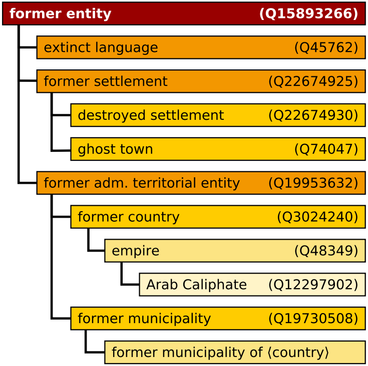


[SDR+16]

Wikidata Challenges: In Times Gone By

Subclasses of former entity:

- discretize time
- hard-code temporal information
- create classes that are perpetually in the past



[SDR+16]

Summary: Wikidata Supported NER in Wikipedia

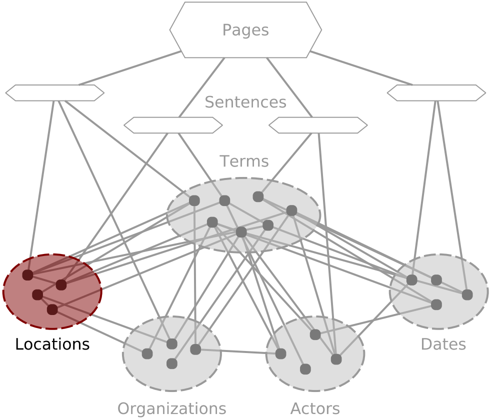
Challenges:

- complicated, evolving hierarchies
- hard-coded, discretized information
- achieving full coverage in NER is difficult
- limited to Wikipedia as a source of text

Benefits:

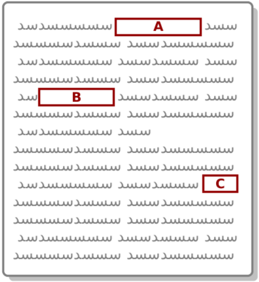
- easy entity extraction
- easy entity linking
- creates a language-agnostic LOAD network from Wikipedia

Location Subnetwork

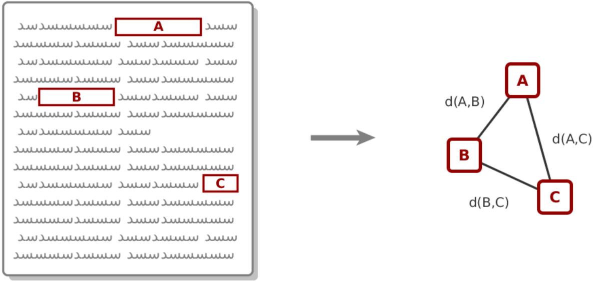


[SGG16, GSSG15]

Graph Extraction from Text



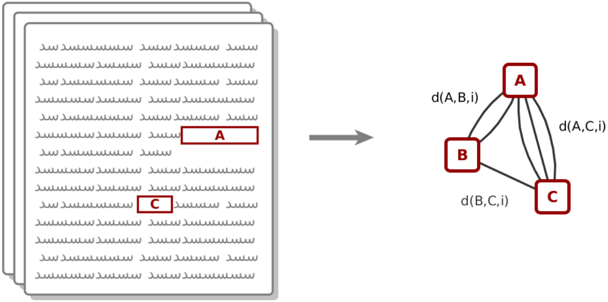
Graph Extraction from Text



$s(v, w) :=$ distance in sentences between toponyms v and w

$$d(v, w) := \exp\left(-\frac{s(v, w)}{2}\right)$$

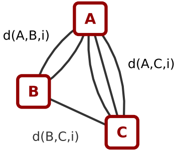
Graph Extraction from Text



$s(v, w)$:= distance in sentences between toponyms v and w

$$d(v, w) := \exp\left(-\frac{s(v, w)}{2}\right)$$

Edge Aggregation



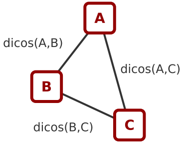
instances i

	1	2	3	4	5	6
A	d_1	d_2	d_3	d_4	d_5	0
B	d_1	d_2	0	0	0	d_6
C	0	0	d_3	d_4	d_5	d_6

dicos

Distance-based cosine for nodes v and w:

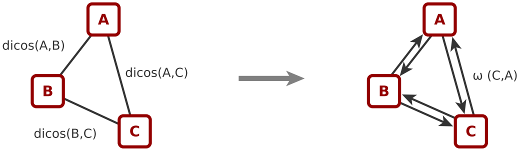
$$dicos(v,w) := \frac{\sum_i d_i(v)\, d_i(w)}{\sqrt{\sum_i d_i(v)^2}\sqrt{\sum_i d_i(w)^2}}$$



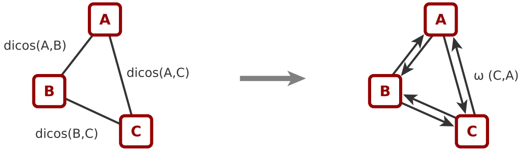
Nonreciprocal Relationships

Dirk Beyer, Wikimedia Commons

Inducing Edge Directions



Inducing Edge Directions



Normalize weights of outgoing edges:

$$\omega(v \rightarrow w) := \frac{dicos(v,w)}{\sum_{x \in V} dicos(v,x)}$$

Network Overview

Network statistics:

$\lvert V \rvert$	$\lvert E \rvert$	density	clustering coefficient
$723,779$	$178,890,238$	$6.8 \cdot 10^{-4}$	0.56

Node types:

Location types:

- continent
- country
- river
- city
- mountain range
- POI
- mountain
- other

Network Overview

Network statistics:

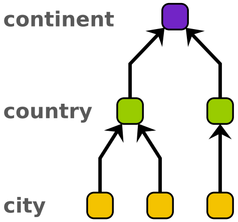
$\|V\|$	$\|E\|$	density	clustering coefficient
$723,779$	$178,890,238$	$6.8 \cdot 10^{-4}$	0.56

Node types:

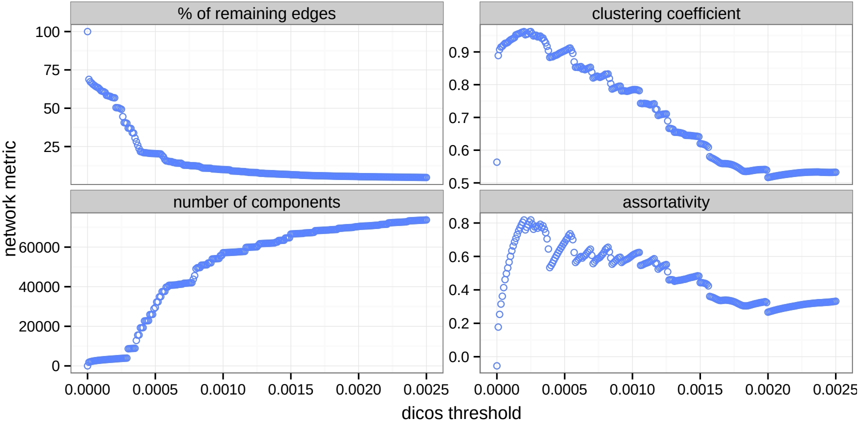
Location types:

- continent
- river
- mountain range
- mountain
- country
- city
- POI
- other

Wikidata location hierarchy:



Network Properties

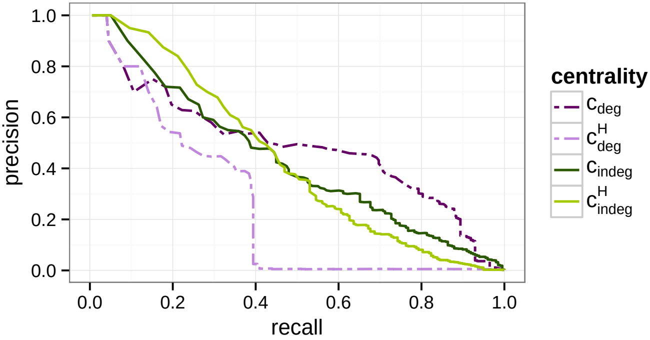


Network Centrality

city	c_{deg}	c_{indeg}	c_{deg}^{H}	c_{indeg}^{H}
Paris	63,150	89.87	8,064	7.56
New York City	79,398	71.74	9,294	12.12
Chicago	54,217	51.84	8,074	7.70
Los Angeles	49,961	51.47	7,276	7.76
Washington, D.C.	62,858	51.05	8,138	8.65
Boston	45,895	50.43	6,121	6.08
Philadelphia	51,237	45.19	6,372	5.03
Vienna	35,724	44.55	4,827	7.44
Moscow	29,026	43.77	4,644	19.47
San Francisco	43,759	40.87	6,029	4.76

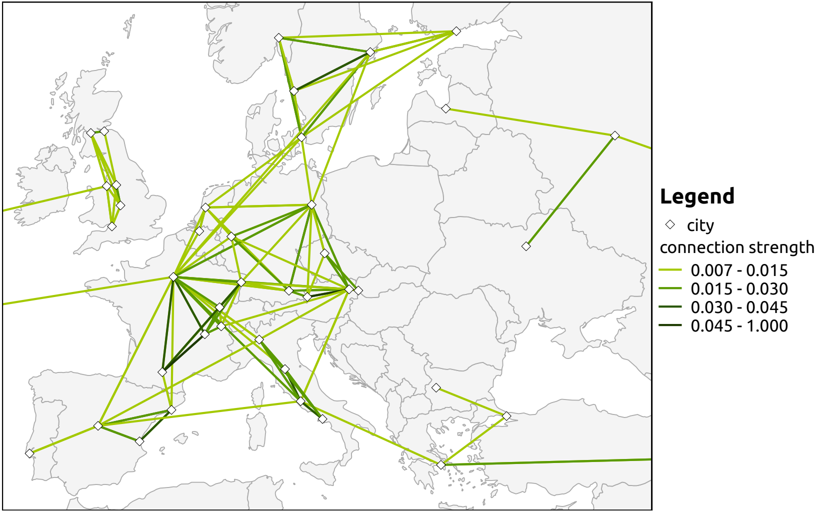
Network between the top 10 European cities by in-degree centrality.

Centrality-Based Hierarchy Classification



Classification into classes *country* and *city* based on centrality.

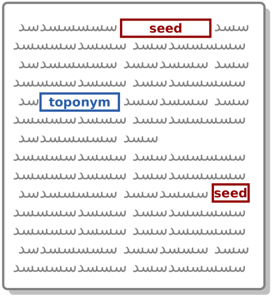
Geographically Embedded Network



Disambiguation Problem

Locations of towns and cities with the name *Heidelberg*.

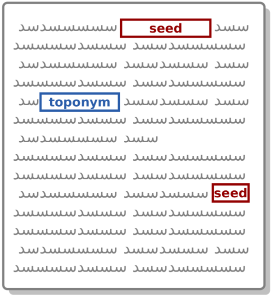
Network-based Toponym Disambiguation



Given a document with toponyms, the following information is available:

- a set of locations L in the network
- a set of seeds $S \subseteq L$ in the document (unambiguous toponyms)
- an ambiguous toponym t in the document with candidates $l \in L$

Network-based Toponym Disambiguation



Given a document with toponyms, the following information is available:

- a set of locations L in the network
- a set of seeds $S \subseteq L$ in the document (unambiguous toponyms)
- an ambiguous toponym t in the document with candidates $l \in L$

Resolve toponyms by their neighbourhood in the network:

$$\text{resolve}(t) := \arg\max_{l \in L} \sum_{s \in S} \omega(l, s)$$

Evaluation on AIDA CoNLL-YAGO data set

	Precision in %			mean distance in km		
	all	seeds	ambig.	all	seeds	ambig.
WLND	**85.7**	86.0	**85.6**	**327.5**	522.9	**179.1**
AIDA	84.9	86.0	83.2	*120.4*	*87.7*	*142.3*
B_{DIST}	81.6	86.0	78.5	683.1	522.9	800.8
B_{MIN}	81.4	86.0	78.8	650.9	522.9	745.0

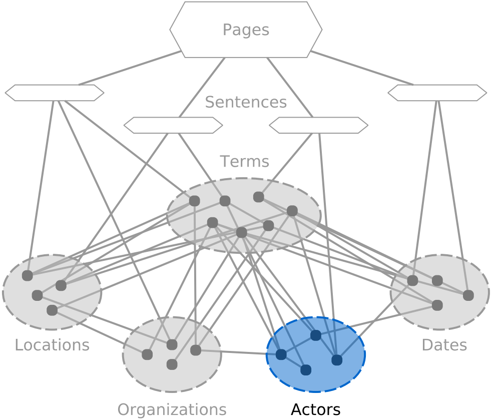
WLDN	Wikipedia Location Network disambiguation
AIDA	AIDA named entity disambiguation
B_{DIST}	Baseline using minimum geographic distance
B_{MIN}	Baseline using lowest Wikidata ID

Location Network Summary

Refined method for implicit network extraction:

- improves the weighting scheme (dicos),
- includes direction for edges,
- supports disambiguation and entity linking,
- is language-agnostic and supports alternative spellings

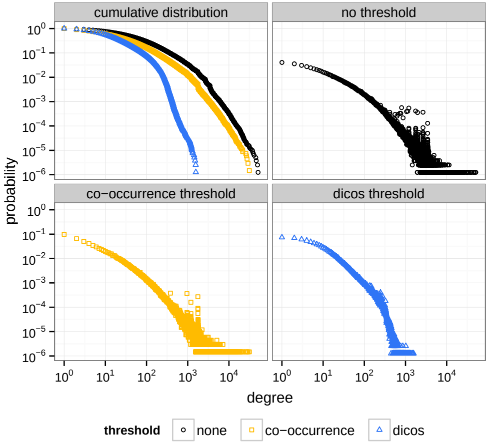
Social Subnetwork



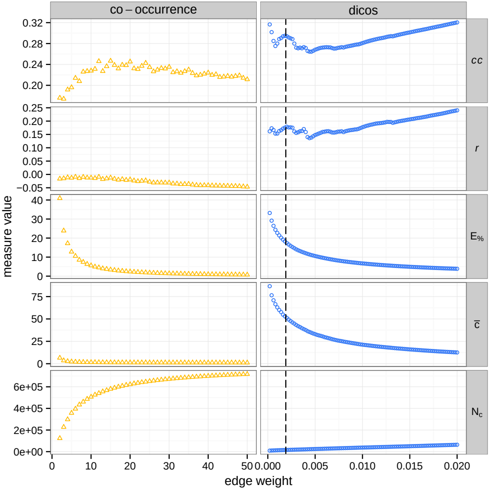
[GSG15]

(Un-) Availability of Social Network Data

Wikipedia Social Network: Topology



Wikipedia Social Network: Metrics

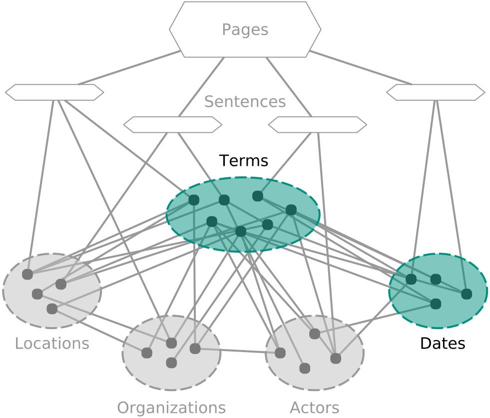


Summary Social Network

Benefits of an implicit social network from Wikipedia:

- large-scale social network based on real persons
- entity linking adds personal information
- stand-in data set for unavailable online social networks

Temporal Subnetwork



[SSBG15]

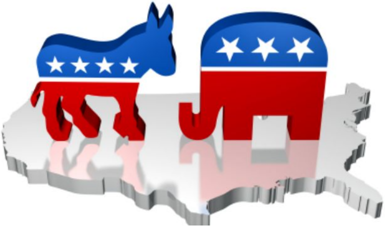
Date Similarity: U.S. Elections Days

Date similarities:

- can we recognize dates with similar content?

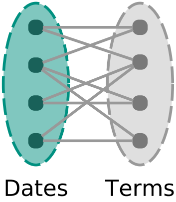
Example: U.S. Election days

- Always on the Tuesday after the first Monday in November
- Every four years: presidential Election Day



Predicting U.S. Elections Days

Model: bipartite graph

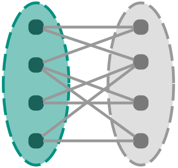


Prediction:

- Collaborative Filtering
- For example: cosine similarity of adjacencies

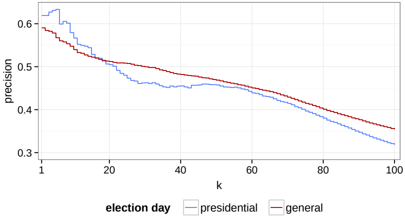
Predicting U.S. Elections Days

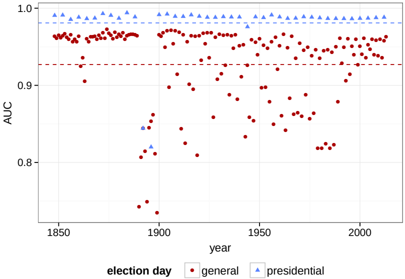
Model: bipartite graph



Prediction:

- Collaborative Filtering
- For example: cosine similarity of adjacencies





Summary: Implicit Textual Networks

LOAD network:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and fast
- language-agnostic with entity linking

Entity-based subnetworks of LOAD:

- flexible selection / extraction for individual tasks
- allow more involved weighting (edge direction, dicos)

Summary: Implicit Textual Networks

LOAD network:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and fast
- language-agnostic with entity linking

Entity-based subnetworks of LOAD:

- flexible selection / extraction for individual tasks
- allow more involved weighting (edge direction, dicos)

LOAD your data for entity-based analyses.

Available for download:

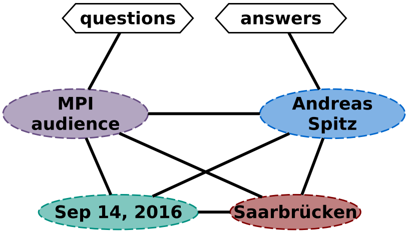
- Wikipedia LOAD networks
- Social and location subnetworks
- Code for generating LOAD networks
- Code for LOAD query interface

http://dbs.ifi.uni-heidelberg.de/index.php?id=load
http://dbs.ifi.uni-heidelberg.de/index.php?id=data

Available for download:

- Wikipedia LOAD networks
- Social and location subnetworks
- Code for generating LOAD networks
- Code for LOAD query interface

http://dbs.ifi.uni-heidelberg.de/index.php?id=load
http://dbs.ifi.uni-heidelberg.de/index.php?id=data



Bibliography I

Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman.
Corpora for topic detection and tracking.
In *Topic Detection and Tracking*. Springer, 2002.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning.
Incorporating non-local information into information extraction systems by Gibbs sampling.
In *ACL*, 2005.

Johanna Geiß, Andreas Spitz, and Michael Gertz.
Beyond friendships and followers: The Wikipedia social network.
In *ASONAM*, 2015.

Johanna Geiß, Andreas Spitz, Jannik Strötgen, and Michael Gertz.
The Wikipedia location network - overcoming borders and oceans.
In *GIR*, 2015.

François Rousseau and Michalis Vazirgiannis.
Graph-of-word and TW-IDF: new approach to ad hoc IR.
In *CIKM*, 2013.

Bibliography II

Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz, and Johanna Geiß.
State of the union: A data consumer's perspective on Wikidata and its properties for the classification and resolution of entities.
In *WikiWorkshop with ICWSM*, 2016.

Jannik Strötgen and Michael Gertz.
Multilingual and cross-domain temporal tagging.
Language Resources and Evaluation, 47(2):269–298, 2013.

Andreas Spitz and Michael Gertz.
Terms over LOAD: Leveraging named entities for cross-document extraction and summarization of events.
In *SIGIR*, 2016.

Andreas Spitz, Johanna Geiß, and Michael Gertz.
So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks.
In *GeoRich*, 2016.

Andreas Spitz, Jannik Strötgen, Thomas Bögel, and Michael Gertz.
Terms in time and times in context: A graph-based term-time ranking model.
In *TempWeb*, 2015.