

The Wikipedia Location Network: Overcoming Borders and Oceans

**Johanna Geiß¹, Andreas Spitz¹,
Jannik Strötgen^{1,2}, and Michael Gertz¹**

¹Heidelberg University, Institute of Computer Science
Database Systems Research Group, Heidelberg

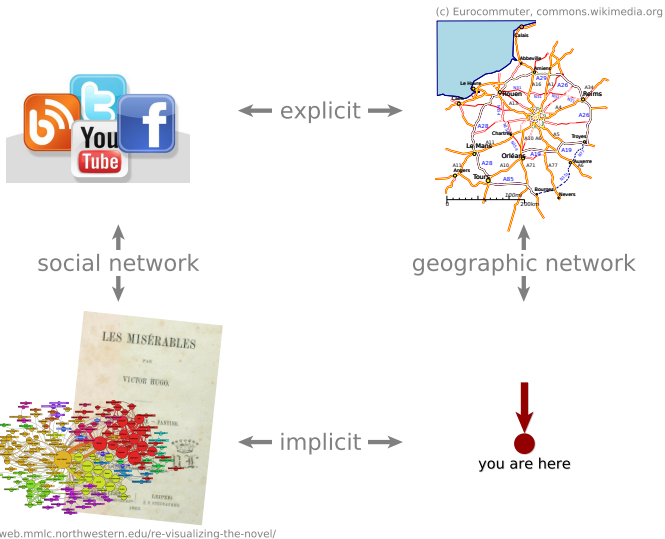
²Max-Planck-Institute for Informatics
Databases and Information Systems, Saarbrücken

{geiss, spitz, stroetgen, gertz}@informatik.uni-heidelberg.de

9th GIR Workshop
Paris, November 26, 2015

What's the difference between France and Illinois?

Implicit Networks



Overview

- ① Motivation
- ② Network Construction
- ③ Properties and Applications
- ④ Summary

Foundations of Implicit Networks

*“Most of the circuits currently in use are specially constructed for competition. The current street circuits are **Monaco, Melbourne, Montreal, Singapore and Sochi**, although races in other urban locations come and go (**Las Vegas and Detroit**, for example) and proposals for such races are often discussed – most recently **New Jersey**.”*

en.wikipedia.org/wiki/Formula_One

Multi-Graph Extraction



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

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

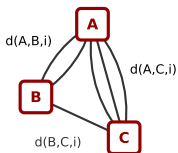
Multi-Graph Extraction



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

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

Edge Aggregation

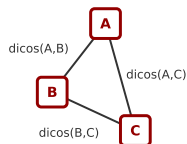


	instances i					
	1	2	3	4	5	6
A	d_1	d_2	d_3	d_4	d_5	0
B	d_1	d_2	0	0	0	d_6
C	0	0	d_3	d_4	d_5	d_6



Distance-based cosine for nodes v and w :

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



Toponym Extraction in Wikipedia



WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikipedia store

Interaction

- Help

Article [Talk](#)

1984 Summer Olympics

From Wikipedia, the free encyclopedia

The **1984 Summer Olympics**, officially known as the **Games of the XXIII Olympiad**, was an international **multi-sport event** held in **Los Angeles, California, United States** in 1984. When **Tehran**, the only other interested city on the international level, declined to bid due to the concurrent Iranian political and social changes, the IOC awarded Los Angeles the Games for the second occasion Los Angeles hosted the Olympics, having first hosted in 1932.

Los Angeles

From Wikipedia, the free encyclopedia

Coordinates: 34°03′N 118°15′W

This article is about the U.S. city. For the county in California, see *Los Angeles County*. For other uses, see *Los Angeles (disambiguation)*.
 "L.A." and "City of Los Angeles" redirect here. For other uses, see *L.A. (disambiguation)* and *City of Los Angeles (disambiguation)*.

Los Angeles (/loʊˈændʒələs/ *and jəl*) or **Los Angeles** or **Los An-ju-liss**),^[14] officially the **City of Los Angeles** and often known by its initials **L.A.**, is the second-largest city in the United States, the most populous city in the U.S. state of California, and the county seat of Los Angeles County. Situated in Southern California, Los Angeles is known for its Mediterranean climate, ethnic diversity, sprawling metropolis, and as a major center of the American entertainment industry. Los Angeles lies in a large coastal basin surrounded on three sides by mountains reaching up to and over 10,000 feet (3,000 m).

Los Angeles, California

Charter city

City of Los Angeles



Tehran

From Wikipedia, the free encyclopedia

Coordinates: 35°41′46″N 51°29′23″E

This article is about the Iranian capital city. For other uses, see *Tehran (disambiguation)*.

Tehran (/tɛhˈrɑːn/; Persian: تهران - *Tehrān*) is the capital of Iran and Tehran Province. With a population of around 9 million in the city and 16 million in the wider metropolitan area,^[5] Tehran is the largest city and urban area of Iran, the second largest city in Western Asia, and the 3rd largest in the Middle East. It is ranked 29th in the world by the population of its metropolitan area.^[6]

In the Classical era, part of the present-day city of Tehran was occupied by a Median city which in the *Avesta* occurs as *Rhaga*.^[7] It was destroyed by the Mongols in the early 13th century, and remains now as a city in Tehran Province, located towards the south end of the modern-day city of Tehran.

Tehran

تهران

Metropolis









کلانشهر تهران · Tehran Metropolis





Network Overview

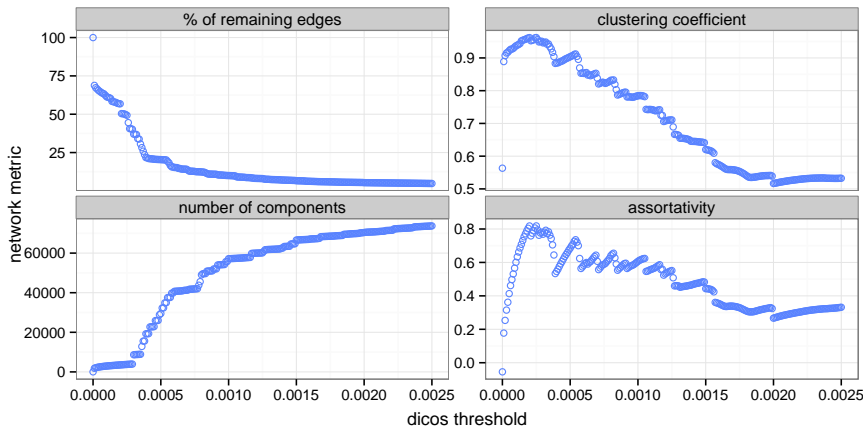
Node types:

Location types:	
 continent	 country
 river	 city
 mountain range	 POI
 mountain	 other

Network statistics:

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

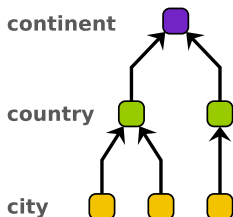
Network Properties



Hierarchical Evaluation

Does the network contain classic geographical relations?

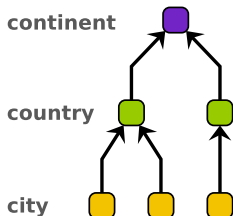
1. Extract hierarchical relations from Wikidata:



Hierarchical Evaluation

Does the network contain classic geographical relations?

1. Extract hierarchical relations from Wikidata:



2. Correspondence of highest weighted incident edge in network with the link to parent in hierarchy:

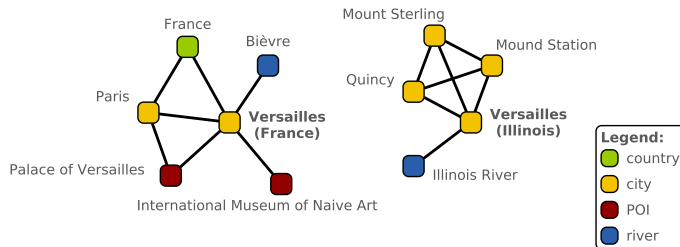
- cities: 81.6% precision for link to parent country
- countries: 80.3% precision for link to parent continent

The Network at a Glance

What's the difference between France and Illinois?

The Network at a Glance

What's the difference between France and Illinois?



Applications in NLP

Support for NLP tasks:

- Disambiguation
- Coreference analysis
- Cross- and multilingual analysis

Data analysis by clustering of the network:

- Finding place similarities
- Categorization of places
- Extraction of hierarchies

Applications in Event Analysis



The network supports spatial components of:

- Event detection
- Event extraction
- Event correlation
- Event similarity

Summary

New method for implicit network extraction that is

- based on text distances of toponyms,
- applicable to any geo-tagged corpus.

Application to Wikipedia / Wikidata results in

- negligible number of mistags,
- accurate and reliable network,
- useful resource for NLP tasks.

Thank you!

Questions?

The Wikipedia Location Network
is available for download.



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

Bibliography



Johanna Geiß, Andreas Spitz, and Michael Gertz.
Beyond Friendships and Followers: The Wikipedia Social Network.
In *ASONAM'15*, 2015.



Yu Liu, Fahui Wang, Chaogui Kang, Yong Gao, and Yongmei Lu.
Analyzing Relatedness by Toponym Co-Occurrences on Web Pages.
T. GIS, 18(1), 2014.



Gianluca Quercini and Hanan Samet.
Uncovering the Spatial Relatedness in Wikipedia.
In *SIGSPATIAL '14*, 2014.



Denny Vrandečić and Markus Krötzsch.
Wikidata: A Free Collaborative Knowledgebase.
C. ACM, 57(10), 2014.