

A Two-Stage Framework for Computing Entity Relatedness in Wikipedia

Marco Ponza, Paolo Ferragina and Soumen Chakrabarti

University of Pisa



IIT Bombay



Menu

1. Introduction
 - Motivation
 - Our Contributions
2. Terminology
3. Known Methods for Entity-Relatedness Computation
4. Our Two-Stage Framework
5. Experiments
 - Accuracy of Relatedness Methods
 - Space and Time Efficiency
6. Conclusion & Future Work

Introduction

Motivation

Proliferation of the usage of Knowledge Graphs



Customers

- ▷ Retrieval of Information (Blanco, WSDM '15), (Cornolti, WWW '16)
- ▷ Entity Linking (Mihalcea, CIKM '07), (Meij, WSDM '12), (Ganea, WWW '16)
- ▷ Document Clustering, Classification and Similarity (Scaiella, WSDM '12), (Vitale, ECIR '12), (Ni, WSDM '16)



Need for *computing relatedness* between entities

Computing how much two entities are related

Relatedness : Entities × Entities → Float

Nodes of the Knowledge Graph



Introduction

Our Contributions

1

New dataset WiRe

- Human-assigned scores
- 503 Wikipedia entity pairs
- Sampled from New York Times (Dunietz, EACL '14)

2

Thorough and systematic study of *all known relatedness measures*

- WiRe (our introduced dataset)
- WikiSim (Milne, AAAI '08)

3

Proposal of a **Two-Stage Framework**

- Space-efficient
- Computationally lightweight
- More accurate than previous proposals

4

Extrinsic evaluation of our proposal

- Domain of **Entity Linking**
- Increase of accuracy and robustness of **TAG-ME** (Scaiella, CIKM '10)

5

Publicly available **WiRe dataset** and the **code of all algorithms!**



Terminology

- ▶ Our Knowledge Graph (KG): **WIKIPEDIA**
The Free Encyclopedia



Terminology

- ▶ Our Knowledge Graph (KG): **WIKIPEDIA**
The Free Encyclopedia
 - Entity?





WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction
[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

Tools
[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Wikidata item](#)
[Cite this page](#)

Article [Talk](#)

Read [View source](#) [View history](#)

Leonardo da Vinci



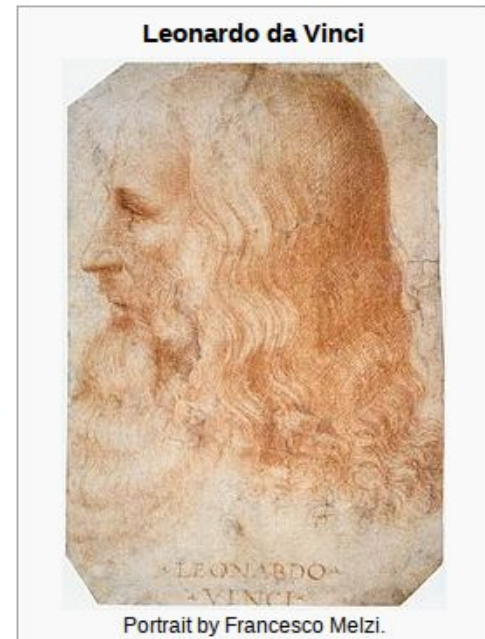
From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see [Da Vinci \(disambiguation\)](#).

*This is a [Renaissance Florentine](#) name. The name *da Vinci* is an indicator of birthplace, not a family name; this person is properly referred to by the given name *Leonardo*.*

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] ⓘ[ⓘ] listen[ⓘ]), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,^{[1][2][3]} he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "[Universal Genius](#)" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".^[4] According to art historian [Helen Gardner](#), the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote".^[4] Marco Rosci notes that while there is much speculation regarding his life and personality, his view of the world was logical rather than mysterious, and that the empirical methods he employed were unorthodox for his time.^[5]



▷ **Entity** = Wikipedia Page = Node of our KG



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction

[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

Tools

[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Wikidata item](#)
[Cite this page](#)

Article [Talk](#)

Read

[View source](#)

[View history](#)

Leonardo da Vinci



From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see [Da Vinci \(disambiguation\)](#).

*This is a [Renaissance Florentine](#) name. The name *da Vinci* is an indicator of birthplace, not a family name; this person is properly referred to by the given name [Leonardo](#).*

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,^{[1][2][3]} he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".^[4] According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote".^[4] Marco Rosci notes that while there is much speculation regarding his life and personality, his view of the world was logical rather than mysterious, and that the empirical methods he employed were unorthodox for his time.^[5]

Leonardo da Vinci



Portrait by Francesco Melzi.

- ▷ **Entity** = Wikipedia Page = **Node** of our KG
- ▷ **Label** of an **Entity** = **Textual Description** of a Wikipedia Page

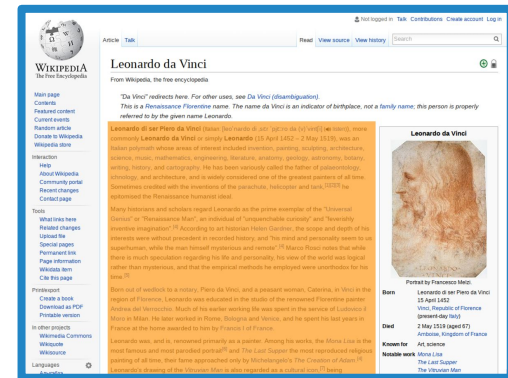
Terminology

▷ Our Knowledge Graph (KG):

- Entity = Wikipedia Page (a node of KG)
- Label = Textual Description of the Wikipedia Page
- Edges?

WIKIPEDIA

The Free Encyclopedia





WIKIPEDIA
The Free Encyclopedia

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

- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact page

- Tools
 - What links here

Science
From Wikipedia, the free encyclopedia

This article is about the general term. For other uses, see Science (disambiguation).

Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.

Contemporary science is typically subdivided into the natural sciences, which study the material universe; the social sciences which study people and societies; and the formal sciences, such as mathematics. The formal sciences are often excluded as they do not depend on empirical observations. Disciplines which use science like engineering and medicine may also be considered to be applied sciences.

During the Middle Ages in the Middle East, foundations for the scientific method were laid by Ibn al-Haytham in his *Book of Optics*. From classical antiquity through the 19th century, science as a type of knowledge was more closely linked to philosophy than it is now and, in fact, in the Western world, the term "natural philosophy" encompassed fields of study that are today associated with science, such as astronomy, medicine, and physics. While the classification of the material world of the ancient Indians and Greeks into air, earth, fire and water was more philosophical, medieval Middle Eastern scientists used practical, experimental observation to classify materials.

In the 17th and 18th centuries, scientists increasingly sought to formulate knowledge in terms of *laws of nature*. Over the course of the 19th century, the word "science" became increasingly associated with the scientific method itself, as a disciplined way to study the natural world. It was in the 19th century that scientific disciplines such as biology, chemistry, and physics reached their modern shapes. The same time period also included the origin of the terms "scientist" and "scientific community," the founding of scientific institutions, and increasing significance of the interactions with society and other aspects of culture.

- Wikimedia Commons
- Wikiquote
- Wikisource

- Languages
 - Адыгэбзэ

Article [Talk](#)

Read [View source](#) [View history](#)

Search

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).

This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,^{[1][2][3]} he epitomised the Renaissance humanist ideal.

Many historians and scholars regard him as a "Genius" or "Renaissance Man".

Invention
From Wikipedia, the free encyclopedia

"Inventor" and "Invented" redirect here. For other uses, see Invention (disambiguation). For more details on inventions throughout history, see Timeline of historic inventions. For the CAD design software, see Autodesk Inventor.

An **invention** is a unique or novel device, method, composition or process. The invention process is engineering and product development process. It may be an improvement upon a machine, process, system, or object or a result. An invention that achieves a completely unique function or result may be novel and **not obvious to others skilled in the same field**. An inventor may be taking a big step in this process. Some inventions can be patented. A patent legally protects the intellectual property rights of an inventor. The rules and requirements for patenting an invention vary by country. The process of obtaining a patent is often expensive.

Another meaning of invention is **cultural invention**, which is an innovative set of useful social practices passed on to others.^[1] The Institute for Social Inventions collected many such ideas in magazines and books. An important component of artistic and design creativity. Inventions often extend the boundaries of human capability.



Astronomy
From Wikipedia, the free encyclopedia

This article is about the scientific study of celestial objects. For other uses, see Astronomy (disambiguation).

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planets, moons, and nebulae) and processes (such as supernovae explosions, gamma ray bursts, and cosmic microwave background radiation), and evolution of such objects and processes, and more generally all phenomena observed from Earth or in outer space. A related but distinct subject, physical cosmology, is concerned with studying the universe as a whole.

Astronomy is the oldest of the natural sciences. The early civilizations in recorded history, such as the Egyptians, Nubians, Iranians, Chinese, and Maya performed methodical observations of the night sky. The early disciplines as diverse as astrometry, celestial navigation, observational astronomy and professional astronomy is nowadays often considered to be synonymous with astrophysics.^[2]

During the 20th century, the field of professional astronomy split into observational and theoretical astronomy. The former is focused on acquiring data from observations of astronomical objects, which is then analyzed using computer models to understand the physical processes that govern the objects. Theoretical astronomy is oriented toward the development of computer or analytical models to understand the physical processes that govern the objects. The two fields complement each other, with theoretical astronomy seeking to explain observations being used to confirm theoretical results.

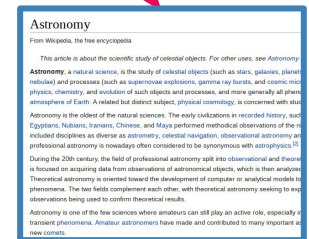
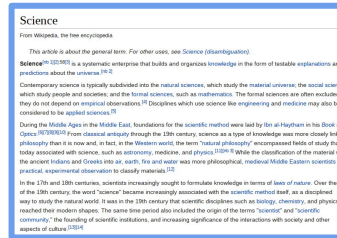
Astronomy is one of the few sciences where amateurs can still play an active role, especially in the discovery of transient phenomena. Amateur astronomers have made and contributed to many important astronomical discoveries, including new comets.

Leonardo was, and is, renowned primarily as a painter. Among his works, the *Mona Lisa* is his most famous and most parodied portrait^[6] and *The Last Supper* the most reproduced religious painting of all time, their fame approached only by Michelangelo's *The Creation of Adam*. Leonardo's drawing of the *Vitruvian Man* is also regarded as a cultural icon,^[7] being

The Vitruvian Man

Terminology

- ▷ Our Knowledge Graph (KG): **WIKIPEDIA**
The Free Encyclopedia
 - **Entity** = Wikipedia Page (a node of KG)
 - **Label** = Textual Description of the Wikipedia Page
 - **Edge** = Wikipedia Hyperlinks

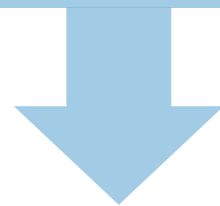


Known Relatedness Methods

A large number of methods proposed in literature...

- **Personalized Web Search** (Haveliwala, WWW '02)
- **Link Prediction** (Liben-Nowell, JAIST '07)
- **Word and Document Similarity** (Gabrilovich, IJCAI '07)
- **Document Annotation** (Piccinno, SIGIR '14)
- **Machine Translation** (Rothe, ACL '14)
- **Document Classification** (Perozzi, KDD '14), (Tan, WWW '15)

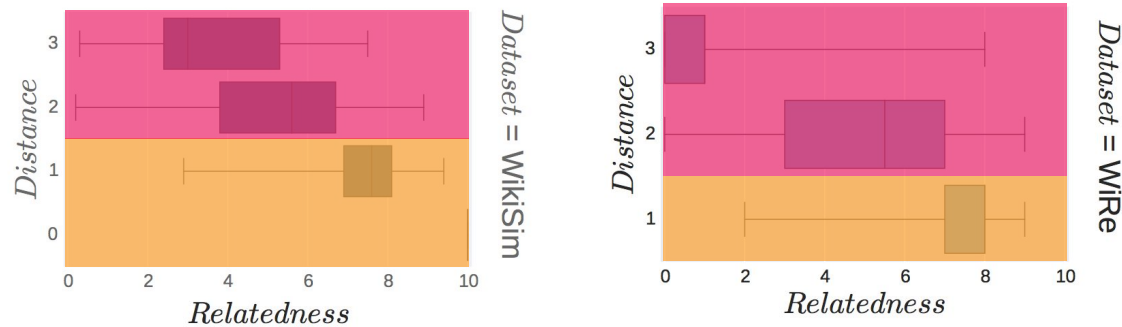
...that have been applied or are similar to our problem



We have experimented them
on the Entity Relatedness task

Our Two-Stage Framework

Why we need a Two-Stage Framework?



- ▷ Both **close** and **far** entities can be both **lowly** and **highly** related
- ▷ Hence distance-based methods are not (always) good predictors
- ▷ Most of known relatedness methods *ignore space* and *time efficiency*

Our Two-Stage Framework

- ▷ Built on the top of existing relatedness algorithms
- ▷ **Improves** current approaches
 - More **accurate** relatedness scores
 - **Fast** at query time
- ▷ The two stages of our framework:

1

A **small** and **weighted subgraph** is dynamically grown around the two *query entities*

2

Computing the **relatedness** between the two *query entities* according with the generated subgraph

▷ Motivations

- Wikipedia **edges** are **noisy** (introduced for **citation, explanation, ...**)
- Subgraph **nodes** are **strongly related** to the query entities (they are good bridges)
- Subgraph **edges** are **less noisy** (confined to few meaningful bridge nodes)

Our Two-Stage Framework

1

A **small** and **weighted subgraph** is dynamically grown around the two *query entities*



Tiger



Cat



Our Two-Stage Framework

1

A **small** and **weighted subgraph** is dynamically grown around the two *query entities*



Tiger



Cat

How can we populate the subgraph?

Our Two-Stage Framework

1

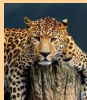
A **small** and **weighted subgraph** is dynamically grown around the two *query entities*



Tiger



Siberian_tiger



Leopard



Jaguar



European_cat



Cat_anatomy



Felidae



Cat

Populating the subgraph. Choosing the **top-k** nodes **most related** to the *query entities*

Our Two-Stage Framework

1

A **small** and **weighted subgraph** is dynamically grown around the two *query entities*

How?

Various Algorithms

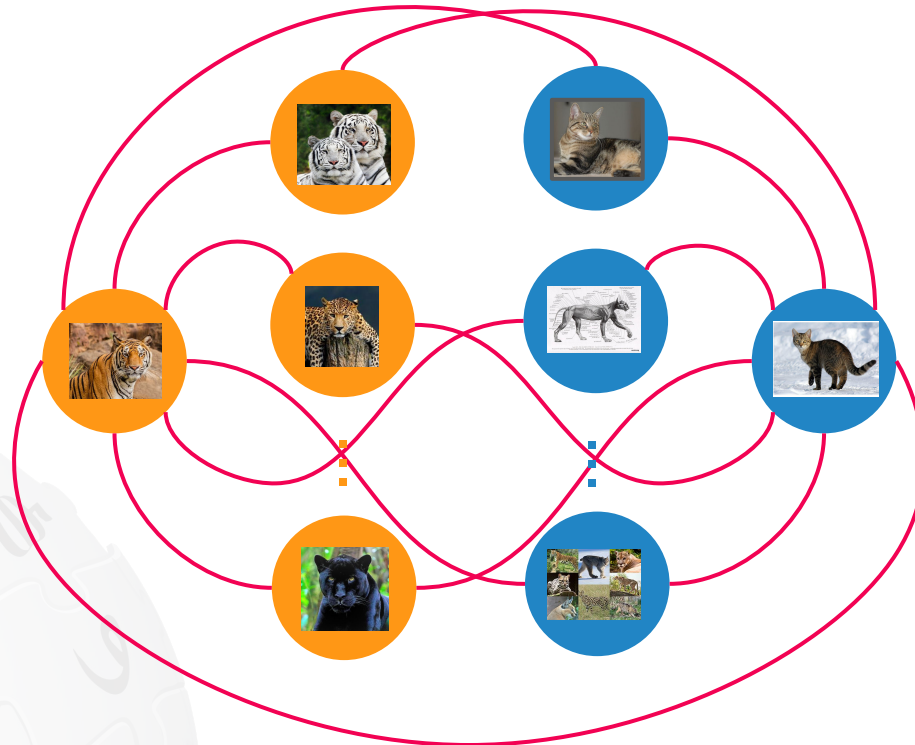
- **ESA** (Gabrilovich, IJCAI '07)
- **Milne&Witten** (Milne, AACL '08)
- **DeepWalk** (Perozzi, KDD '14)
- **Entity2Vec** (Ni, WSDM '16)

Populating the subgraph. Choosing the **top-k** nodes **most related** to the *query entities*

Our Two-Stage Framework

1

A **small** and **weighted subgraph** is dynamically grown around the two *query entities*



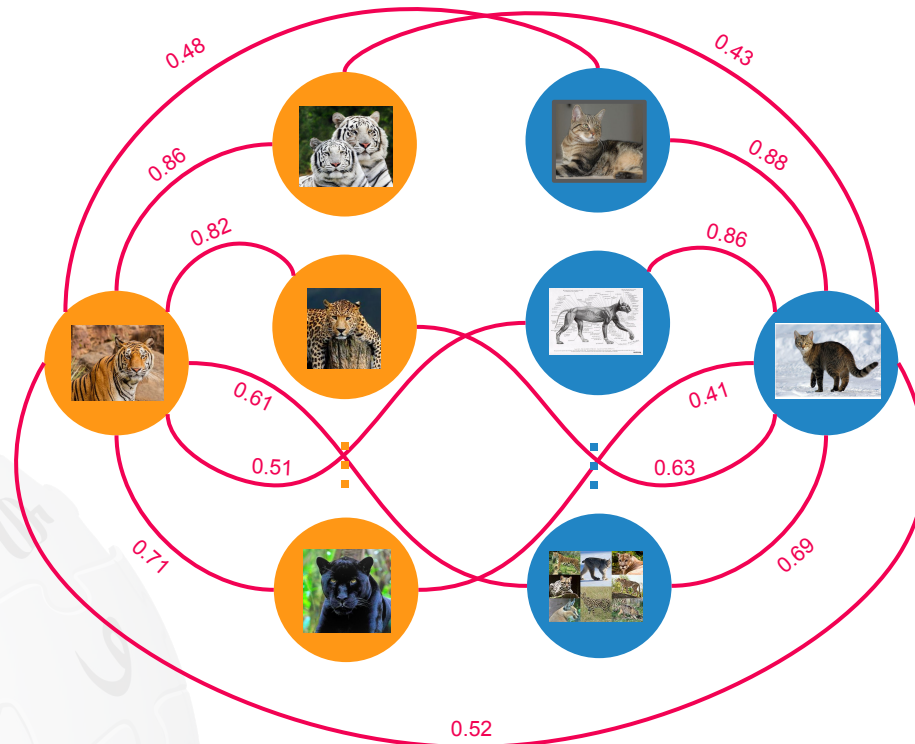
Creating the edges. Each query entity is linked to

- the **other** query entity
- its **top-k related** entities
- the **other top-k related** entities

Our Two-Stage Framework

1

A **small** and **weighted subgraph** is dynamically grown around the two *query entities*



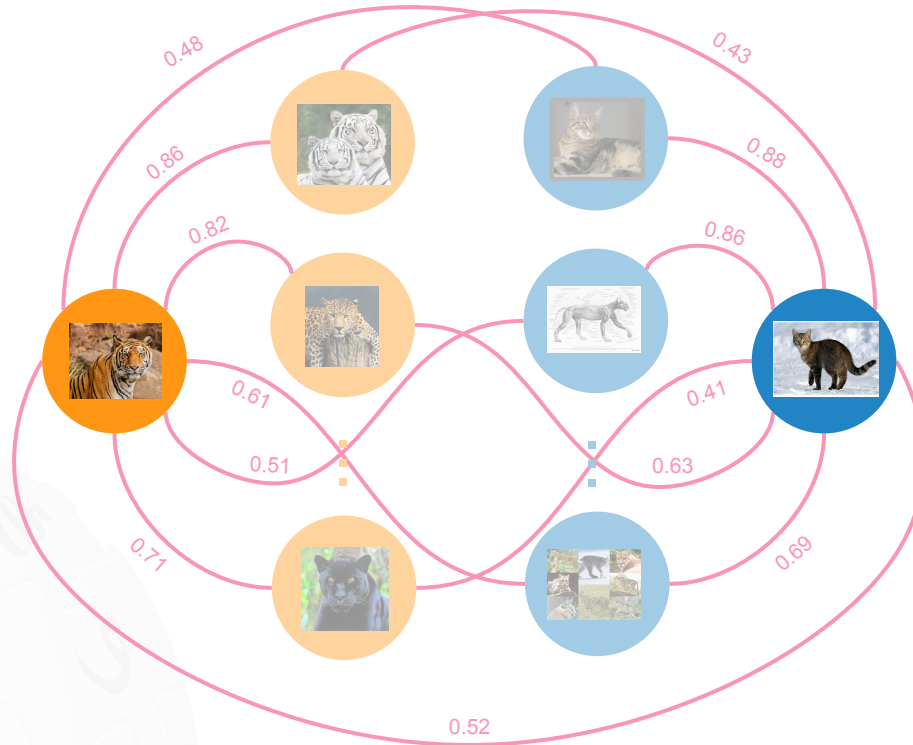
Weighting the edges. How?

- Milne&Witten (Milne, AAAI '08)
- DeepWalk (Perozzi, KDD '14)
- Entity2Vec (Ni, WSDM '16)

Our Two-Stage Framework

2

Computing the **relatedness** between the two *query entities* according with the generated subgraph



Computing Relatedness

CoSimRank (Rothe, ACL '14)

$$\text{relatedness} \left(\text{tiger_image}, \text{cat_image} \right) = 0.65$$

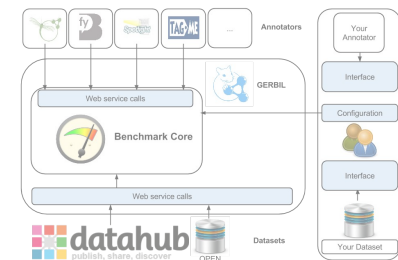
Experiments

- ▷ **Intrinsic** evaluation on **pairs** of **Wikipedia Entities**

Dataset	WikiSim (Milne, AAAI '08)	WiRe
Size	268	503
Pair Type	Common Nouns	Named Entities
Ground-Truth	Crowdsourcing	Human Experts

- ▷ **Extrinsic** evaluation

- Domain of **Entity Linking**
- On **four** different **datasets** (Usbeck, WWW '15)



- ▷ **Optimizations and time efficiency**
 - Compressed vs uncompressed

Experiments

Intrinsic Evaluation

- ▷ **Two-Stage Framework** instantiated with
 - Milne&Witten as Top-k Retrieval
 - Weights = Milne&Witten and DeepWalk
- ▷ Evaluation as (Hassan, AAAI '11):
 - Pearson, Spearman and their Harmonic Mean

Method	WikiSim			WiRe			AVG
	Pearson	Spearman	Harmonic	Pearson	Spearman	Harmonic	
ESA	0.61	0.72	0.67	0.60	0.63	0.62	0.645
Milne&Witten	0.62	0.65	0.63	0.77	0.69	0.72	0.675
DeepWalk	0.71	0.70	0.71	0.74	0.68	0.71	0.710
Entity2Vec	0.68	0.70	0.69	0.74	0.70	0.72	0.705
Two-Stage Framework	0.74	0.75	0.74	0.83	0.75	0.79	0.765

- ▷ *More experiments* in the paper (comparison between *more than 15 methods!*)

Experiments

Intrinsic Evaluation

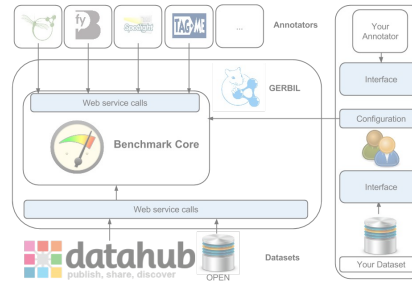
- ▷ **Two-Stage Framework** instantiated with
 - Milne&Witten as Top-k Retrieval
 - Weights = Milne&Witten and DeepWalk
- ▷ Evaluation as (Hassan, AAAI '11):
 - Pearson, Spearman and their Harmonic Mean

Method	WikiSim			WiRe			AVG
	Pearson	Spearman	Harmonic	Pearson	Spearman	Harmonic	
ESA	0.61	0.72	0.67	0.60	0.63	0.62	0.645
Milne&Witten	0.62	0.65	0.63	0.77	0.69	0.72	0.675
DeepWalk	0.71	0.70	0.71	0.74	0.68	0.71	0.710
Entity2Vec	0.68	0.70	0.69	0.74	0.70	0.72	0.705
Two-Stage Framework	0.74	0.75 +3%	0.74	0.83	0.75 +7%	0.79 +5%	0.765

- ▷ *More experiments* in the paper (comparison between *more than 15 methods!*)

Experiments

Extrinsic Evaluation

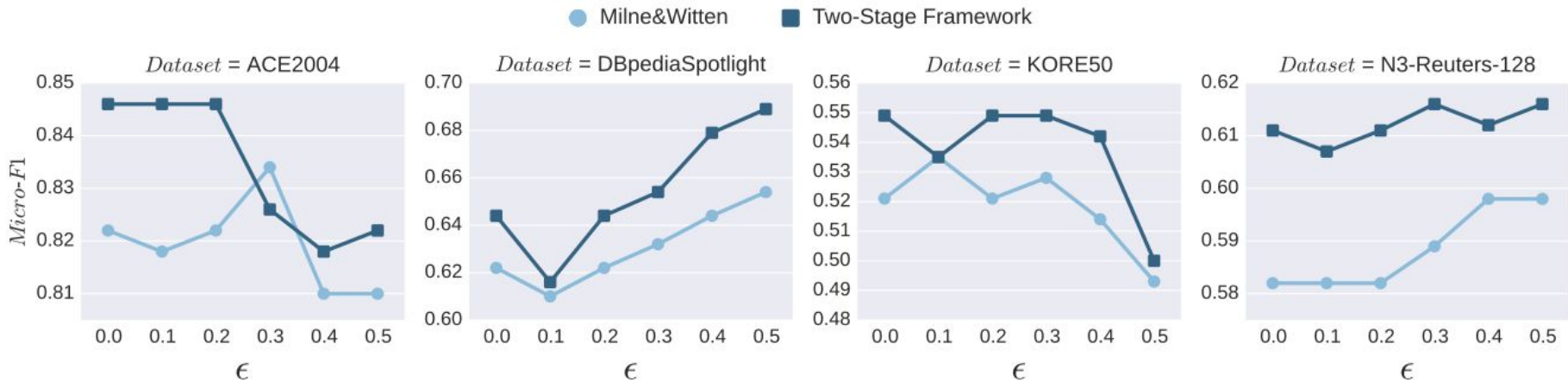


▷ Domain of Entity Linking

- Annotating short but meaningful sequence of words with proper Wikipedia Entities

▷ Entity Linker used for experiments: **TAG-ME**

- We replaced the relatedness method used in TagMe (e.g. [Milne&Witten](#)) with our **Two- Stage Framework**



- ### ▷ Our relatedness measure not only improves TagMe, but also makes it more insensitive to choices of the ϵ -parameter in TagMe

Experiments

Optimizations & Efficiency




- ▷ **Top-k preprocessing** of Milne&Witten on the entities' out-neighbors
- ▷ **Compression** of
 - *Wikipedia Graph* with Webgraph (Boldi, WWW '04)
 - *DeepWalk embeddings* with FEL (Blanco, WSDM '15)

	Uncompressed	Compressed	
Average Time	0.5 ms	3 ms	6x slower
Space	5 GB	445 MB	10x space-saving!

Our framework fits in few hundred of MB and the computation of the relatedness is still sufficiently fast at query time!

Conclusion & Future Work

Several *open issues* are there.

- Extending our framework to **other KGs**:
 - YAGO (Suchanek, WWW '07) 
 - WikiData 
 - ... 
- How can we **further speedup** our framework?
 - LSH (Gionis, VLDB '99)
 - Sketches (Akiba, KDD '16)
 - ...
- Impact of our framework to **other domains**?
 - Query understanding (Cornolti, WWW '16)
 - Document similarity (Ni, WSDM '16)
 - ...*any suggestions?*

CODE & DATA



<http://github.com/mponza/WikipediaRelatedness>

ACKNOWLEDGEMENTS

- **Bloomberg** Data Science Research Grant 2017
- **SIGIR** Student Travel Grant for CIKM 2017
- **SoBigData** Social Mining & Big Data Ecosystem EU Grant

Thanks!
Any questions?