A Two-Stage Framework for Computing Entity Relatedness in Wikipedia

Marco Ponza, Paolo Ferragina and Soumen Chakrabarti

University of Pisa





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 \times Entities \rightarrow Float



Nodes of the Knowledge Graph

Introduction Our Contributions

- New dataset WiRe
 - Human-assigned scores
 - 503 Wikipedia entity pairs
 - Sampled from New York Times (Dunietz, EACL '14)
- Thorough and systematic study of all known relatedness measures
 - WiRe (our introduced dataset)
 - WikiSim (Milne, AAAI '08)
- Proposal of a Two-Stage Framework
 - Space-efficient
 - Computationally lightweight
 - More accurate than previous proposals
- Extrinsic evaluation of our proposal
 - Domain of Entity Linking
 - Increase of accuracy and robustness of (Scaiella, CIKM '10)

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



No description or websit	e pravided.			
whipeds restores to	anage topica			
© 266 commits	i/ € branches	C D releases	AL 1 centributor	& Apache-2.0
Statch also v New pulls	es int		Greate new file Upload files /	we file Game or America
Toposta REACHE description	ation of japh response		Later	et commit Masiaca 2 hours
In datajdatasets	Added head of WikiSin			3 hours
Bit project	First commit.			a pear
All ord	Twosking			3 months
E atignere	Datasets added.			3 hours
E CONSC	License added			tt days
E READAE.nd	READARE description of	f jaan response		2 hours
E suitcast	Some more statistics			6 months
R leasing	Logo file added			11 days
Wil	cipedi	A		
Wil	3 52 72	A		
Wil	on	13 維		
Wil-	3 52 72	1 a repository stores		
Introducti Water to be W pager A he-Stopp code cheering will	on kipedia Relatedress libery. Tr	nia rapository stores city Relatedness In		
Introduction We provide the year	On Property Designed a Relatedness (Dary Th Property Agency of the Constitute of the be done in the next months.	nia repository stores etty felatedress in API present in the Whipe	Vikipetta . More document	ation and some
Introduction We provide the po- the American	on hopedia Relatectives a Baary. Th Prosecurk for Consecting on be done in the next months. PSS Webservice solution to use the algorithms of	Internet and the second s	Vicipetta . More document	ation and some
Introduction We provide the po- the American	Appedia Palatechasa Ibary, Tr Pramerick far: Consisting Di to does in the seat montal. EXERCISED STATES EXERCISED STATES Statistics of the seat montal of the site of the seat montal of the site of the seat montal of the site of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the seat of the se	Internet and the second s	Vicipetta . More document	ation and some
Introduction Marchaeter and a marchaeter Marchaeter and a marchaeter and a marchaeter Marchaeter and a marchaeter and a marchaeter and a marchaeter Marchaeter and a marchaeter	A second se	An repeation y stores and the repeation of the API present in the Wikipe ass.	vestepetta . More docurrent dia Ralabedraza library as a l	ation and some

Terminology

Our Knowledge Graph (KG):





Terminology

Our Knowledge Graph (KG):



• Entity?



Q

(+)

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

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: [leo'nardo di ,sɛr 'pjɛ:ro da (v)'vintʃi] (📣 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]



Portrait by Francesco Melzi.

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

Search

Q

(+) 🔒

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation). This is a Renaissance Elerentine name. The name da Vinci is an indicator o

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: [leo'nardo di ,sɛr 'pjɛ:ro da (v)'vintʃi] (🐗 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



1.000			& Not logged	in Talk Contributions Create acc	count Log in
a "	Article Talk	Read View so	surce View history	Search	Q
WIKIPEDIA	Leonardo da Vinci				•
The Free Encyclopedia	From Wikipedia, the free encyclopedia				
Main page Contents Featured content Current events	"Da Vinci" redirects here. For other uses, sea This is a Renaissance Florentine name. The referred to by the given name Leonardo.		inthplace, not a fan	ily name; this person is prope	nly
Random article Donate to Wikipedia Wikipedia store	Leonardo di ser Piero da Vinci (tatas: Jeo'naro commonly Leonardo da Vinci or simply Leonar Italian polytisch whose areas of interest included	rdo (15 April 1452 – 2 May 1519).	was an	Leonardo da Vinci	
Interaction Help About Wikipedia Community portal Recent changes Contact page	science, music, mathematics, engineering, iterat writing, history, and cartography. He has been w ichnology, and architecture, and is widely consid Sometimes credited with the invention of the pa epitomised the Renaissance humanist ideal.	ariously called the father of palaes leved one of the greatest painters	ontology, of all time.	2.92	
Tools What links here Related charges Upload file Special pages Permanent link Page information Wikidata item Cite this page	Many historians and schalars regard Leonardo a Consula [®] or Parainsance Mani, in minidual of inventive imagination [®] ¹⁸ According to an historia interests were without pre-order is in recorded hat superformance, which the mash historial model there is much speculation regarding his life and marker than mysterious, and that the empirical moderne time ¹⁸ .	"unquenchable curiosity" and "leve an Helen Gardner, the scope and tory, and "his mind and personalit and remote", ¹⁴ Marco Rosci notes personality, his view of the world v	arishly depth of his y seem to us that while kas logical	Print by Pancesco Melo	
Printlexport Create a book Download as PDF Printable version	Bern out of wedlock to a notary, Piero da Vinci, a region of Piorence, Leonardo was educated in th Andrea del Verrocchio. Much of his earlier worke Moro in Man: He later worked in Rome, Boloon	te studio of the renowned Florenti ng Ele was spent in the service of	ne painter Ludovico il	m Leonardo di ser Piero o 15 April 1452 Vind, Republic of Flore (present-day taly)	
In other projects Wikimedia Commons Wikipuote	France at the home awarded to him by Francis I Leonardo was, and is, renowned primarily as a p	l of France. painter. Among his works, the Mor	to Line is the Ki	ed 2 May 1519 (aged 67) Ambolae, Kingdom of I own for Art, science	France
Willisource	most famous and most parodied portrait ⁽¹⁾ and 7 painting of all time, their fame approached only b			table work Mona Lisa The Last Supper The Vituxian Man	



Artic

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^{[10} 비견(3563)</sup> is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and h predictions about the universe.^[ND 2]

Contemporary science is typically subdivided into the natural sciences, which study the material universe; the social science 9 which study people and societies; and the formal sciences, such as mathematics. The formal sciences are often excluded a they do not depend on empirical observations.¹⁴ Disciplines which use science like engineering and medicine may also be considered to be applied sciences.⁸⁹

During the Middle Ages in the Middle East, foundations for the scientific method were laid by Ibn al-Haytham in lis *Book of Optics*.^{[807](809,019)} From classical antiquity through the 19th century, science as a type of knowledge was more closely linked philosophy than it is now and, in fact, in the Western world, the term "natural philosophy" encompassed fields of study that today associated with science, such as astronomy, medicine, and physics.^{[11](10)} ³⁰ While the classification of the material world the ancient Indians and Greeks into air, earth, fire and water was more philosophical, medieval Middle Eastern scientists us practical, experimental observation to classify materials.^[12]

In the 17th and 18th centuries, scientists increasingly sought to formulate knowledge in terms of *laws of nature*. Over the co of the 19th century, the word "science" became increasingly associated with the scientific method itself, as a disciplined way to study the natural word! 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 IN314

Wikimedia Co	mmons
Wikiquote	
Wikisource	
anguages	¢
Алыгабаа	

Article Talk

Read View source View history

Search

Q

Leonardo da Vinci

From Wikipedia, the free encyclopedia

lany historians and scholars

Senius" or "Renaissance Man

"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: [leo'nardo di ,sɛr 'pjɛ:ro da (v) 'vintʃi] (4) 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 palasontology, 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.

Invention

:0

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.

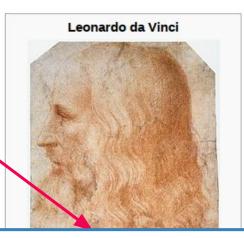
An **invention** is a unique or novel device, method, composition or process. The invention p engineering and product development process. It may be an improvement upon a machine an object or a result. An invention that achieves a completely unique function or result may novel and not obvious to others skilled in the same field. An inventor may be taking a big st

Some inventions can be patented. A patent legally protects the intellectual property rights claimed invention is actually an invention. The rules and requirements for patenting an inve process of obtaining a patent is often expensive.

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

ed in Rome, Bologna and Venice, and he spent his last y d to him by Francis I of France.

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



Astronomy

of his

From Wikipedia, the free encyclopedia

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

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planet nebulae) and processes (such as supernovae explosions, gamma ray bursts, and cosmic mici physics, chemistry, and evolution of such objects and processes, and more generally all phene atmosphere of Earth. A related but distinct subject, physical cosmology, is concerned with study

Astronomy is the oldest of the natural sciences. The early civilizations in recorded history, such Egyptians, Nubians, Iranians, Chinese, and Maya performed methodical observations of the ni included disciplines as diverse as astrometry, celestial navigation, observational astronomy an 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 theored is focused on acquiring data from observations of astronomical objects, which is then analyzed Theoretical astronomy is oriented toward the development of computer or analytical models to phenomena. The two fields complement each other, with theoretical astronomy seeking to exp observations being used to confirm theoretical results.

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

The Vitruvian Man

Terminology

- \triangleright
 - Entity = Wikipedia Page (a node of KG)
 - Label = Textual Description of the Wikipedia Page
 - Edge = Wikipedia Hyperlinks \bigcirc

Science

Our Knowledge Graph (KG): WIKIPEDIAThe Free Encyclopedia



Invention

- or more details on inventions throughout history, see For the CAD design software, see Autodesk I
- ention is a unique or novel device, method in that achieves a l
- ion is actually an invention. The rules and cess of obtaining a patent is often expensiv Another meaning of invention is cultural invention, which is an in
- The III The

Known Relatedness Methods

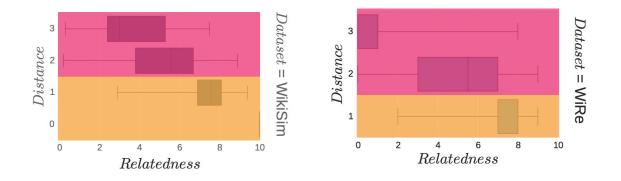
A large number of methods proposed in literature...

- 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

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

- 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:

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

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)

Tiger

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

Cat

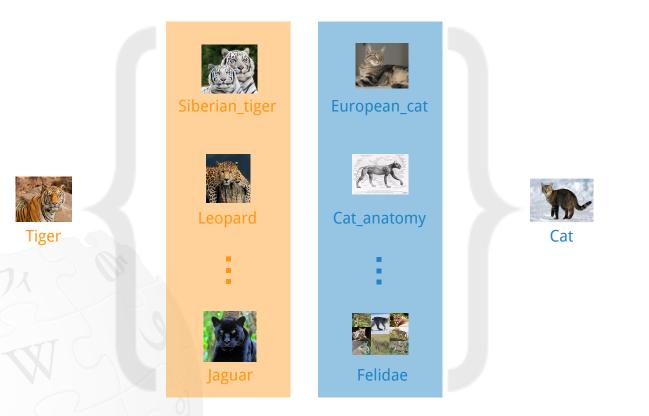
Tiger

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

Cat

How can we populate the subgraph?

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



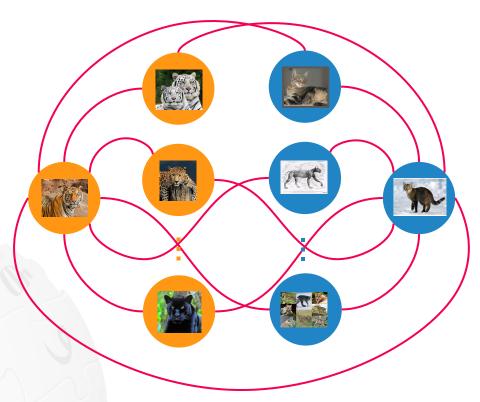
Populating the subgraph. Choosing the top-k nodes most related to the query entities

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



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

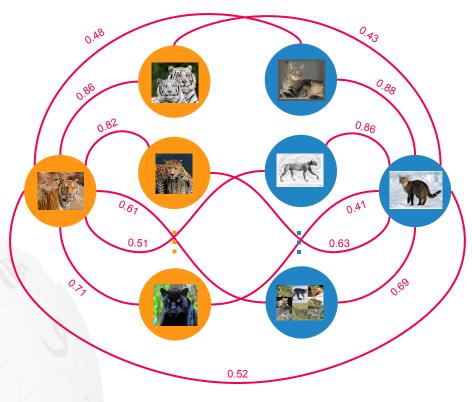
A small and weighted subgraph is dynamically grown around the two *query entities*



Creating the edges. Each query entity is linked to

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

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



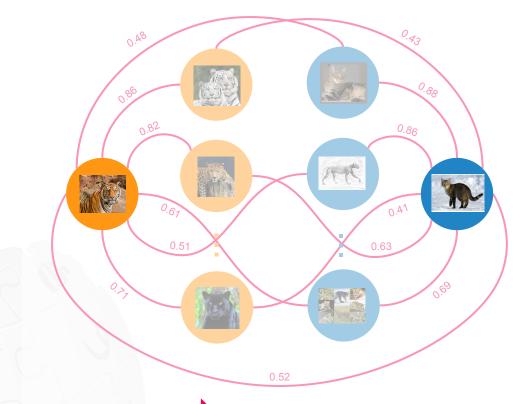
Ο

• Milne&Witten (Milne, AAAI '08)

Weighting the edges. How?

- DeepWalk (Perozzi, KDD '14)
- Entity2Vec (Ni, WSDM '16)

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



Computing Relatedness

CoSimRank (Rothe, ACL '14)

relatedenss (



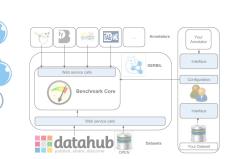
= 0.65

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
MGLIIUU	Pearson	Spearman	Harmonic	Pearson	Spearman	Harmonic	AVU
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!)

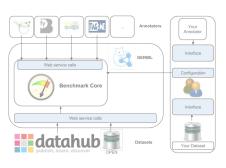
Experiments Intrinsic Evaluation

- 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
MGLIIUU	Pearson	Spearman	Harmonic	Pearson	Spearman	Harmonic	AVU
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	8% 0.74	0.83	0.75 +7	7 % 0.79 +5	<mark>% 0.765</mark>

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

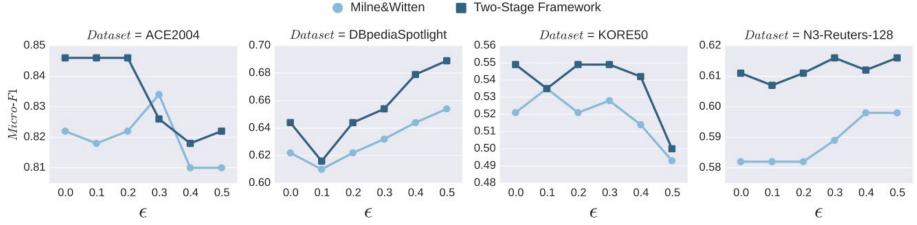
Experiments Extrinsic Evaluation



- Domain of Entity Linking
 - Annotating short but meaningful sequence of words with proper Wikipedia Entities
- Entity Linker used for experiments:



 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	6× slower
Space	5 GB	445 MB	10× space-sa

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, W/W/W '07)
 - WikiData
- Freebase
- How can we **further speedup** our framework?
 - LSH (Gionis, VLDB '99)
 - Sketches (Akiba, KDD '16)
 - 0 ...

0

- Impact of our framework to other domains?
 - Query understanding (Cornolti, WWW '16) Ο
 - **Document similarity** (Ni, WSDM '16)
 - ...any suggestions? 0









A			
	cipediA		
Wil	cipediA		
	The second second second		
Wil	The second second second		
Wil	ion		
Wil-	ion Ion Ion Ioner / Products for the second second Index / Products of the second second Index / Products of the second		
Wil-	ion Internet in the second sec		
References for an and a second	Ion Traditional sectors and a sector material sectors and a sector and a material sector and a sector sector a sector and a sector and a sector and a sector and a sector and a		
Wil-	In the second se	1 1013010 1013 1014 1015 1016 1017 1017 101 101	

http://github.com/mponza/WikipediaRelatedness

ACKNOWLEDGEMENTS

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

Thanks! Any questions?