Algorithms for Knowledge and Information Extraction in Text with Wikipedia

Marco Ponza
University of Pisa

Supervisor
Prof. Paolo Ferragina

Menu

0. Introduction

Knowledge and Information Extraction

1. Algorithms for Entity Relatedness
2. Algorithms for Entity and Fact Salience

Applications

3. Algorithms for Expert Finding

4. Future Research Directions

0

Introduction

Introduction

- Enhancing the humankind progress with new intelligent technologies

→ Tools that can afford general- or specific-domain tasks with performance close or better than humans

- Machines need of access, read and understand information stored in data archives

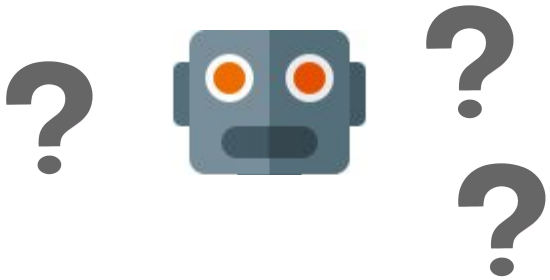
→ The dominant form on which information is produced every day by humans is still Natural Language

Introduction

Text Understanding

Easy for humans

Hard for machines



Introduction

Text Understanding

Leonardo is the scientist who painted Mona Lisa

1

Map ambiguous words into the real-world entities they refer to as well as *contextualize* them together with related entities

Introduction

Text Understanding

Leonardo is the scientist who painted Mona Lisa

("Leonardo", "is", "scientist")

("Leonardo", "painted", "Mona Lisa")

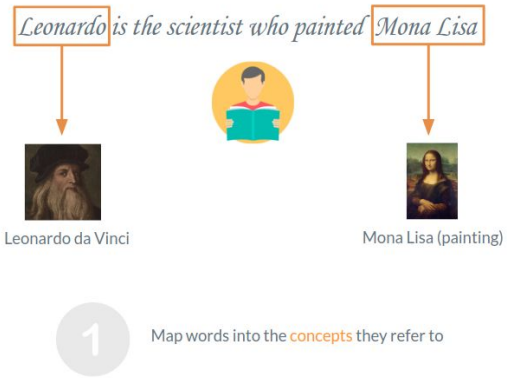
2

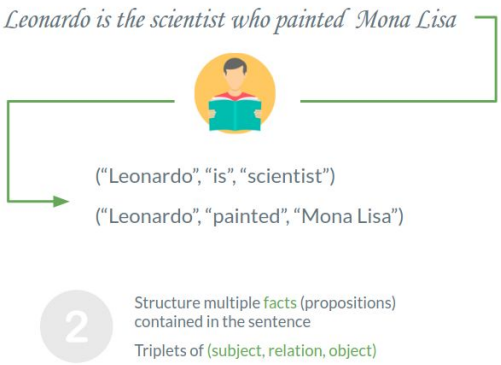
Structure multiple facts (propositions) contained in the sentence

Triplets of (subject, relation, object)

Introduction

Text Understanding





How can we do that?

Humans can interpret words in a *larger context* hinging onto their background and linguistic knowledge (Gabrilovich, SIGIR'16)

Detect (1) unambiguous entities (2) facts, (3) quantifying how much their are related, and (4) efficiently retrieve related entities

Introduction

Text Understanding

- Literature currently offers a number of solutions based on BoW (Harris, Word'54) : A text is a vector of *ambiguous keywords*

Limitations
- Curse of Dimensionality
- Synonymy and Polysemy problems of keywords
- No understanding of real-world entities
- Structure of the sentence is lost (no facts)

Introduction

Text Understanding

- Literature currently offers a number of solutions based on BoW (Harris, Word'54) : A text is a vector of *ambiguous keywords*

  Limitations
  - Curse of Dimensionality
  - Synonymy and Polysemy problems of keywords
  - No understanding of real-world entities
  - Structure of the sentence is lost (no facts)

- LDA/LSI (Huffman, NIPS'10) and Word Embeddings (Mikolov, NIPS'13) overcome some limitations

  A text is mapped into a *latent space* (vector of floating-points)

Introduction

Text Understanding

- Literature currently offers a number of solutions based on BoW (Harris, Word'54) : A text is a vector of *ambiguous keywords*

Limitations

- Curse of Dimensionality ✓
- Synonymy and Polysemy problems of keywords ✓
- No understanding of real-world entities ✗ ✓
- Structure of the sentence is lost (no facts) ✗

- LDA/LSI (Huffman, NIPS'10) and Word Embeddings (Mikolov, NIPS'13) overcome some limitations

  A text is mapped into a *latent space* (vector of floating-points)

Introduction

Text Understanding

- Literature currently offers a number of solutions based on BoW (Harris, Word'54) : A text is a vector of *ambiguous keywords*

  - Curse of Dimensionality
  - Limitations
  - words
  - s
  - Structure of the sentence is lost (no facts)

Need for a more efficient and effective semantic paradigm

- LDA/LSI (Huffman, NIPS'10) and Word Embeddings (Mikolov, NIPS'13) overcome some limitations

  A text is mapped into a *latent space* (vector of floating-points)

Introduction

- Need for a more efficient and effective semantic paradigm

Exploiting two different resources

- World Knowledge — WIKIPEDIA The Free Encyclopedia
- Linguistic Knowledge — Language Grammar

...thanks to recent advancements in the field of Natural Language Processing:

- Entity Linking (Bunescu, EACL'06), (Scaiella, IEEE'12), (Piccinno, SIGIR'14)

Introduction

- Need for a more efficient and effective semantic paradigm

Exploiting two different resources

- World Knowledge — WIKIPEDIA The Free Encyclopedia
- Linguistic Knowledge — Language Grammar

...thanks to recent advancements in the field of Natural Language Processing:

- Entity Linking (Bunescu, EACL'06), (Scaiella, IEEE'12), (Piccinno, SIGIR'14)

Leonardo painted Mona Lisa

Introduction

- Need for a more efficient and effective semantic paradigm

Exploiting two different resources

- World Knowledge — WIKIPEDIA The Free Encyclopedia
- Linguistic Knowledge — Language Grammar

...thanks to recent advancements in the field of Natural Language Processing:

- Entity Linking (Bunescu, EACL'06), (Scaiella, IEEE'12), (Piccinno, SIGIR'14)

Leonardo *painted* *Mona Lisa*

Entities

Leonardo da Vinci

Mona Lisa (painting)

Introduction

- Need for a more efficient and effective semantic paradigm

Exploiting two different resources
- World Knowledge — WIKIPEDIA The Free Encyclopedia
- Linguistic Knowledge — Language Grammar

...thanks to recent advancements in the field of Natural Language Processing:

- Entity Linking (Bunescu, EACL'06), (Scaiella, IEEE'12), (Piccinno, SIGIR'14)
- Open Information Extraction (Banko, IJCAI'07),(Del Corro, WWW'13), (Gashteovski, EMNLP'17)

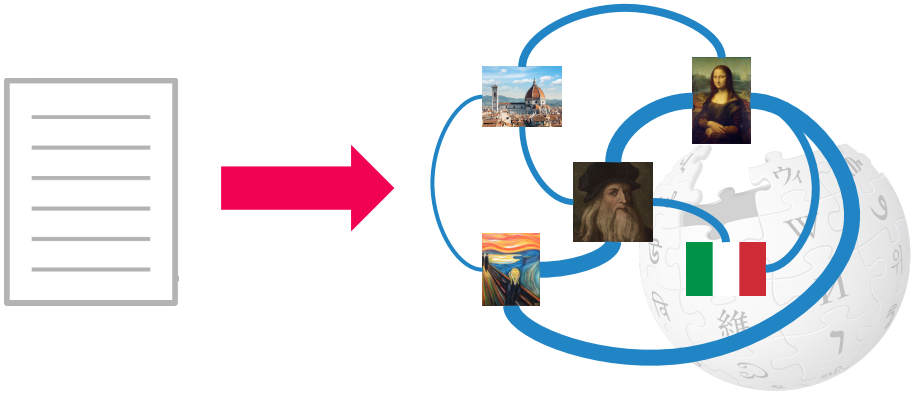
Leonardo, the scientist, painted Mona Lisa

Open Facts
- *("Leonardo", "is", "scientist")*
- *("Leonardo", "painted", "Mona Lisa")*

Introduction

- Need for a more efficient and effective semantic paradigm

...how? Applying Graph Theory to entity linking and open information extraction



Model a text as a ***Small Wikipedia Graph!***

- Curse of Dimensionality ✓ → The graph is small
- Synonymy and Polysemy problems of keywords ✓ →
- Understanding of real-world entities ✓ → Wikipedia entities are unique and they represent a real-world concept
- Structured facts ✓ → OpenIE preserves subject-relation-object structure

Contributions

1. Entity Relatedness

How much two Wikipedia entities are related?

2. Entity Salience

Summarize document's subject matter with its Salient Wikipedia Entities

3. Fact Salience

Summarize document's subject matter with its Salient Open Facts

4. Expert Finding

Who are the experts of a given topic?

Contributions

1 Entity Relatedness

A Two-Stage Framework for Computing Entity Relatedness in Wikipedia
Marco Ponza, Paolo Ferragina and Soumen Chakrabarti

CIKM 2017 SINGAPORE

2 Entity Salience

Document Aboutness via Sophisticated Syntactic and Semantic Features
Marco Ponza, Paolo Ferragina and Francesco Piccinno

NLdb 2017 Liège, Belgium

SWAT: A System for Detecting Salient Wikipedia Entities in Texts
Marco Ponza, Paolo Ferragina and Francesco Piccinno

WILEY Computational Intelligence 2019

3 Fact Salience

Facts That Matter
Marco Ponza, Luciano Del Corro and Gerhard Weikum

EMNLP 2018 Brussels

4 Expert Finding

WISER: A Semantic Approach for Expert Finding in Academia based on Entity Linking
Paolo Cifariello, Paolo Ferragina and Marco Ponza

ELSEVIER *Information Systems 2019*

1

Algorithms for Entity Relatedness

Entity Relatedness

Motivation

Proliferation of the usage of Knowledge Graphs

Consumers

- 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 *entity relatedness*

Compute how much two entities are related

Relatedness : Entities x Entities →Real

The Wikipedia Knowledge Graph

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

The Wikipedia Knowledge Graph

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

Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search

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

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

Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search

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

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]

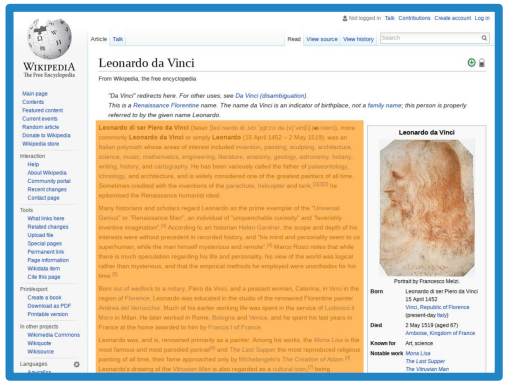
Leonardo da Vinci

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): WIKIPEDIA The Free Encyclopedia
  - Entity = Wikipedia Page (a node of KG)
  - Label = Textual Description of the Wikipedia Page
  - Edges?



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

Wikimedia Commons
Wikiquote
Wikisource

Languages
Адыгэбзэ

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 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 ... of his "Genius" or "Renaissance Man ...

... 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 ...

Leonardo da Vinci

The Vitruvian Man

Science

From Wikipedia, the free encyclopedia

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

Science[nb 1][2]:58[3] is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.[nb 2]

Contemporary science is typically subdivided into the natural sciences, which study the material universe; the social science 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.[4] Disciplines which use science like engineering and medicine may also be considered to be applied sciences.[5]

During the Middle Ages in the Middle East, foundations for the scientific method were laid by Ibn al-Haytham in his *Book of Optics*.[6][7][8][9][10] 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][nb 3] While the classification of the material wo 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 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.[13][14]

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 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 c 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.

Astronomy

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, planets nebulae) and processes (such as supernovae explosions, gamma ray bursts, and cosmic micr physics, chemistry, and evolution of such objects and processes, and more generally all phenc atmosphere of Earth. A related but distinct subject, physical cosmology, is concerned with stud

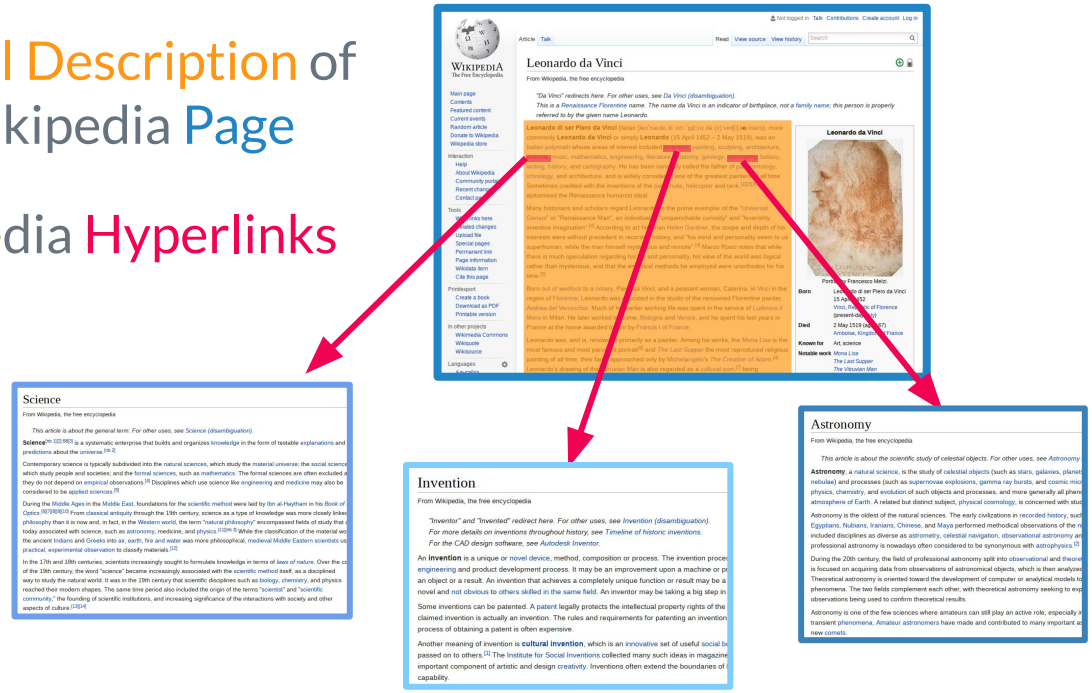
Astronomy is the oldest of the natural sciences. The early civilizations in recorded history, such Egyptians, Nubians, Iranians, Chinese, and Maya performed methodical observations of the ni included disciplines as diverse as astrometry, celestial navigation, observational astronomy an professional astronomy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

During the 20th century, the field of professional astronomy split into observational and theoret 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 Wikipedia Knowledge Graph

- 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

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 TAGME (Scaiella, CIKM '10)

5

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

GitHub

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

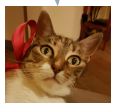
Cat

Our Two-Stage Framework

1

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

Tiger



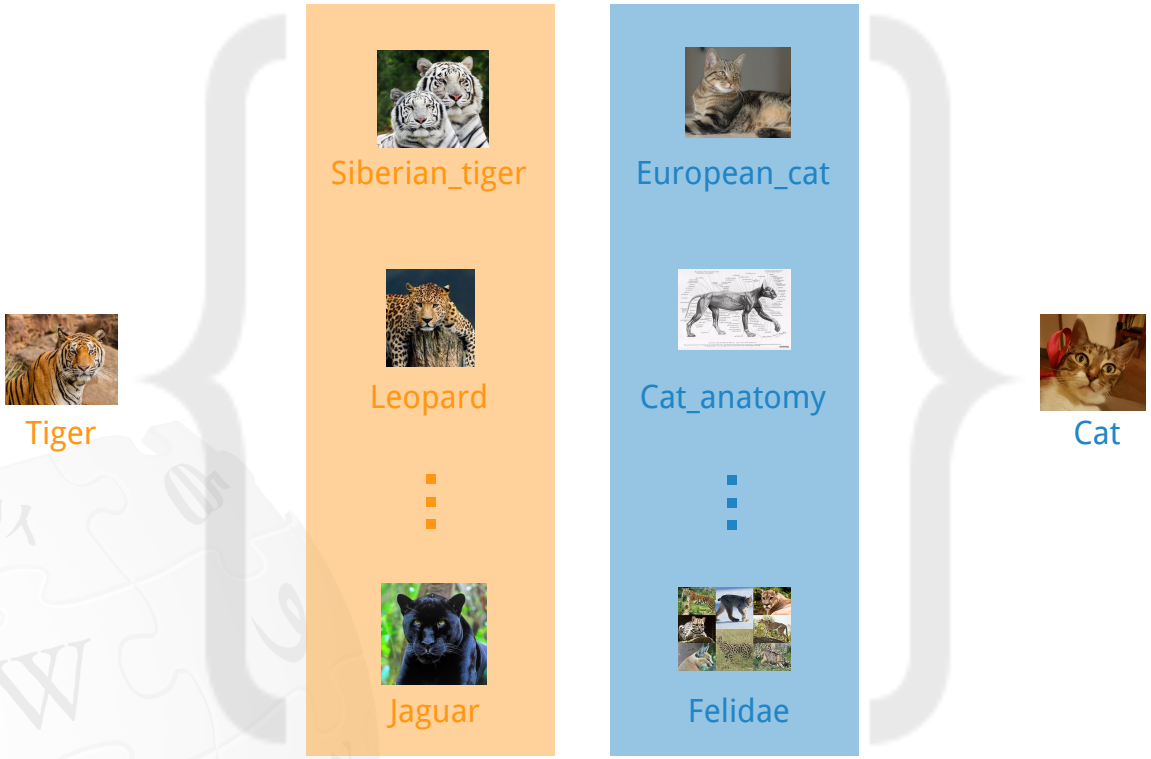
Cat

How can we populate the subgraph?

Our Two-Stage Framework

1

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

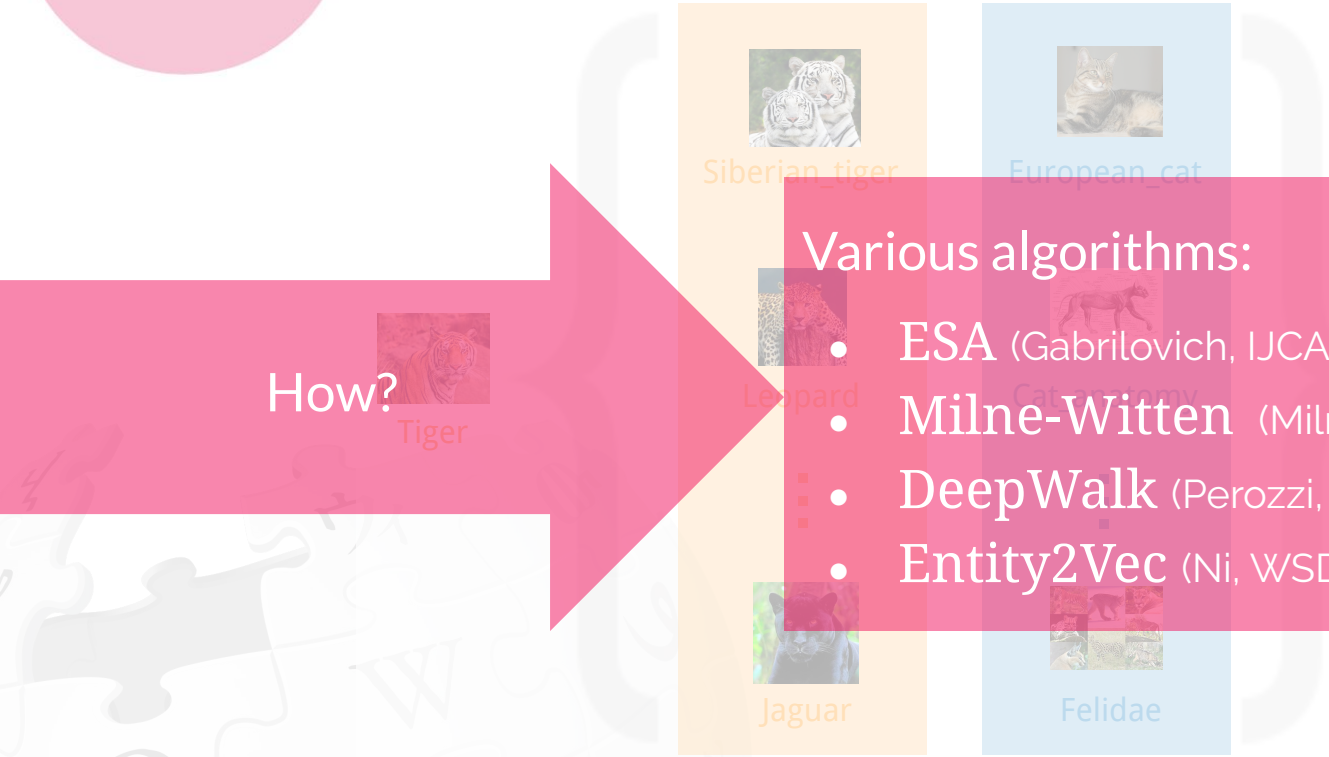


Populating the subgraph. Choosing the *top-k* nodes *most related* to the query entities

Our Two-Stage Framework

1

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



How?

Various algorithms:

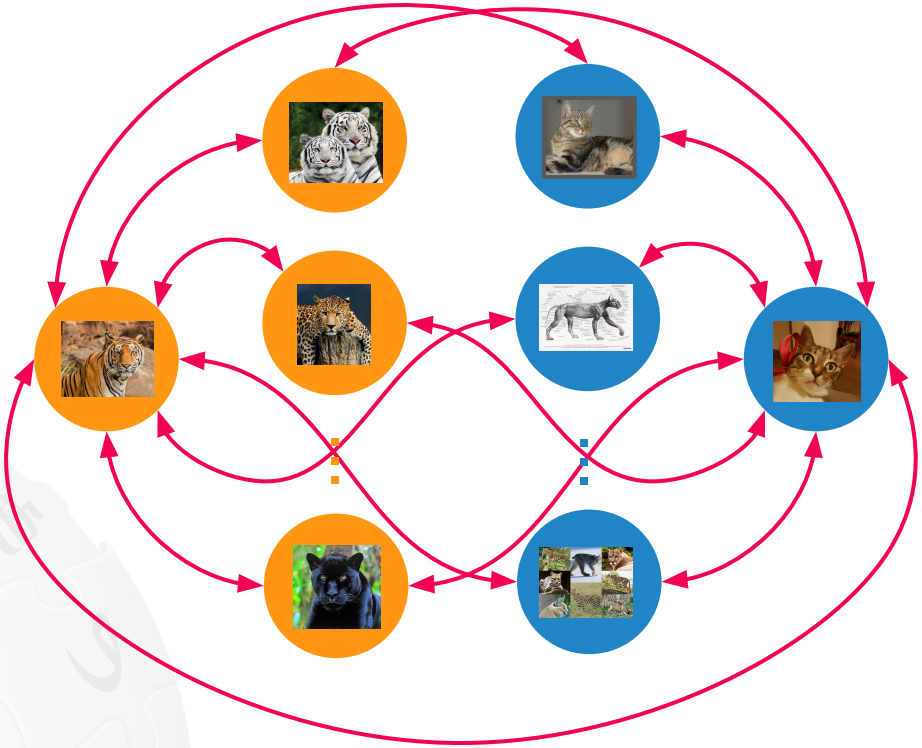
- ESA (Gabrilovich, IJCAI '07)
- Milne-Witten (Milne, AAAI '08)
- DeepWalk (Perozzi, KDD '14)
- Entity2Vec (Ni, WSDM '16)

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*



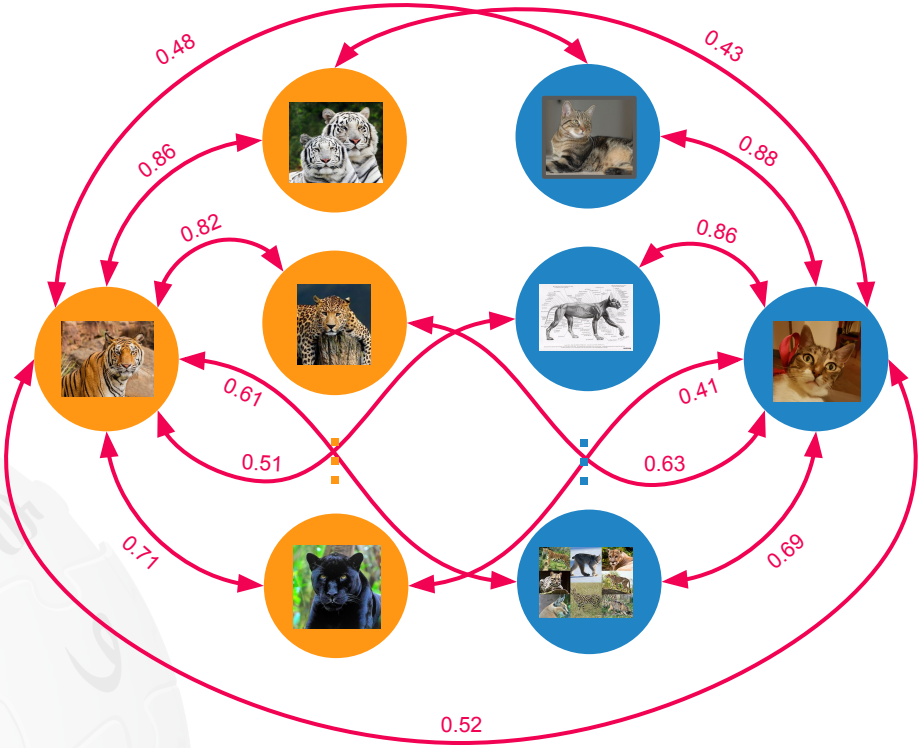
Creating the edges. Each query entity is linked to

- *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*



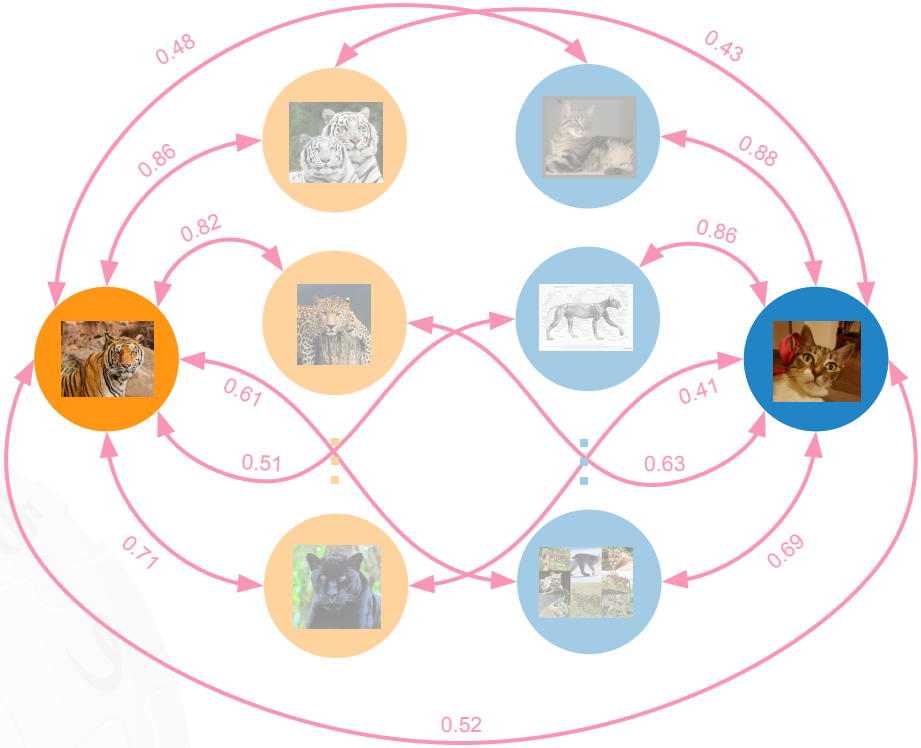
Weighting the edges. How?

- 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)

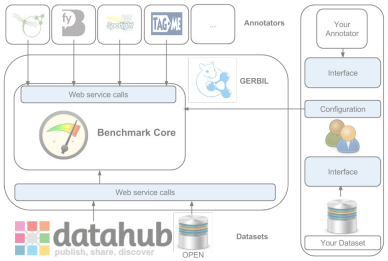
relatedenss ( , ) = *0.65*

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)



Experiments

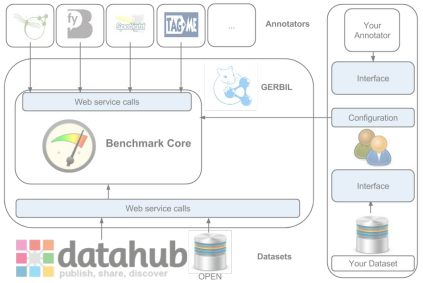
Intrinsic Evaluation

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**

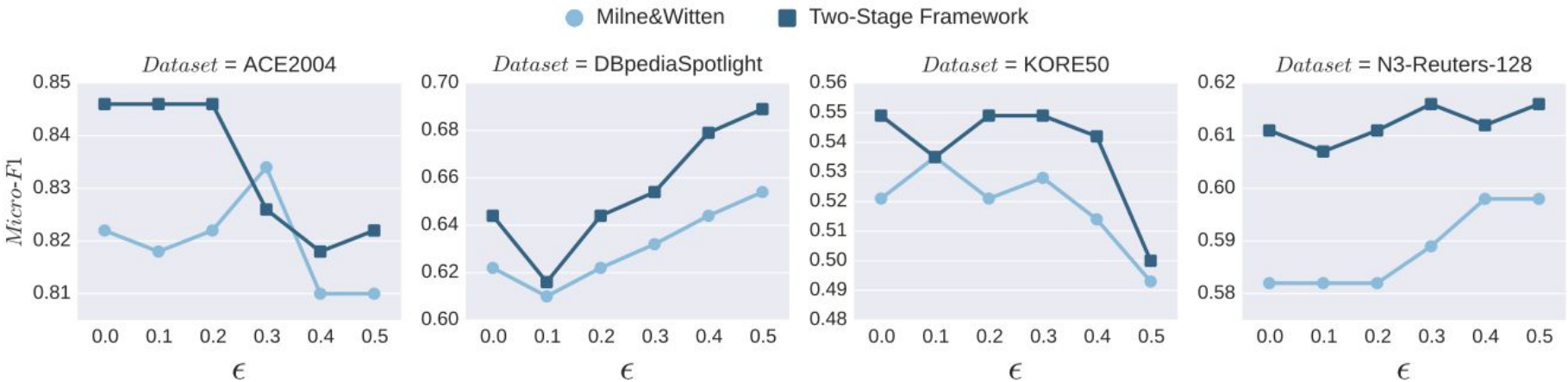
- Pearson measures predicted-vs-correct scores
- Spearman focuses on the ranking order among entity pairs
- Two-Stage Framework instantiated with
  - Milne-Witten as Top-k Retrieval
  - Weights are the average between Milne-Witten and DeepWalk
- More experiments in the paper (first known comparison among *more than 15 methods!*)

Experiments

Extrinsic Evaluation



- Domain of *Entity Linking*
  - Linking 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!

2

Algorithms for Entity and Fact Salience

Document Aboutness via Sophisticated Syntactic and Semantic Features
Marco Ponza, Paolo Ferragina, and Francesco Piccinno

NLdb 2017
Liège, Belgium

SWAT: A System for Detecting Salient Wikipedia Entities in Texts
Marco Ponza, Paolo Ferragina, and Francesco Piccinno

WILEY
Computational Intelligence
AN INTERNATIONAL JOURNAL
2019

Facts That Matter
Marco Ponza, Luciano Del Corro, and Gerhard Weikum

EMNLP 2018
Brussels

Introduction

Automatic Document Summarization

- Succinct representation of the Document's Subject Matter (Bruza, AIR '96)

- Condensing salient information from an input text into a summary
  - Enable fast and accurate document search
  - Help reader to identify relevant topics

(Hasan, ACL'14)

Introduction

Automatic Document Summarization

- Succinct representation of the Document's Subject Matter (Bruza, AIR '96)
- Condensing salient information from an input text into a summary

Input Document

Summary

Introduction

Automatic Document Summarization

- Succinct representation of the Document's Subject Matter (Bruza, AIR '96)
- Condensing salient information from an input text into a summary

- Sentences
- Keywords
- Proper Nouns
- *... their combinations!*

Introduction

Automatic Document Summarization

- Succinct representation of the Document's Subject Matter (Bruza, AIR '96)
- Condensing salient information from an input text into a summary

Input Document

Summary

Our Context

We attack the *summarization* problem from *two different points of view*

- Salient Wikipedia Entities
- Salient Open Facts

2.1
Algorithms for Entity Salience

Document Aboutness via Sophisticated Syntactic and Semantic Features
Marco Ponza, Paolo Ferragina and Francesco Piccinno

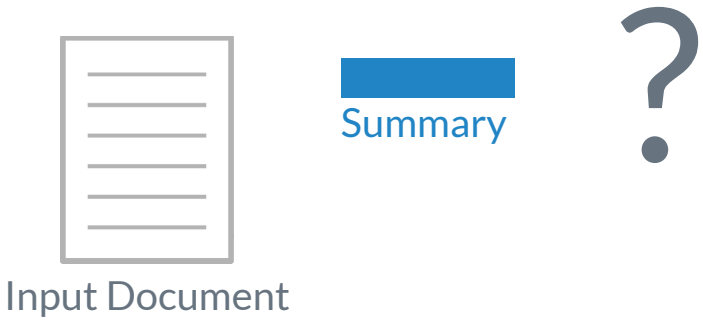
NLdb 2017
Liège, Belgium

SWAT: A System for Detecting Salient Wikipedia Entities in Texts
Marco Ponza, Paolo Ferragina and Francesco Piccinno

WILEY
Computational Intelligence
AN INTERNATIONAL JOURNAL
2019

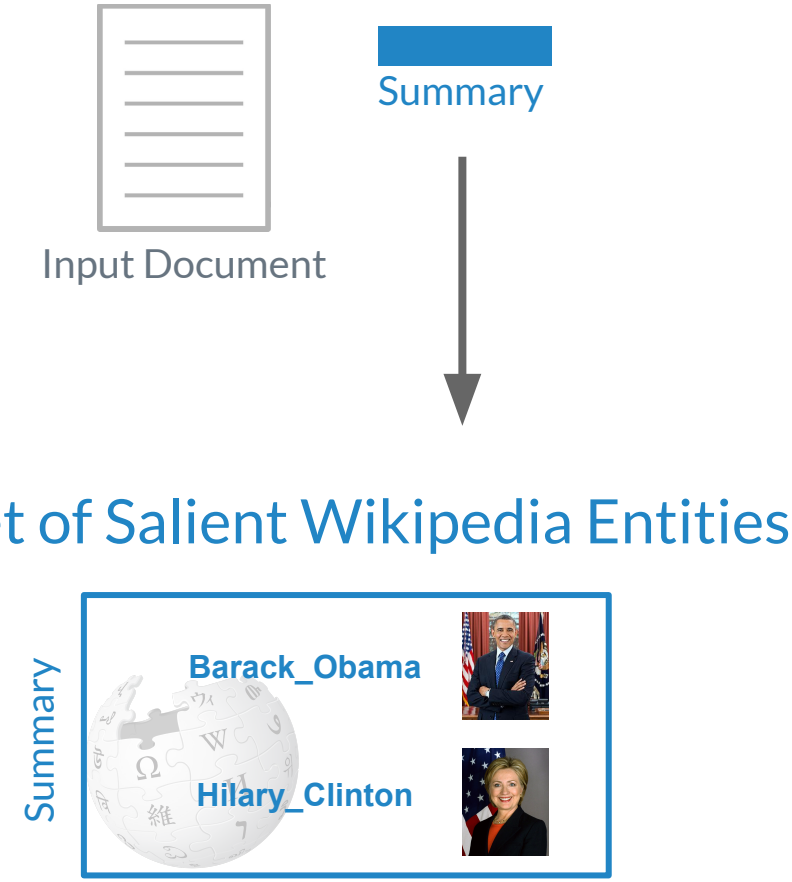
Entity Salience

Summarization via Salient Wikipedia Entities



Entity Salience

Summarization via Salient Wikipedia Entities



Entity Salience

Contributions: Our Solution

- Three-Stage System for Entity Salience Extraction
- In-Depth Feature Engineering:
  - Syntactic:
    - Sentence Ranking
    - Dependency Trees
  - Semantic:
    - Entity Annotations
    - Relatedness Graph
- Improves current solutions
  - From +1.9% up to +14%
- The first publicly available API

The Salient Wikipedia Annotator of Text

SWAT

So Big Data Research Infrastructure

Entity Salience

General Structure

Input Document

The Salient Wikipedia Annotator of Text

Document Enrichment

1. CoreNLP
2. WAT

Feature Generation

Features

1. Basics
2. Syntactic
3. Semantic

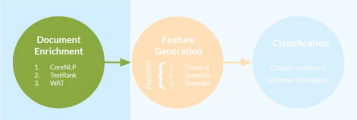
Classification

Classify entities in *Salient* or *Non-Salient*

Salient Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

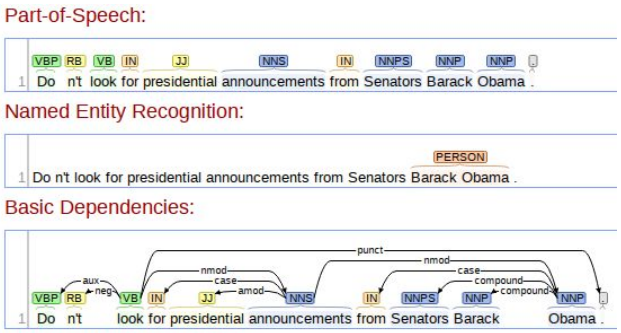
1. Document Enrichment

- ▷ CoreNLP (Manning, ACL '14)

Entity Salience

Three-Stage System

Images via
http://corenlp.run



1. Document Enrichment

▷ **CoreNLP** (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging
Named Entity Recognition
Dependency Parsing
Coreference

Entity Salience

Three-Stage System

The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

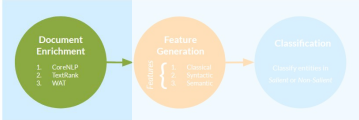
By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS



1. Document Enrichment

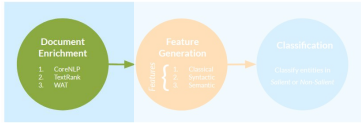
- ▷ CoreNLP (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ WAT (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



1. Document Enrichment

- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ **WAT** (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System

The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

1. Document Enrichment

- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ **WAT** (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System

The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

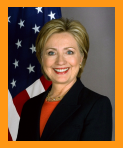
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

Barack_Obama



Hilary_Clinton

George_Walker Bush

1. Document Enrichment

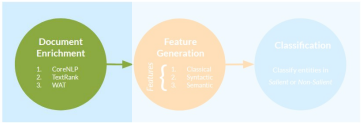
- CoreNLP (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- WAT (Piccinno, SIGIR '14)
  - Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

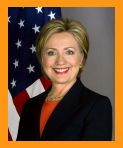
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS


Barack_Obama


Hilary_Clinton


United_States Capitol


George_Walker Bush

1. Document Enrichment

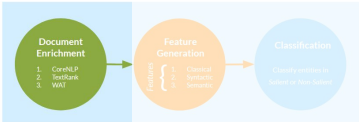
- ▷ CoreNLP (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ WAT (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

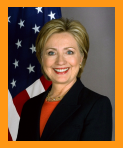
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS


Barack_Obama


Hilary_Clinton


United_States Capitol


Hawaii


George_Walker Bush

1. Document Enrichment

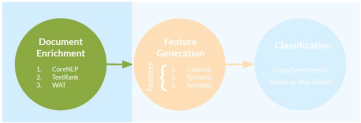
- ▷ CoreNLP (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ WAT (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

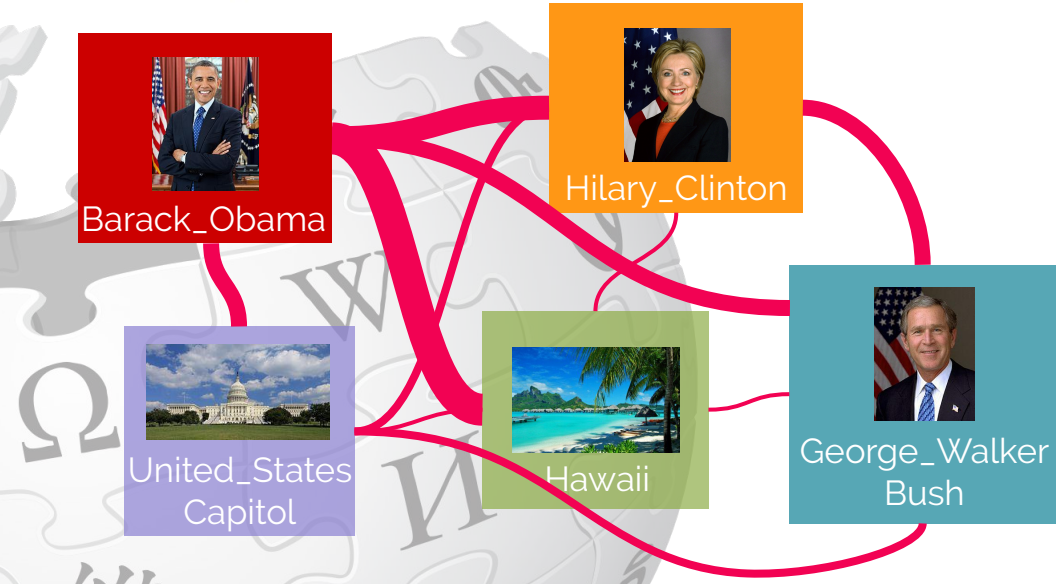
By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS



1. Document Enrichment

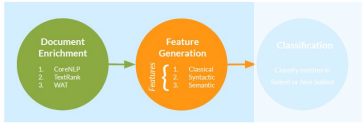
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

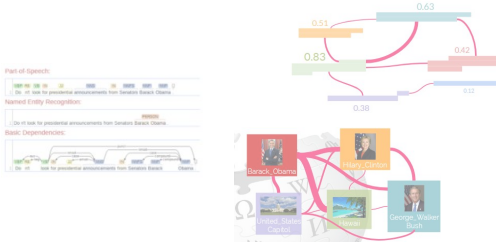
- ▷ **WAT** (Piccinno, SIGIR '14)
  - Annotates them with Wikipedia Entities
  - Relatedness Graph
    - Nodes = Entities
    - Weights are defined with two different relatedness scores:
      (i) Wikipedia Jaccard In-Links
      (ii) Cosine between Entity Embeddings

Entity Salience

Three-Stage System



1. Document Enrichment



Tokens, POS Tags, Dependency Relations, Coreference Chains, Wikipedia Entities and their Relatedness

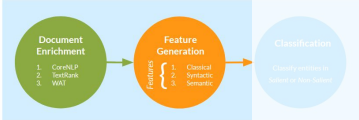
2. Feature Generation

2. Feature Generation

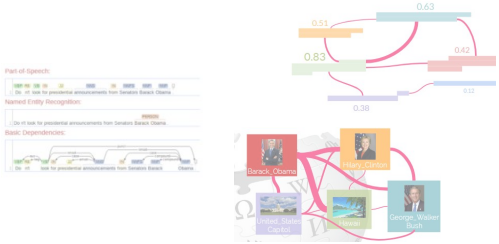
- Standard Entity Features
  - Frequency
  - Positions
  - ...
- CMU-Google Features
  - POS-Tags, Coreference Freq.
  - PageRank on a graph whose weights are based on co-occ.
  - ...
- Syntactic Features
  - Statistics on Sentence Ranks
  - Frequency/Positions of Dependency Relations
  - ...
- Semantic Features
  - Statistics on annotations (coherence, commonness)
  - Graph Centralities on Relatedness Graph
  - Relatedness over Positions
  - ...

Entity Salience

Three-Stage System

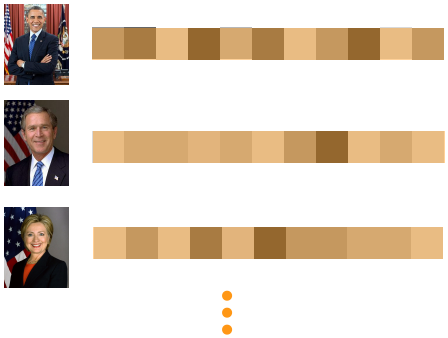


1. Document Enrichment



Tokens, POS Tags, Dependency Relations, Coreference Chains, Sentence Ranks, Wikipedia Entities and their Relatedness

2. Feature Generation



Entity Feature Vectors

3. Salience Classification

dmlc XGBoost

(Chen, SIGKDD '16)

Salient Entities

Experiment

Competitors & Benchmarks

- ▷ CMU-Google System (Dunietz, EACL '14)

  Limitations

  Re-implemented

  - Proprietary modules substituted with open-source tools
    - WAT + scikit learn

- ▷ SEL (Trani, DocEng '16)

  Limitations

  - No comparison with CMU-Google System
  - Benchmark on small dataset
  - Not publicly available

Experiments

Results

System	New York Times			Wikinews		
	Micro			Macro		
	P	R	F1	P	R	F1
CMU-Google (Dunietz, EACL '14)	60.5	63.5	61.5	-	-	-
CMU-Google-ours	58.8	62.6	60.7	42.3	61.0	46.0
SEL (Trani, DocEng '16)	-	-	-	**61.0**	50.0	52.0
SWAT	**62.4**	**66.0**	**64.1**	57.7	**67.0**	**58.3**
	+1.9%	+2.5%	+2.6%	-3.3%	+6.0%	+6.3%

Experiments
Results

Independence from position of salient entities

+14%

2.2
Algorithms for Fact Salience

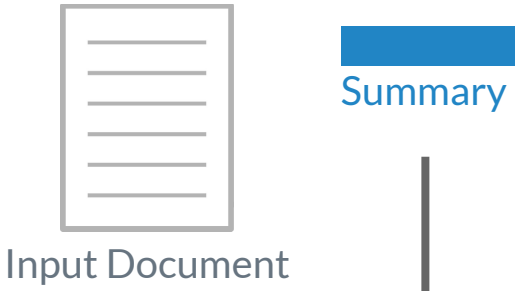
Facts That Matter *
Marco Ponza, Luciano Del Corro and Gerhard Weikum

*work done during an internship at max planck institut informatik

EMNLP 2018
Brussels

Fact Salience

Summarization via Salient Open Facts



Set of Salient Open Facts

Summary

("Abrams", "had been stabbed to death in", "apartment")

("Remains", "were discovered at", "soccer field")

Fact Salience

Contributions

1 We introduce a NEW **Research Task** called **Fact Salience**

- Extraction of relevant information from an input document expressed in the smallest number of facts

2 Proposal of the 1st Fact Salience system:

SalIE

- Fully unsupervised
- Based on PageRank and Clustering
- Public available at `https://github.com/mponza/SalIE`

3 Experiments show that open facts are an effective way to compress information

SalIE Salient Information Extraction

Overview

Input Document

Open Information Extraction

MinIE (Gashteovski, EMNLP 2017)

Open Facts

- *("Abrams", "was 56-years-old native of", "Pittsburgh area")*
- *("Abrams", "had been stabbed to death in", "apartment")*
- *("Apartment", "tending wounds at time of", "murder")*
- *("Cousin of husband", "had gone into", "business")*
- *("Remains", "were discovered at", "soccer field")*
- *("Abrams", "got more involved in", "real estate")*

SalIE

1. Fact Relevance
2. Diversification

Salient Open Facts

SalIE Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score

"Most relevant facts are the ones more central in the input document"

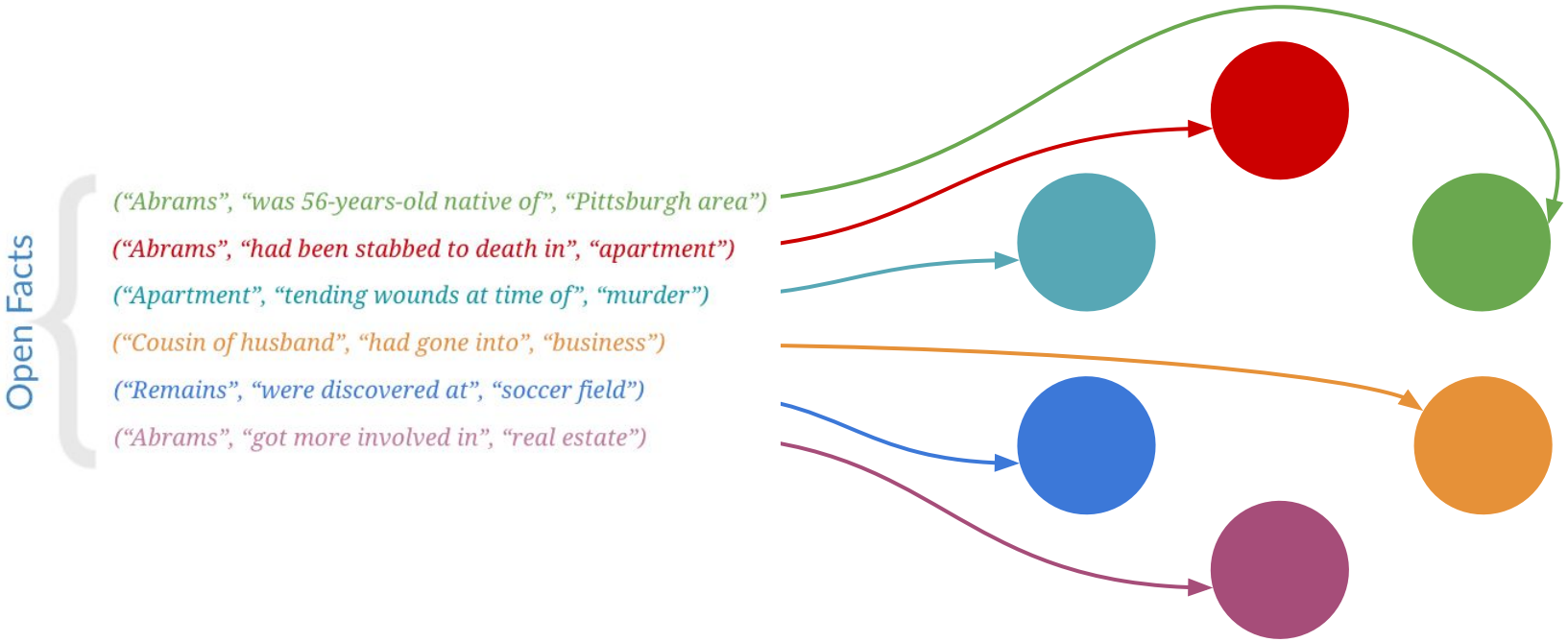
We can use PageRank!

1. How do we define the graph structure?
2. How do we weight the edges?
3. How do we instantiate the teleport vector?

Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score



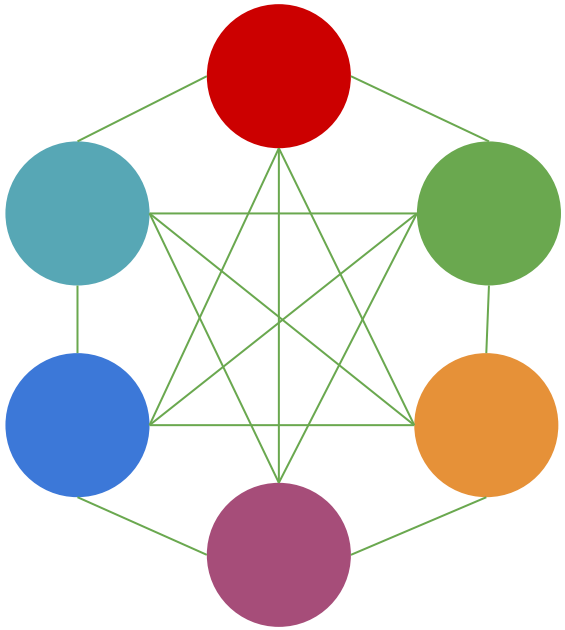
SalIE Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score

Open Facts

- *("Abrams", "was 56-years-old native of", "Pittsburgh area")*
- *("Abrams", "had been stabbed to death in", "apartment")*
- *("Apartment", "tending wounds at time of", "murder")*
- *("Cousin of husband", "had gone into", "business")*
- *("Remains", "were discovered at", "soccer field")*
- *("Abrams", "got more involved in", "real estate")*



1. How do we define the graph structure?
   We can grow a *fully connected graph of facts*!

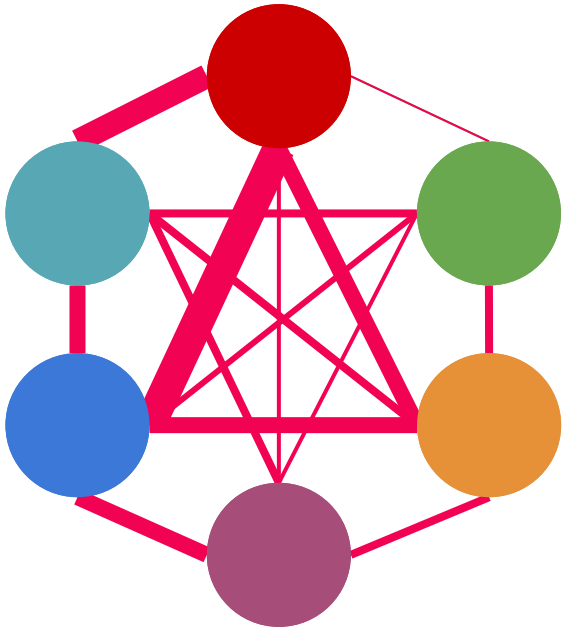
SalIE Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score

Open Facts

- *("Abrams", "was 56-years-old native of", "Pittsburgh area")*
- *("Abrams", "had been stabbed to death in", "apartment")*
- *("Apartment", "tending wounds at time of", "murder")*
- *("Cousin of husband", "had gone into", "business")*
- *("Remains", "were discovered at", "soccer field")*
- *("Abrams", "got more involved in", "real estate")*



2. How do we weight the edges?
   We can use the cosine similarity between facts' embeddings vectors

Salient Information Extraction

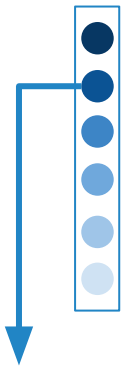
First Stage: Fact Relevance

Provide, for each open fact, a relevance score

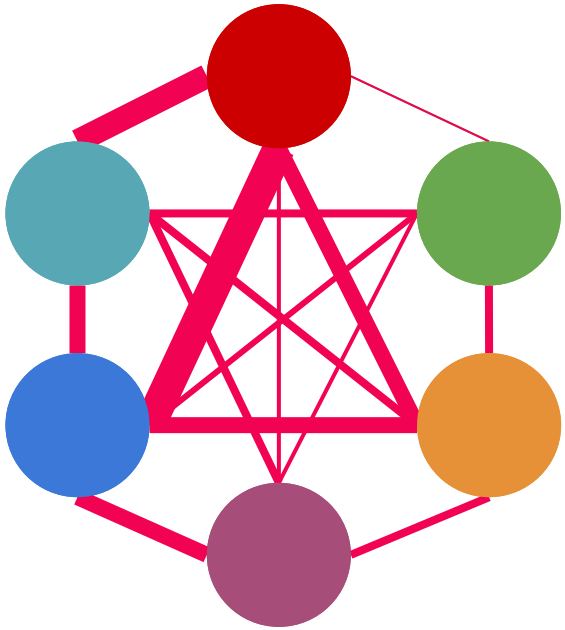
Open Facts

- ("Abrams", "was 56-years-old native of", "Pittsburgh area")
- ("Abrams", "had been stabbed to death in", "apartment")
- ("Apartment", "tending wounds at time of", "murder")
- ("Cousin of husband", "had gone into", "business")
- ("Remains", "were discovered at", "soccer field")
- ("Abrams", "got more involved in", "real estate")

Teleport Vector



Fact Position!



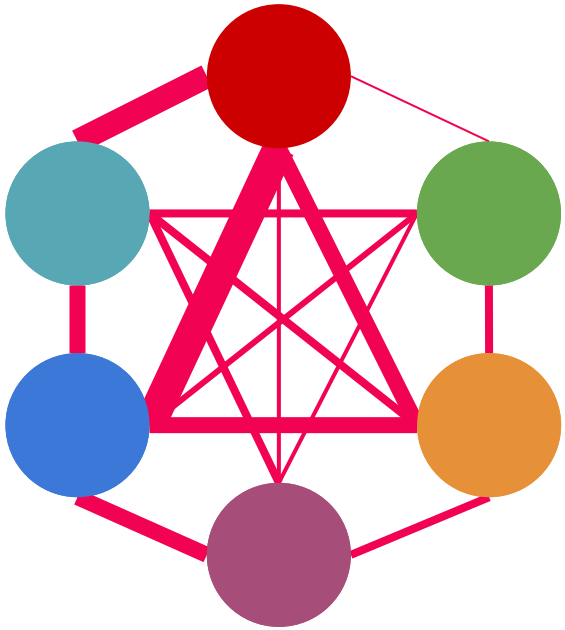
3. How do we instantiate the teleport vector? Each facts' entry in the teleport vector is scored wrt facts' positional information

SalIE Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score

1. Fully connected graph
2. Edges weighted with cosine between word embeddings
3. Teleport vector instantiated as a function of the position



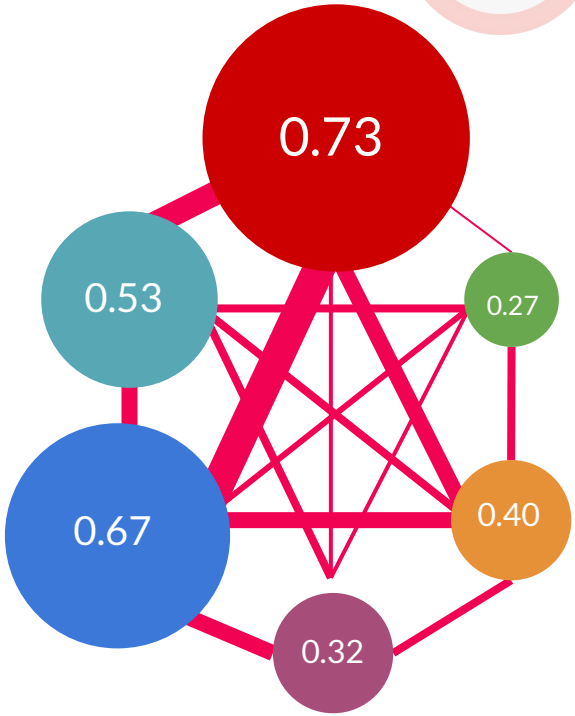
Yes! Now we can run PageRank!

SalIE Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score

1. Fully connected graph
2. Edges weighted with cosine between word embeddings
3. Teleport vector instantiated as a function of the position



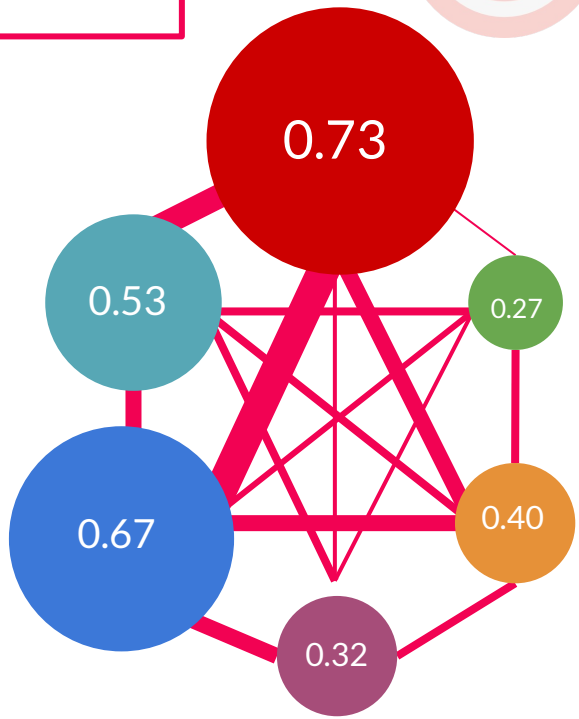
Yes! Now we can run PageRank!

SalIE Salient Information Extraction

First Stage: Fact Relevance

Provide, for each open fact, a relevance score

("Abrams", "was 56-years-old native of", "Pittsburgh area")	0.27
("Abrams", "had been stabbed to death in", "apartment")	0.73
("Apartment", "tending wounds at time of", "murder")	0.53
("Cousin of husband", "had gone into", "business")	0.40
("Remains", "were discovered at", "soccer field")	0.67
("Abrams", "got more involved in", "real estate")	0.32



Salient Information Extraction

Second Stage: Fact Diversification

Maximize the information in the *smallest number of facts*

- Salient open facts should provide a wide spectrum of the information in the document content

↓

Salient Fact Diversification with Clustering

SalIE Salient Information Extraction

Second Stage: Fact Diversification

Maximize the information in the *smallest number of facts*

("Abrams", "was 56-years-old native of", "Pittsburgh area")	0.27
("Abrams", "had been stabbed to death in", "apartment")	0.73
("Apartment", "tending wounds at time of", "murder")	0.53
("Cousin of husband", "had gone into", "business")	0.40
("Remains", "were discovered beside warehouse at edge of", "cinder-topped soccer field on outskirts of Panama City")	0.67
("Abrams", "got more involved in", "real estate")	0.32

Ranked Facts
(from the First Stage)

→

Clustering Rule

Facts are clustered together wrt their subject

0.73 - ("Abrams", "had been stabbed to death in", "apartment")

0.32 - ("Abrams", "got more involved in", "real estate")

0.27 - ("Abrams", "was 56-years-old native of", "Pittsburgh area")

0.67 - ("Remains", "were discovered at", "soccer field")

0.53 - ("Apartment", "tending wounds at time of", "murder")

0.40 - ("Cousin of husband", "had gone into", "business")

SalIE Salient Information Extraction

Second Stage: Fact Diversification

Maximize the information in the *smallest number of facts*

0.73 - ("Abrams", "had been stabbed to death in", "apartment")
0.32 - ("Abrams", "got more involved in", "real estate")
0.27 - ("Abrams", "was 56-years-old native of", "Pittsburgh area")

0.67 - ("Remains", "were discovered beside warehouse at edge of", "cinder-topped soccer field on outskirts of Panama City")

0.53 - ("Apartment", "tending wounds at time of", "murder")

0.40 - ("Cousin of husband", "had gone into", "business")

→

Diversification Rule

From each cluster, select the fact with the strongest PageRank score

Final Output

("Abrams", "had been stabbed to death in", "apartment")

("Remains", "were discovered at", "soccer field")

("Apartment", "tending wounds at time of", "murder")

("Cousin of husband", "had gone into", "business")

Experiments
Evaluation

- Dataset: New York Times
  - ~4K news each one with a manually annotated human summary
- Methodology: Evaluate how salient are top-*k* facts
  - We evaluate 5 configurations: From top-1 to top-5 facts
- Metrics: How close are top-*k* facts to a human summary?
  - ROUGE (Lin et al. 2004) metrics
    - % of unigrams, bigrams, subsequences between generated summary and ground-truth

Experiments

Baselines

- Position Baseline
  - Returns facts wrt their order in the document
  - Standard for saliency and summarization tasks
- TextRank (Mihalcea, EMNLP'04)
  - Graph-based summarizer based on token-overlap between sentences
  - Re-implemented to work at fact-level
- Berkeley (Durrett, ACL'16)
  - Supervised summarizer based on handcrafted features and SVM

Experiments

Results

Method	ROUGE-1					ROUGE-L				
	1	2	3	4	5	1	2	3	4	5
Position	13.9	20.4	24.8	27.8	29.7	12.8	18.1	21.8	24.4	26.0
TextRank	15.2	21.5	24.5	26.1	26.8	13.0	17.5	19.8	21.3	22.0
Berkeley	8.5	18.0	25.4	**30.4**	**34.1**	8.00	16.3	22.5	**26.7**	**29.7**
SalIE	**17.1**	**24.2**	**28.0**	30.0	30.9	**15.3**	**21.2**	**24.3**	26.0	26.8
	+1.9	**+2.7**	**+2.6**	**+1.9**	**+0.5**	**+2.3**	**+3.1**	**+1.9**	**+1.5**	**+0.2**

- General improvements over all metrics
- Facts are an effective way to compress information!

3
Algorithms for Expert Finding

WISER: A Semantic Approach for Expert Finding in Academia based on Entity Linking

Paolo Cifariello, Paolo Ferragina, and Marco Ponza

ELSEVIER

Information Systems 2019

Expert Finding

- Searching for experts with respect to an input topic

  - Extremely challenging task: *Who is an expert?*

    The notion of expertise is hard to formalize as well as to be modeled (Balog, FTIR'12)

...so difficult that literature refers to expertise as "tacit knowledge"!

  - Expertise is actually carried by people in their minds
  - Machines have only one way to access to people expertise

    → Artifacts (e.g., papers, emails, ...) people write to share their expertise!

Experiments
Contributions

- New Expert Finding system

WISER
Wikipedia Expertise Ranking

- Fully unsupervised
- Jointly combines classical retrieval techniques with the Wikipedia KG via Entity Linking

elastic mongoDB

TAG ME

WikipediA RELATEDNESS

Indexing

Every authors' profile is modeled through a small Wikipedia graph...

Query Time

...used to design new profile-centric scoring strategies for the retrieval of experts!

Wikipedia Expertise Ranking

Indexing

Wikipedia Expertise Ranking

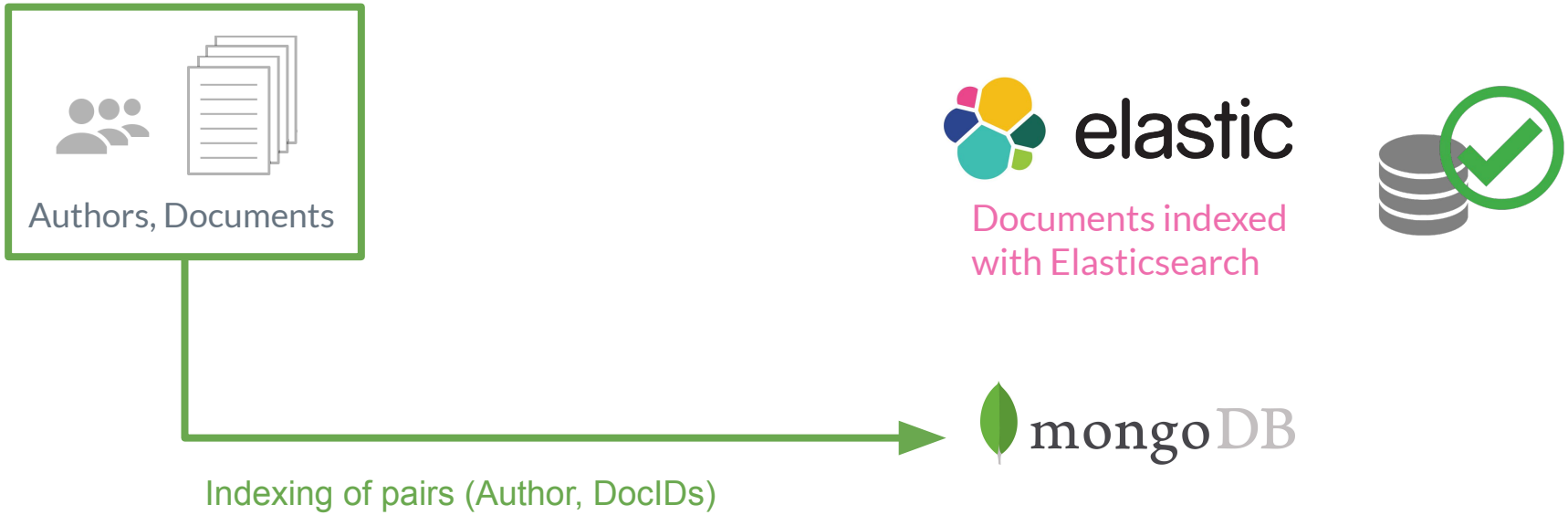
Indexing

Authors, Documents

Documents indexed with Elasticsearch

Wikipedia Expertise Ranking

Indexing



Wikipedia Expertise Ranking

Indexing

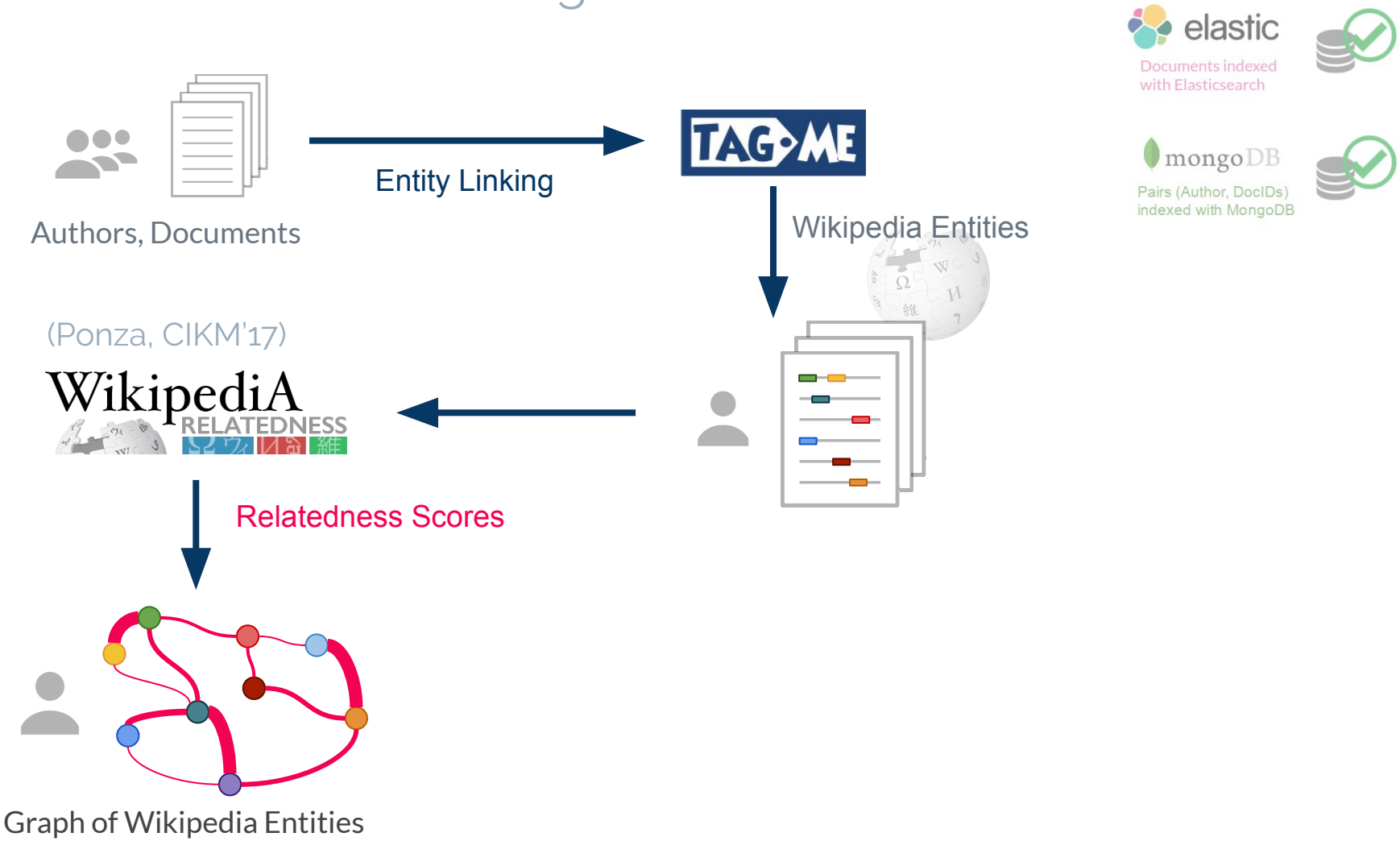
Authors, Documents

Documents indexed with Elasticsearch

Pairs of (Author, DocIDs)

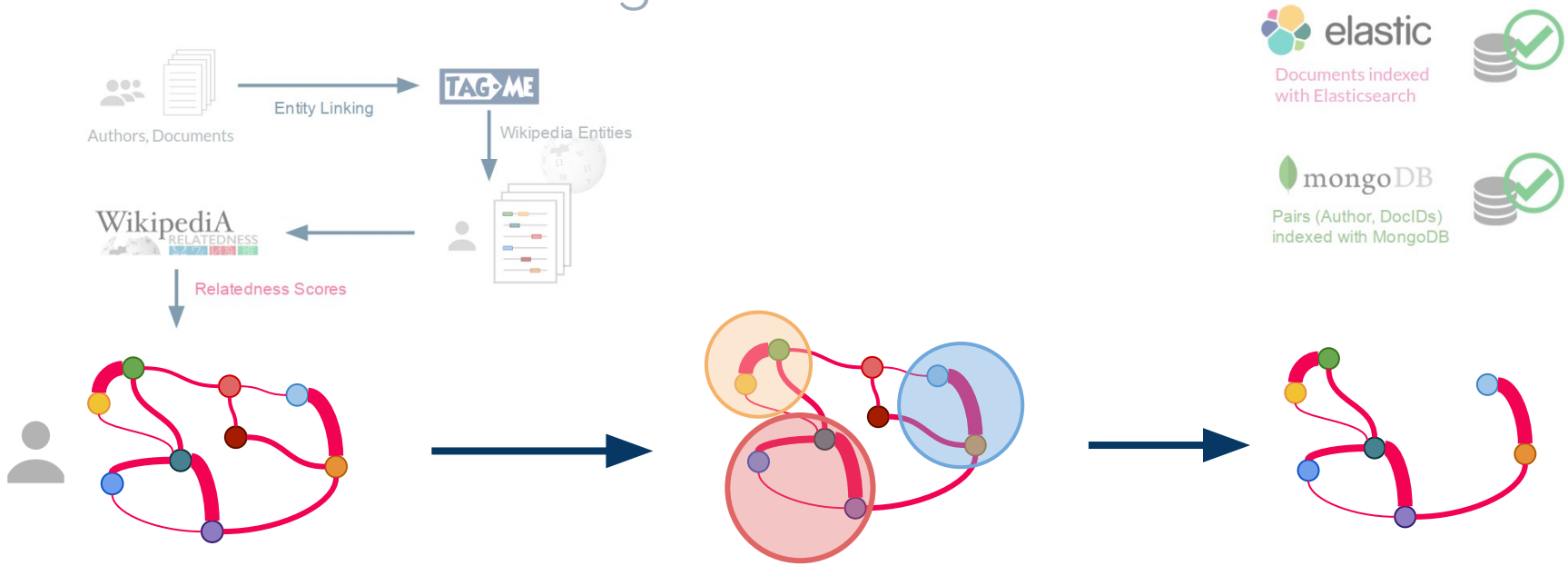
Wikipedia Expertise Ranking

Indexing



Wikipedia Expertise Ranking

Indexing



Graph of Wikipedia Entities

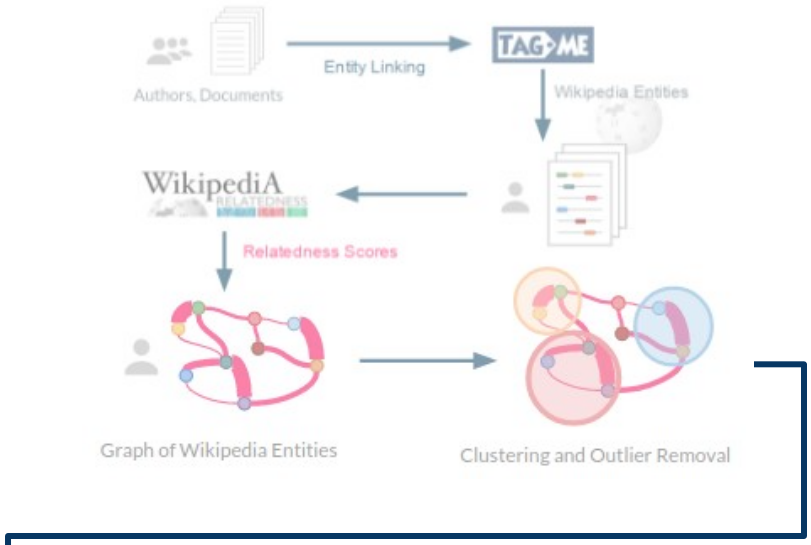
Clustering and Outlier Removal of Wikipedia Entities

Cleaned Graph of Wikipedia Entities

- HDBScan Algorithm (McInnes, IEEE'17)
- Conservative Approach:
  >=20% of nodes marked as outliers implies no cleaning

Wikipedia Expertise Ranking

Indexing

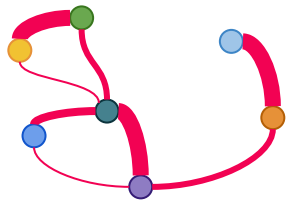


elastic

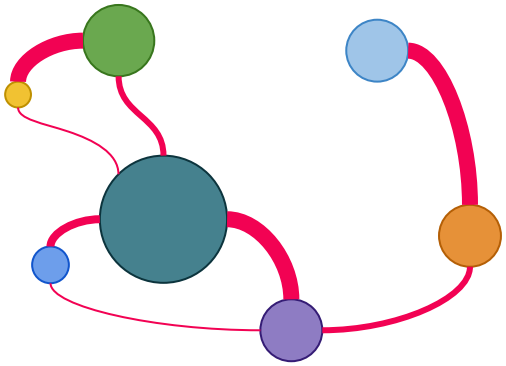
Documents indexed with Elasticsearch

mongoDB

Pairs (Author, DocIDs) indexed with MongoDB



Cleaned Graph of Wikipedia Entities

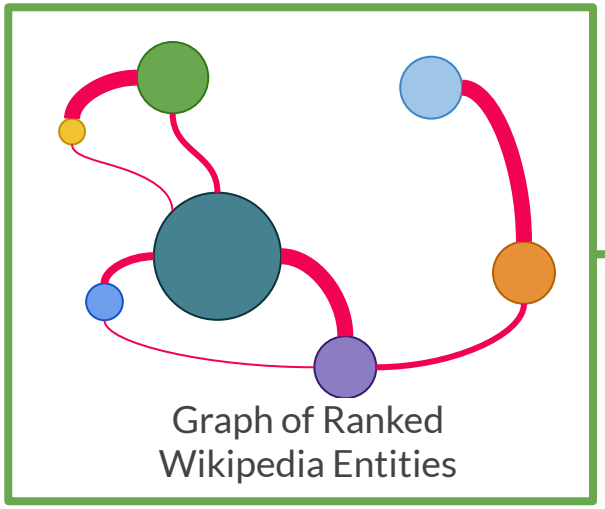


Ranking of Wikipedia Entities

- PageRank Algorithm
- Teleport vector instantiated by taking into account the frequency of an entity annotated in the documents of the authors

Wikipedia Expertise Ranking

Indexing



elastic

Documents indexed
with Elasticsearch

mongoDB

Pairs of (Author, DocIDs)

Wikipedia Expertise Ranking

Indexing

Indexing Completed!

elastic

Documents indexed with Elasticsearch

mongoDB

Pairs of (Author, DocIDs)
+
Pairs of(Author, Graph of Ranked Wikipedia Entities)

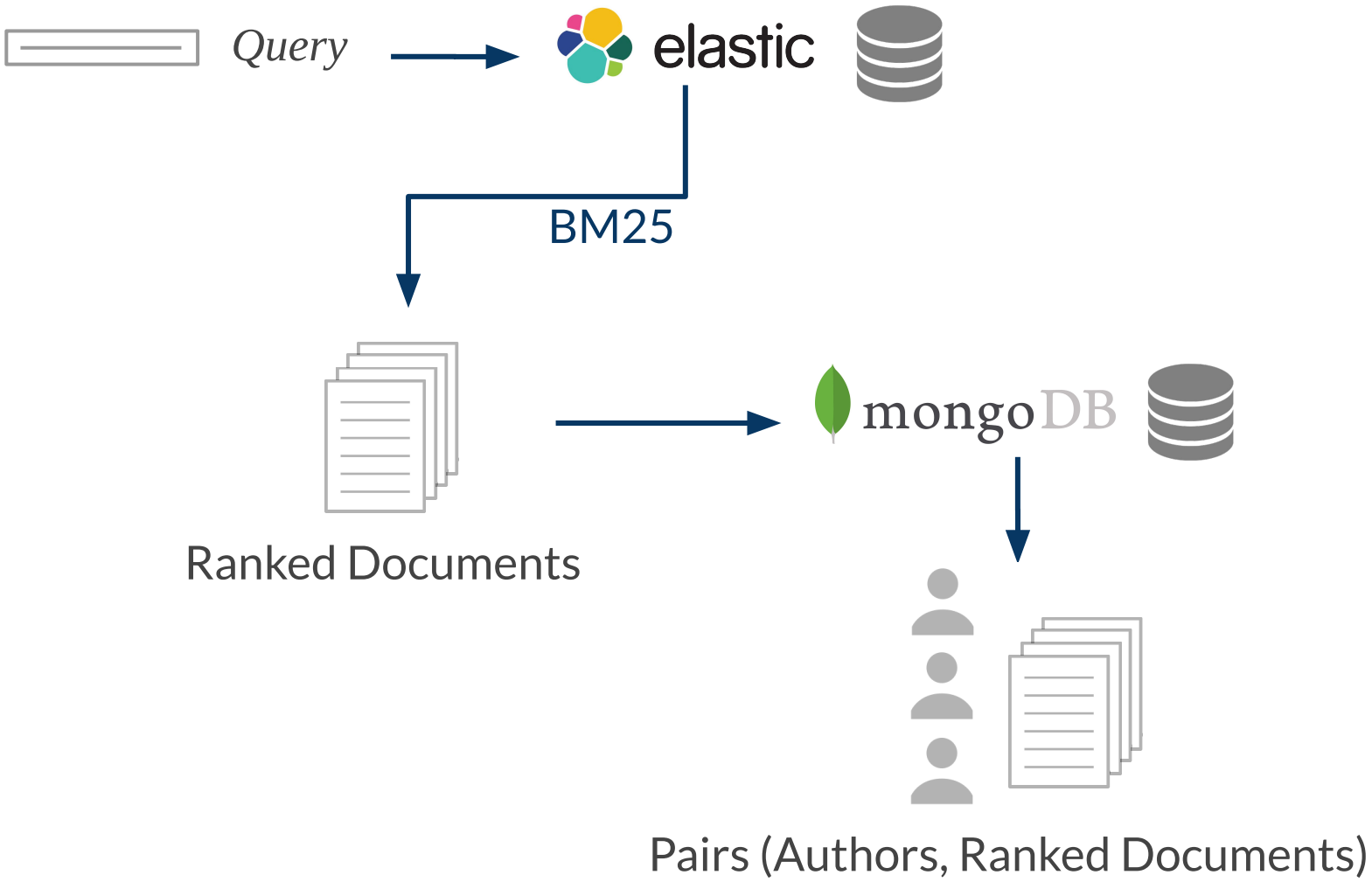
Wikipedia Expertise Ranking

Query Time: Two Strategies

- Jointly combine two different Authors' Scoring Strategies
  - Document-Centric
    1. Retrieve relevant documents
    2. Score each author wrt documents' rank (BM25)
  - Profile-Centric
    1. Retrieve relevant authors (wrt query Wikipedia entities)
    2. Score each author wrt entities relevance

Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy

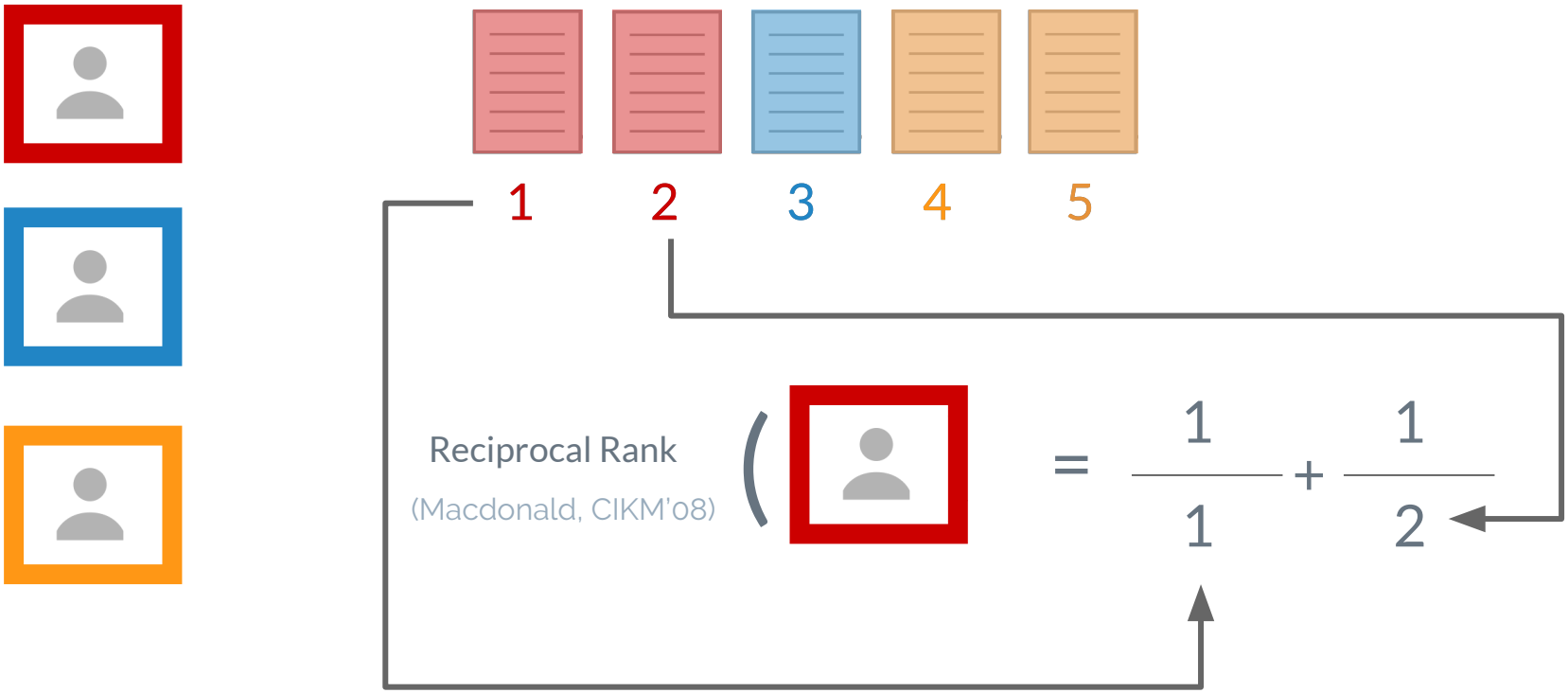


Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy

Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy



Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy

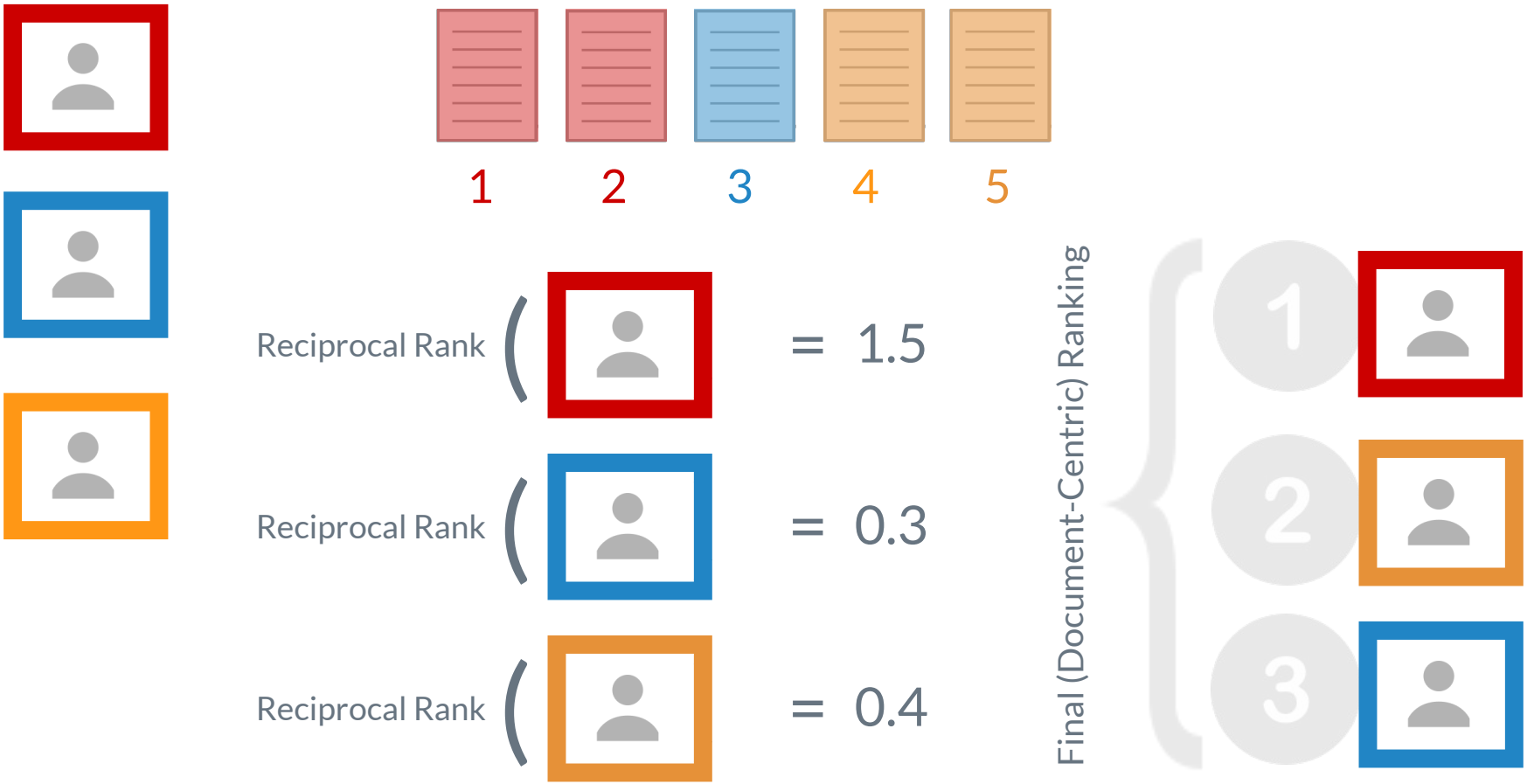
Reciprocal Rank

(Macdonald, CIKM'08)

$$= \underbrace{\frac{1}{1} + \frac{1}{2}}_{1.5}$$

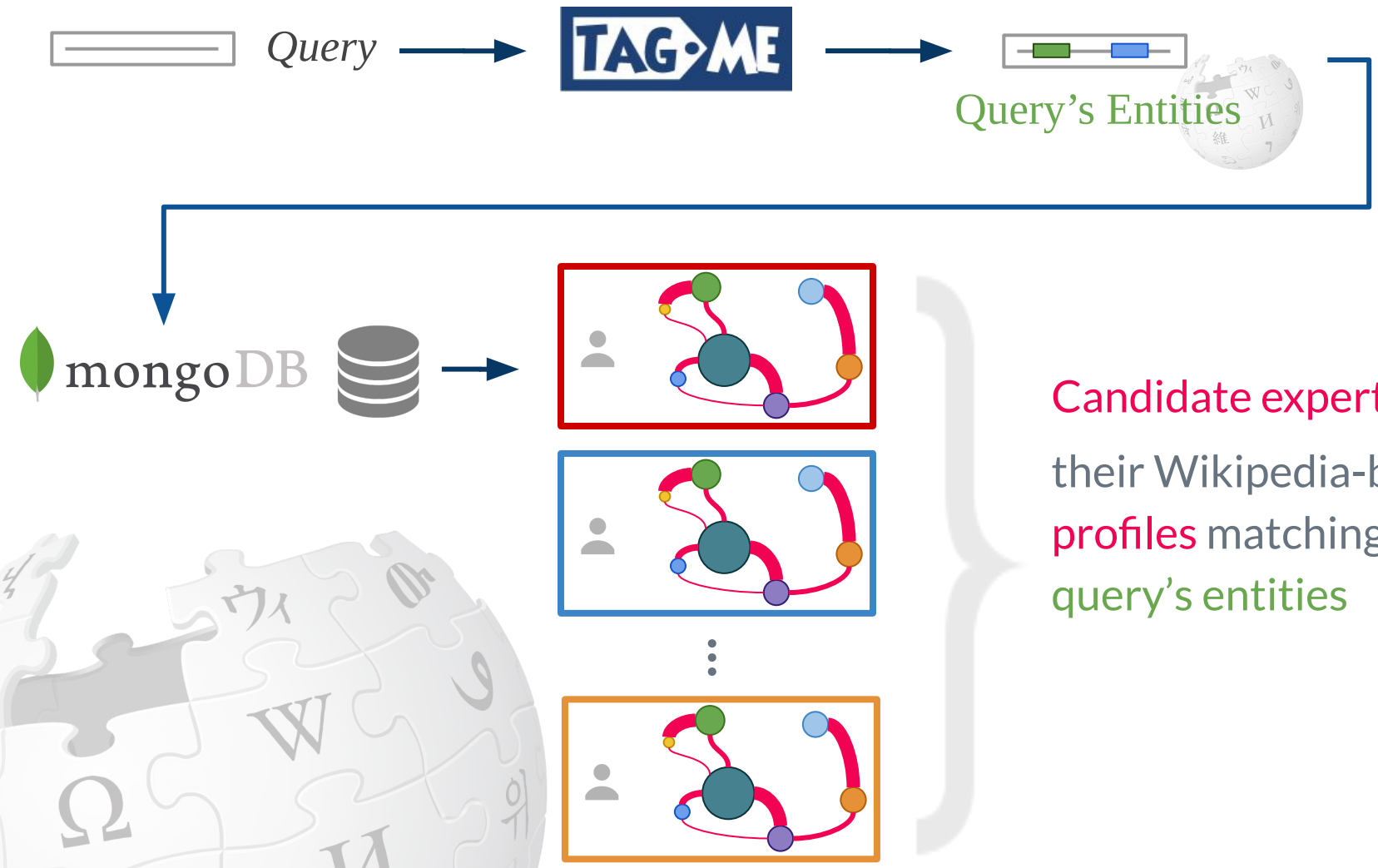
Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy



Wikipedia Expertise Ranking

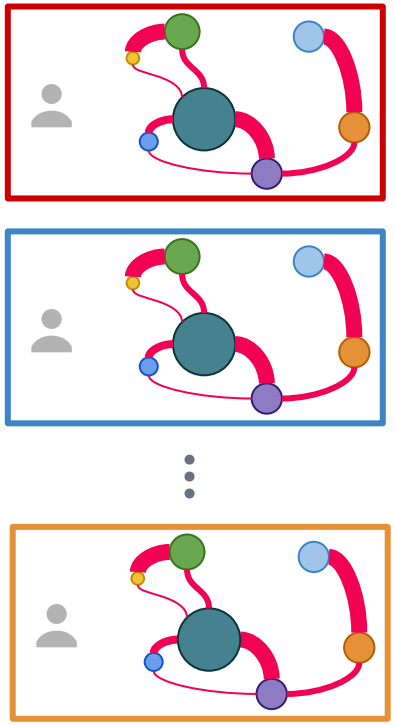
Query Time: Profile-Centric Strategy



Candidate experts and their Wikipedia-based profiles matching the query's entities

Wikipedia Expertise Ranking

Query Time: Profile-Centric Strategy

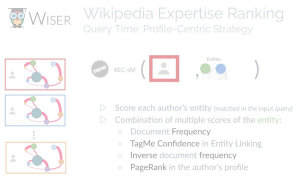


- Score each author's entity (matched in the input query)
- Combination of multiple scores of the entity:
  - Document Frequency
  - TagMe Confidence in Entity Linking
  - Inverse document frequency
  - PageRank in the author's profile

Wikipedia Expertise Ranking

Query Time: Data Fusion

- We have two different rankings
  - Document-Centric Ranking
  - Profile-Centric Ranking



Final ranking of experts is given by the Reciprocal Rank (Macdonald, CIKM'08) between the Product of these two ranking scores

Final Score () $= \dfrac{1}{\text{Doc-Cent Rank}(\square)} \cdot \dfrac{1}{\text{Prof-Cent Rank}(\square)}$

Experiments
Benchmark

- TU Dataset (Berendsen, DBWIR'13)
  - ~31K documents (*largest available*)
  - ~1K researchers
  - ~1K test queries
  - Human-assessed Ground-Truth

- Other systems
  - JM Model (Balog, SIGIR'06)
    - Based on Frequency statistics between (Author, Keywords)
  - Log-Linear (Van Gysel, WWW'16)
    - Based Deep Learning (each author's profile is represented with an embedding vector)
  - Ensemble
    - Product Reciprocal Rank between JM Model and Log-Linear

Experiments

Results

Method	MAP	MRR	P@5	P@10	NDCG@100
JM Model	0.253	0.302	0.108	0.081	0.394
Log-Linear	0.287	0.363	0.134	0.092	0.425
Ensemble	0.331	0.402	0.156	**0.105**	0.477
WISER	**0.385**	**0.459**	**0.163**	**0.105**	**0.513**
	+5.4%	+5.7%	+0.7%		+3.6%

Indexing Experts at the University of Pisa

URL: https://wiser.d4science.org

WISER
Wikipedia Expertise Ranking

Search for expertise...

Search by Expertise | Search by Name | Search by Department

- ~1.5K Authors
- ~65K Documents (papers' abstracts)
- ~35K Research Topics
- More than 1K queries and ~2K profiles view in few months
- Currently used by UniPi's Technology Transfer Office

WISER

Main Topics | Main Areas | Stream Graph | Tag Cloud | Publications | Survey

Select the range of years to analyze.

1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

Ra...	Entity	Count	Doc. count	Years	Wiser score
1	Recurrent neural network	26	15		
2	Artificial neural network	44	24		
3	Tree (data structure)	24	17		
4	Machine learning	16	14		
5	Recursive neural network	16	16		
6	Quantitative structure–activi...	30	16		
7	Echo state network	14	13		
8	Mathematical model	39	28		
9	Prediction	19	17		
10	Generative model	9	7		
11	Group representation	22	14		
12	Empiricism	15	15		
13	Data model	8	8		
14	Data set	20	14		
15	Reservoir computing	8	8		

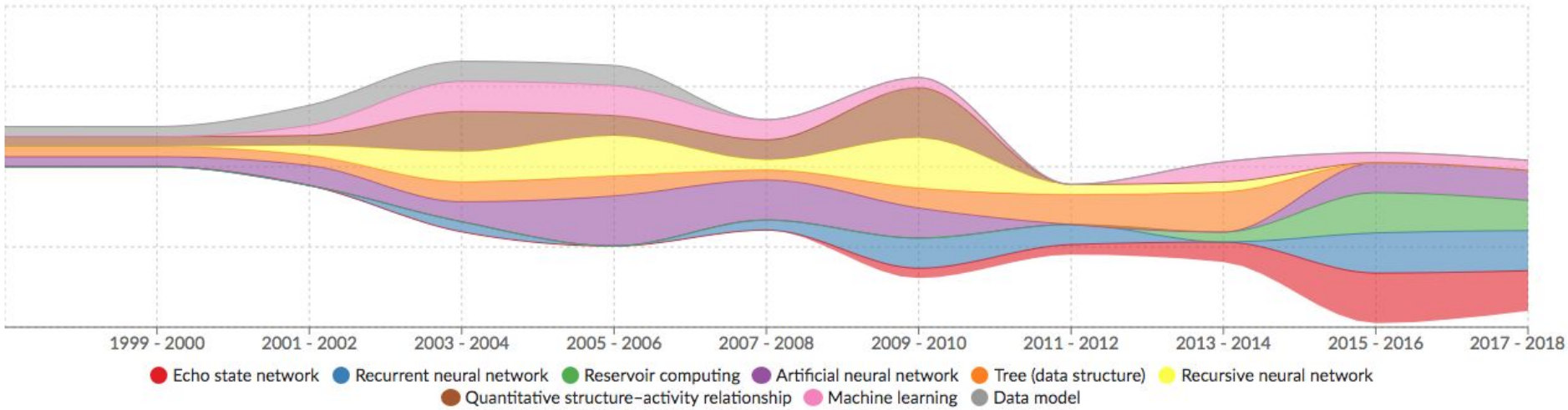
Previous | Page 1 of 24 | 15 rows | Next

WISER

Main Topics | Main Areas | **Stream Graph** | Tag Cloud | Publications | Survey

Select range of years to examine. 6

Select how many topics for each range 4



4
Future Research Directions

Future Research Directions

- Entity Relatedness

  - Apply our Two-Stage Framework over other KGs
  - Extend it to *labels* associated to entities' *relationships*

Leonardo da Vinci

place of birth →

Anchiano

Conclusion and Future Directions

- Entity and Fact Salience
  - Improve quality of Entity Salience annotations for NYT
  - Entity Linking research will start focusing on efficiency (*best solutions are currently extremely slow, especially when applied over large-scale!*)
  - Number of Applications for both Entity and Fact Salience
    - News Credibility (Popat, EMNLP'18)
    - KGs Construction (Nguyen, VLDB'18)
    - Facts Contextualization (Voskarides, SIGIR'18)

Conclusion and Future Directions

- Expert Finding

  - **Fine-grain Clustering** of Wikipedia entities for the visualization of groups of topics of an expert

    Algorithms Computer Science

    Gzip Burrows-Wheeler Transform

    - Classical Clustering Algorithms generate one single cluster
    - Gzip and Burrows-Wheeler should actually belong to a domain-specific cluster

  - Apply our graph-based profiling technique to **other domains**, e.g., recommendation systems, conversational AI

Thanks!

Any questions?

Contributions

1 Entity Relatedness

A Two-Stage Framework for Computing Entity Relatedness in Wikipedia
Marco Ponza, Paolo Ferragina and Soumen Chakrabarti

CIKM 2017 SINGAPORE

2 Entity Salience

Document Aboutness via Sophisticated Syntactic and Semantic Features
Marco Ponza, Paolo Ferragina and Francesco Piccinno

NLdb 2017 Liège, Belgium

SWAT: A System for Detecting Salient Wikipedia Entities in Texts
Marco Ponza, Paolo Ferragina and Francesco Piccinno

WILEY Computational Intelligence 2019

3 Fact Salience

Facts That Matter
Marco Ponza, Luciano Del Corro and Gerhard Weikum

EMNLP 2018 Brussels

4 Expert Finding

WISER: A Semantic Approach for Expert Finding in Academia based on Entity Linking
Paolo Cifariello, Paolo Ferragina and Marco Ponza

ELSEVIER *Information Systems 2019*