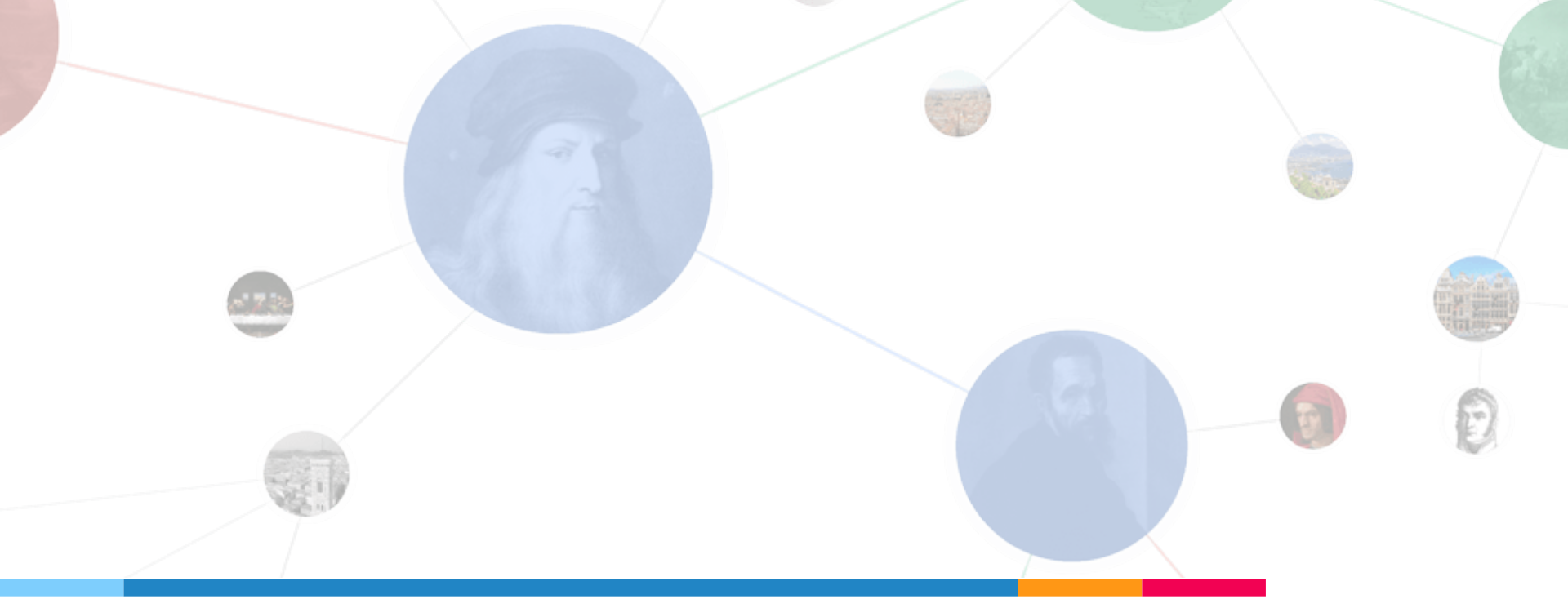


Algorithms for Knowledge and Information Extraction in Text with Wikipedia

Marco Ponza

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

The New York Times

Introduction

Text Understanding

Easy for humans

Hard for machines

Introduction

Text Understanding

Leonardo is the scientist who painted Mona Lisa


Science


Italy


Renaissance


Cartography


Leonardo da Vinci


Art


Louvre


Mona Lisa (painting)


Florence

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”)

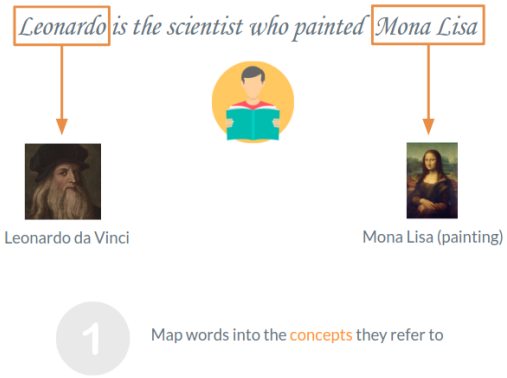
2

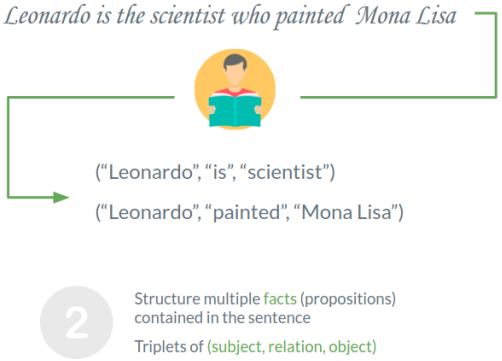
Structure multiple facts (propositions) contained in the sentence

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, and (3) quantifying how much their are related

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*

Limitations

Need for a more structured and efficient algorithmic 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 structured algorithmic 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 structured algorithmic 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 structured algorithmic 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

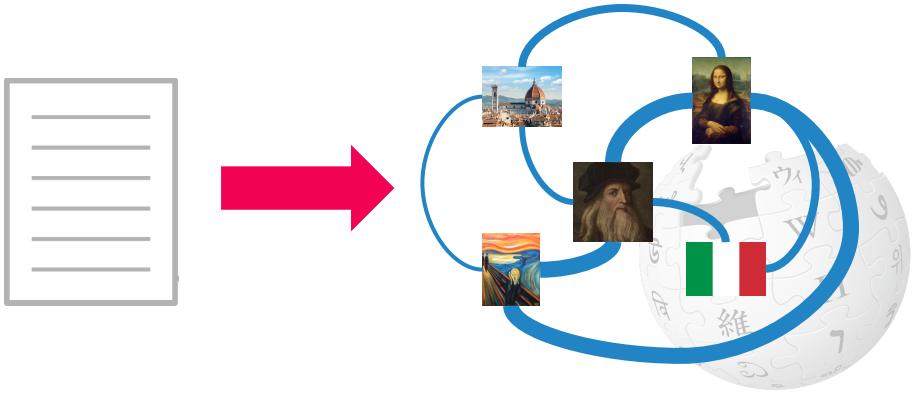
Leonardo da Vinci

Mona Lisa (painting)

Introduction

- Need for a more structured algorithmic 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 ✓ → Wikipedia entities are unique and they represent a real-world concept
- Understanding of real-world entities ✓ → Wikipedia entities are unique and they represent a real-world concept
- Structured facts ✓ → OpenIE preserves subject-relation-object structure

3
Algorithms for Expert Finding

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

Paolo Cifariello, Paolo Ferragina, Marco Ponza

Under review at Information Systems (Elsevier Journal)

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

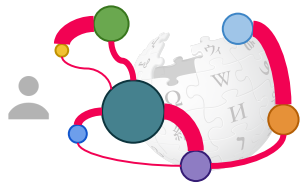
elastic mongoDB

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

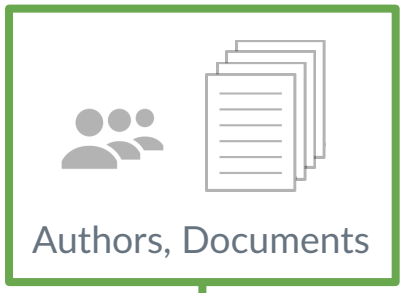
Indexing

Authors, Documents

Documents
indexed
with Elasticsearch

Wikipedia Expertise Ranking

Indexing



elastic

Documents
indexed
with Elasticsearch

mongoDB

Indexing of pairs (Author, DocIDs)

Wikipedia Expertise Ranking

Indexing

WISER

Authors, Documents

elastic

Documents
indexed
with Elasticsearch

mongoDB

Pairs of (Author, DocIDs)

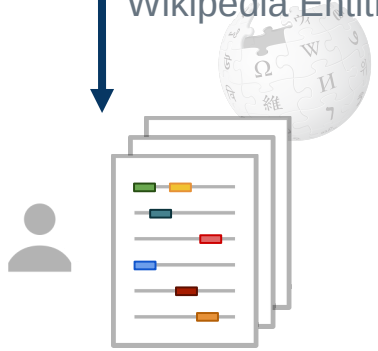
Wikipedia Expertise Ranking

Indexing

Authors, Documents

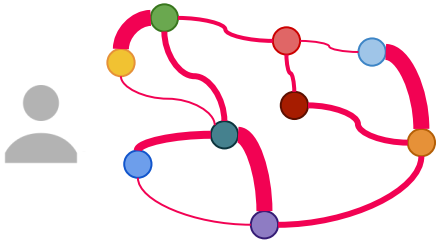
Entity Linking

Wikipedia Entities



(Ponza, CIKM'17)

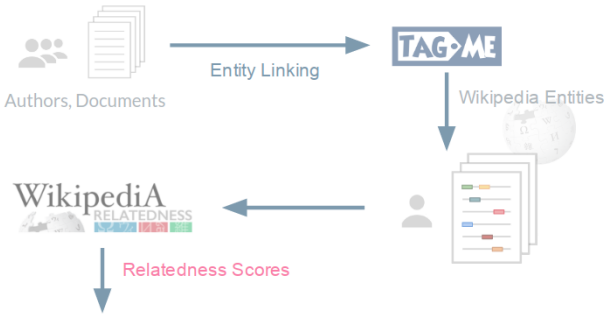
Relatedness Scores

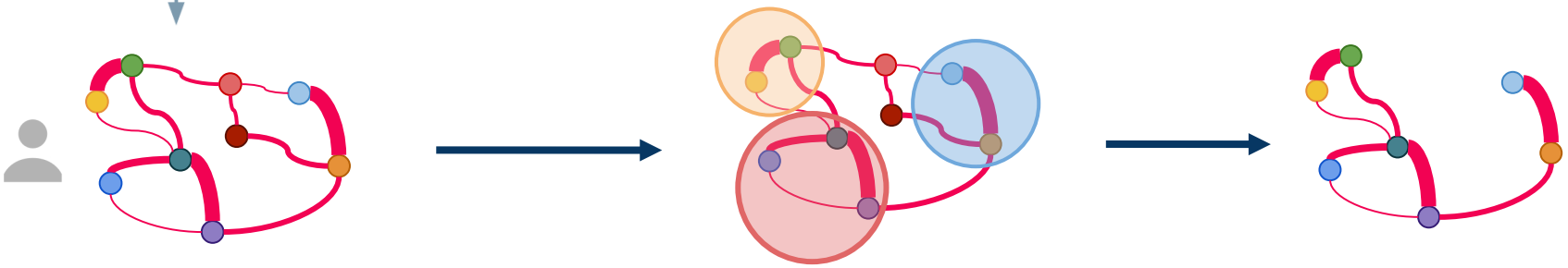


Graph of Wikipedia Entities

Wikipedia Expertise Ranking

Indexing





Graph of Wikipedia Entities

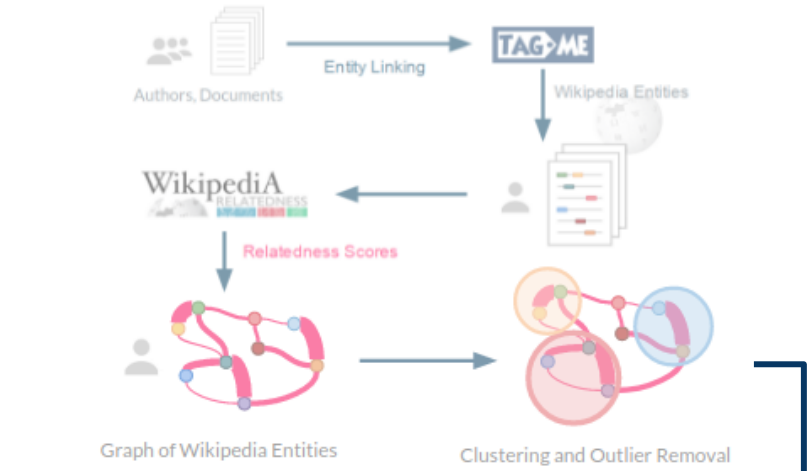
Clustering and Outlier Removal of Wikipedia Entities

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

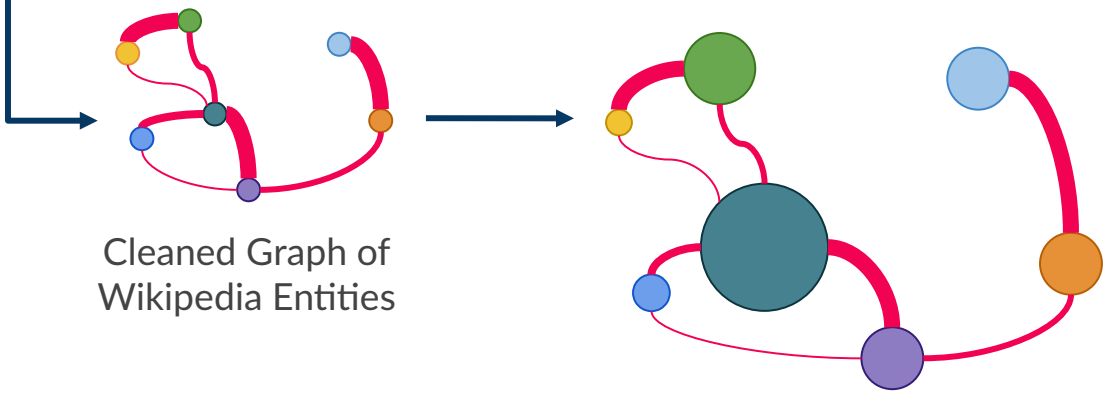


elastic

Documents indexed with Elasticsearch

mongoDB

Pairs (Author, DocIDs) indexed with MongoDB



Cleaned Graph of Wikipedia Entities

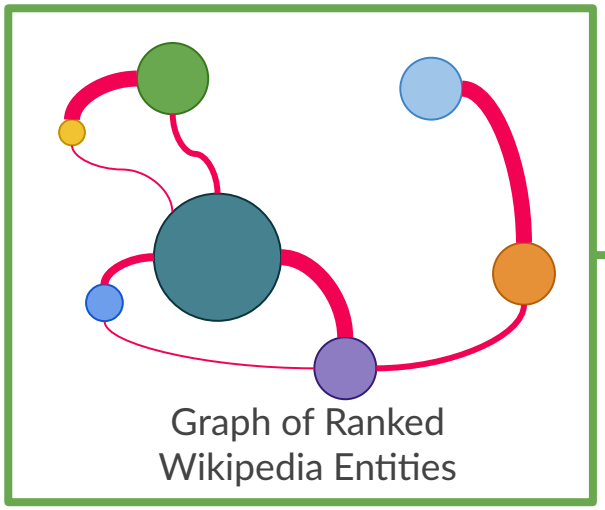
Ranking of Wikipedia Entities

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

Wikipedia Expertise Ranking

Indexing

WISER



elastic

Documents
indexed
with Elasticsearch

mongoDB

Pairs of (Author, DocIDs)

Wikipedia Expertise Ranking

Indexing

WISER

Indexing Completed!

elastic

Documents
indexed
with Elasticsearch

mongoDB

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

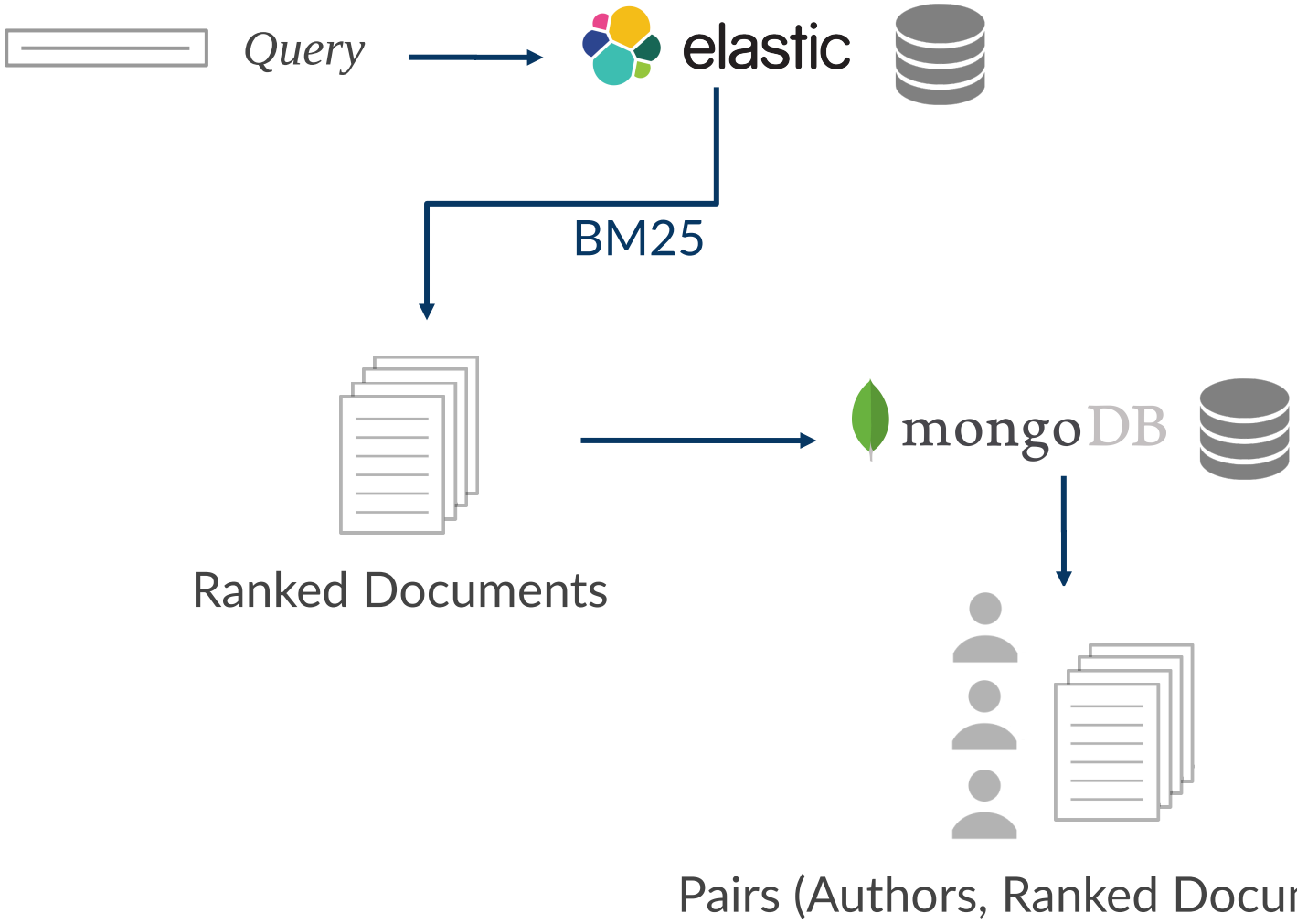
Wikipedia Expertise Ranking

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

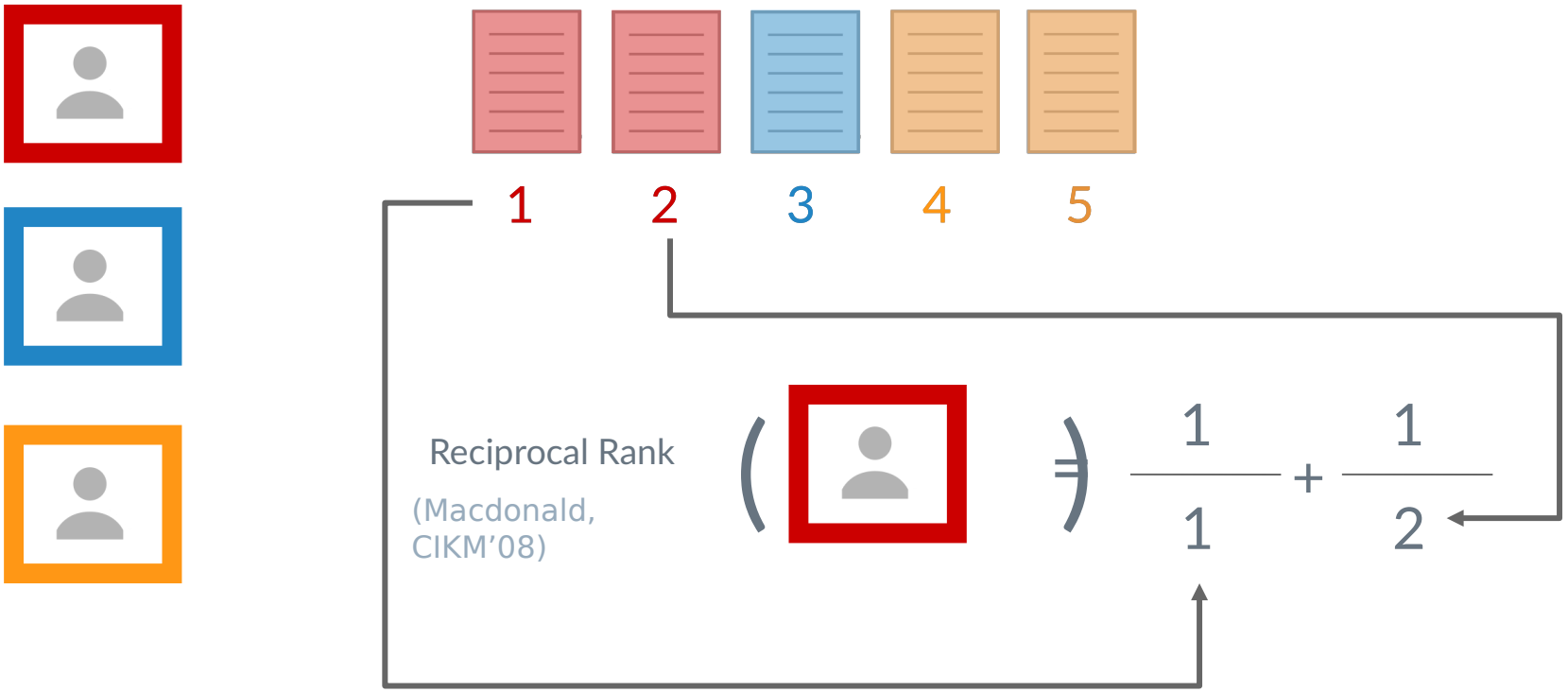


Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy

Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy



Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy

1 2 3 4 5

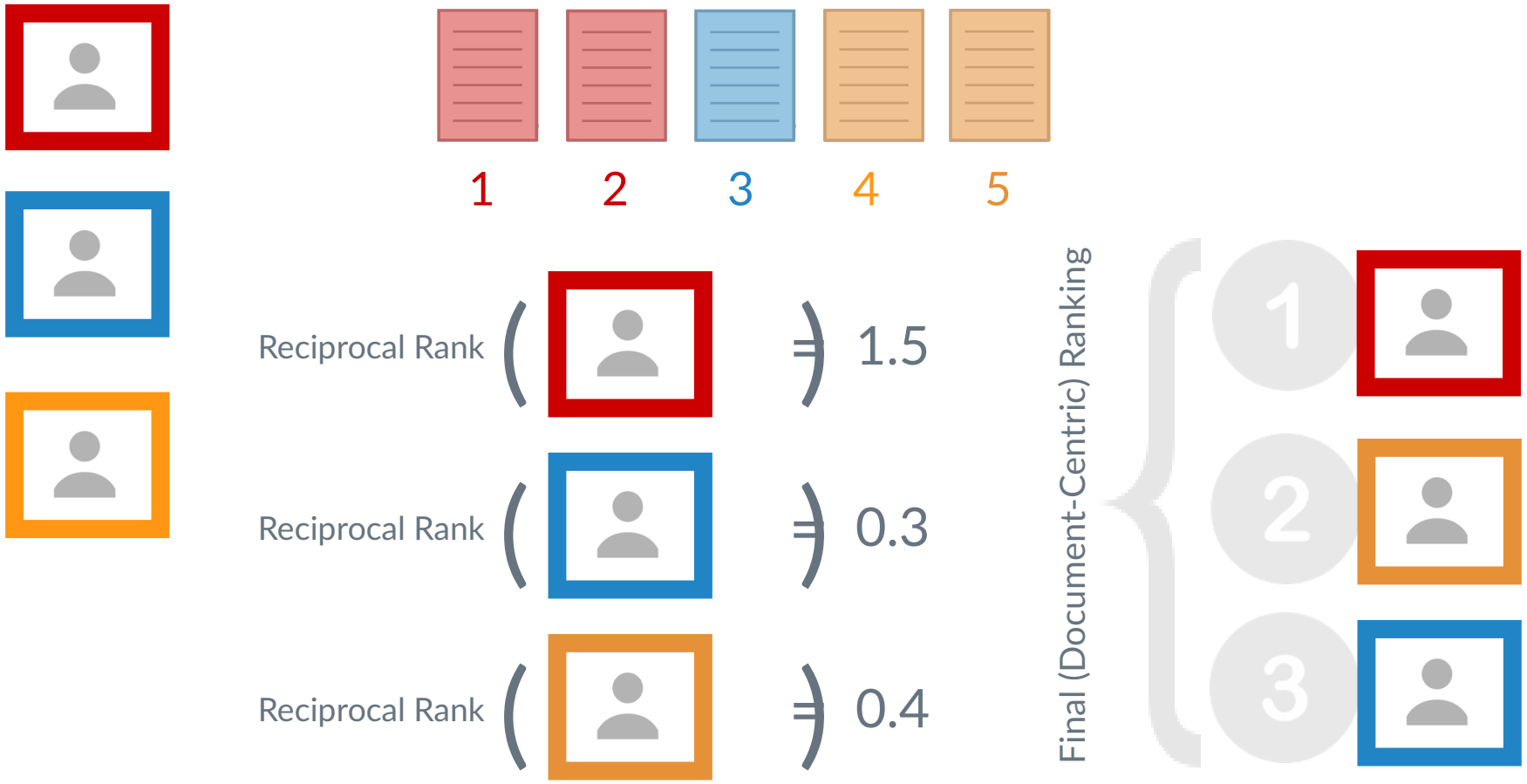
Reciprocal Rank

(Macdonald, CIKM'08)

$$\left(\ \right) = \underbrace{\frac{1}{1} + \frac{1}{2}}_{1.5}$$

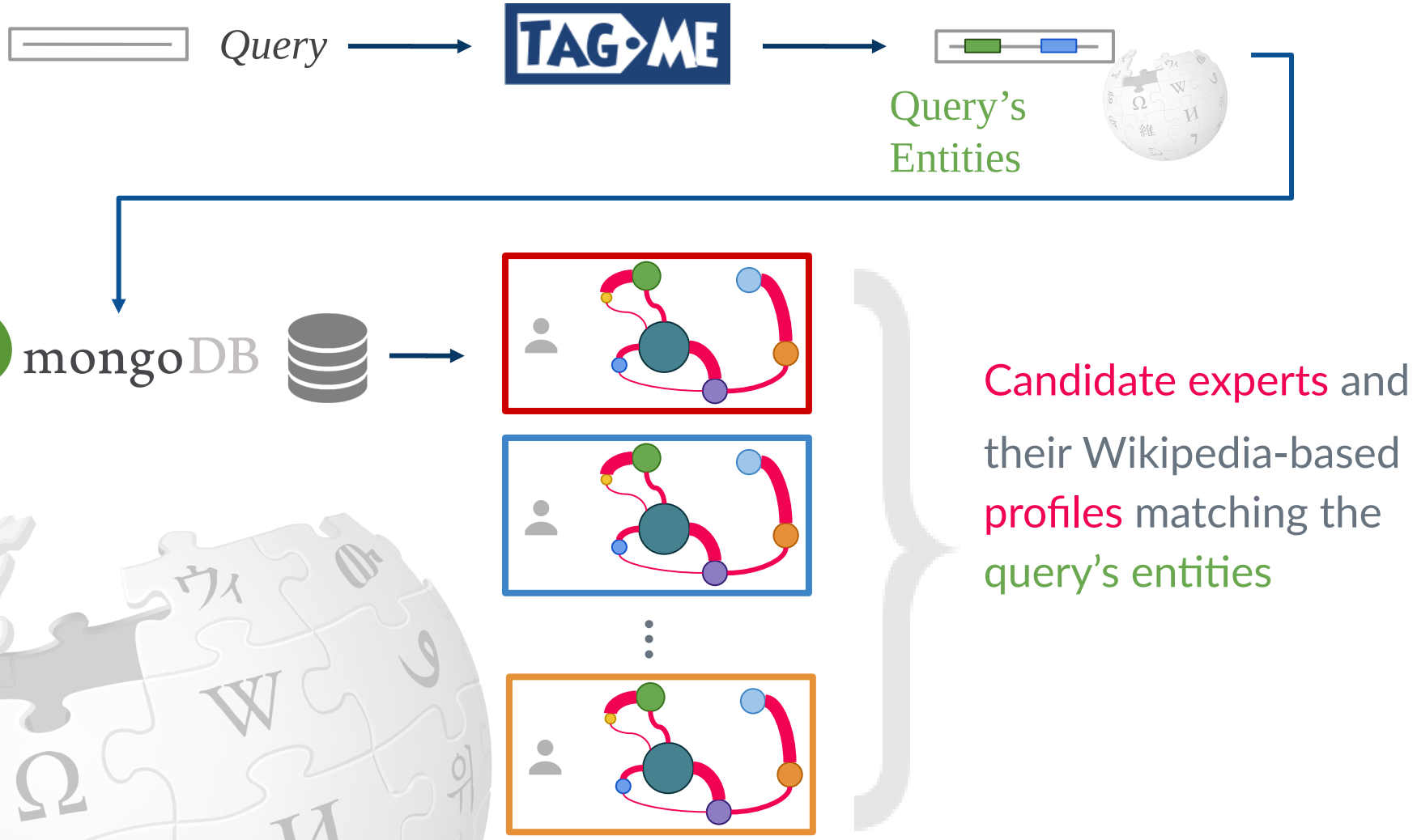
Wikipedia Expertise Ranking

Query Time: Document-Centric Strategy



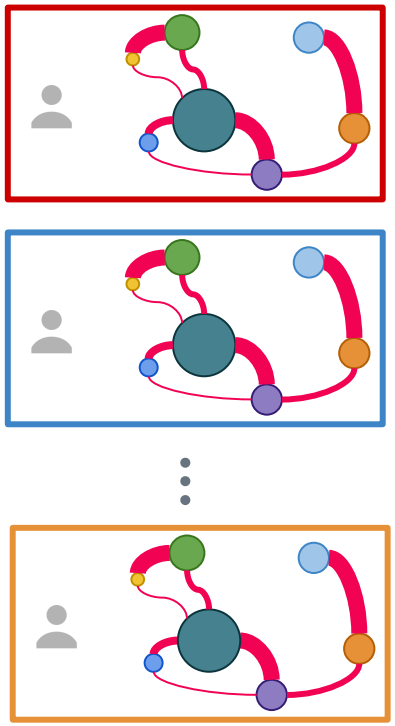
Wikipedia Expertise Ranking

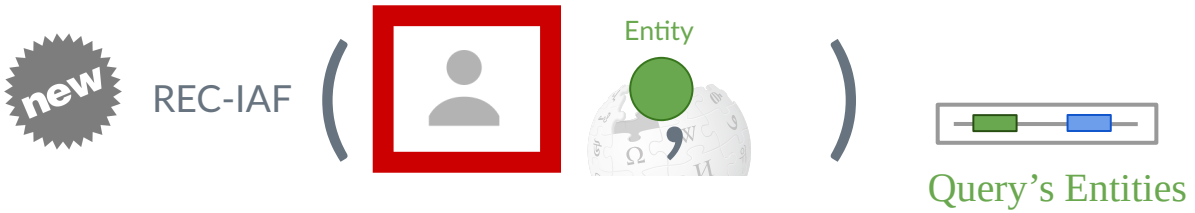
Query Time: Profile-Centric Strategy



Wikipedia Expertise Ranking

Query Time: Profile-Centric Strategy

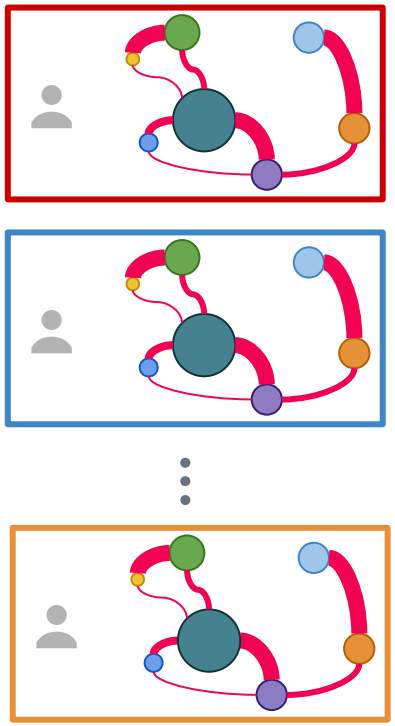


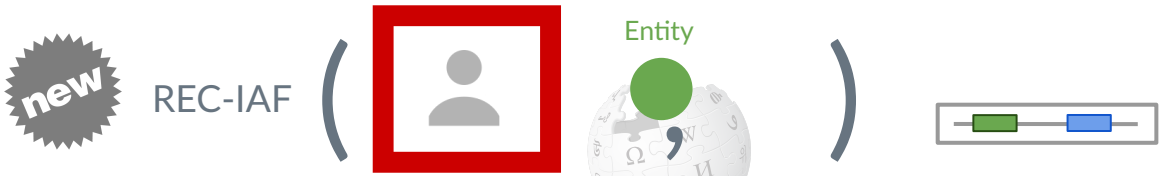


- Score each author's entity (matched in the input query)
- Combination of multiple scores of the entity:
  - Document Frequency
  - Confidence (provided by TagMe)
  - Inverse document frequency
  - PageRank in the author's profile

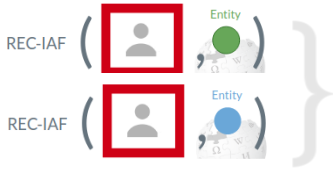
Wikipedia Expertise Ranking

Query Time: Profile-Centric Strategy





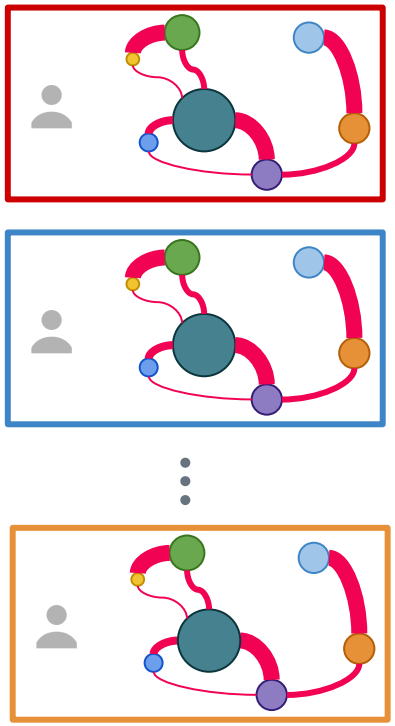
REC-IAF *is computed for each entity!*

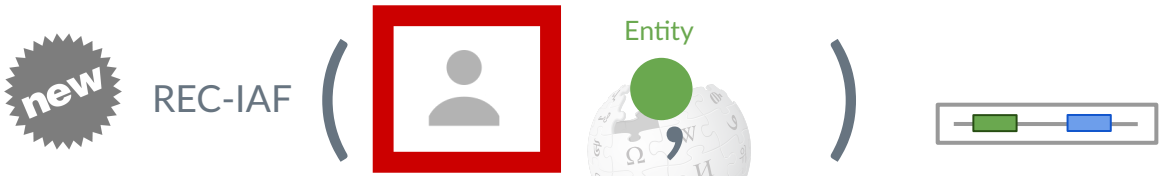


The average is the ranking of the authors!

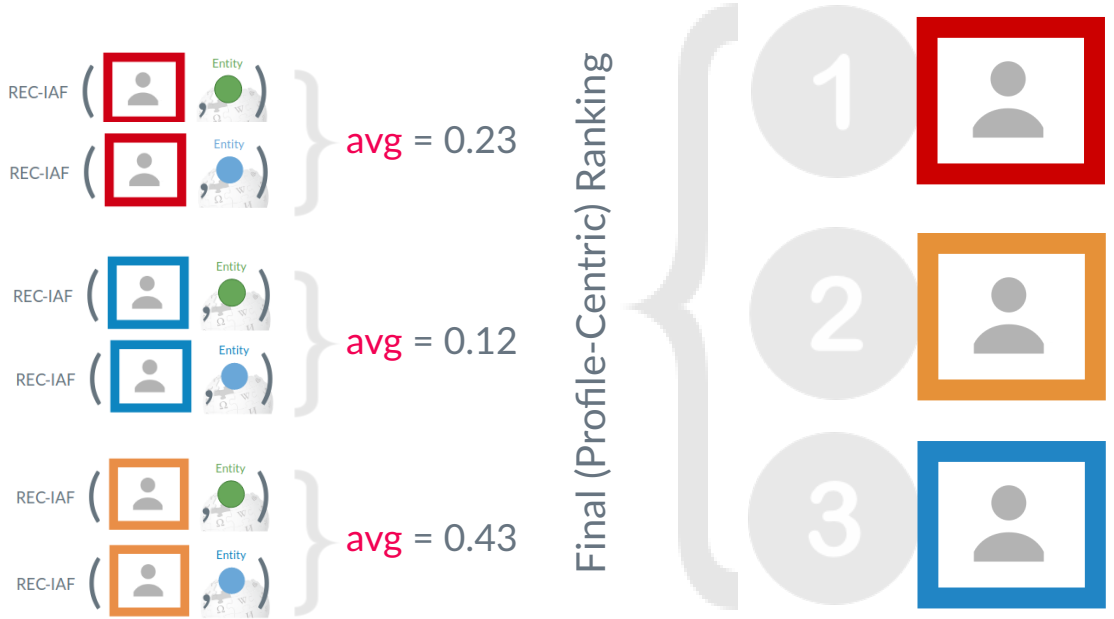
Wikipedia Expertise Ranking

Query Time: Profile-Centric Strategy





REC-IAF *is computed for each entity!*



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

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

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)
  - Ensamble
    - Product Reciprocal Rank between JM Model and Log-Linear

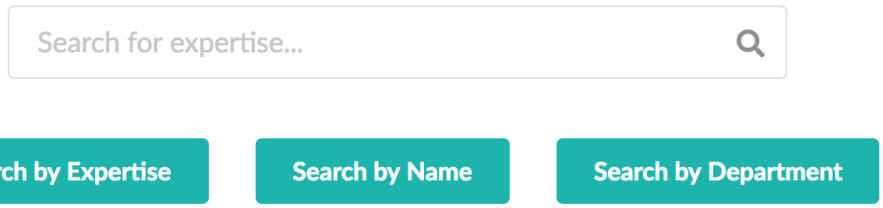
Experiments
Results

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
Ensamble	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%

from the University of Pisa

URL: https://wiser.mkapp.it

WISER
Wikipedia Expertise Ranking



- ~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 Technological 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