Using Knowledge Graphs for Text Retrieval

github.com/laura-dietz/tutorial-utilizing-kg

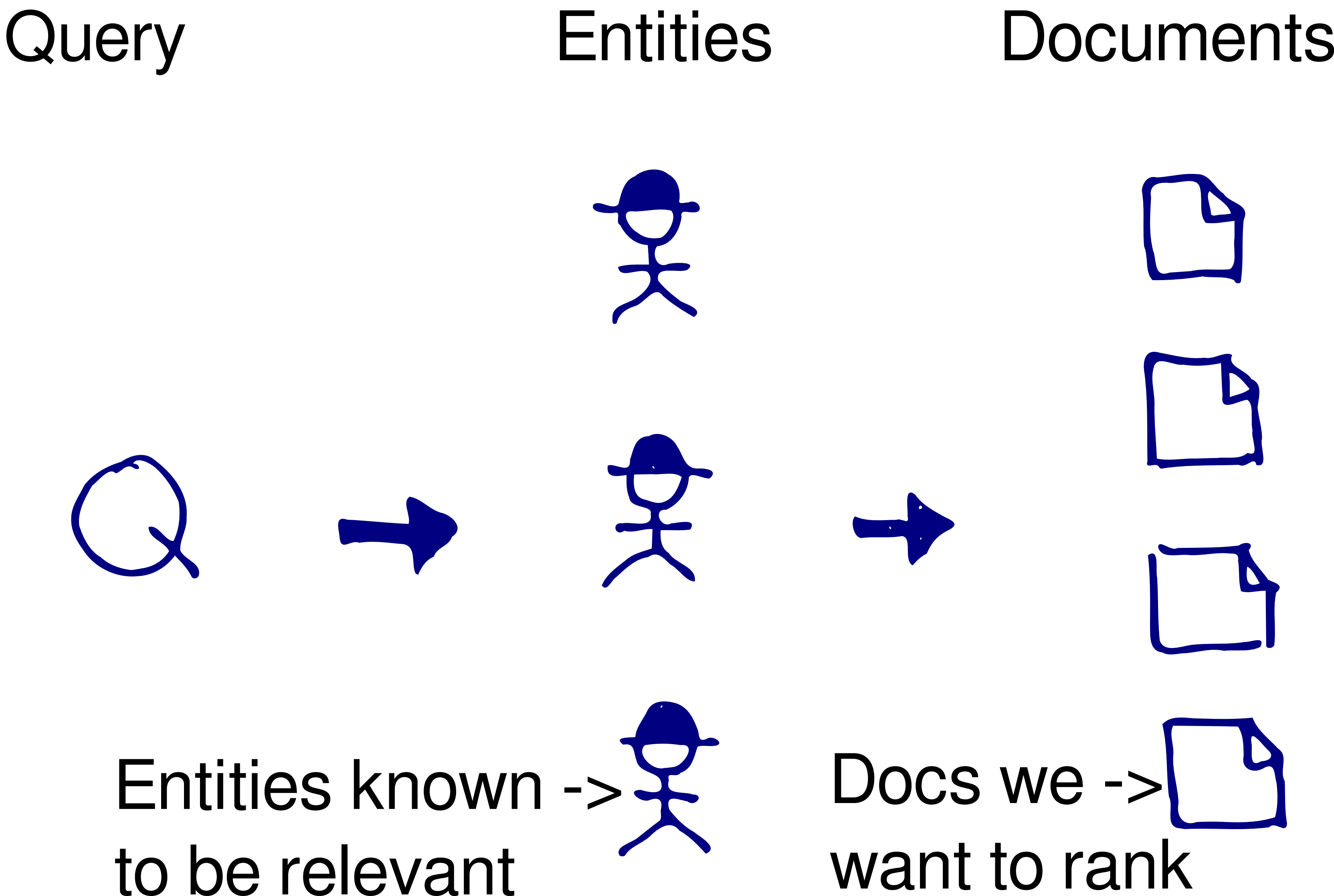
Laura Dietz
University of New Hampshire

Alex Kotov
Wayne State University

Edgar Meij
Bloomberg L.P.

Please take the survey!

Document Retrieval with Entities

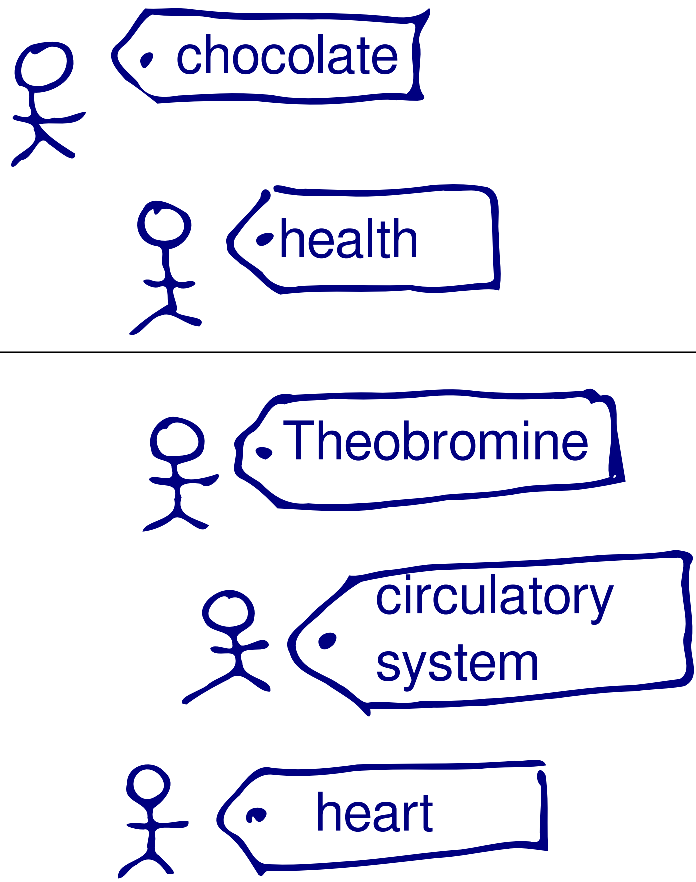


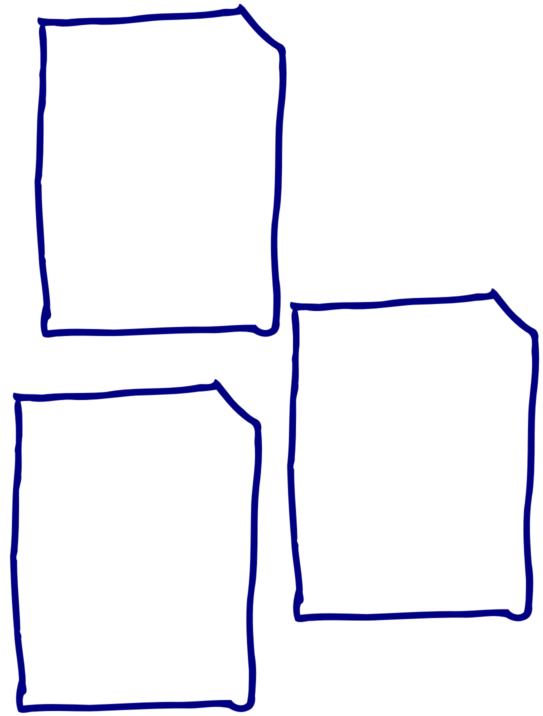
Outline

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Matching Entities in Documents

dark chocolate
health benefits





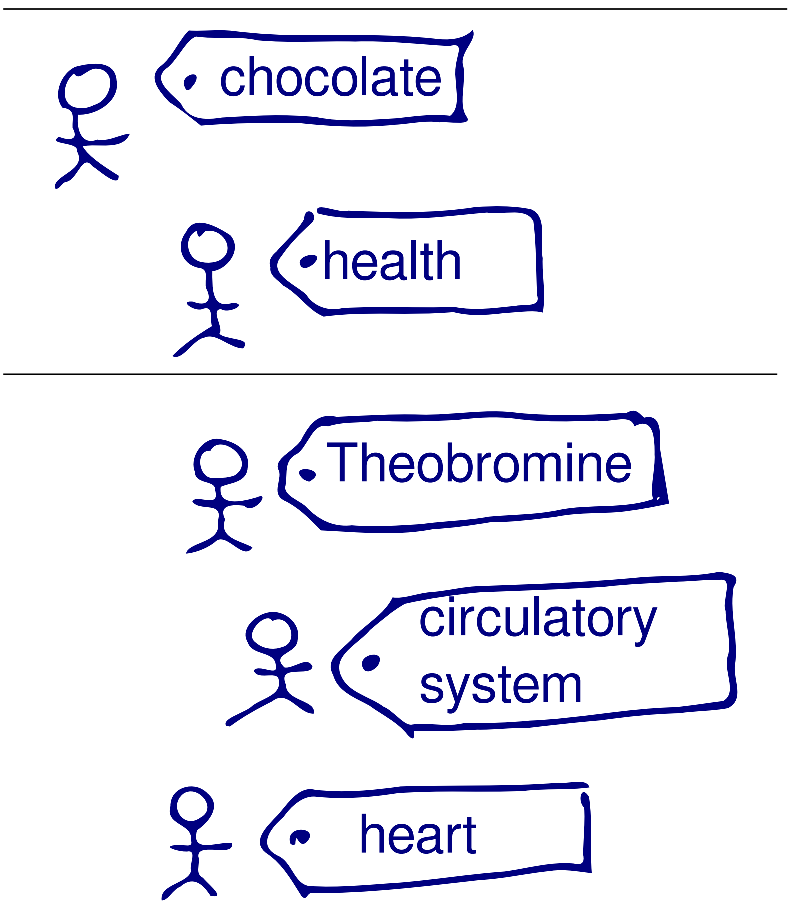
Which doc should be promoted in the ranking?

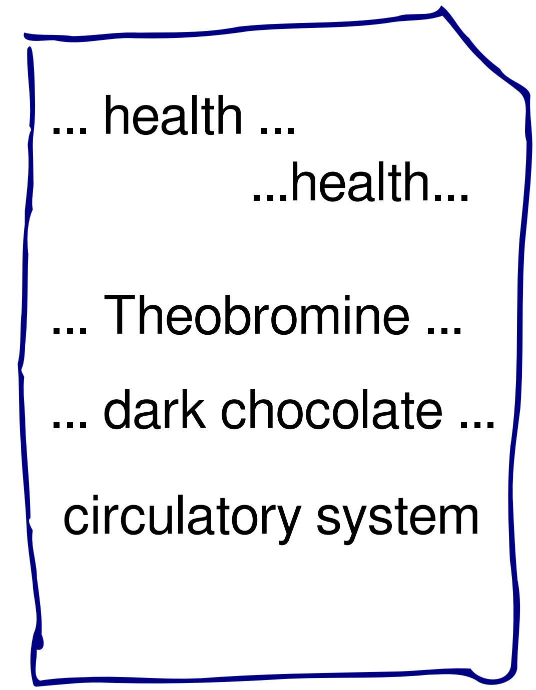
Different Queries - Different Entities

Query	EU UK relations	dark chocolate health benefits
Query entities	EU UK	chocolate health
Latent entities	Brexit Theresa May	Theobromine circulatory system heart
[Hasibi ICTIR16]	**Named Entities**	**Concepts**

Matching Entities in Documents by Name

dark chocolate
health benefits

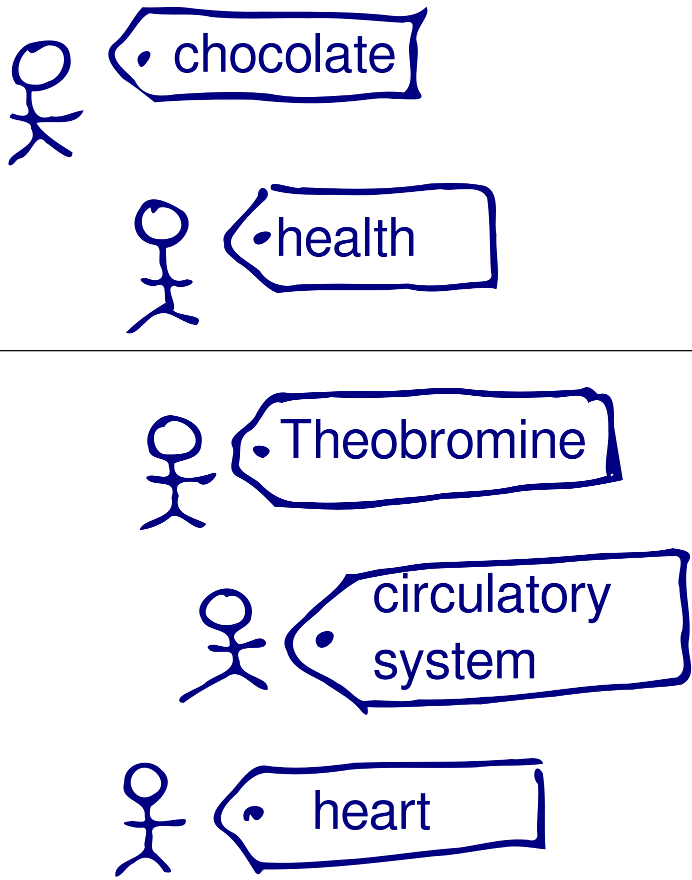


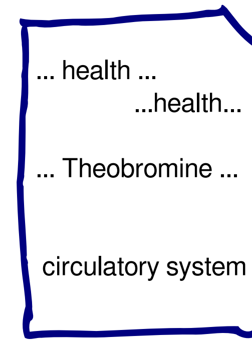


Should this doc be promoted in the ranking?

Matching Entities in Documents by Name

dark chocolate
health benefits

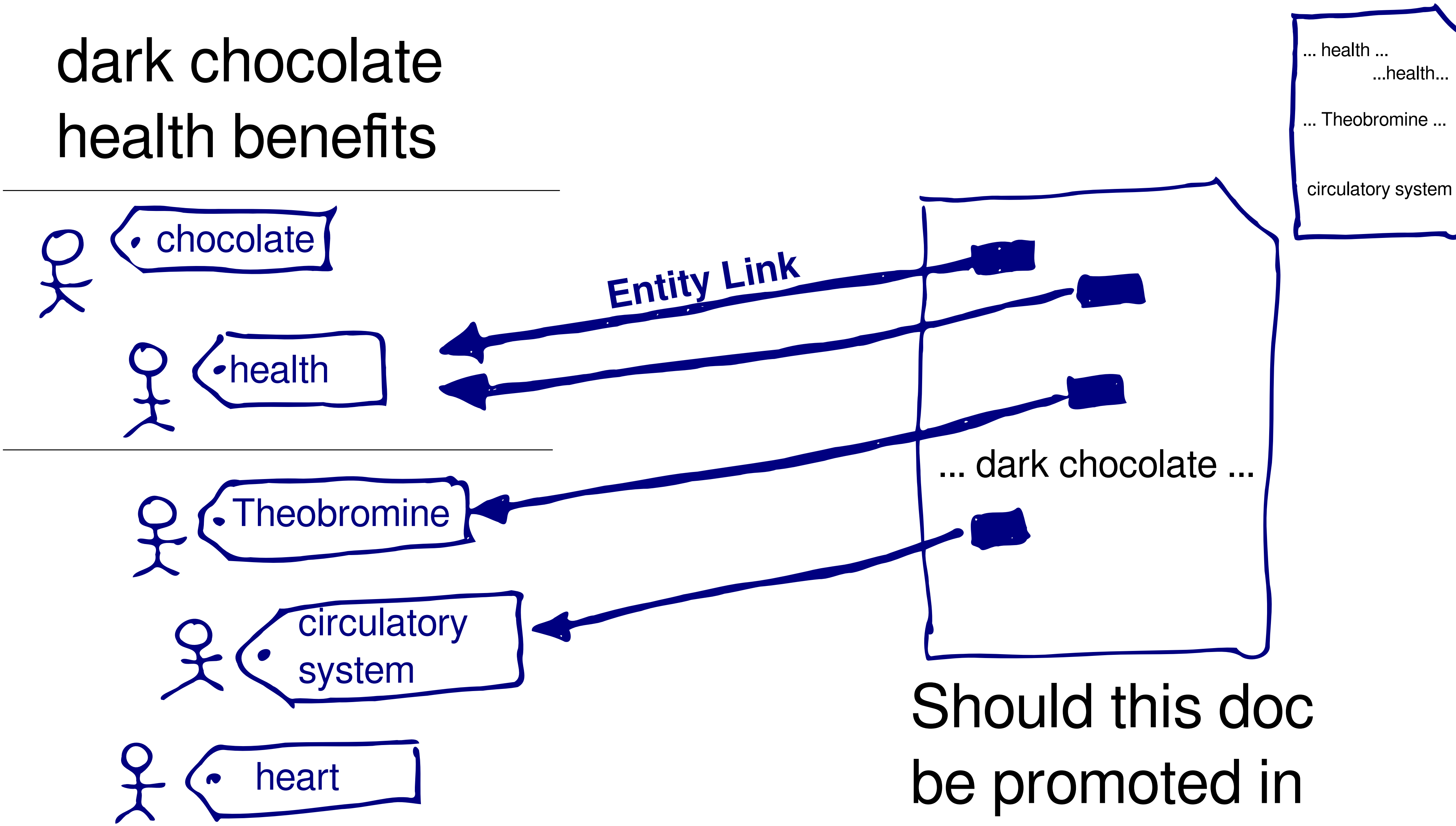




Should this doc be promoted in the ranking?

Matching Entities in Documents by Entity Links

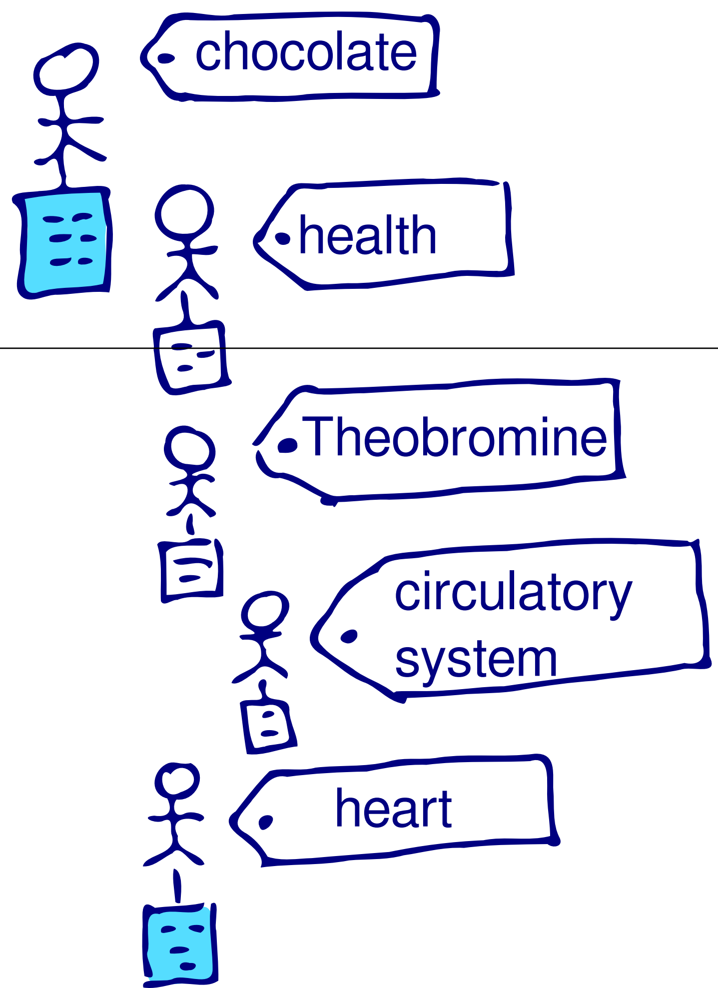
dark chocolate
health benefits



Should this doc
be promoted in
the ranking?

Matching Entities in Documents by Article Terms

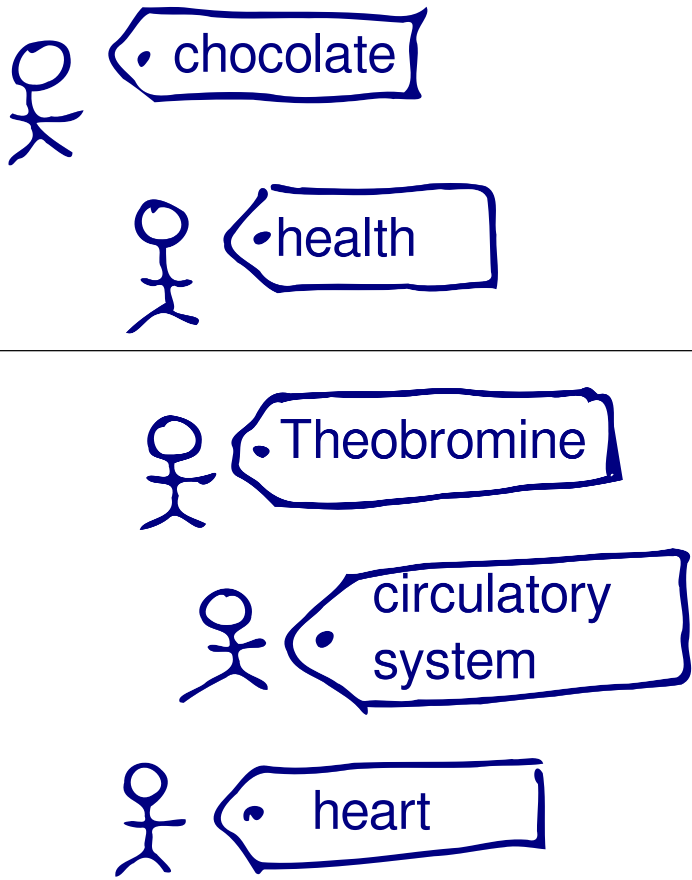
dark chocolate
health benefits

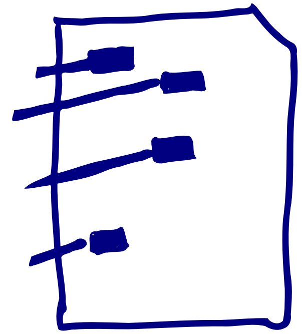


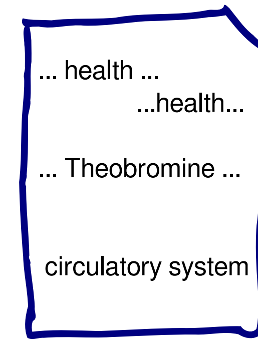
Should this doc be promoted in the ranking?

Matching Entities in Documents by Entity Links

dark chocolate
health benefits

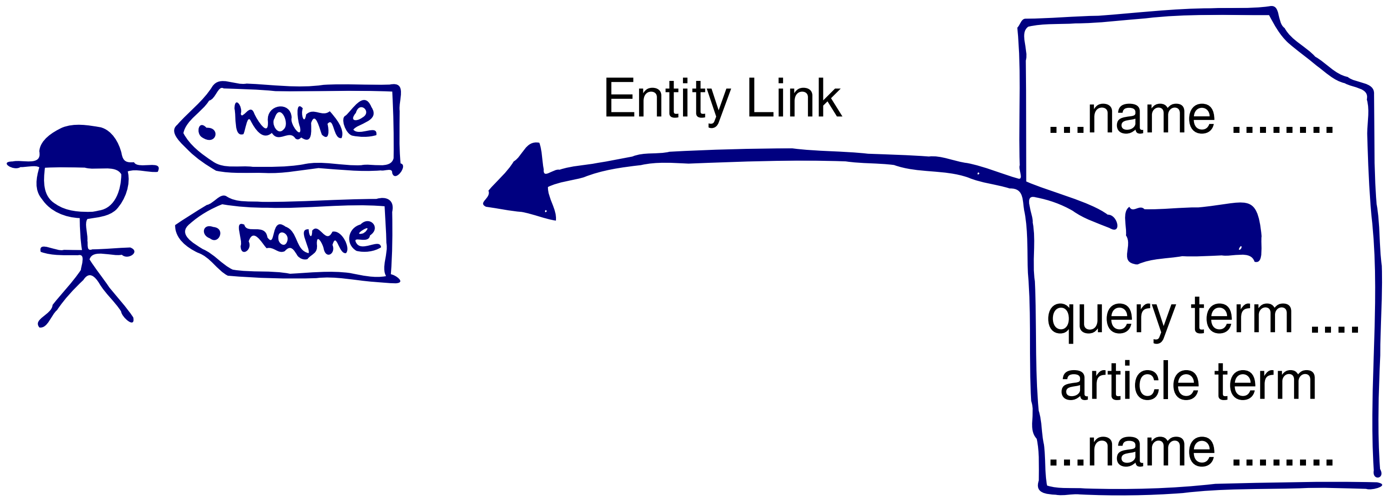






Should this doc be promoted in the ranking?

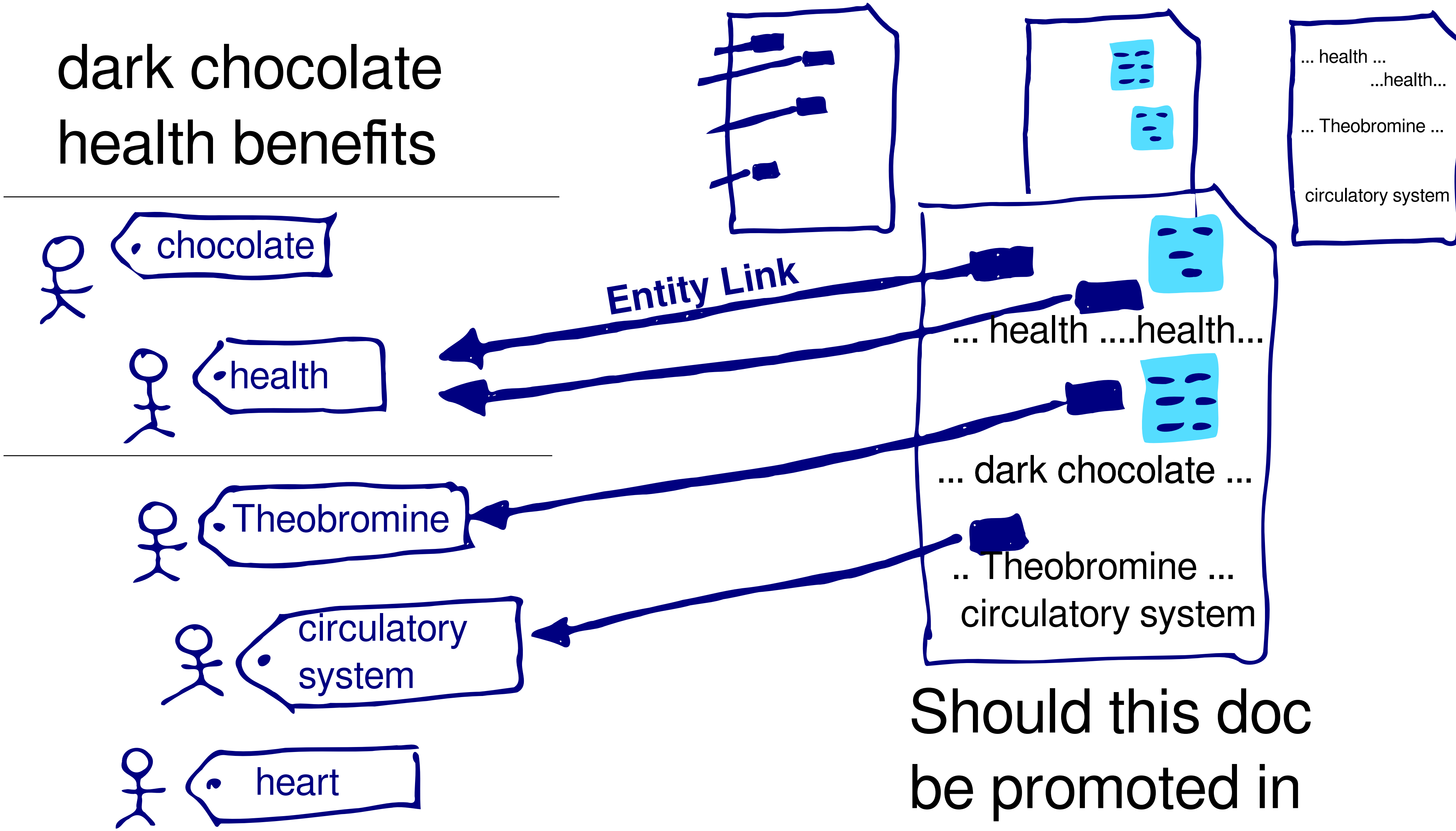
Using Entities as a Vocabulary of Concepts



$$score(\;) = \lambda_1 \text{query terms} + \lambda_2 \text{names} + \lambda_3 \text{entity links} + \lambda_4 \text{article terms} + \ldots$$

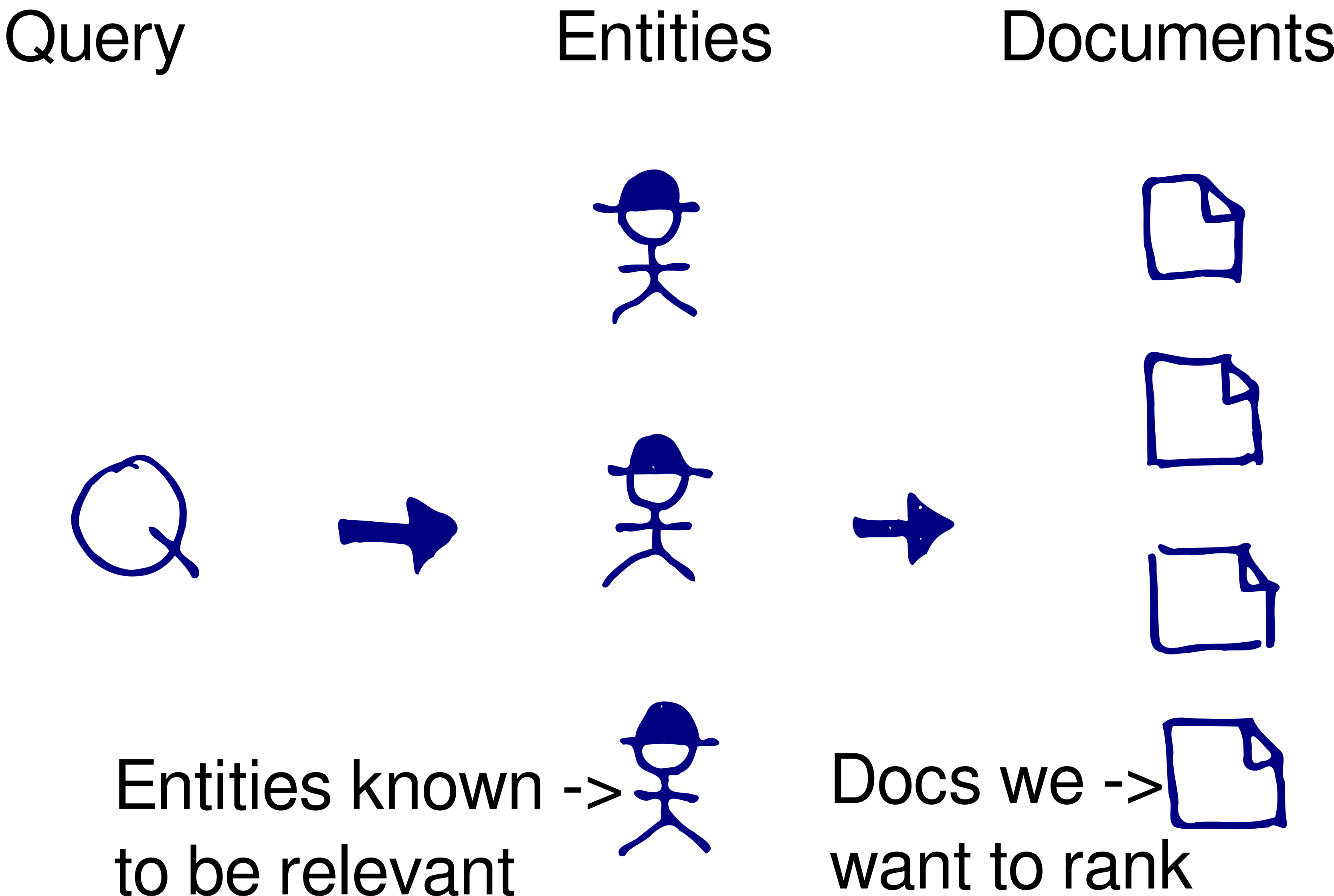
use your favorite retrieval model here!

Combine All Names, Links, Terms



Should this doc be promoted in the ranking?

Document Retrieval with Entities



How to Find Relevant Entities?

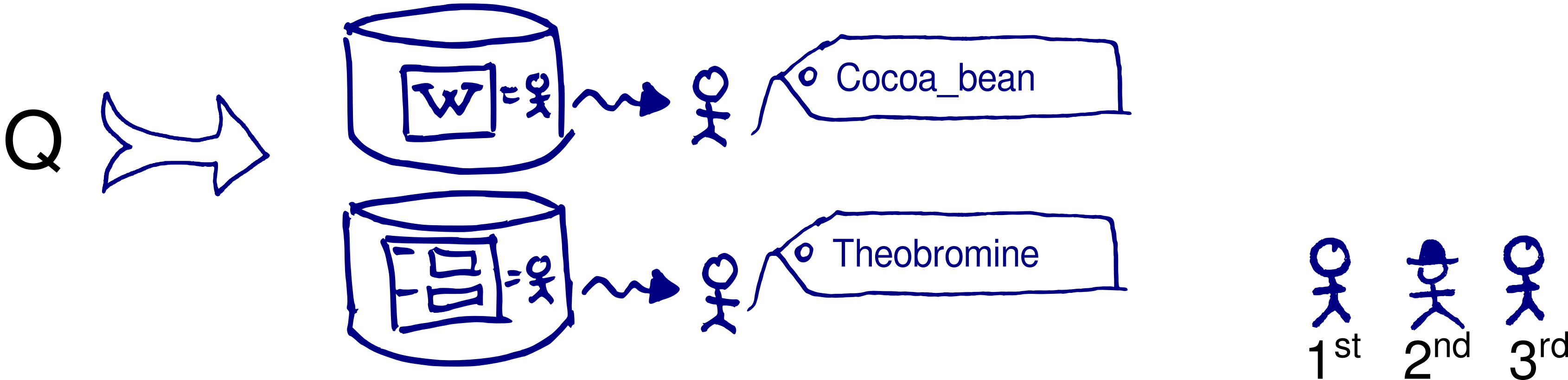
	Named Entities	Concepts
Query	EU UK relations	dark chocolate health benefits
Query entities	EU UK	chocolate health
Latent entities	Brexit Theresa May	Theobromine circulatory system heart

Find Relevant Entities

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

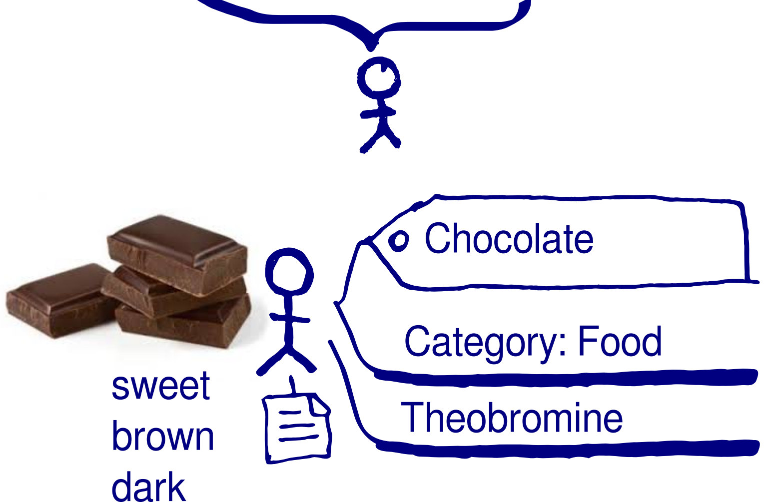
Latent Entities through Retrieval (e.g., Part 3)

Retrieve entities from knowledge base
to obtain ranking of entities E (with score)



Query Entities through Entity Linking

Query: dark chocolate health benefits

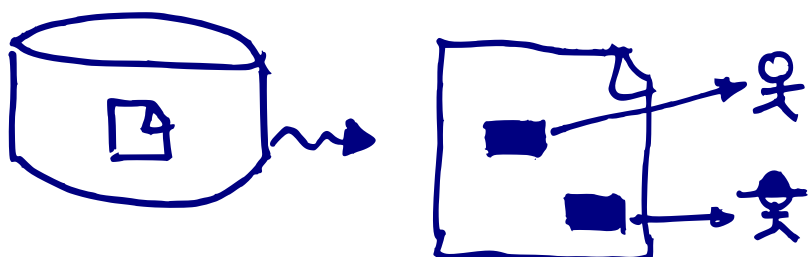


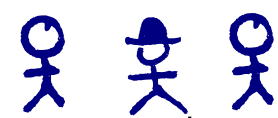
Latent Entities through Pseudo-Relev. Feedback

1. Retrieve preliminary documents
2. Entity link documents
3. Derive distribution over  (bag of entities)

(see Relevance Model / RM3)

[Dalton SIGIR14, Liu IRJ15]



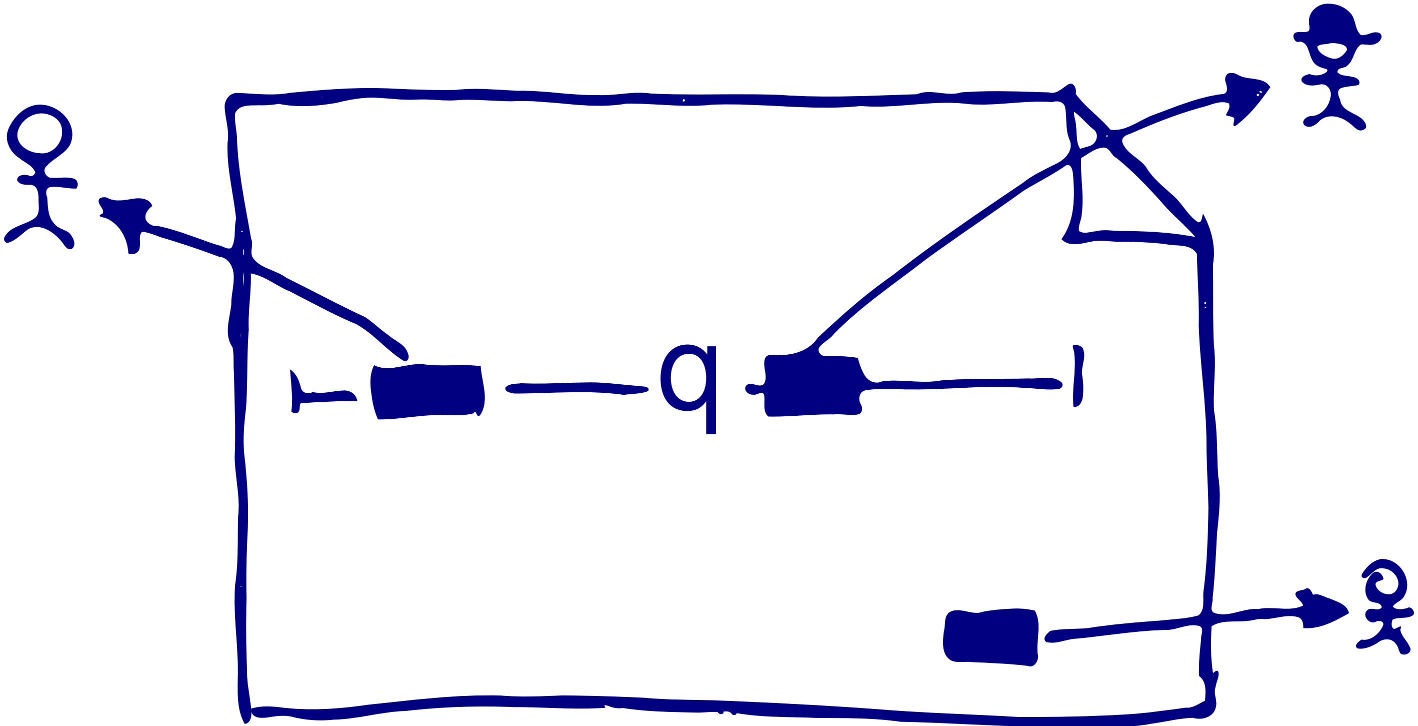


1st 2nd 3rd

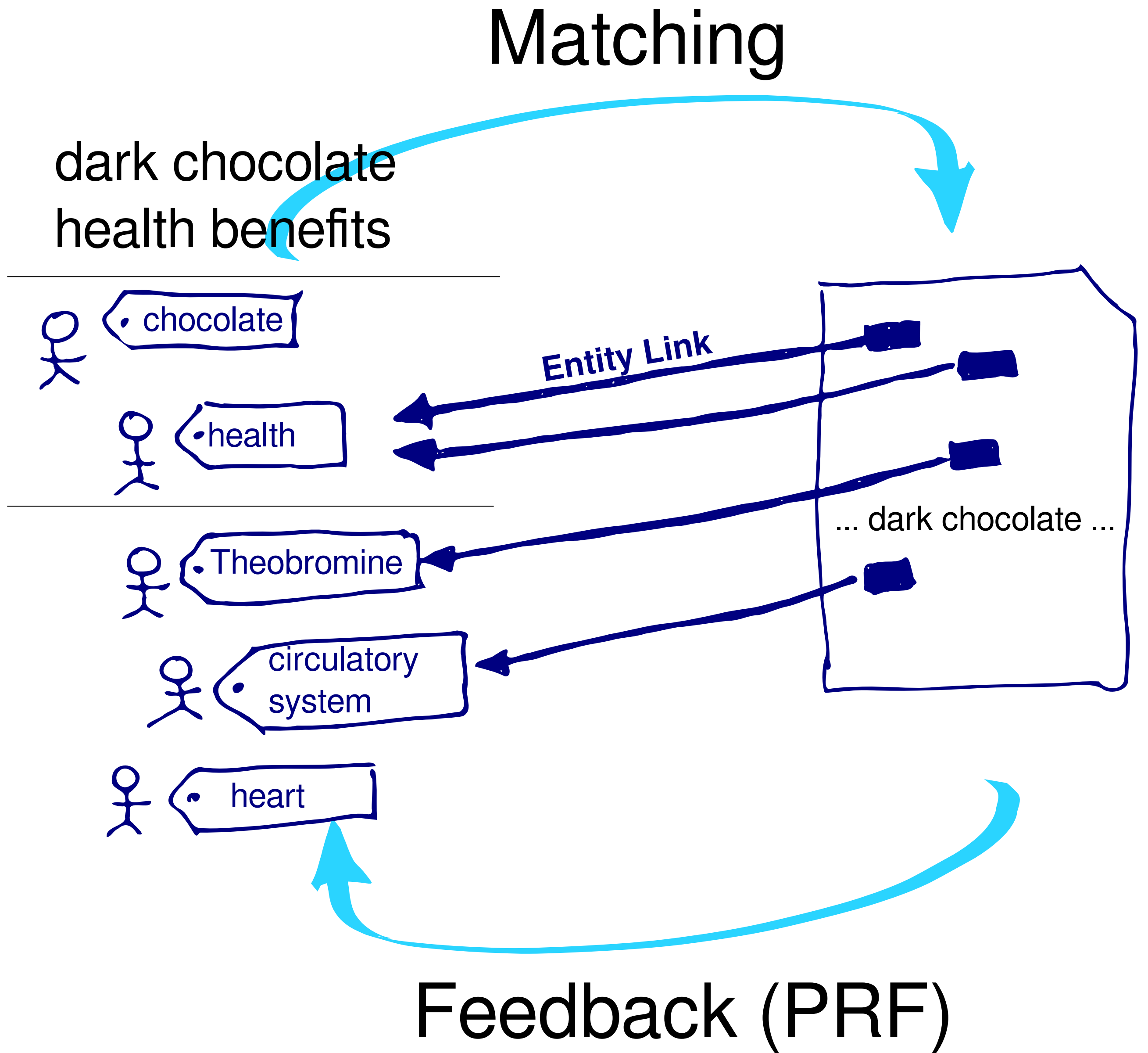
Latent Entities through Proximity to Query Words

Using distance between entity mentions and query words **q** as a measure for relevance.

[Petkova & Croft, 07]

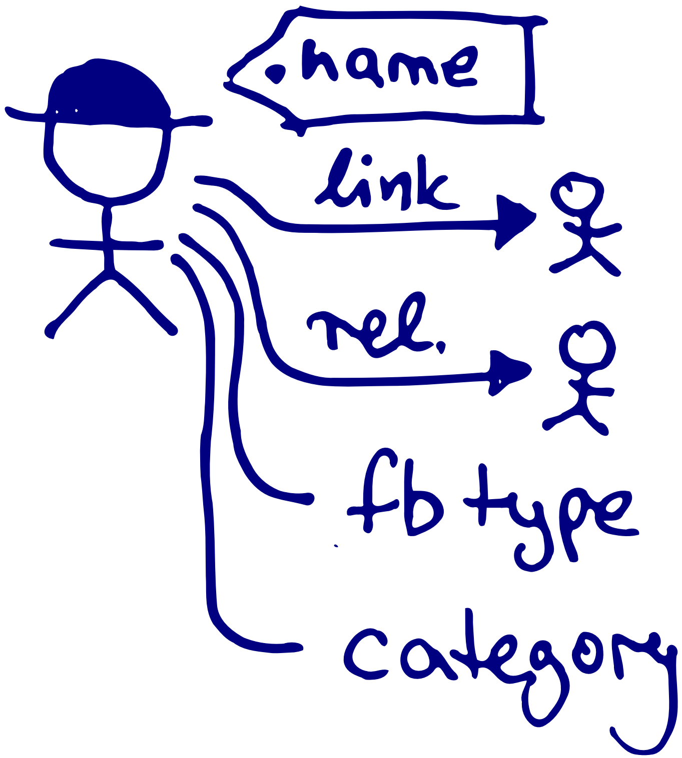


PRF is Inverse of Matching Entity Links



Using More from the Knowledge Graph

So far we used names and entity links.
But KGs have so much more information!



Names

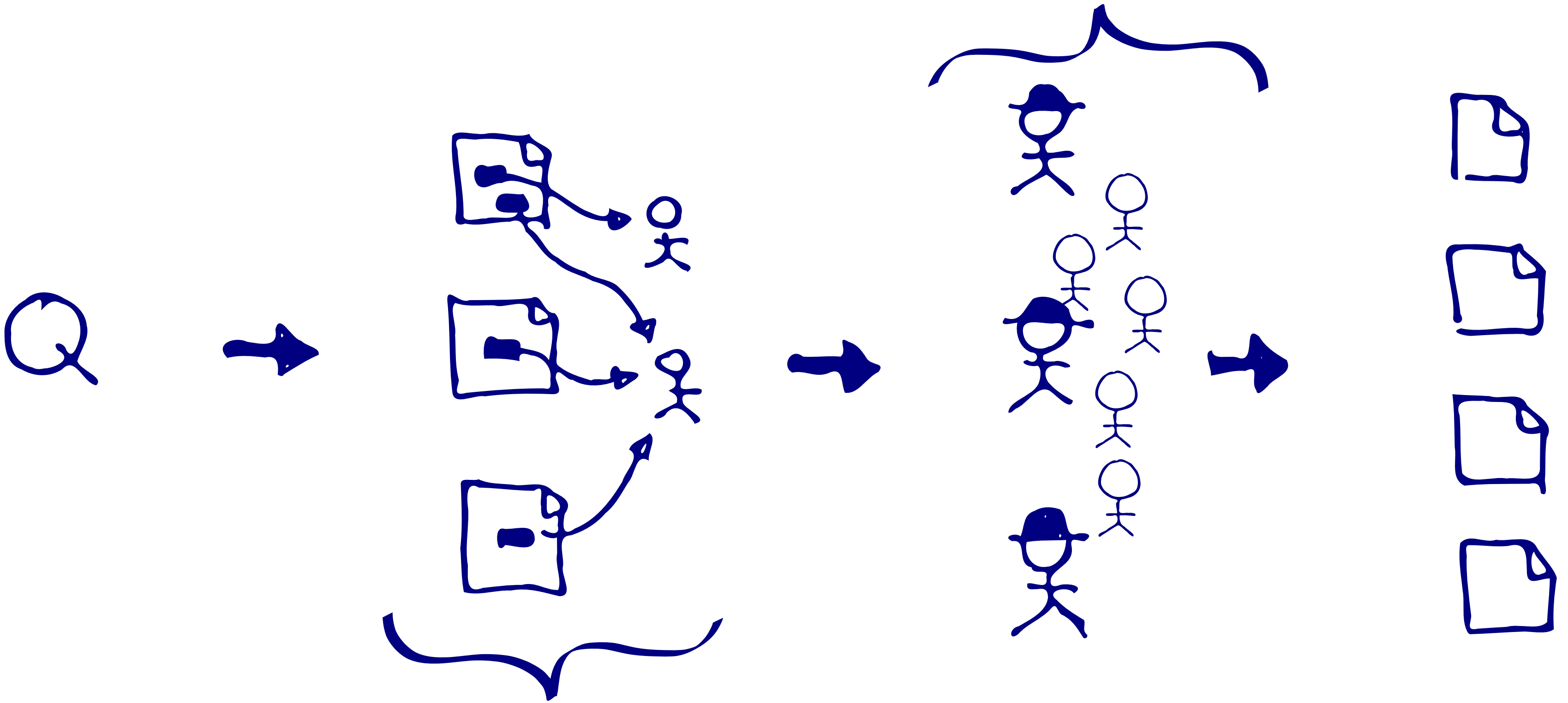
Links and relations

Different taxonomic type systems

How can we make use of it?

Entity Expansion for Document Retrieval

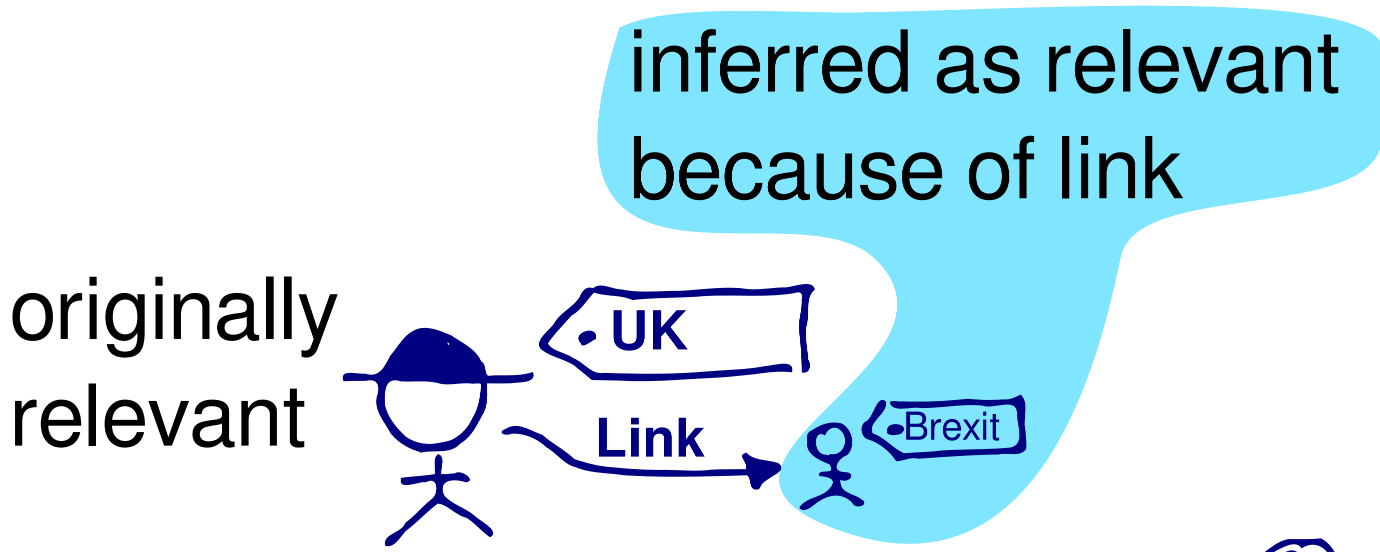
Query entities + Object retrieval (Part 3)

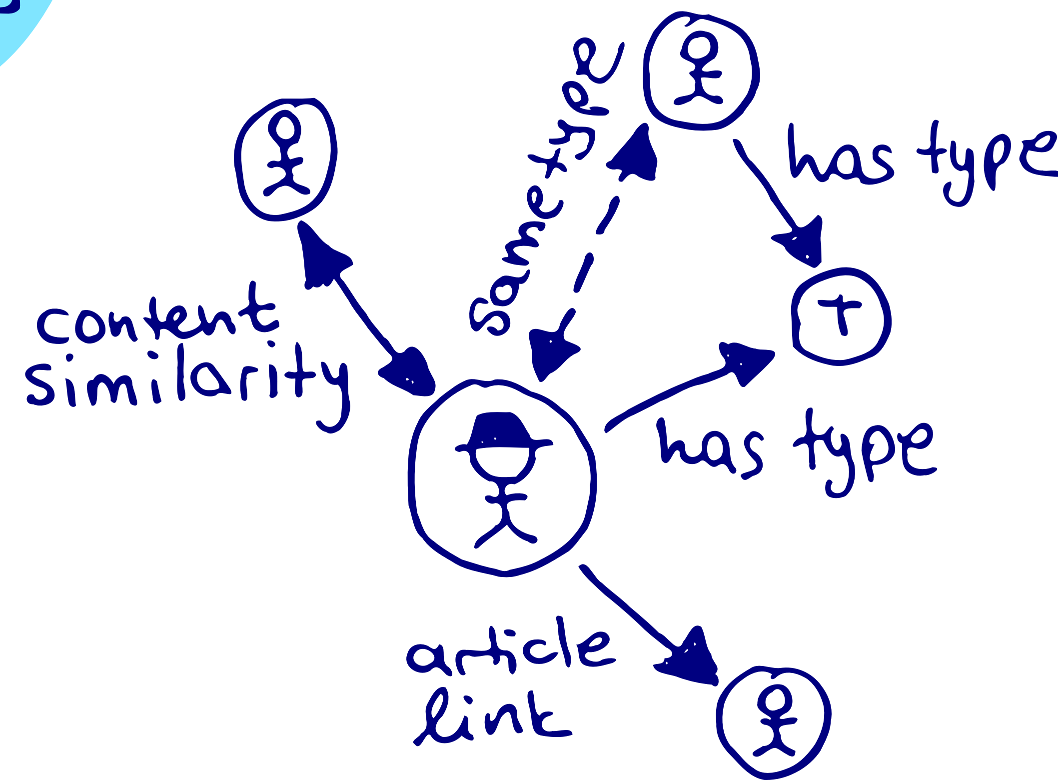


Pseudo-relevance feedback (RM3)

Document = bag of entity links (instead of terms)

Using Relations and Types with Entity Links

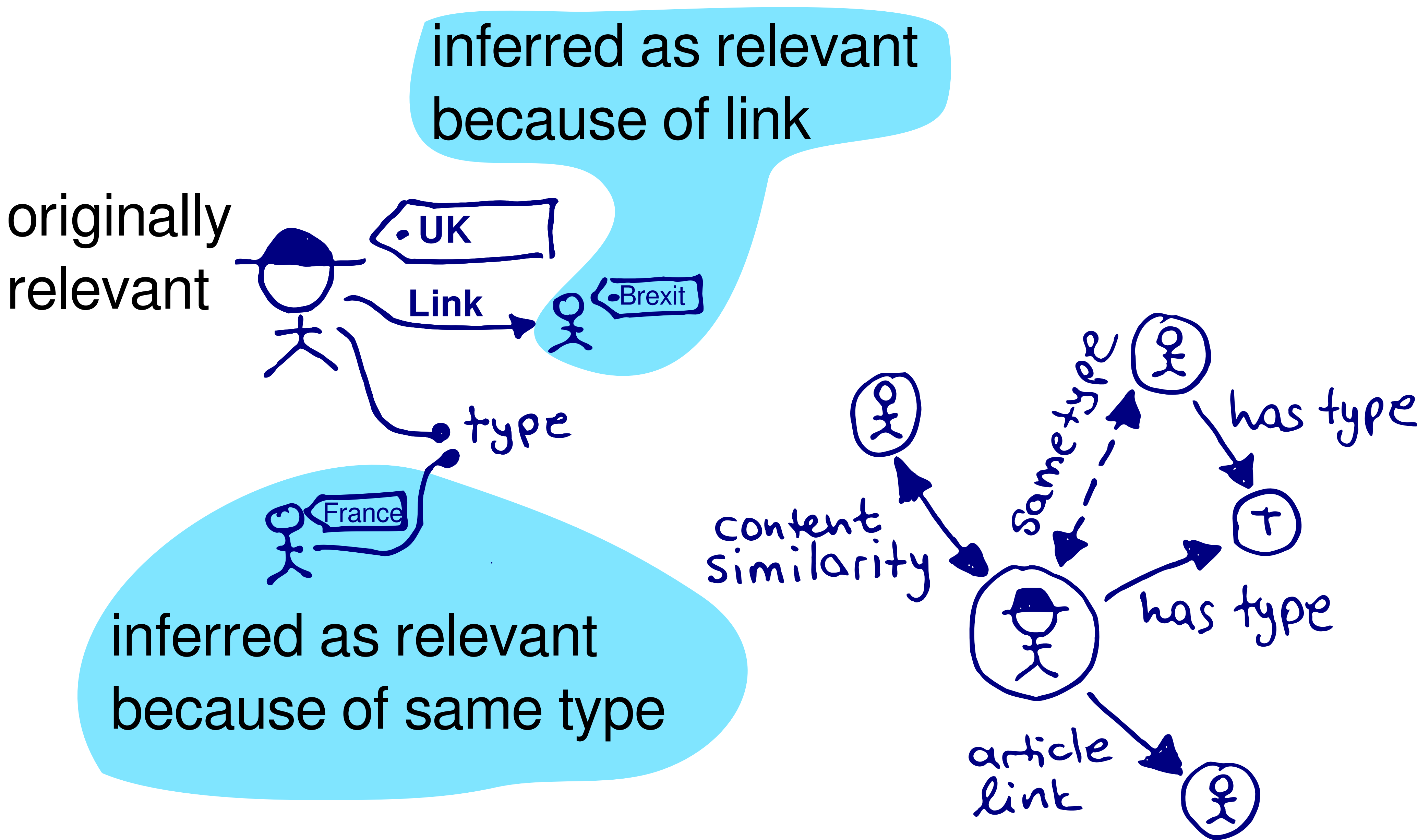




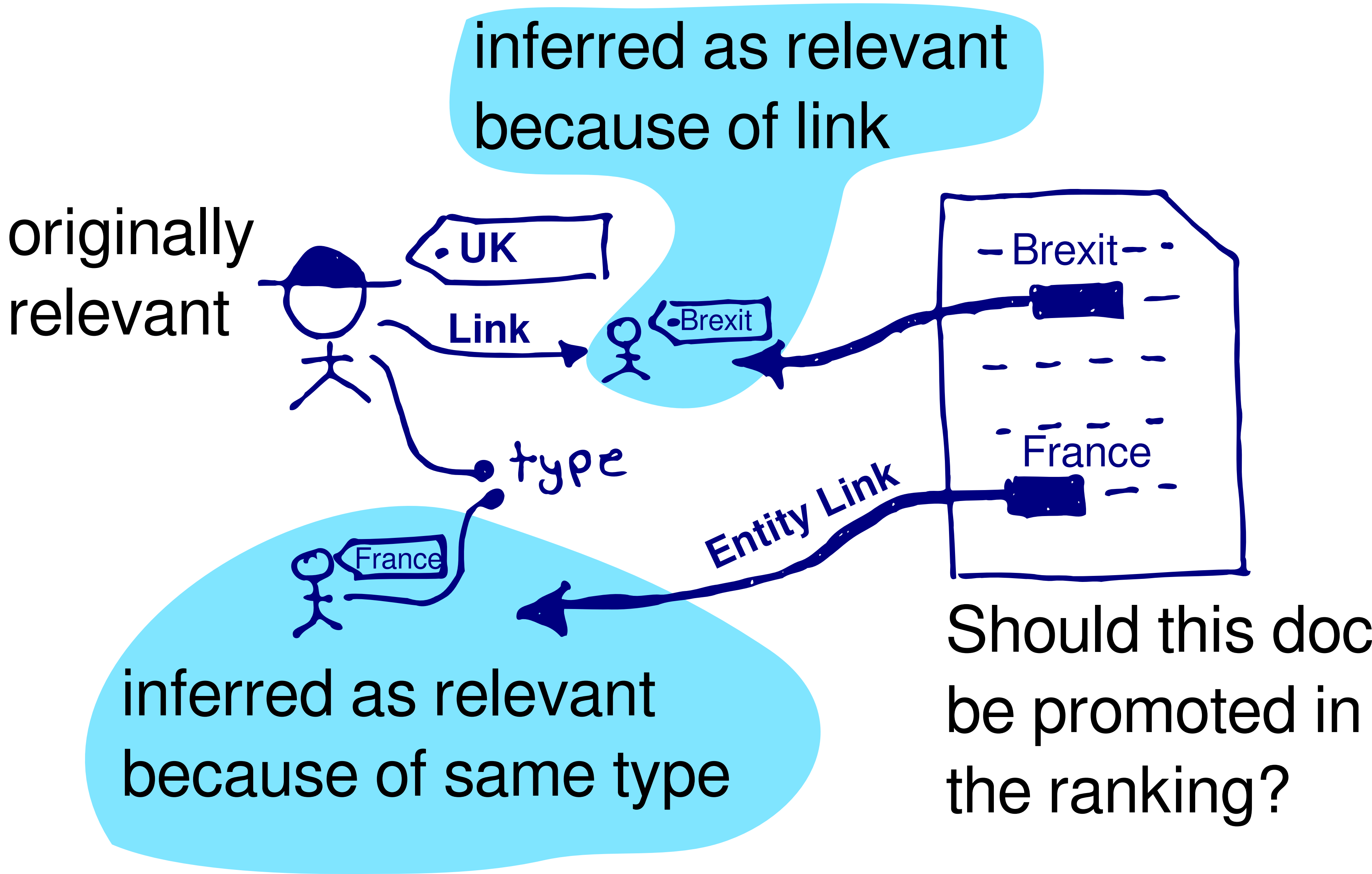
Graph Expansion

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Using Relations and Types with Entity Links

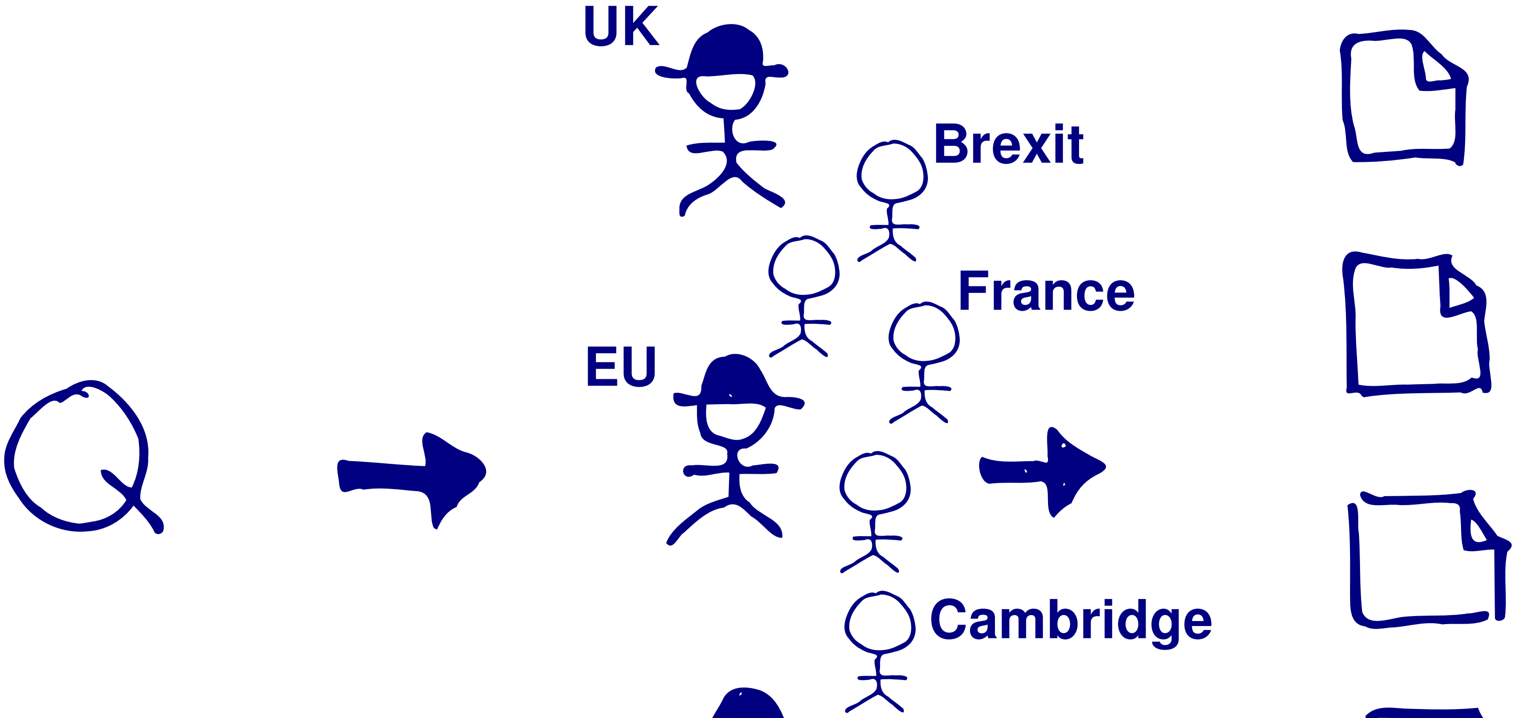


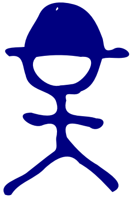
Using Relations and Types with Entity Links

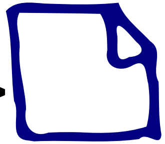


Document Retrieval with (More) Entities

Query | Entities | Documents



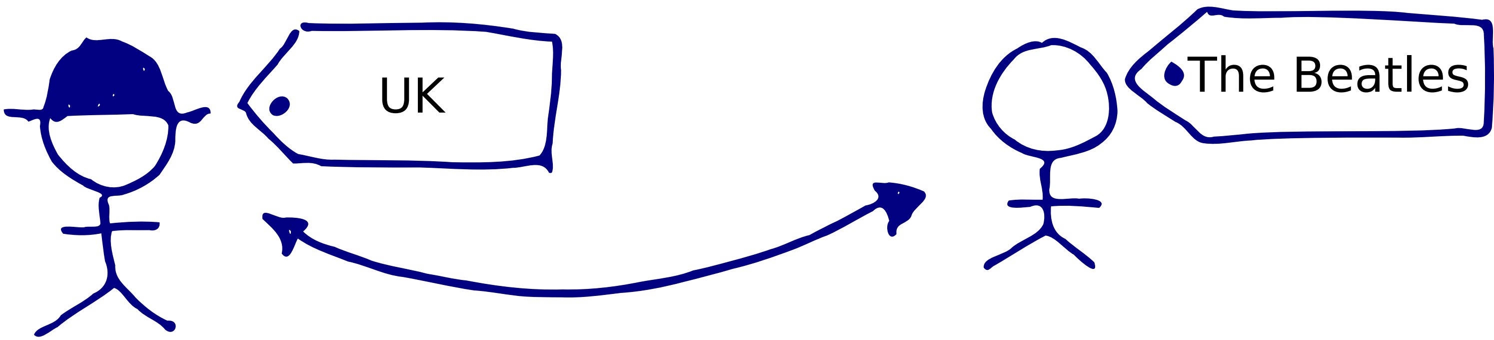
Entities known **or** -> 
assumed to be relevant

Docs we -> 
want to rank

KG expansion: A Potential Issue

Example query: EU UK relations

Consider:



Correct connection, but:

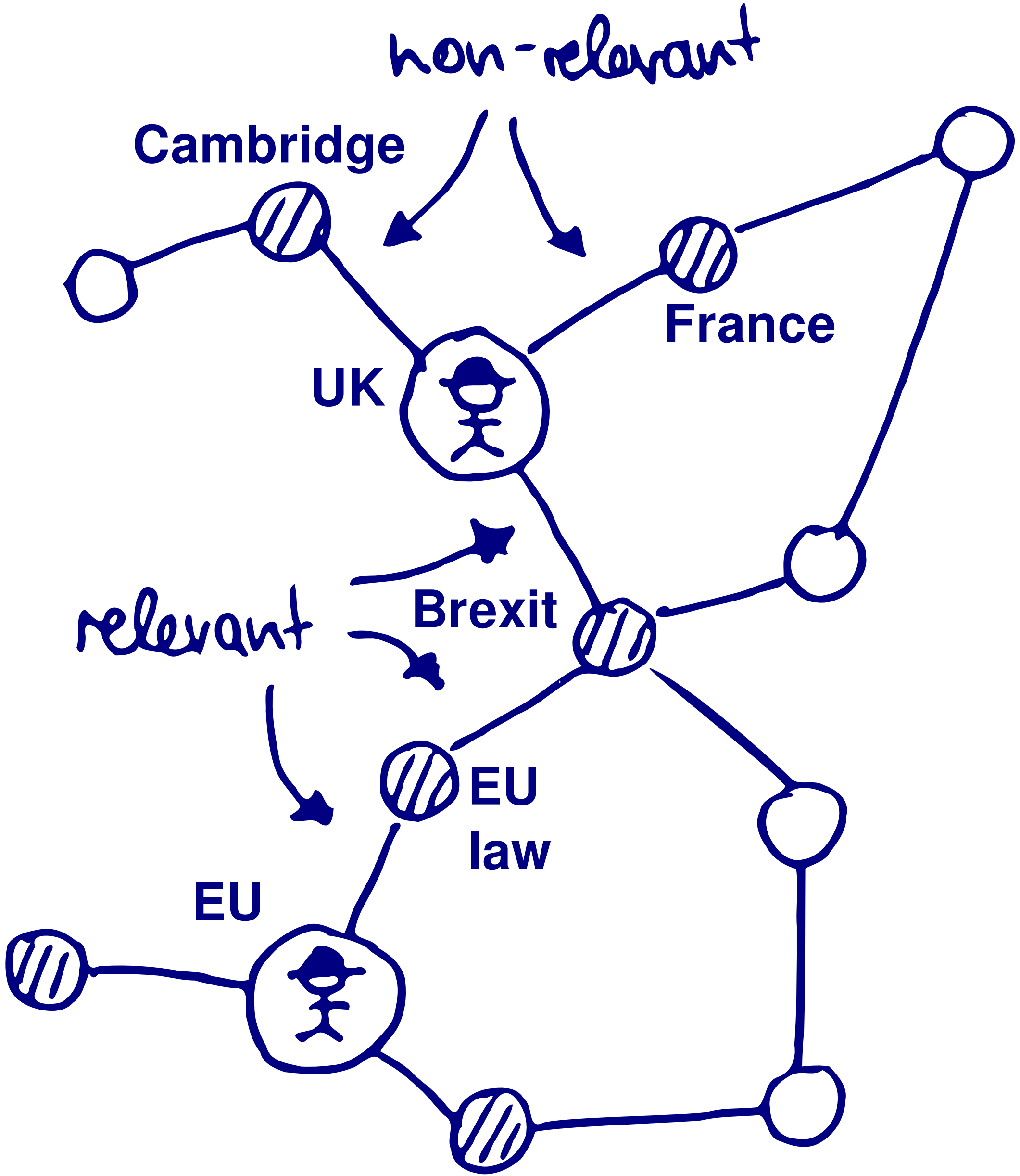
The connection is not relevant in the context of "UK" as in "EU relations".

If we promote docs because they talk about The Beatles, we ruin an okay ranking.

General Approach: Graph Expansion

So many connections in a knowledge graph
- Some are relevant!
- But many are only relevant in a certain (other?) context.

Expanding with non-relevant entities leads to low precision rankings.



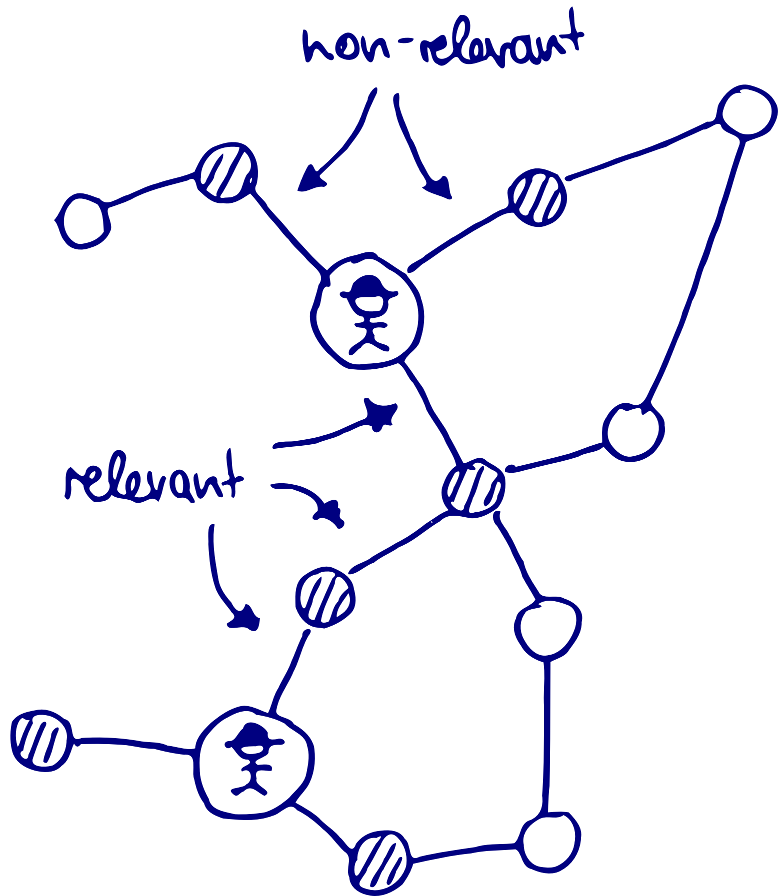
Theresa May

Big Question

How to infer which other connected entities / nodes are relevant for the information need Q?

...and therefore safe for expansion?

Maybe entities in between query entities?



Boston et al 2013: Wikimantic: Toward effective ...

Weight entities by:

M: How well **E**s article content matches the query

MR: How often **E** is linked by others (PageRank)

Method	F1 on TREC QA
M	76.92
M+d*MR	79.47

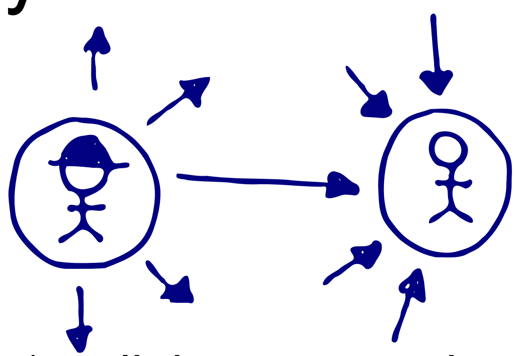
d=0.0001

Weight Edges in the Knowledge Graph

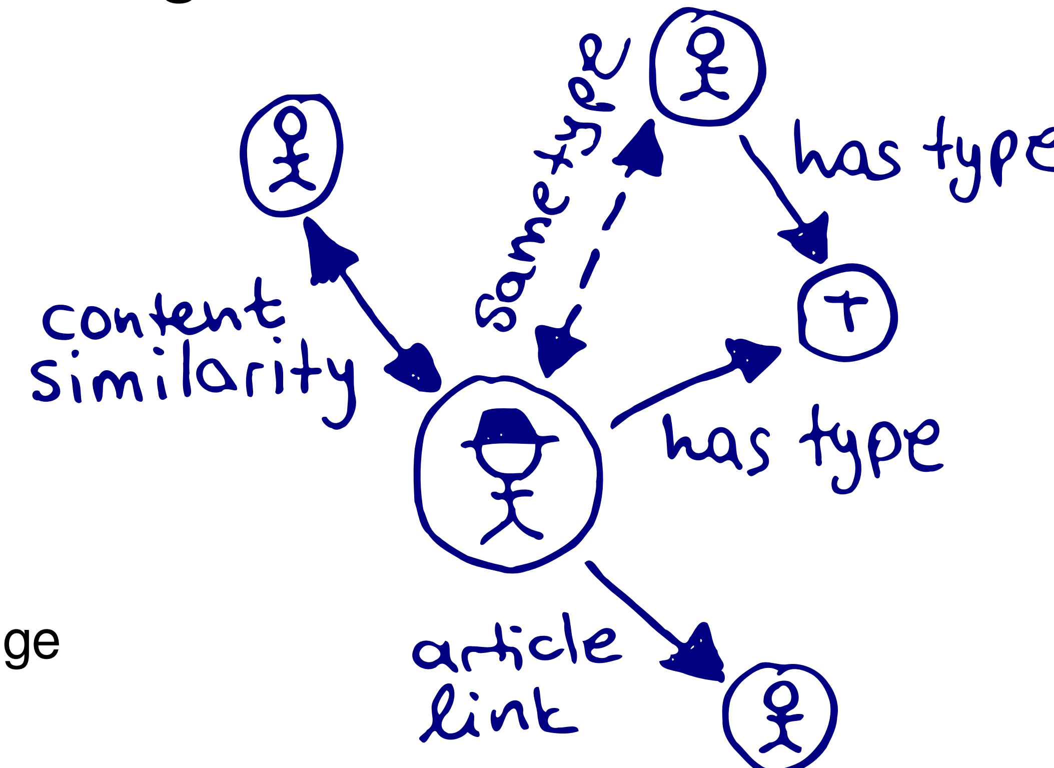
Using seed entity nodes and...

- Graph walks: PageRank / HITS, Shortest Paths
- Different edge types
- Edge weighting + Clustering

Exclusivity-based
Entity Relatedness



fewer in/out links => more important edge
[Hulpus WSDM13, Weiland ICTIR16]

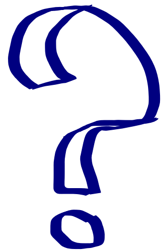


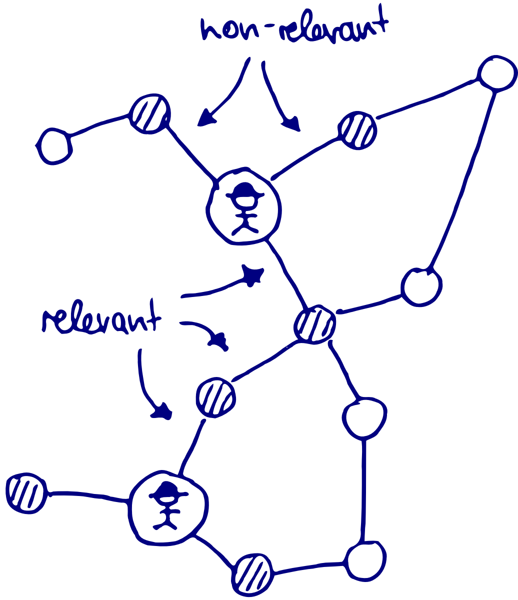
Using Types and Categories

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Entity Aspects and the Graph Structure

Edge weights and random walks help identify popular connections. BUT...

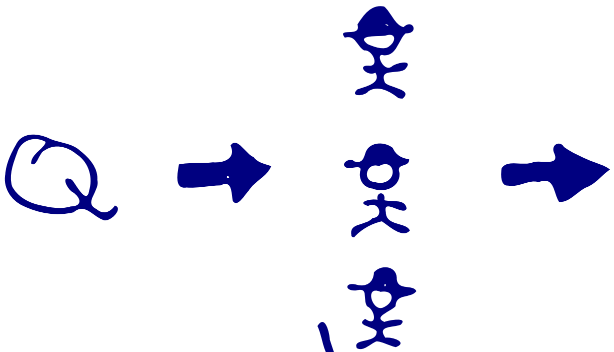




An **open issue** remains:

- Entities have multiple aspects
- Graph structure = overlay of all aspects
- How to identify:
    1. Which aspects of E are relevant for Q?
    2. How to select edges that are relevant?

Entity Types Inferred through Entity Links



Which types are relevant ?

How to match types to documents ?

a) same-type entities

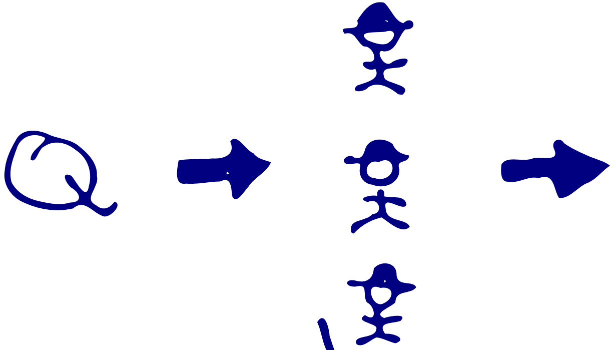
[Kaptein CIKM10]

majority types among entities

prefer docs with entities of this type

Method	MAP on INEX
Full Text	0.03
Link	0.09
Type+Link	0.13

Entity Types (inferred through entities)



Which types are relevant ?

How to match types to documents ?

a) same-type entities

majority types among entities

prefer docs with entities of this type

[Kaptein CIKM10]

Entity Types through Text Classification

Q

Which types
are relevant
?

How to
match types
to documents
?

b) term
classifier
[Xiong CIKM15]

classify query terms
with naive Bayes

classify documents
with naive Bayes

Combination of Multiple Sources

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Complementary Sources

Typical approaches:

1) Use **complementary sources**:

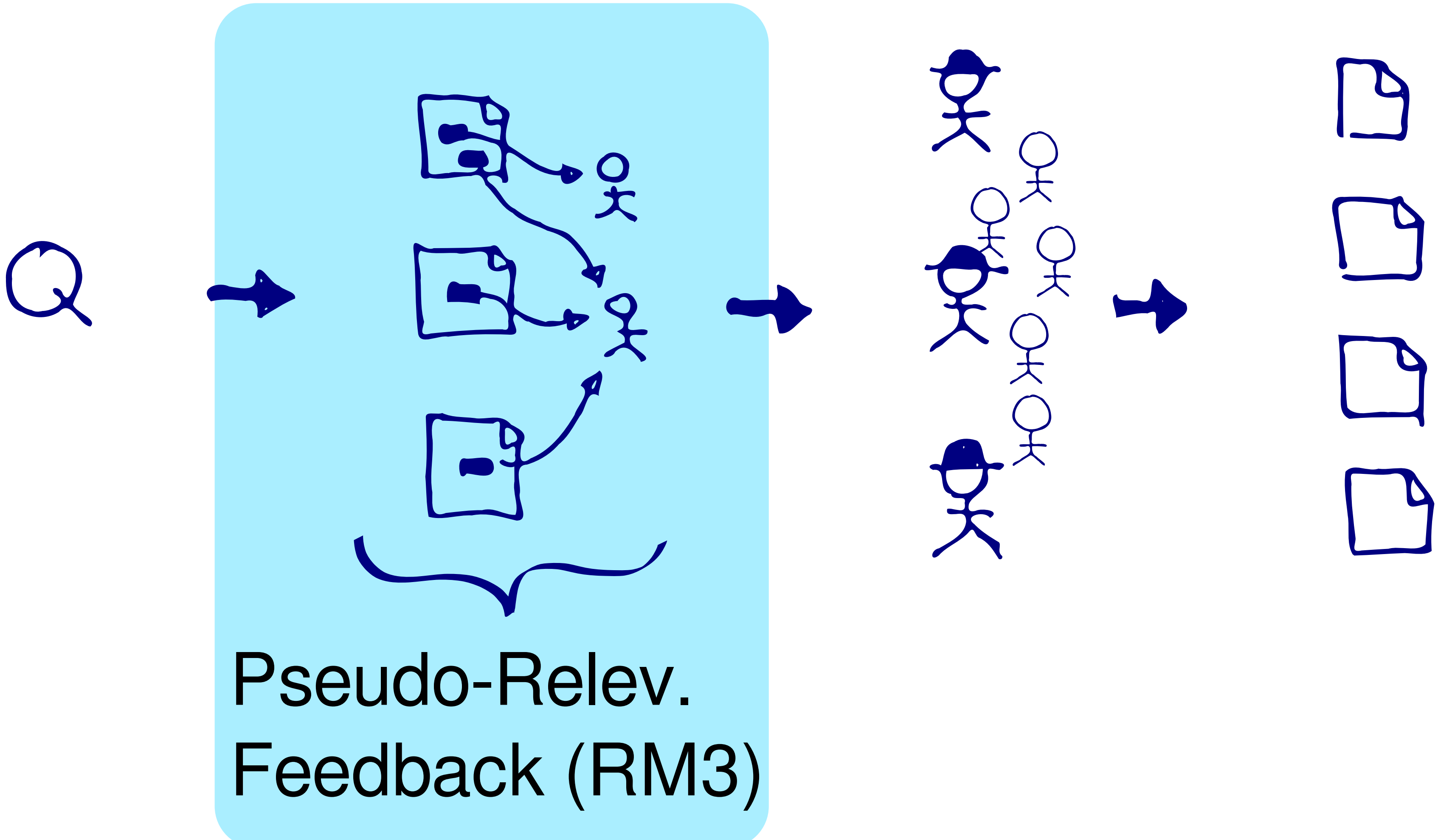
graph, article text, relevance feedback, type info

2) Use **machine learning**:

Train weights for sources on test collection

3) Model relevant **entity aspects**

Source: Relevance Feedback with Entity Links

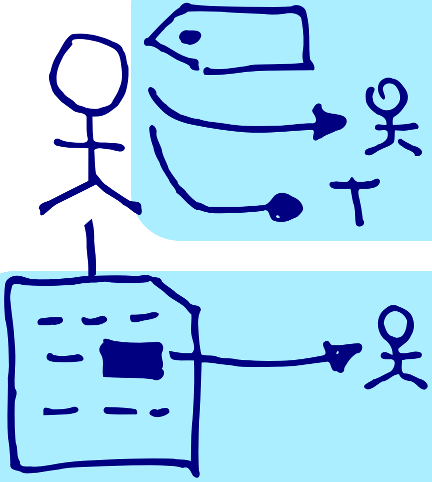


Document = bag of Entity Links

Proximity of query and Entity Links

[Petkova 2007, Dalton SIGIR14, Liu IRJ15]

Source: Object AND Article Content Retrieval



Entities as attribute-structured objects:
Object retrieval (see Part 3 & [Hasibi ICTIR16])

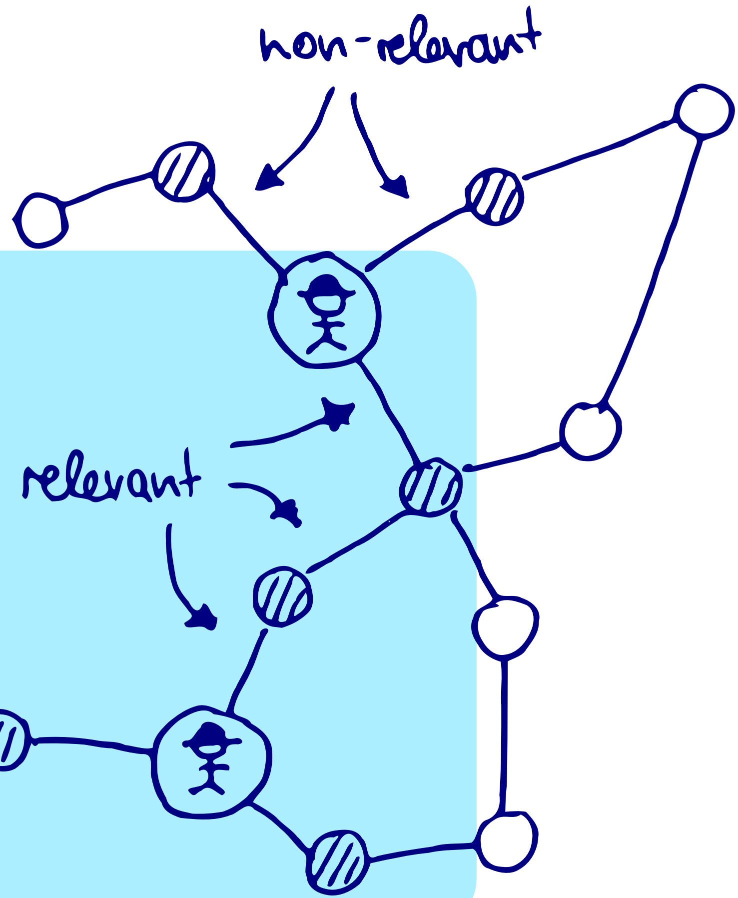
Entities as text:
Each article represents an Entity
Retrieve articles with keyword query Q
=> ranking / score of Entity

[Xiong ICTIR15, Dalton SIGIR14]

Source: Graph Structure and Walks

Graph Walks

[Boston 2014,
Kotov&Zhai 2012]



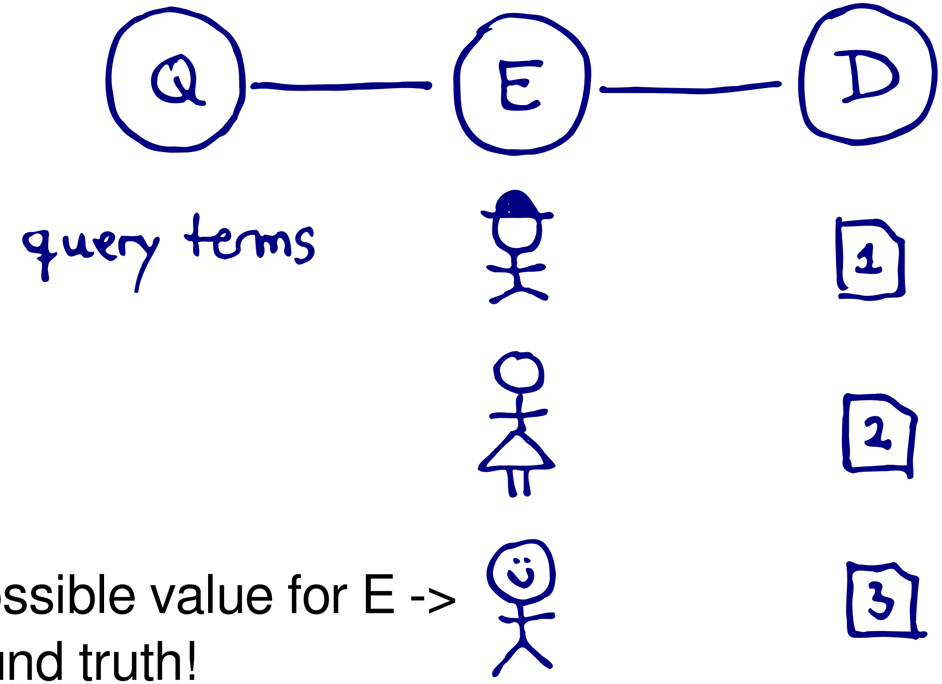
Machine Learning / Probabilistic Models

Three approaches based on similar ideas:

- Dalton: Entity Query Feature Expansion
- Xiong: EsdRank
- Liu: Latent Entity Space

Probabilistic model with random variables Q,E,D.

An edge represents a measure of compatability or similarity.



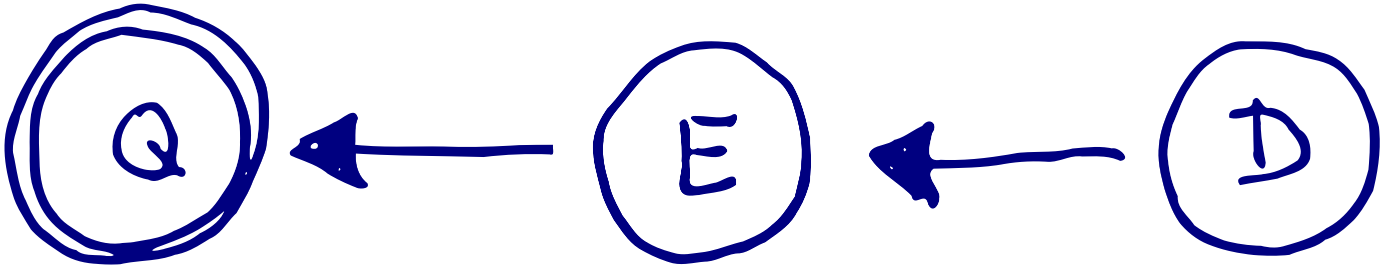
One possible value for E ->
no ground truth!

<- One possible value for D
ground truth available (TREC)

Machine Learning

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Latent Entity Space [Liu IRJ15]



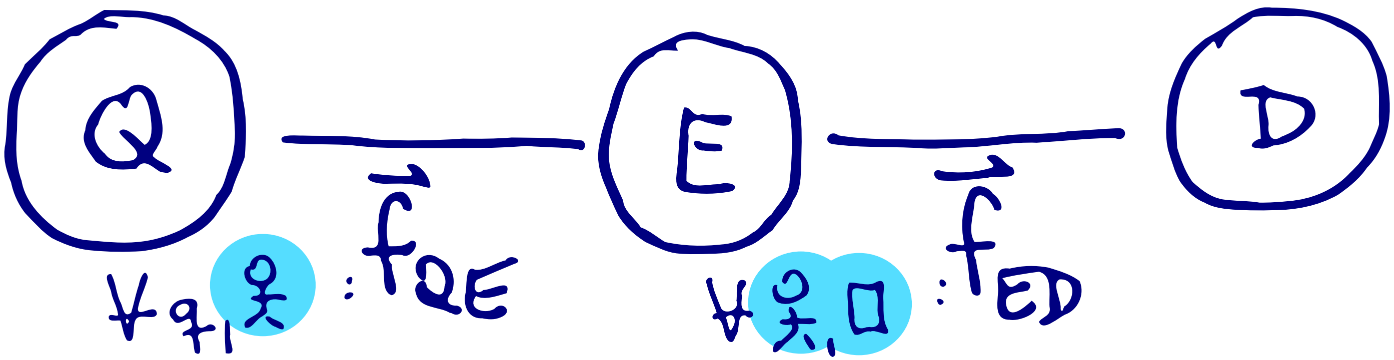
$$p(q|D = d, R = 1) = \sum_{e \in \mathcal{E}} p(q|e) \cdot p(e|d)$$

similarity of LM(q) and LM(e)

similarity of LM(e) and LM(d)

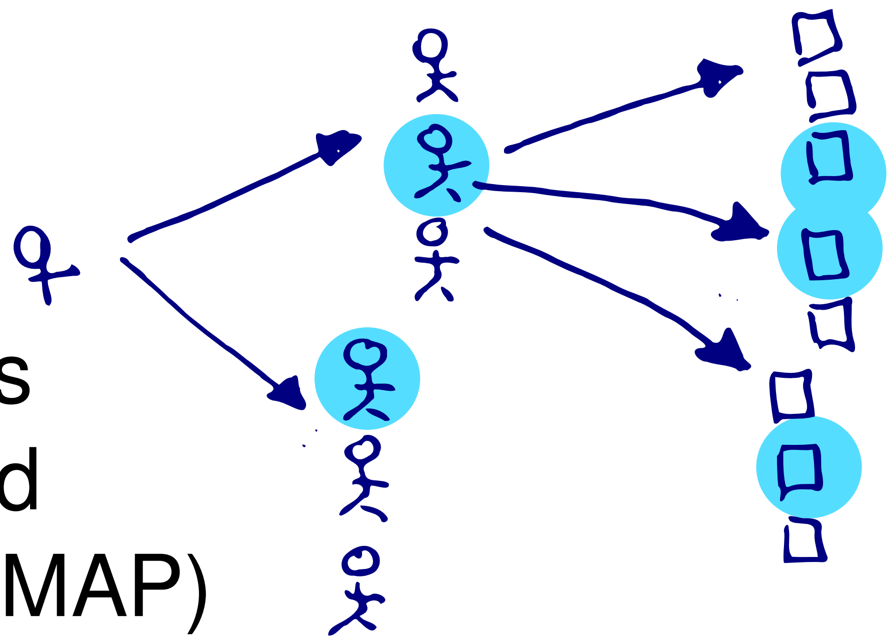
Wide range of experiments on which similarity measure / data source combination works best.

Entity Query Feature Expansion [Dalton SIGIR14]



n different ways to compute p(q|e)

m different ways to compute p(e|d)

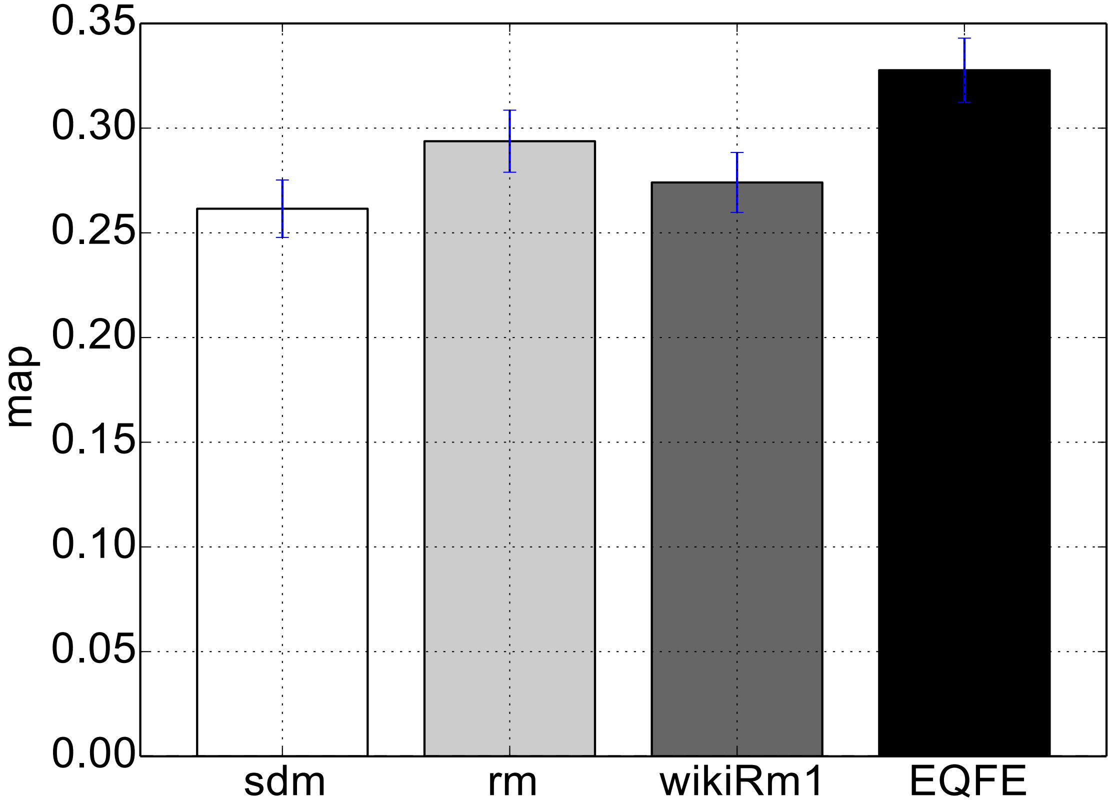


n x m features!

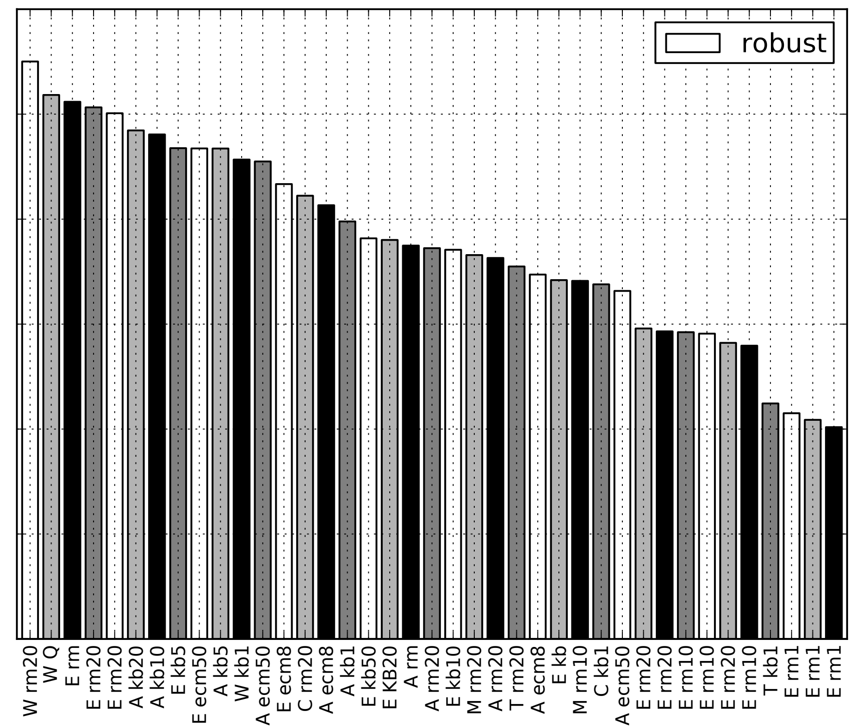
Combine features then use standard learning to rank (MAP)

Entity Query Feature Expansion [Dalton SIGIR14]

Results on Robust04 ad hoc document retrieval.



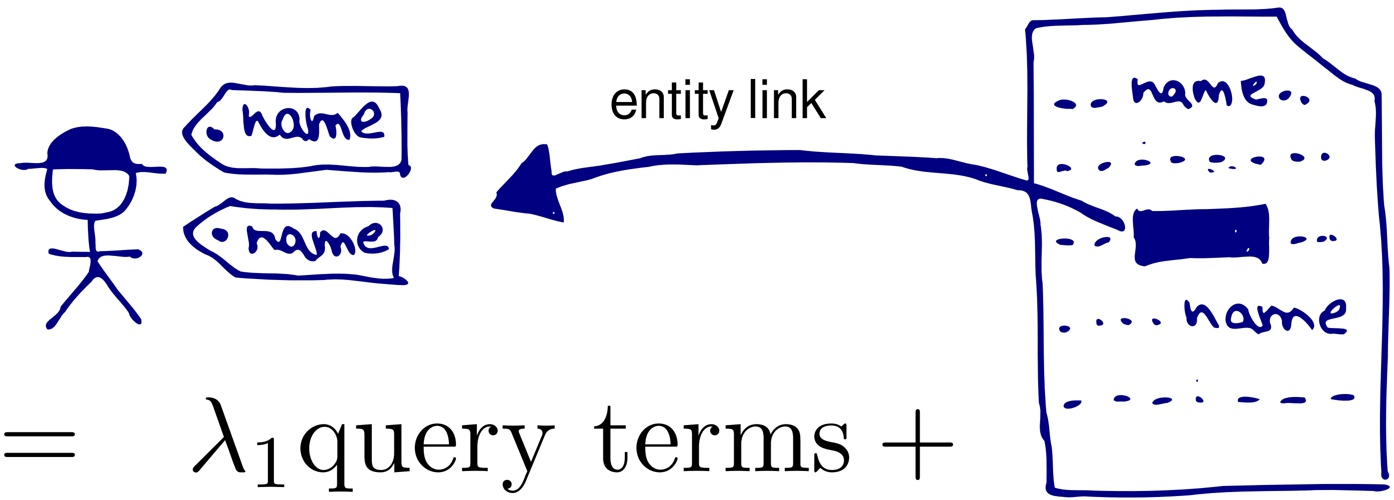
combined



features

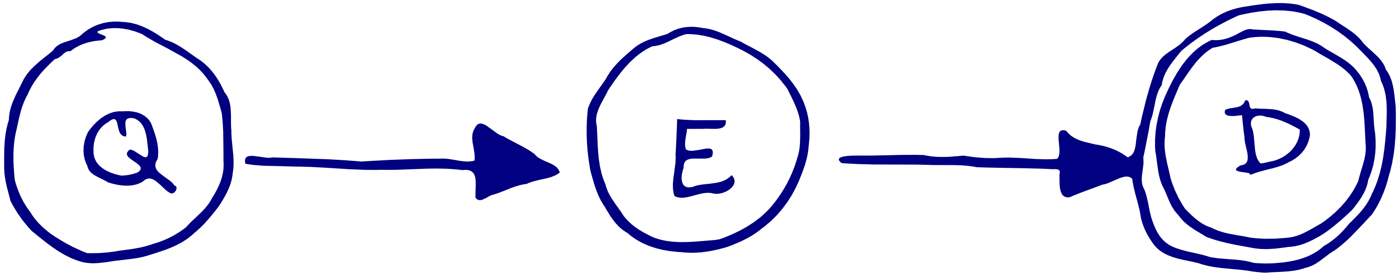
Relation to Query / Latent Concept Expansion

Various vocabularies, but all represented by sets



$$
\begin{aligned}
score(\square) = \quad & \lambda_1 \text{query terms} + \\
& \lambda_2 \text{names} + \\
& \lambda_3 \text{entity links} + \\
& \lambda_4 \text{article terms} + \ldots
\end{aligned}
$$

EsdRank [Xiong CIKM15]



$$p(d_i|q) = \sum_{e \in \mathcal{E}} \underbrace{p(d_i|e)}_{\frac{1}{Z_1}\exp\left\langle \vec{w}_1, \vec{f}_{D,E} \right\rangle} \cdot \underbrace{p(e|q)}_{\frac{1}{Z_2}\exp\left\langle \vec{w}_2, \vec{f}_{E,Q} \right\rangle}$$

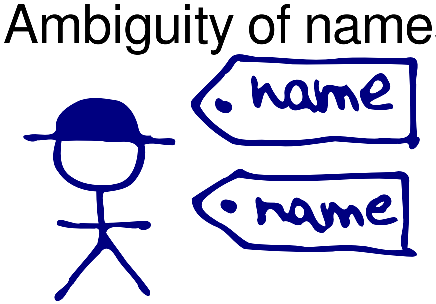
Discriminative probabilistic model based on Generalized linear models + EM Algorithm for learning weights w1, w2.

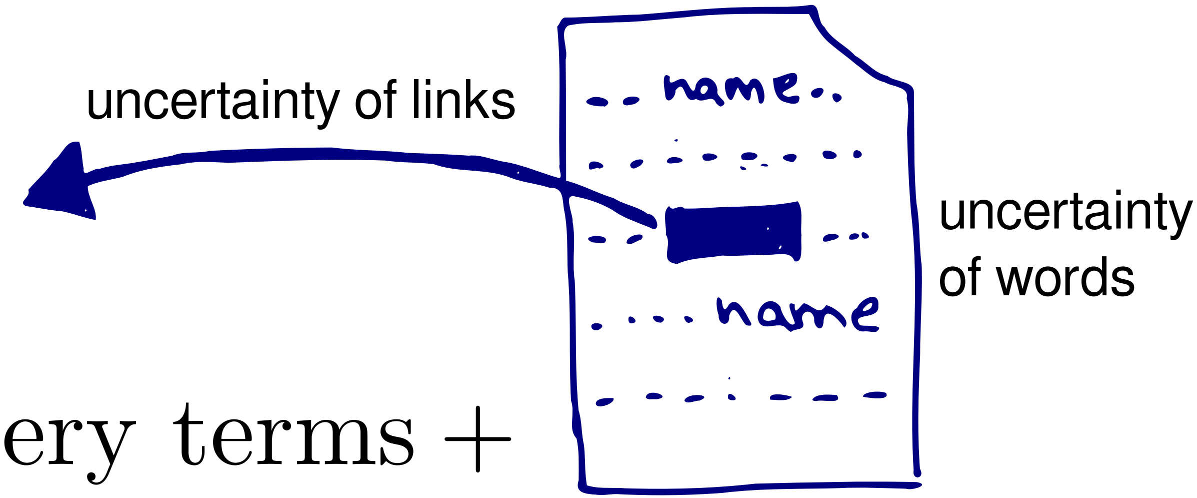
Only n+m features! But needs custom learning code.

Query Expansion with Uncertainties

Taking uncertainty and confidences into account.

[Raviv SIGIR16, Liu IRJ15]





$$
\begin{aligned}
score(\square) = \quad & \lambda_1 \text{query terms} + \\
& \lambda_2 \sum p(\text{names}|e) + \\
& \lambda_3 p(\text{entity link to } e|d) \\
& \lambda_4 KL\left(p(\text{terms}|e) \parallel p(\text{terms}|d)\right)
\end{aligned}
$$

Entity Aspects

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Entity Aspects

Danger: An entity is relevant, but:
only because of one aspect
=> many non-relevant aspects of relevant entities.

Example aspects about UK:
- still a member of the European Union
- is a constitutional monarchy
- the Raspberry Pi was invented in the UK
- there are many great UK bands

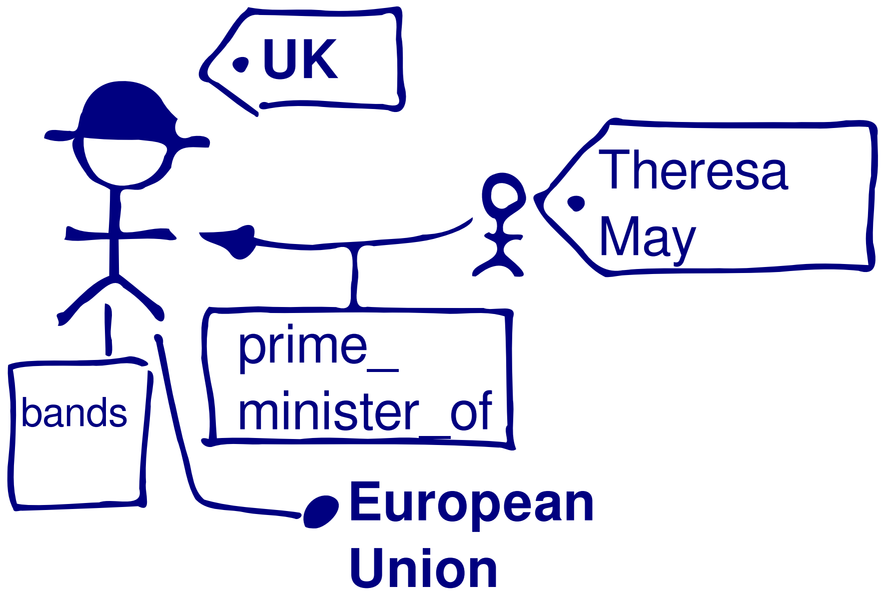
Depending on query, some are relevant, some not.

How to Represent Entity Aspects?

As terms?	UK bands brexit
As types?	UK member of "European Union"
As is-a?	UK as a European country
Related entities?	`UK` `Theresa_May`
Relations?	`Theresa_May` prime_minister_of `UK`
Language Model	p(brexit)=0.4 p(leave)=0.25 p(immigration)=0.10

[Reinanda SIGIR15, Liu IRJ15, Prasojo CIKM15]

Entity Aspects: Using KG ...



UK bands

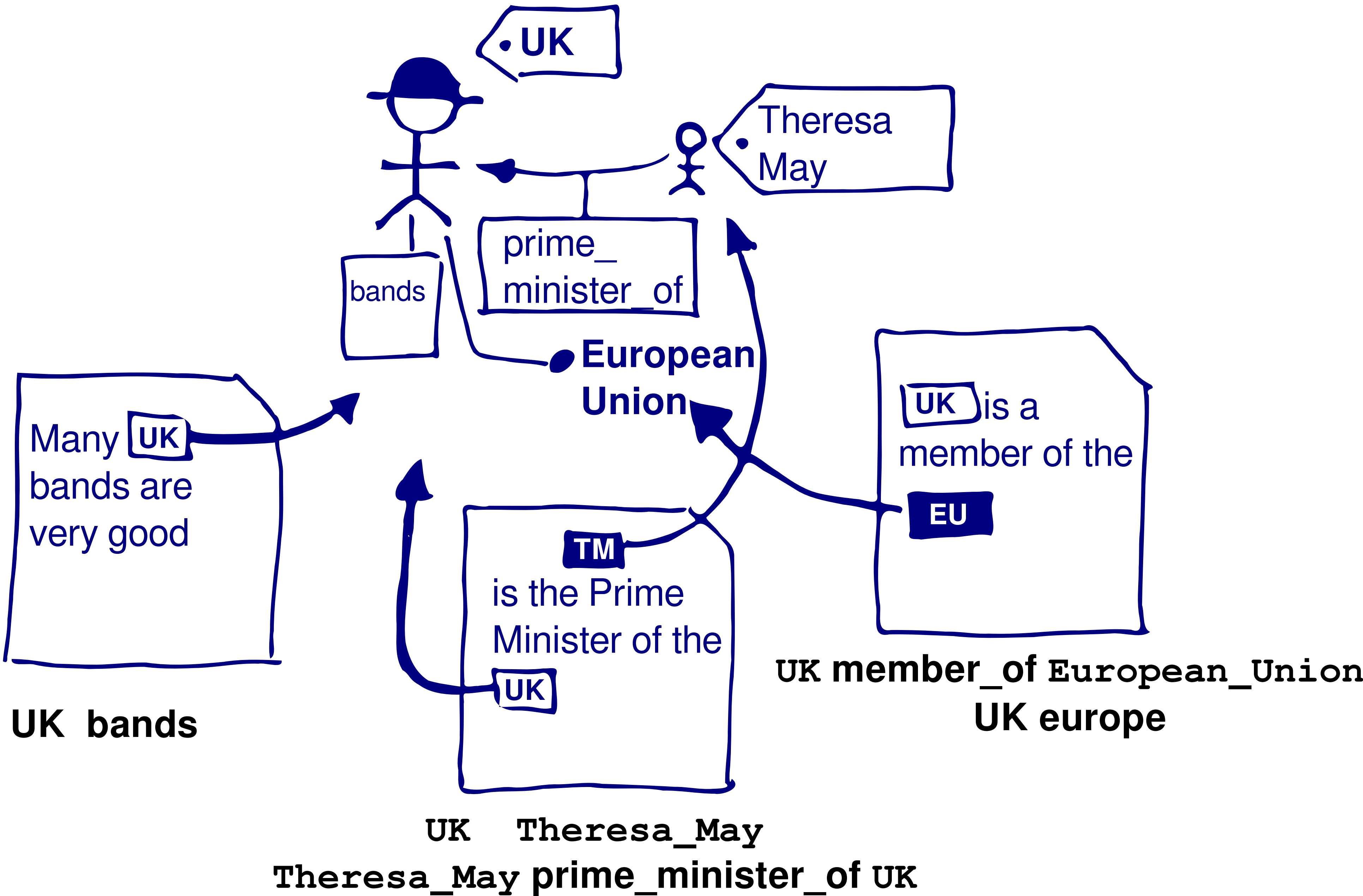
`UK` **member_of** `European Union`

UK europe

`UK Theresa_May`

`Theresa_May` **prime_minister_of** `UK`

Entity Aspects: Using KG and Text

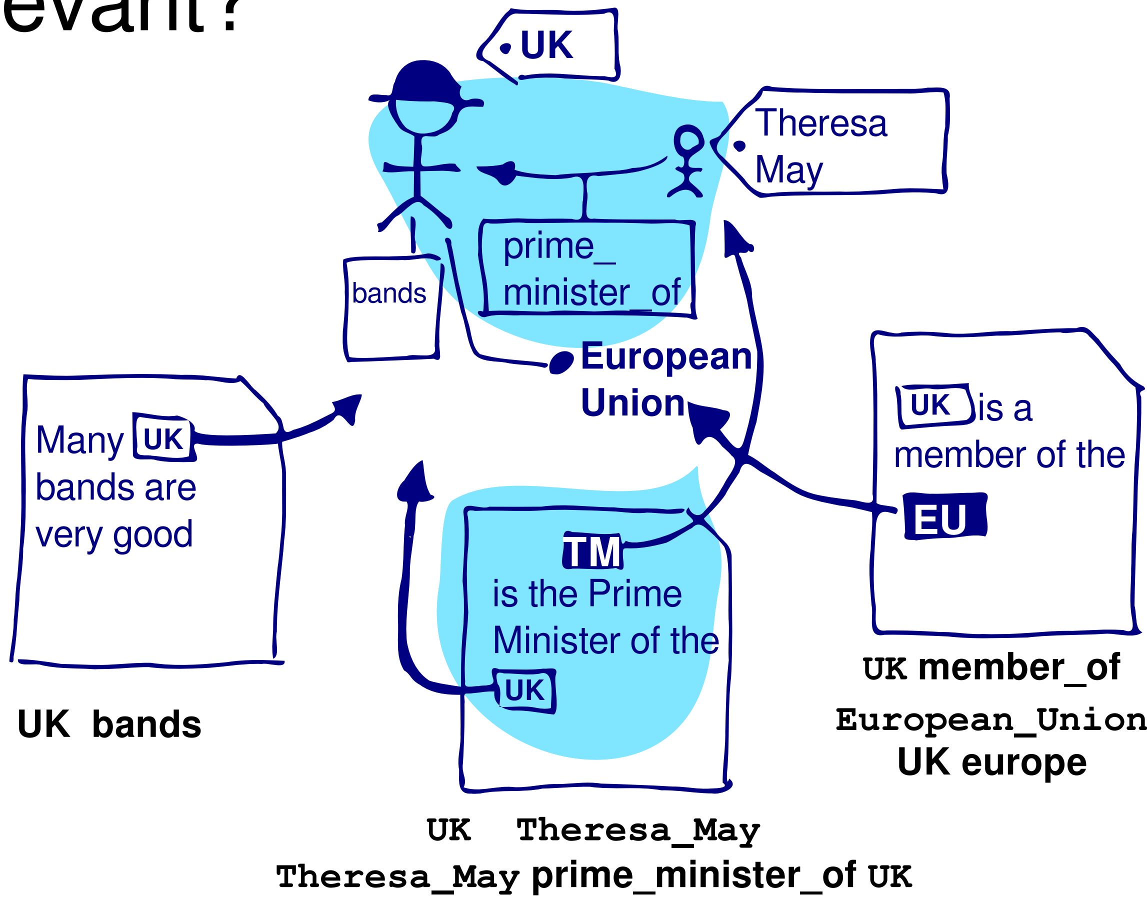


Entity Aspects: Infer Relevance, Match, Extract

1) Relevance:
Which aspects are relevant?

2) Match:
How to match in text?

pseudo relevance feedback

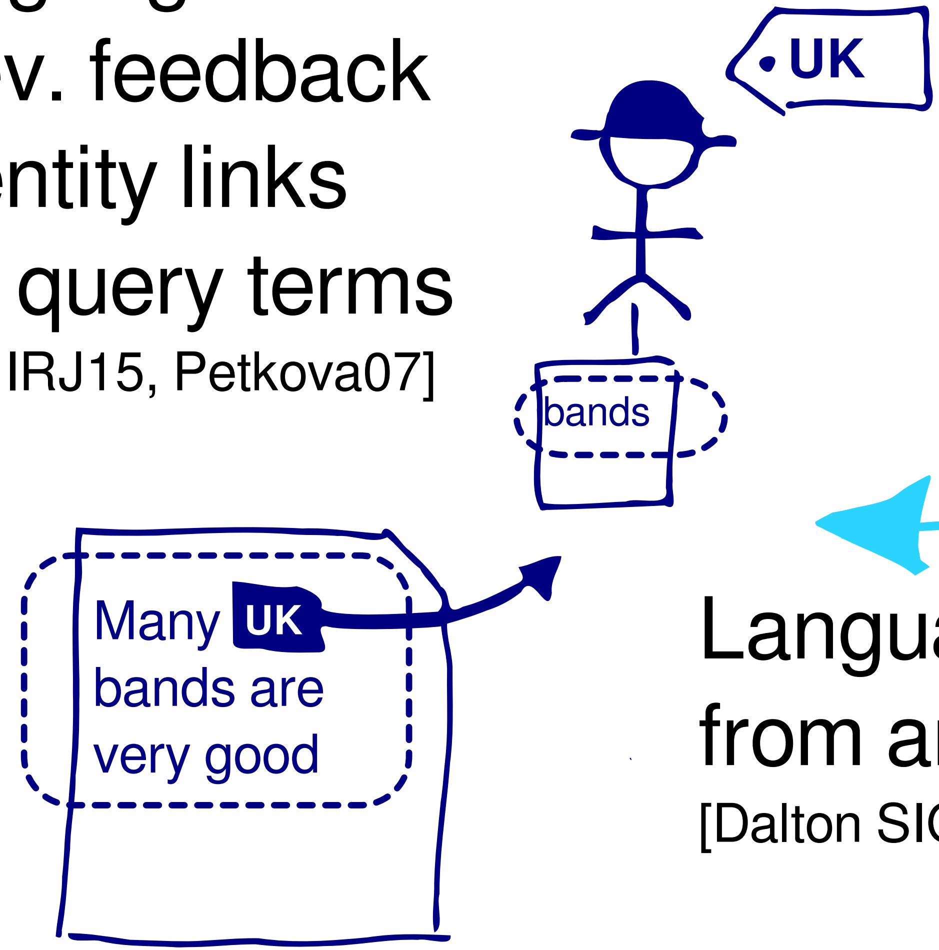


3) Extract:
How to extract new aspects? (KB population)

Entity Aspects as Terms

Passage-Language Model
- Pseudo relev. feedback
- Context of entity links
- Proximity to query terms

[Dalton SIGIR14, Liu IRJ15, Petkova07]



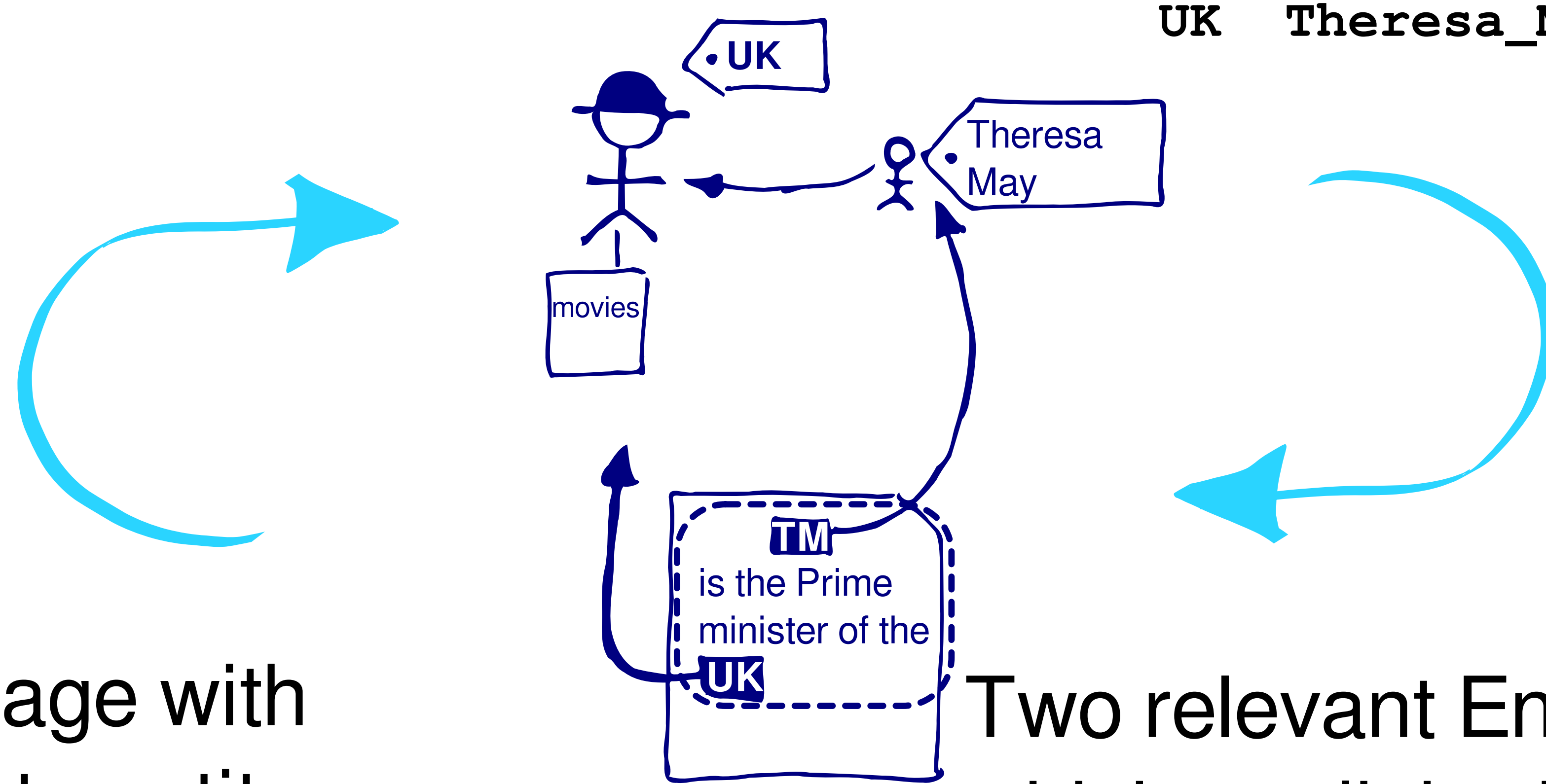
UK movies

Language model from article / descr.

[Dalton SIGIR14, Liu IRJ15]

Entity Aspects through Co-mentioned Entities

`UK   Theresa_May`



Passage with
- link to entity
- matching query terms

⇒ other entities relevant?

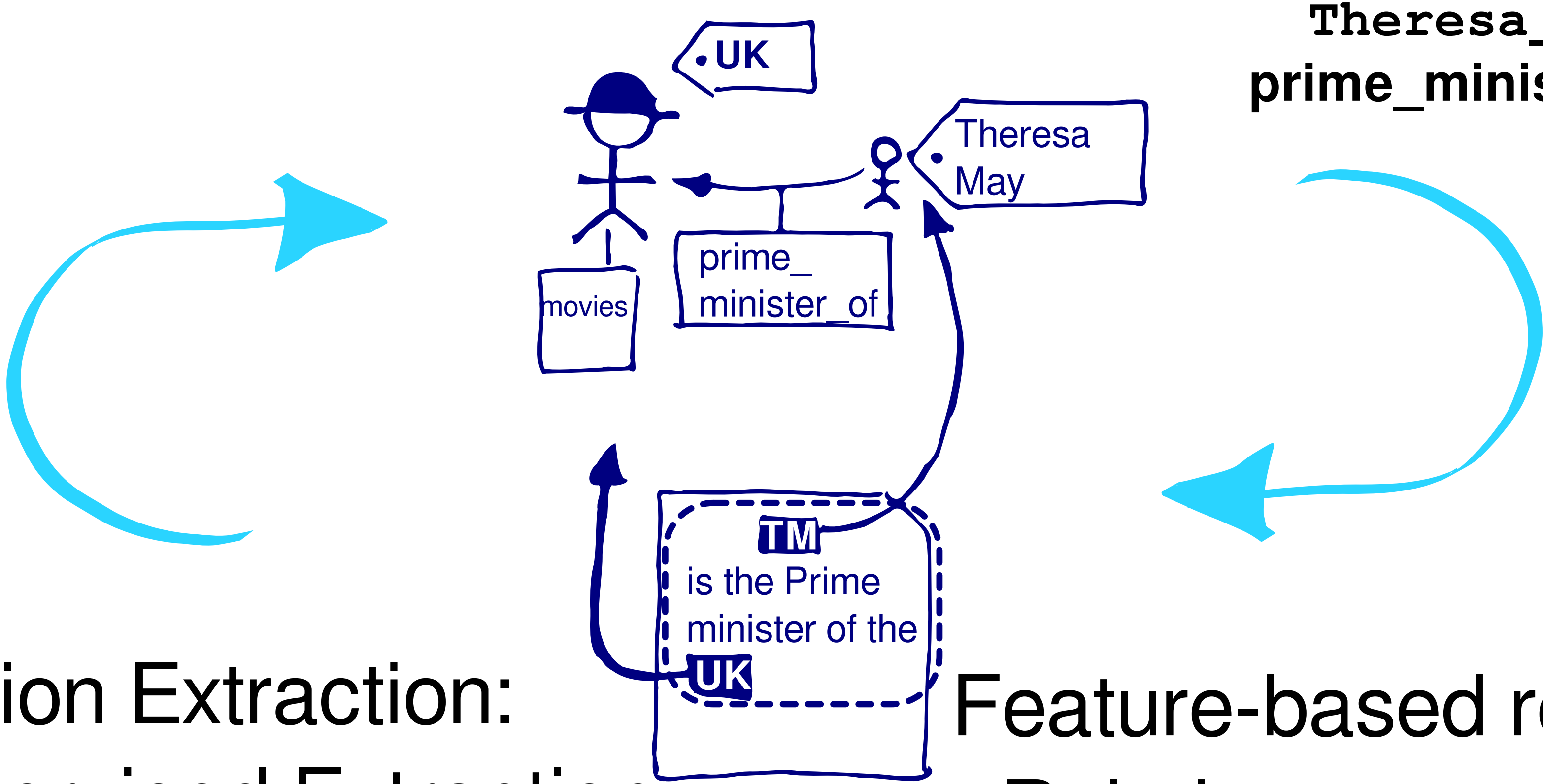
Infer & Extract Aspects

Two relevant Entities which are linked in KG

⇒ Promote documents that mention both

Match Aspects

Entity Aspects through Relations (Triples)



Theresa_May prime_minister_of UK

Relation Extraction:
- Supervised Extraction from Text

[Schuhmacher ECIR16]

Feature-based retrieval:
- Relation terms
- Cosine of word vectors

[Voskarides ACL15]

Infer & Extract Aspects

Match Aspects

Retrieving/Matching relevant Entity Aspects?

- Terms and entity links in documents
- Co-occurrence (AND versus OR)
- Proximity
- Frequency
- Probability under a language model
- Classification (e.g., Naive Bayes for types)
- Information extraction and matching

For your reference

Extract/Infer relevant Entity Aspects?

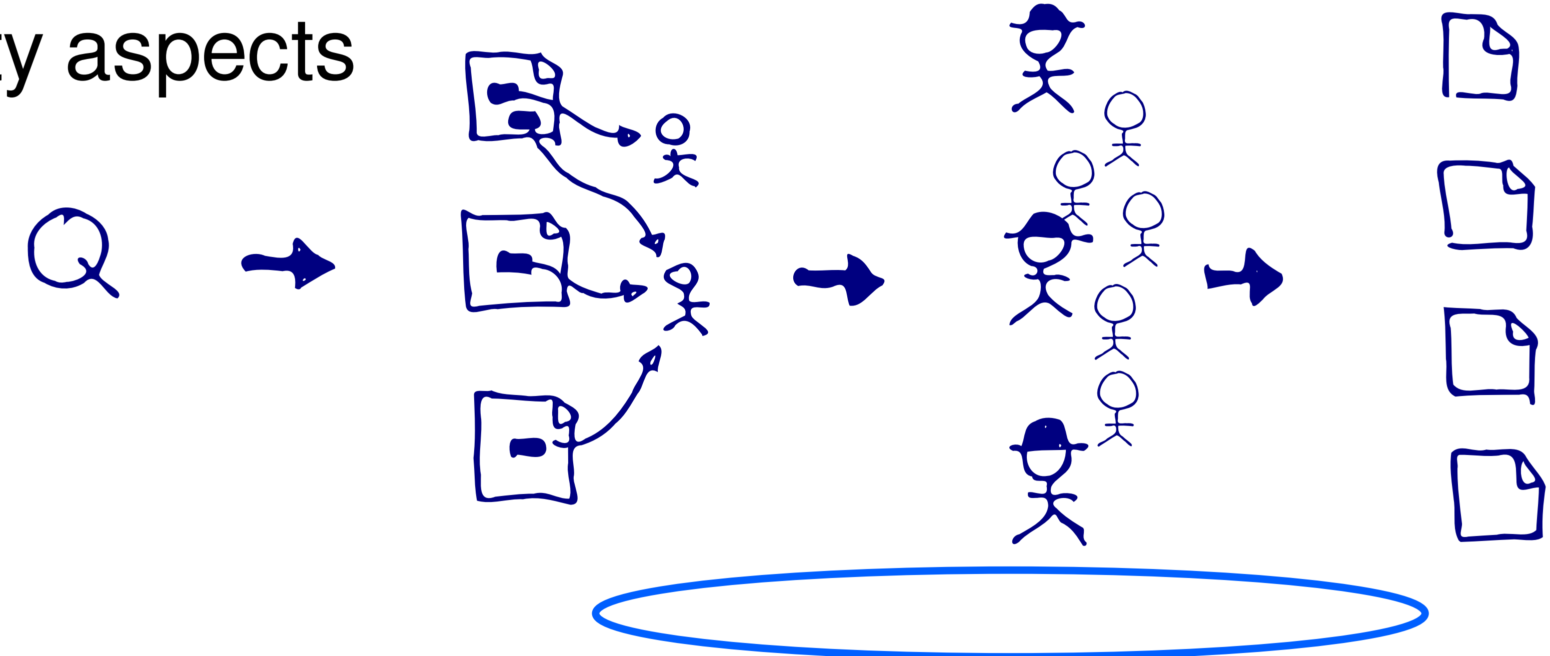
- From collocations in pseudo-relevant documents
- From passages surrounding entity links
- Through graph analysis
- Frequency/proximity of entities in context
- Extracting a language model

For your reference

Summary (Part 4)

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Please take the survey!



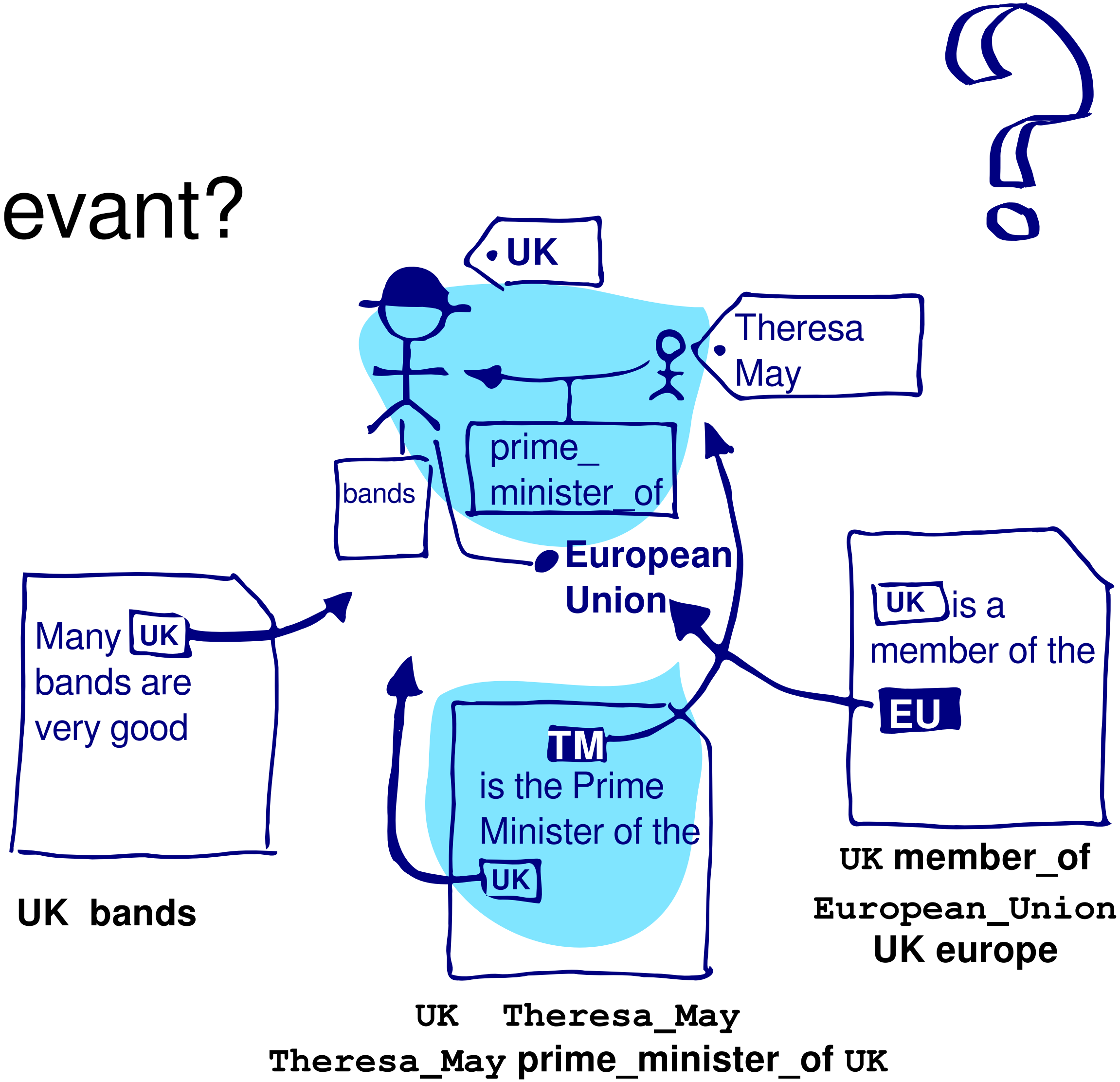
Open Issue: Entity Aspects for Document Ranking

1) Relevance:
Which aspects are relevant?

2) Match:
How to match in text?

pseudo relevance feedback

inverse tasks



3) Extract:
How to extract new aspects? (KB population)