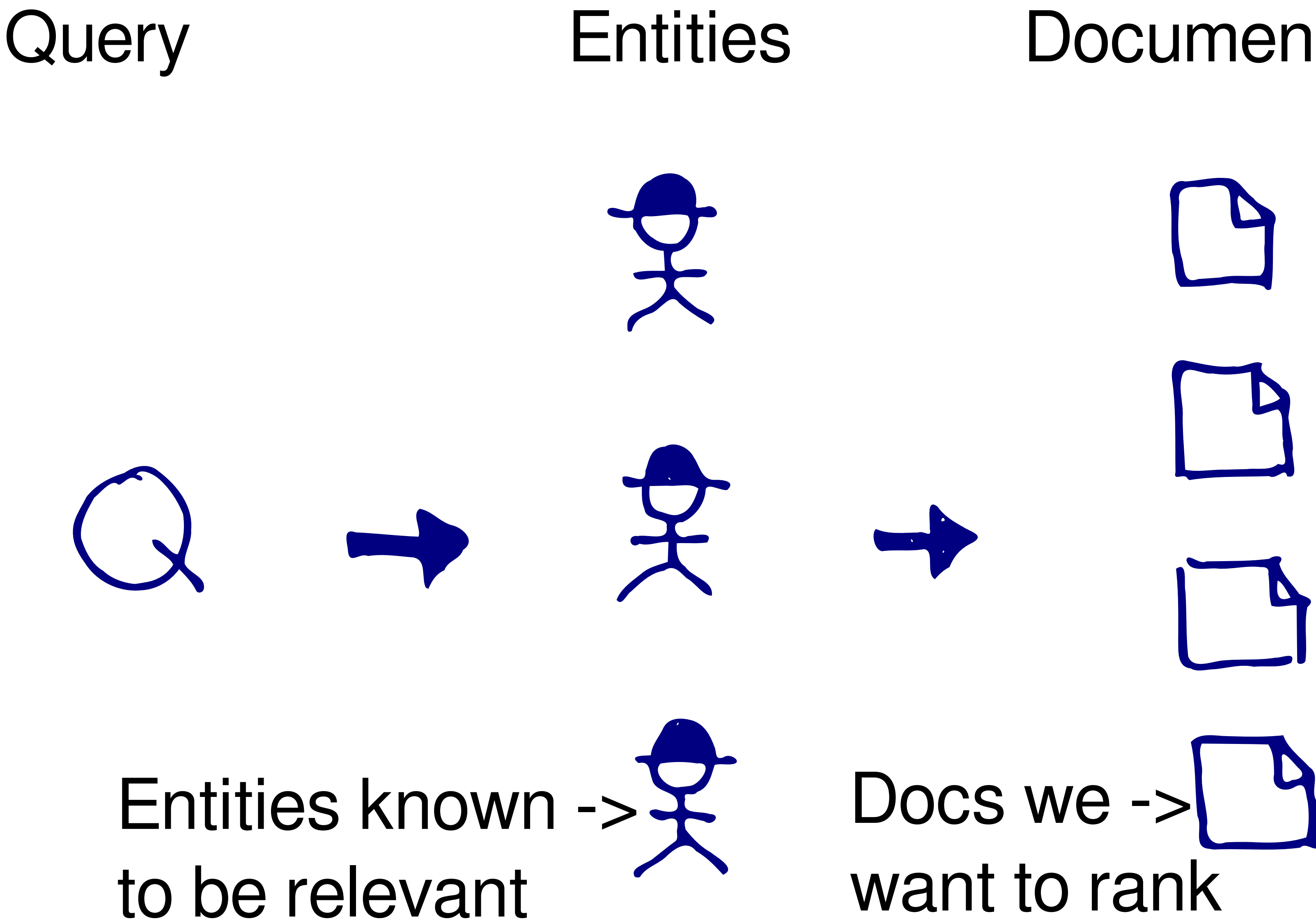
Tutorial - Part 5:
Using Knowledge Graphs for Text Retrieval

Laura Dietz
dietz@cs.unh.edu

Document Retrieval with Entities

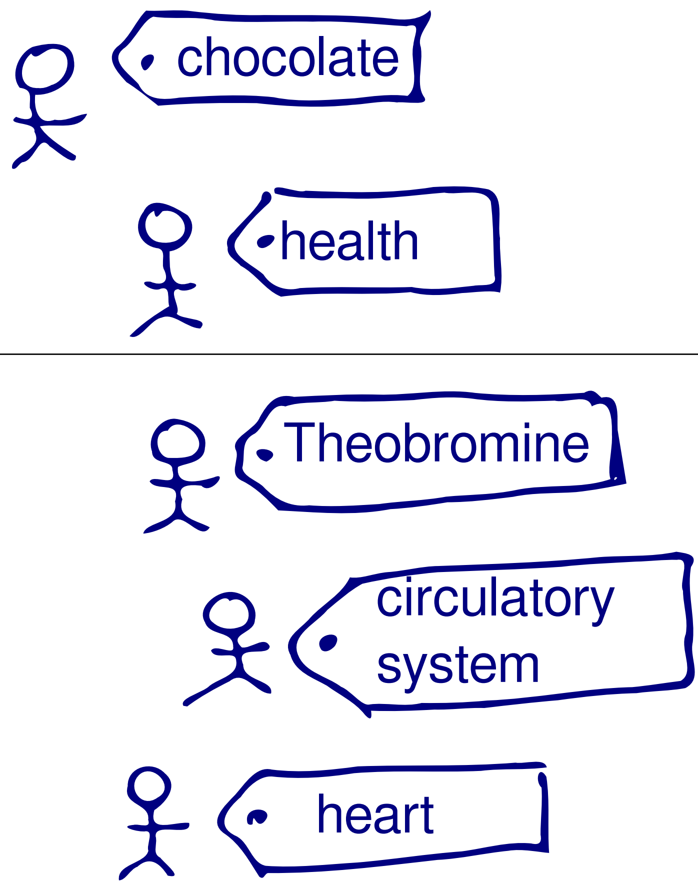


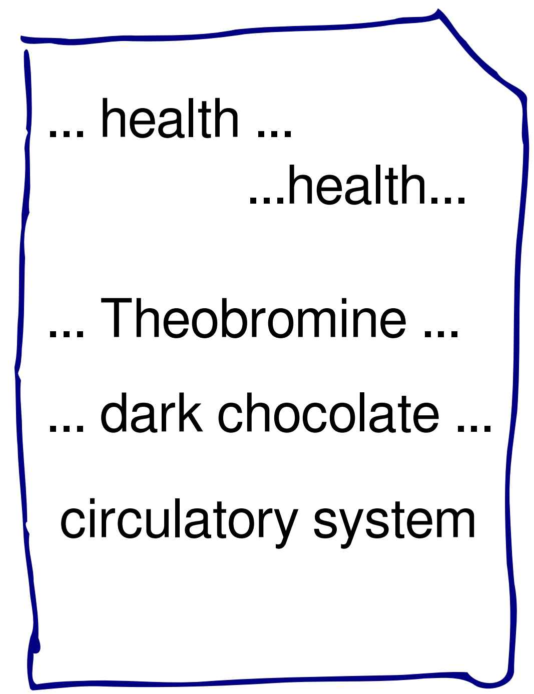
Different Queries - Different Entities

Query	nicolas cage movies	dark chocolate health benefits
Query entities	Nicolas Cage	chocolate health
Latent entities	Left Behind Lea Thompson	Theobromine circulatory system heart
[Hasibi ICTIR16]	**Named Entities**	**Concepts**

Matching Entities in Documents by Name

dark chocolate
health benefits

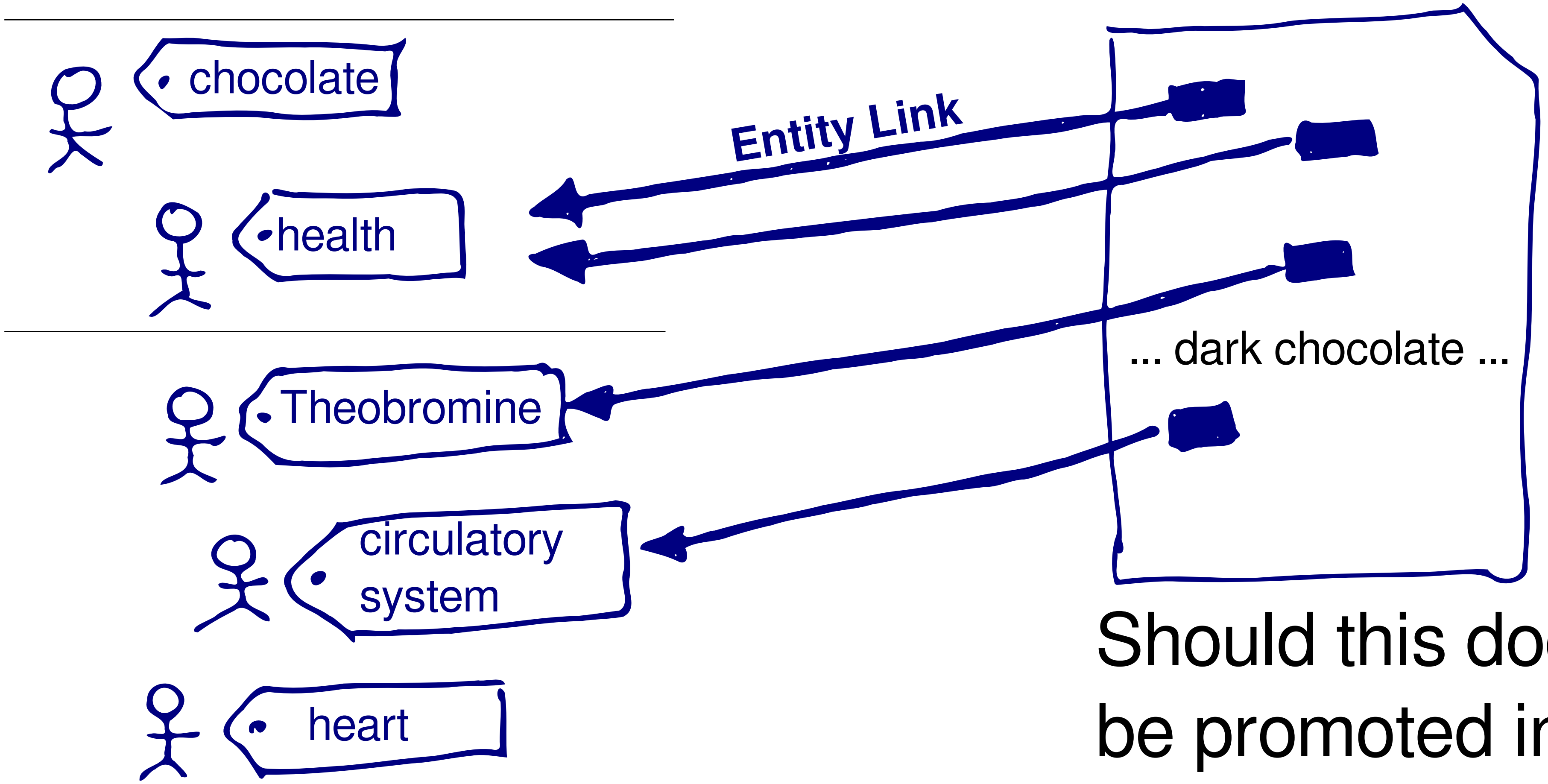




Should this doc be promoted in the ranking?

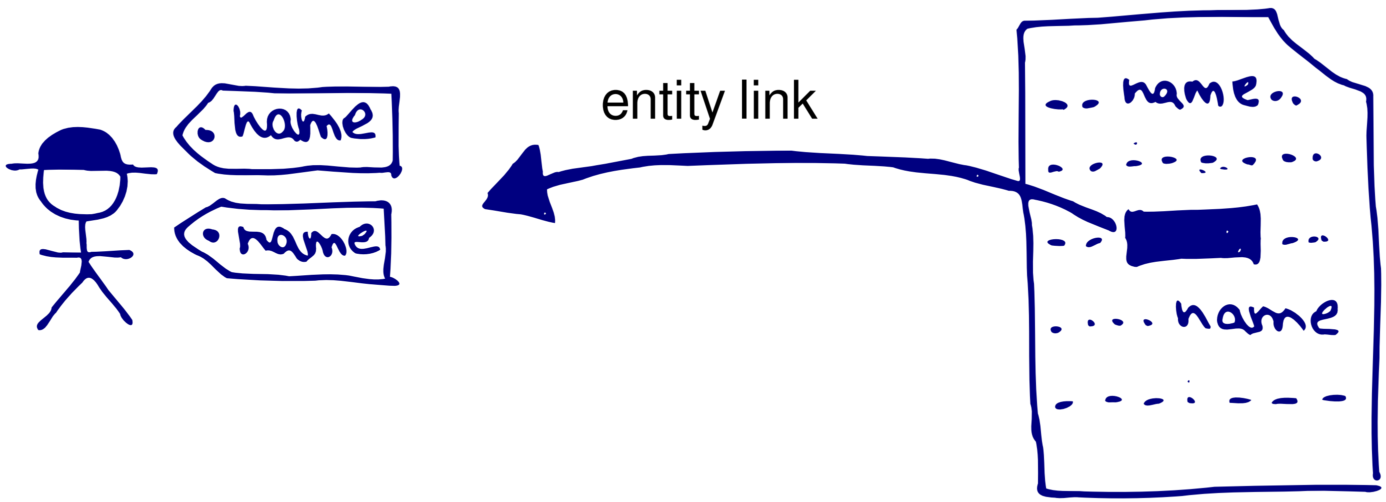
Matching Entities in Documents by Entity Links

dark chocolate
health benefits



Should this doc be promoted in the ranking?

Using Entities as a Vocabulary of Concepts

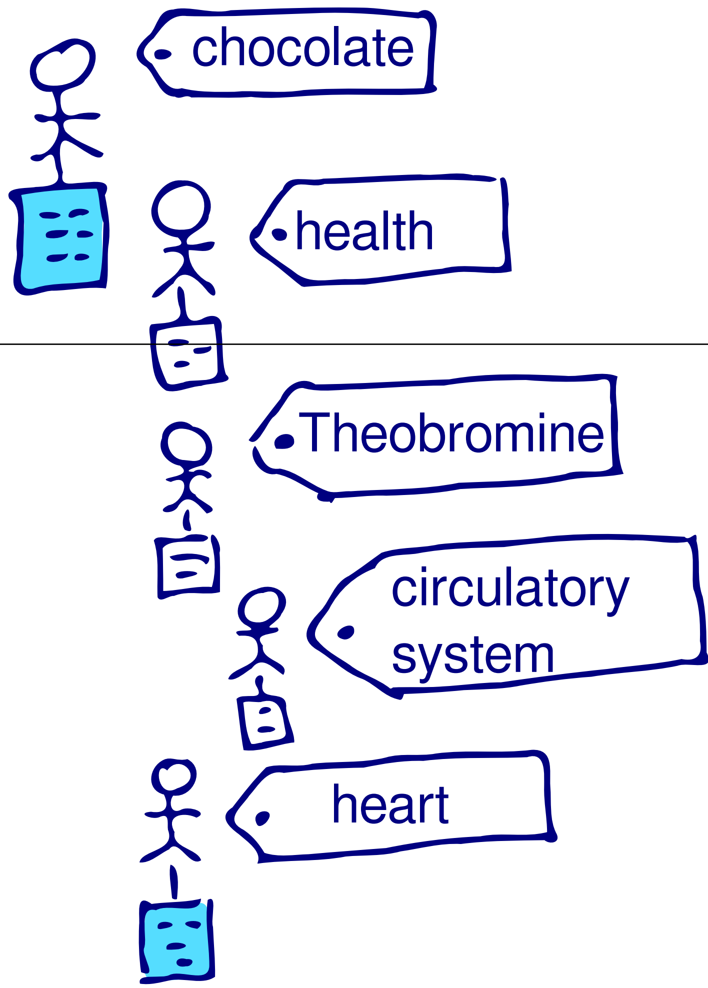


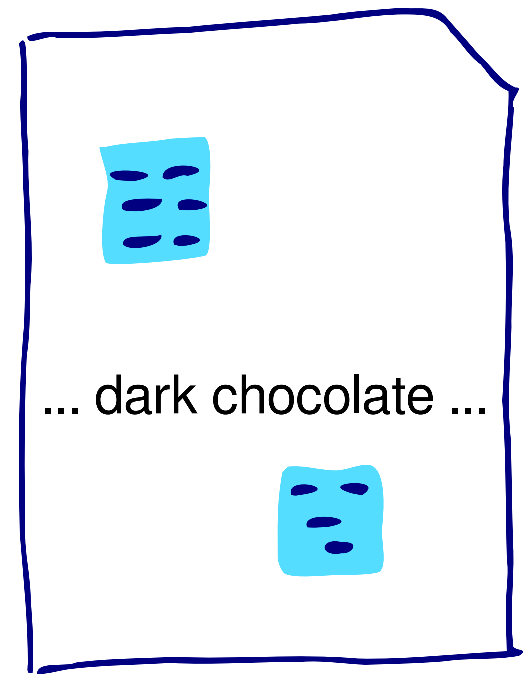
$$score(\square) = \lambda_1 \text{query terms} + \lambda_2 \text{names} + \lambda_3 \text{entity links} + \lambda_4 \text{article terms} + \ldots$$

use your favorite retrieval model here!

Matching Entities in Documents by Article Terms

dark chocolate
health benefits

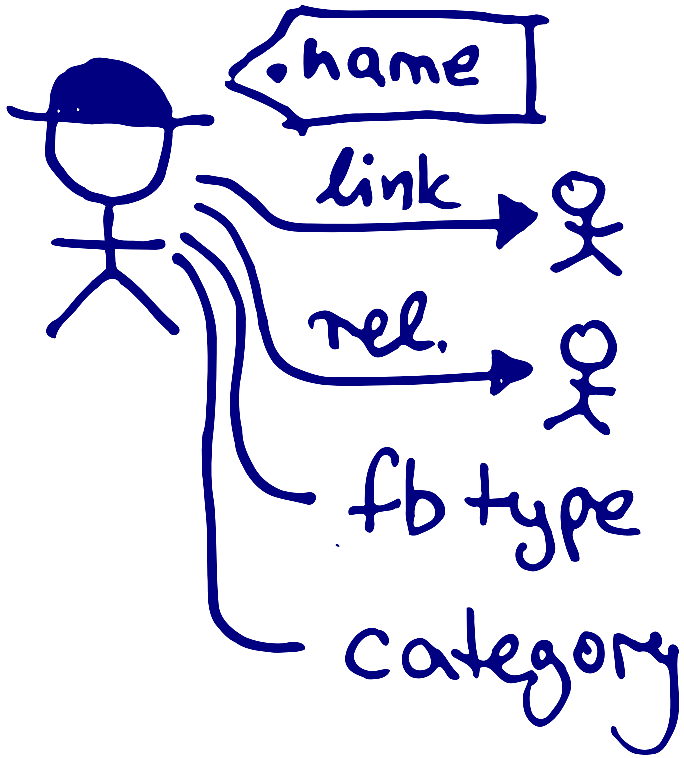




Should this doc be promoted in the ranking?

Using more from the Knowledge Base

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



Names

Links and Relations

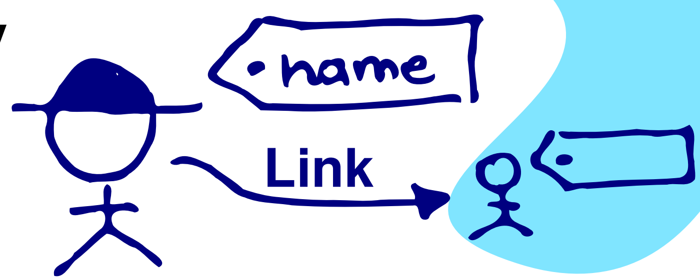
Different taxonomic
Type systems

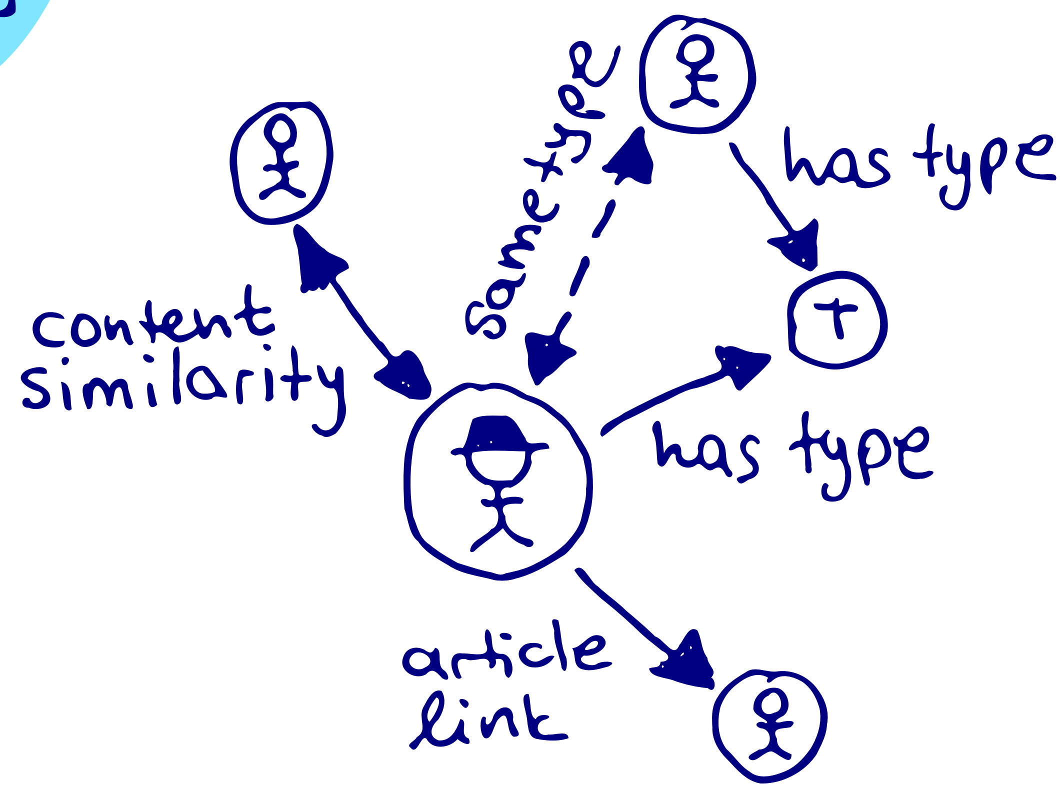
How can we make use of it?

Using Relations and Types with Entity Links

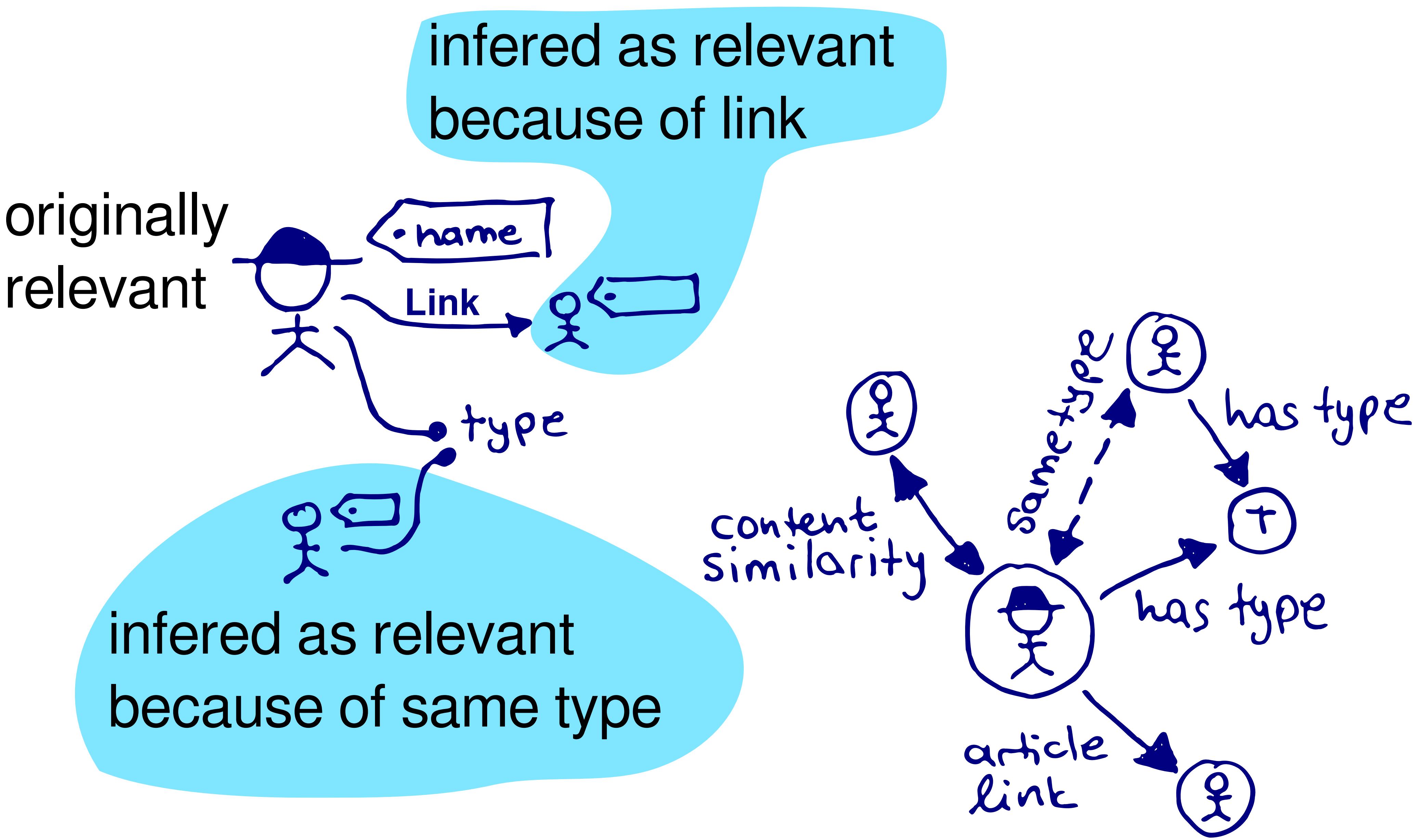
infered as relevant
because of link

originally
relevant

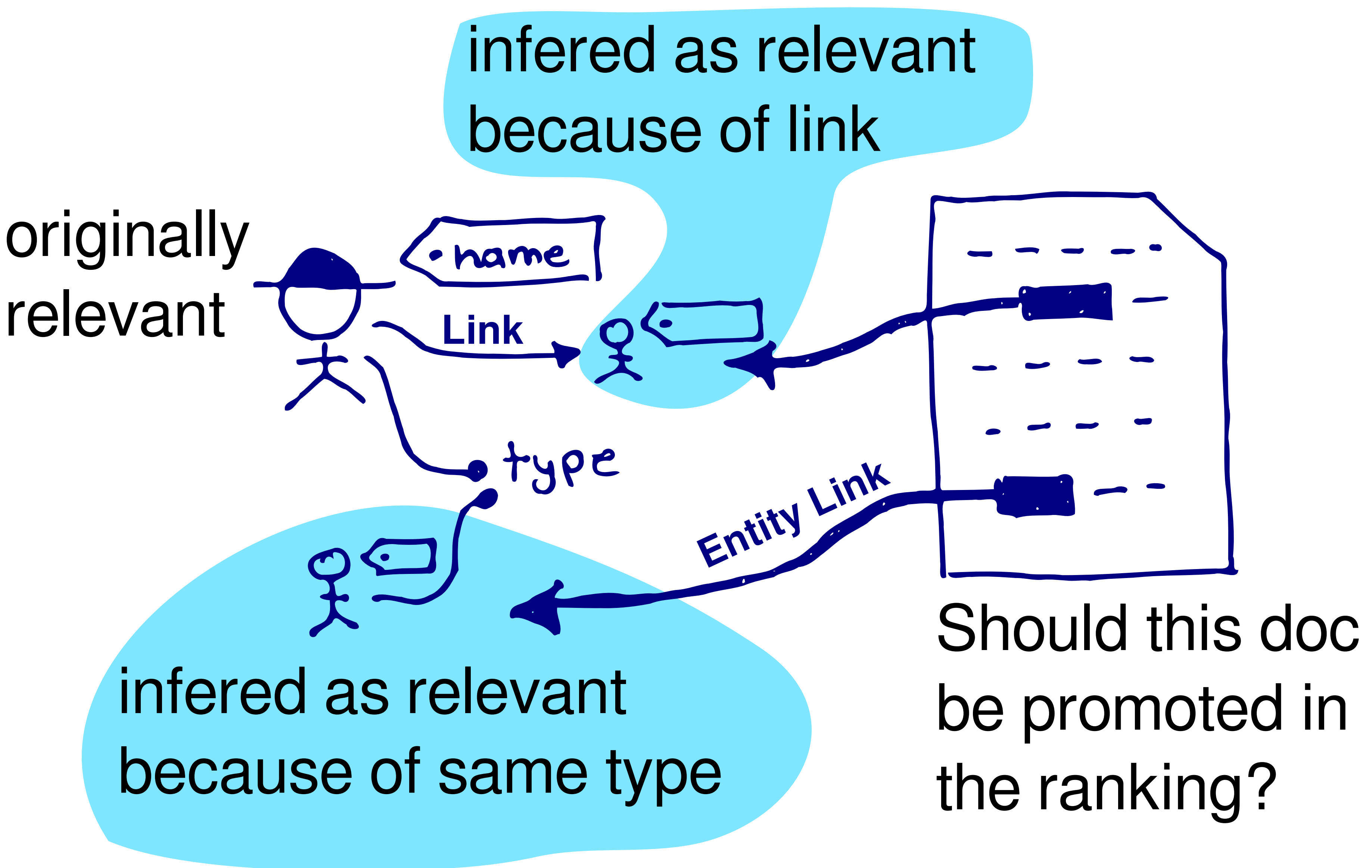




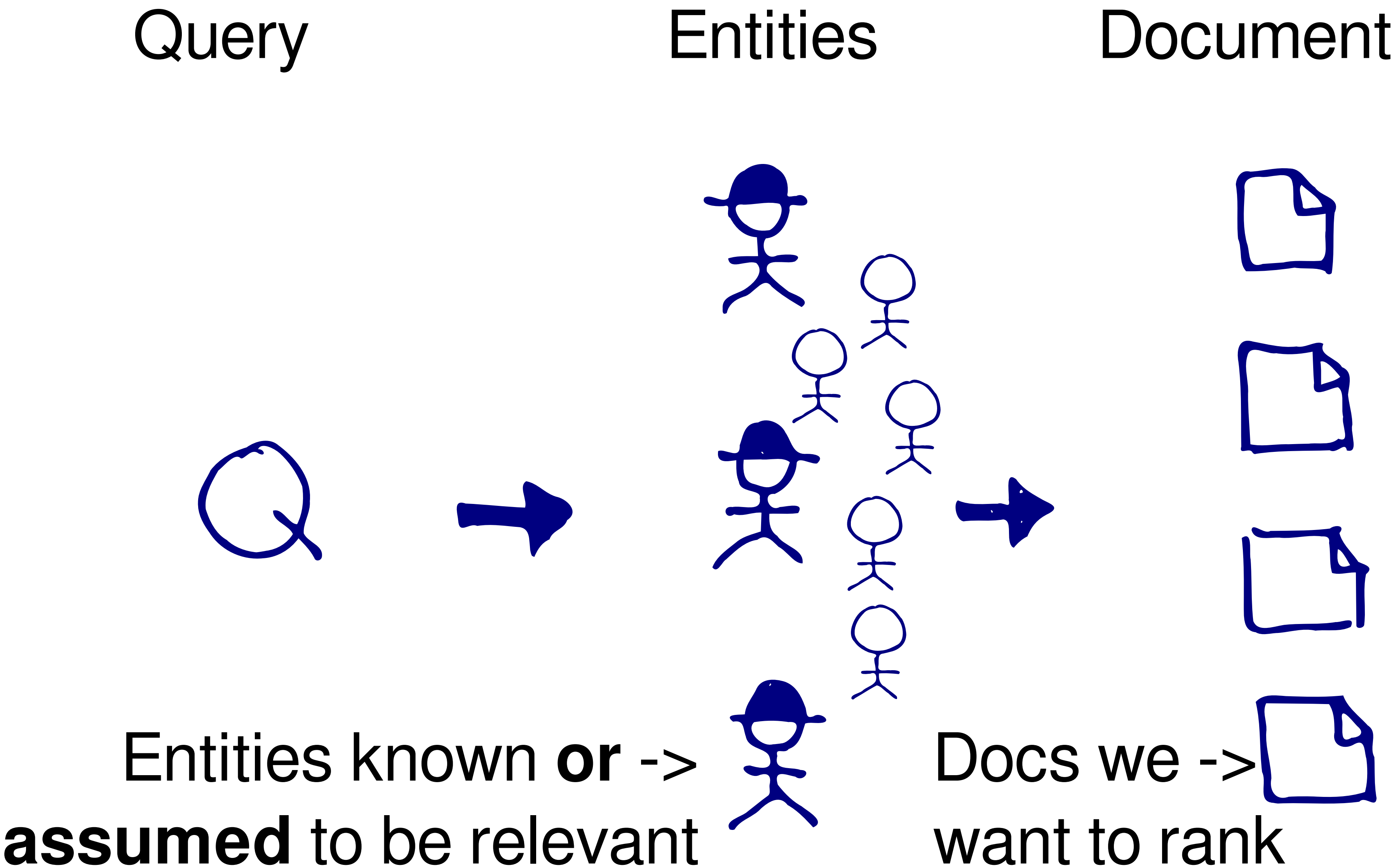
Using Relations and Types with Entity Links



Using Relations and Types with Entity Links

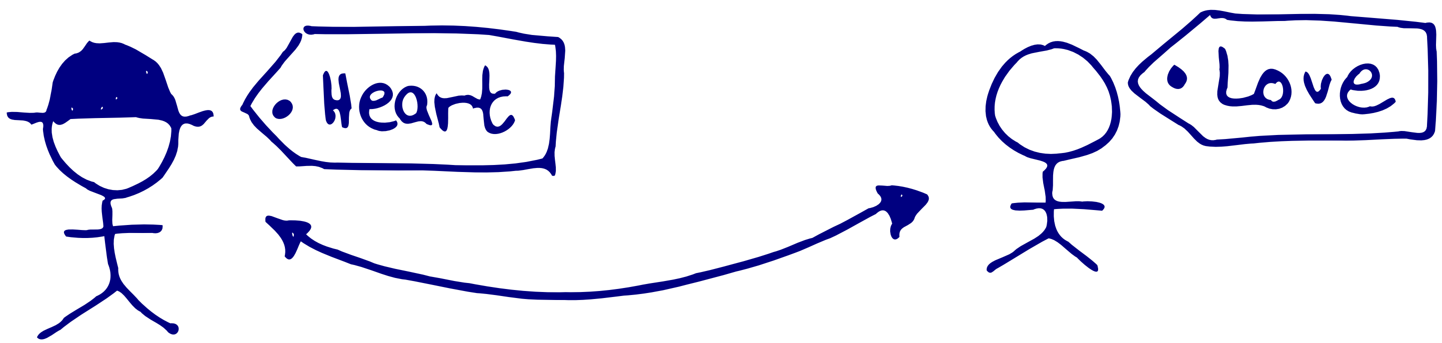


Document Retrieval with (more) Entities



KG expansion: A Potential Issue

Example query: Heart disease
Consider:

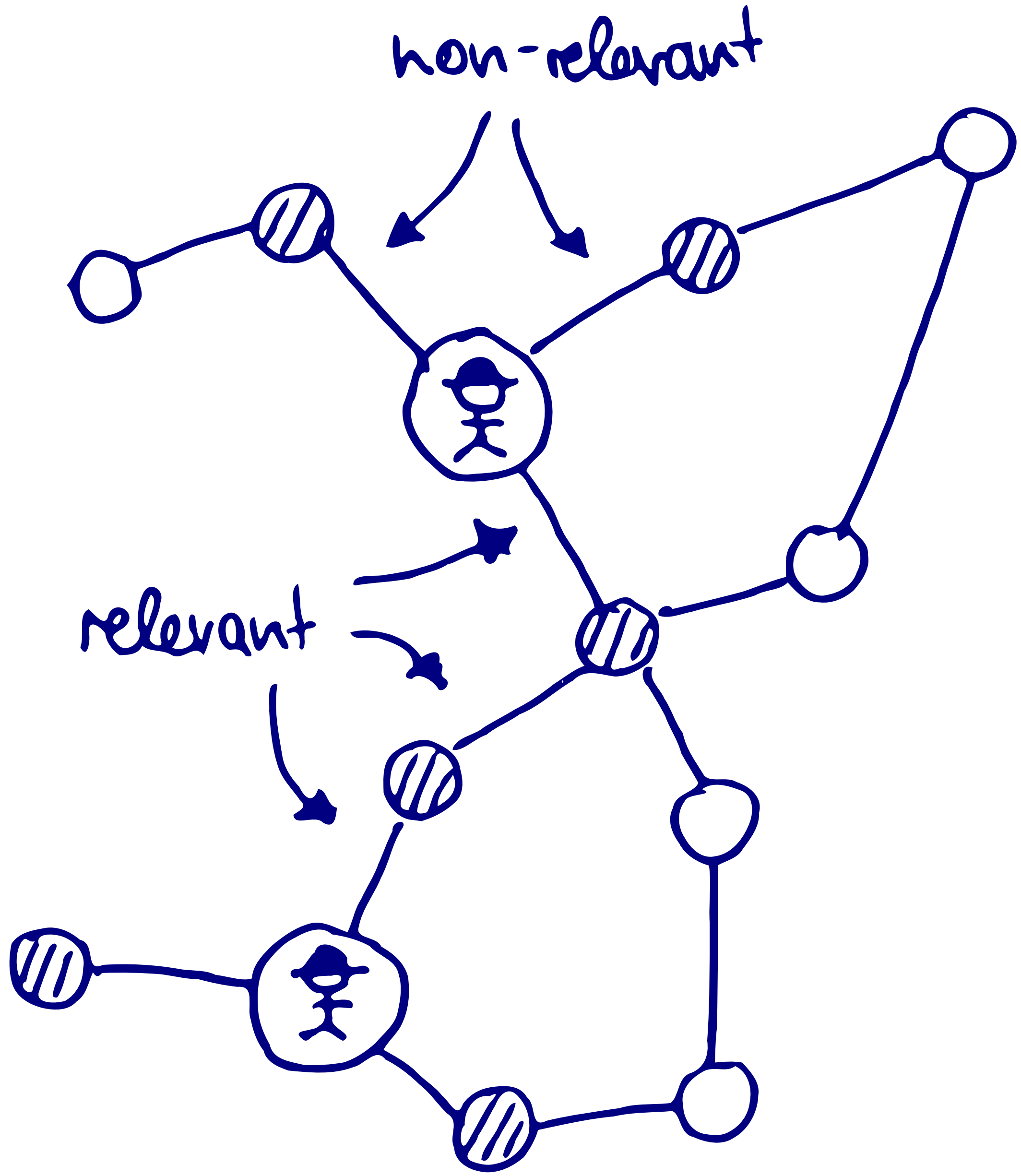


Correct connection, but:
The connection is not relevant in context of "heart" as in "heart disease".
If we promote docs because they talk about love, we ruin a fine ranking on the topic heart disease.

General Approach: Graph Expansion

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

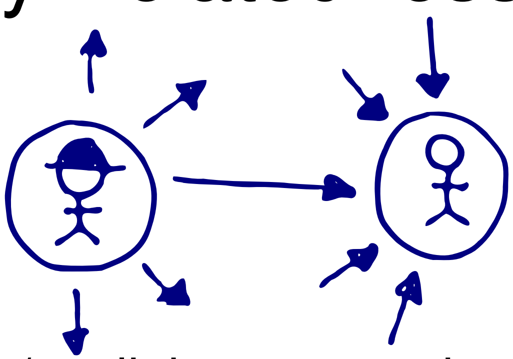
Expanding with non-relevant entities leads to low precision rankings.



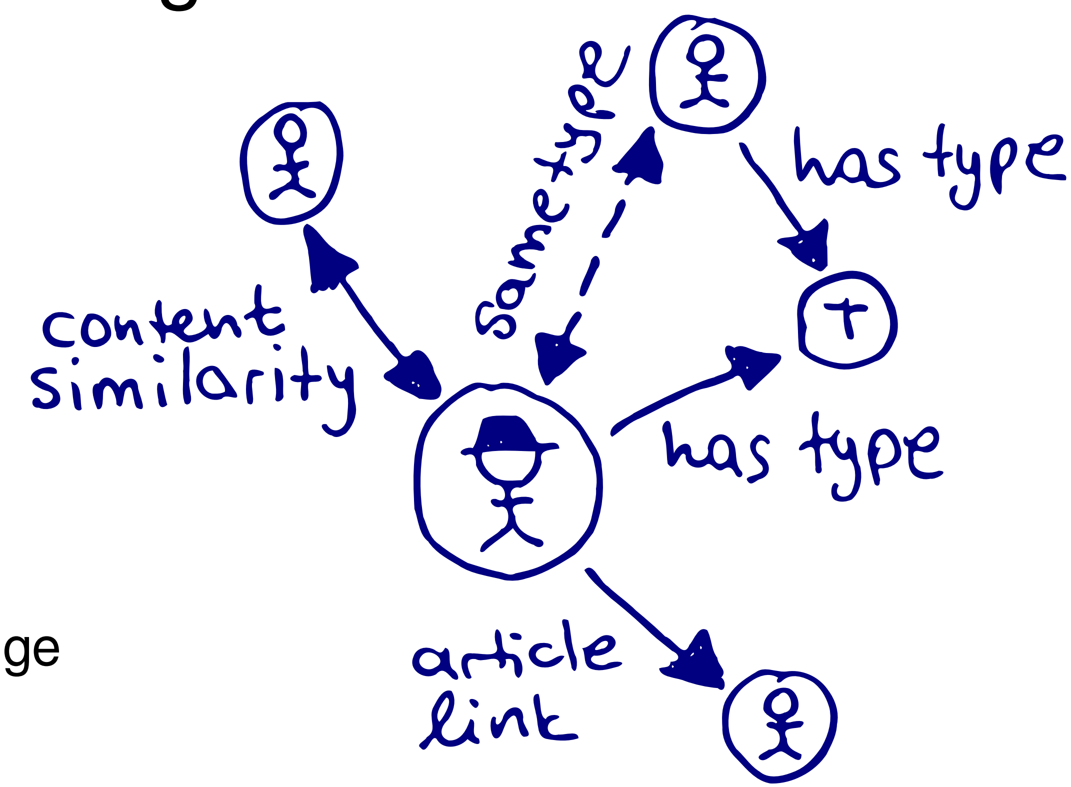
Using the Graph Structure (KG)

Using seed entity nodes and...
- Graph walks: PageRank / HITS
- Different edge types
- Edge weighting + Clustering

Exclusivity-based
Entity Relatedness



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

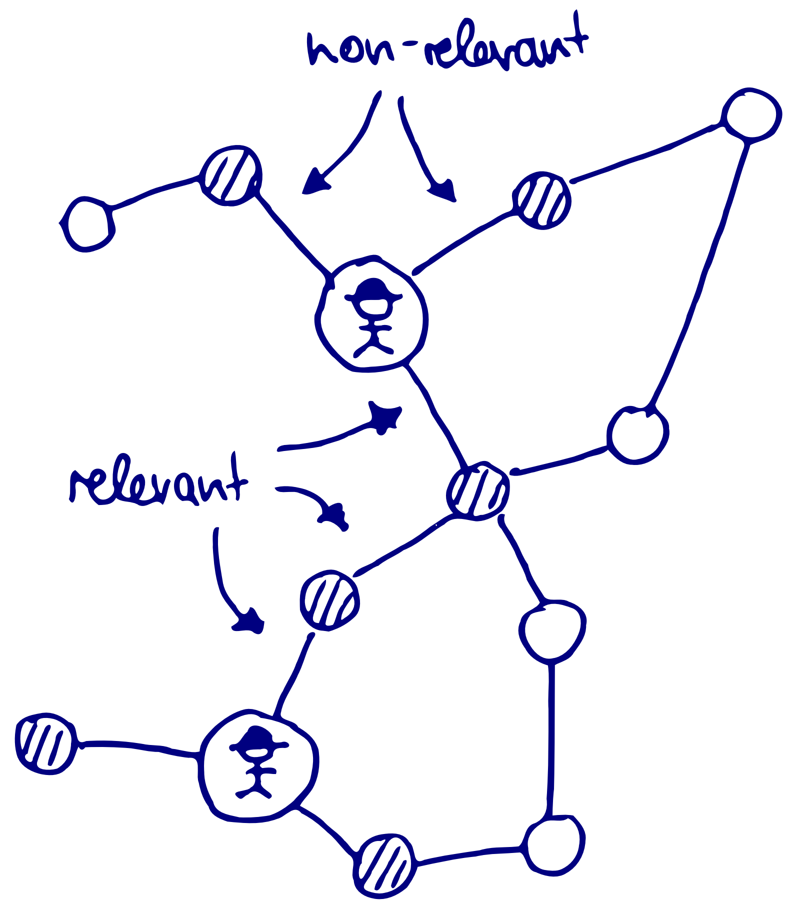


Big Question

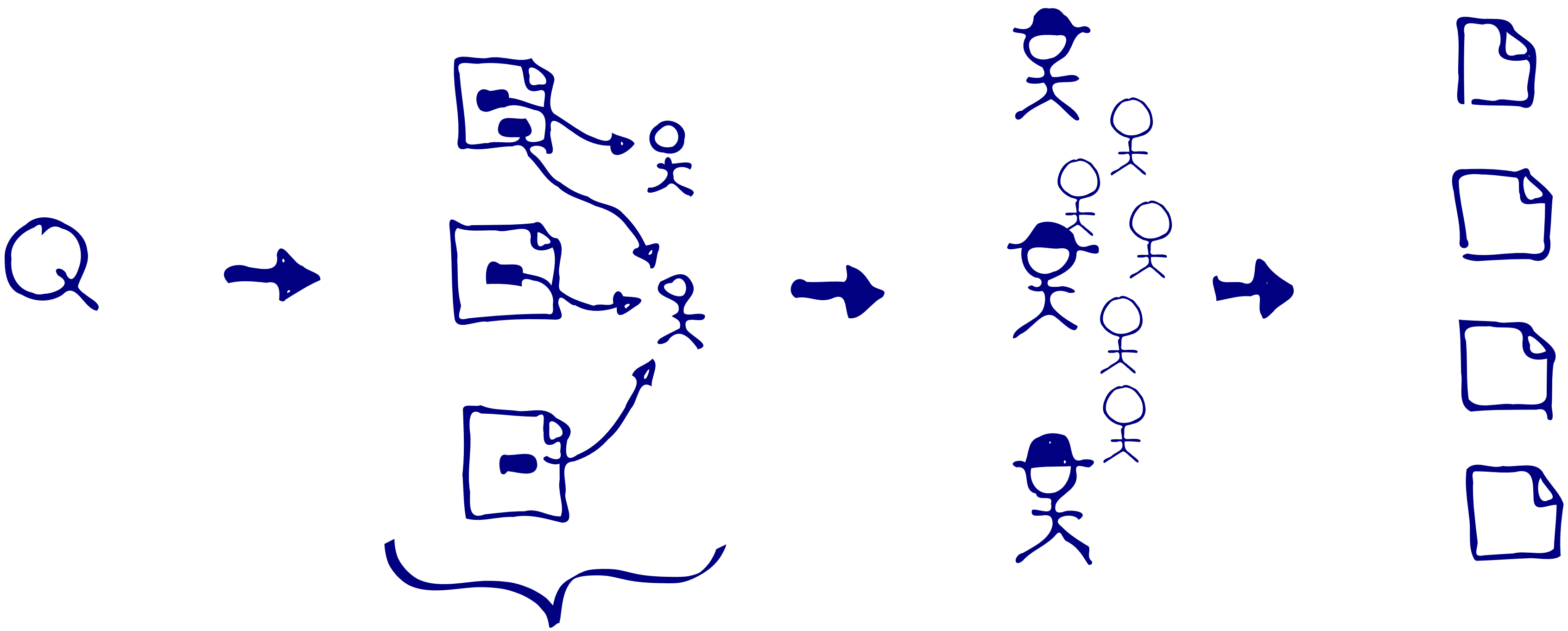
How to infer which other connected entities / nodes are relevant for the information need Q?

...and therefore safe for expansion?

Maybe entities in between relevant entities?



Source: Relevance Feedback with Entity Links



Pseudo-Relevance Feedback (RM3)

Document = bag of Entity Links (instead of terms)

[Dalton SIGIR14, Liu IRJ15]

Beyond the Graph Structure

Why only look at graph structure,
and ignore all the other kinds of information?

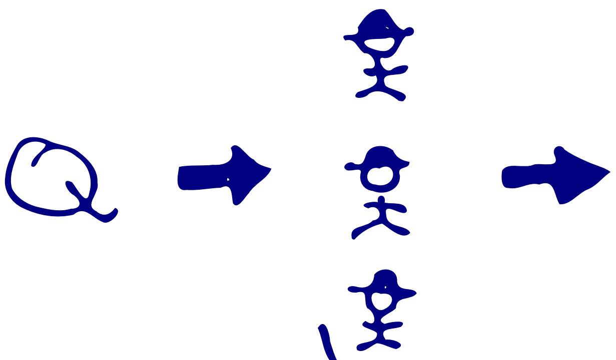
Typical approaches:

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: Entity Types (or Wikipedia Categories)

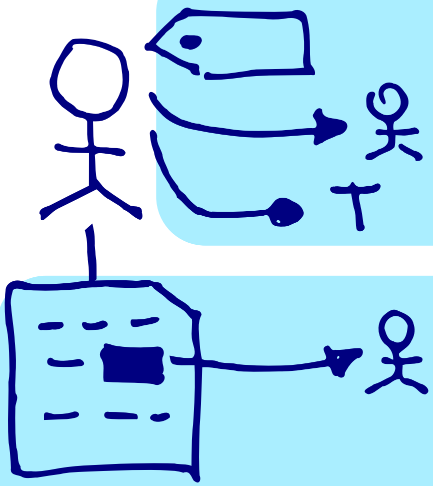


Which types are relevant ?

How to match types (or cat's) to documents ?

a) same-type entities [Kaptein CIKM10, Dalton SIGIR14]	majority types among entities	prefer docs with entities of this type
b) term classifier [Xiong CIKM15]	classify query terms with naive Bayes	classify documents with naive Bayes

Source: Object AND Article Content Retrieval



Entities as attribute-structured objects:
Object retrieval (see Part 4 [Hasibi ICTIR16])

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

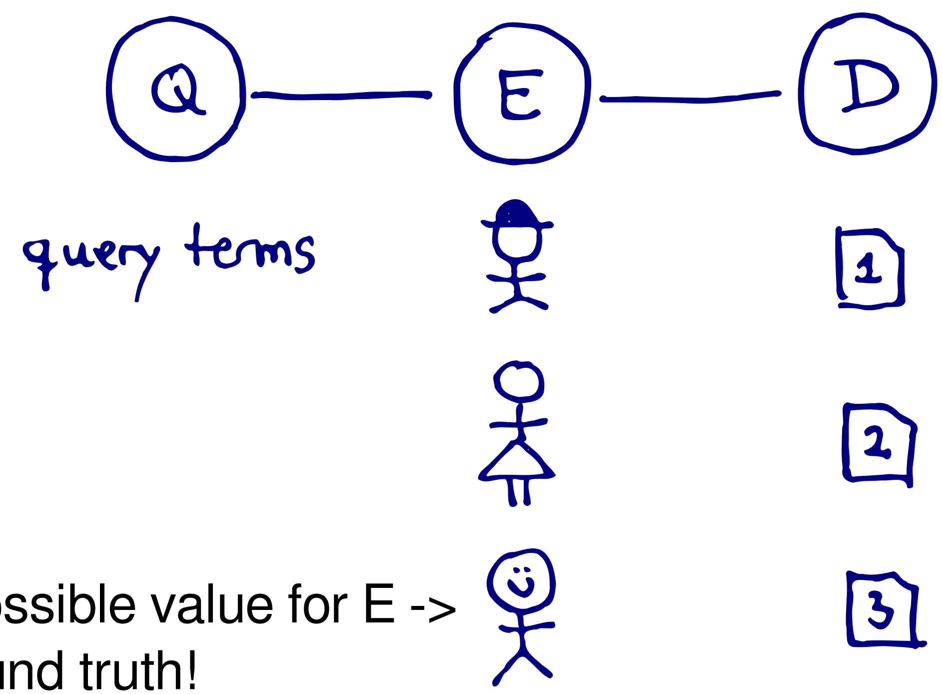
[Xiong CIKM15, Dalton SIGIR14]

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.

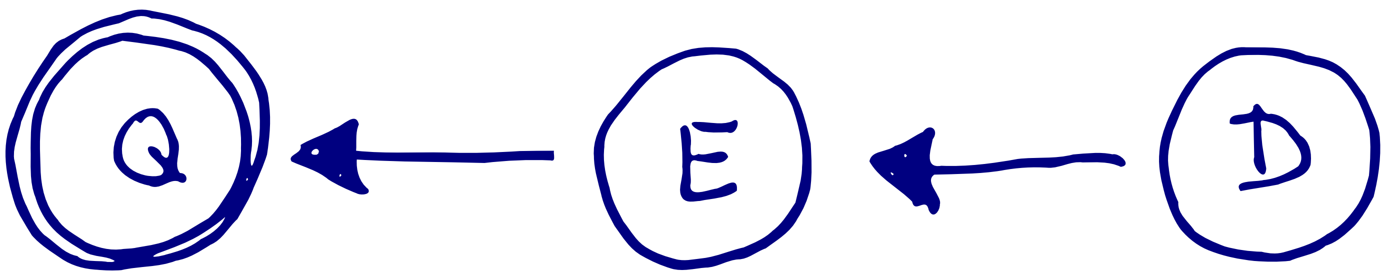
An edge represents a measure of compatability or similarity.



One possible value for E -> no ground truth!

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

Latent Entity Space [Liu IRJ15]



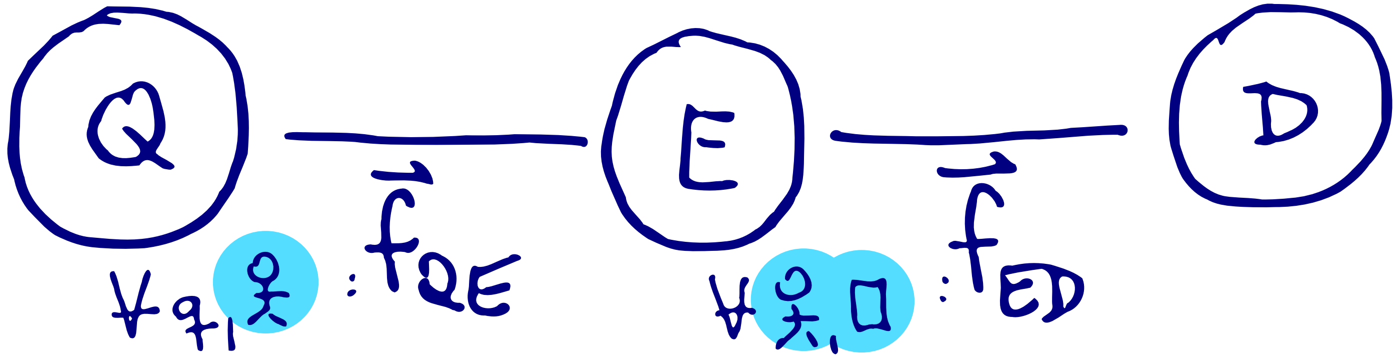
$$p(q|D=d,R=1)=\sum_{e\in\mathcal{E}} p(q|e)\cdot p(e|d)$$

similarity of LM(q) and LM(e)

similarity of LM(e) and LM(d)

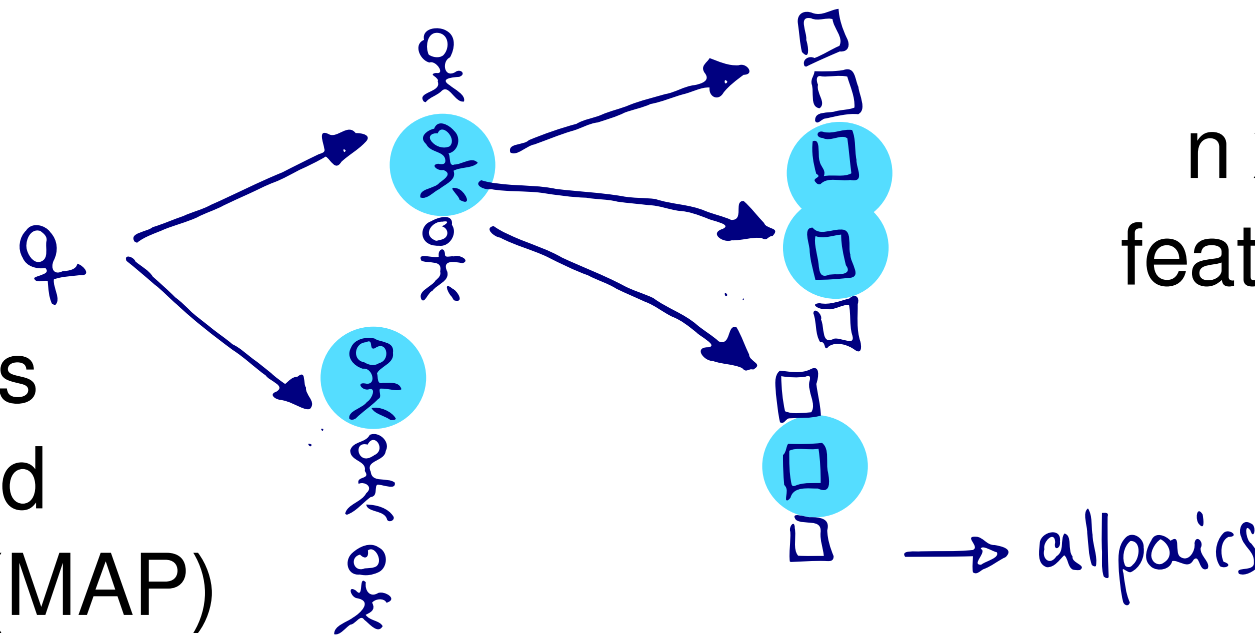
Wide range of experiments on which similarity measure / data source combination works best.

Entity Query Feature Expansion [Dalton SIGIR14]

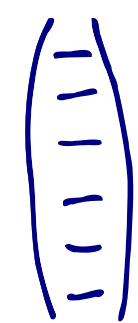


n different ways to compute p(q|e)

m different ways to compute p(e|d)



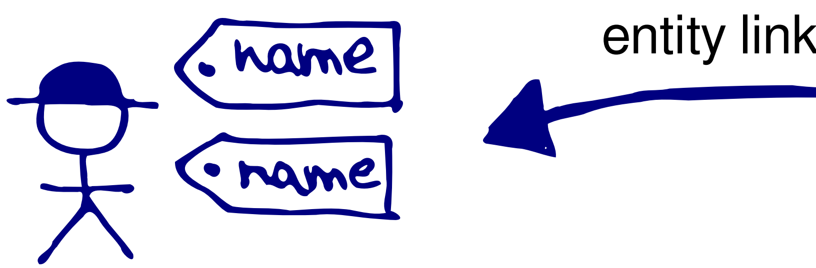
n x m features!

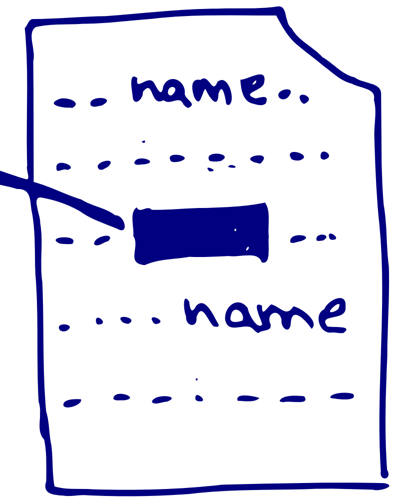


Combine features then use standard learning to rank (MAP)

Relation to Query / Latent Concept Expansion

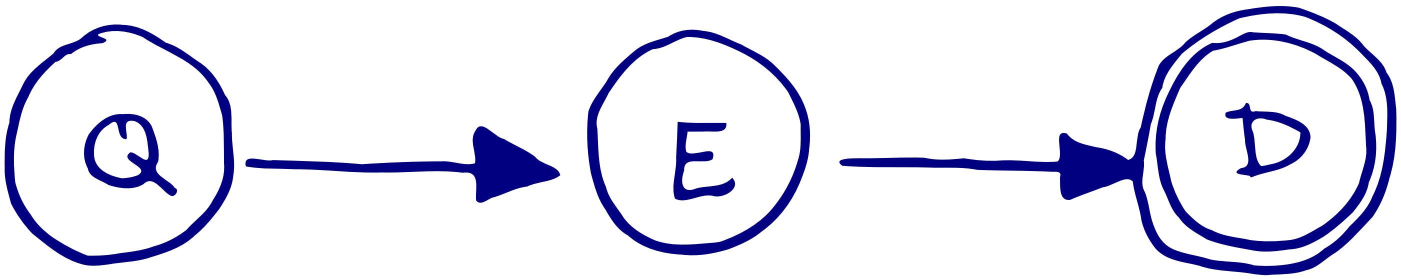
Various vocabularies, but all represented by sets





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

EsdRank [Xiong CIKM15]



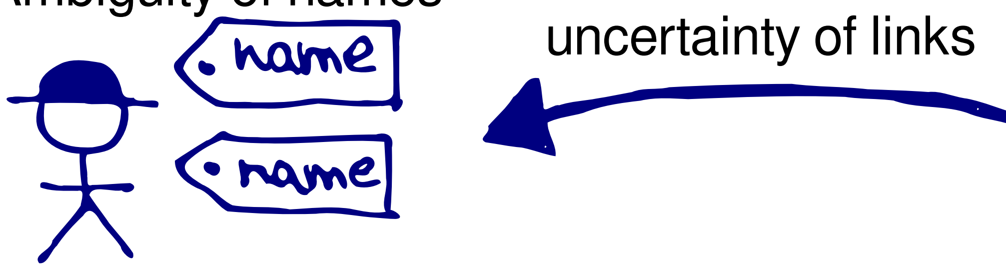
$$p(d_i|q) = \sum_{e \in \mathcal{E}} \underbrace{p(d_i|e)}_{\frac{1}{Z_1}\exp\left\langle \vec{w}_1, \vec{f}_{D,E} \right\rangle} \cdot \underbrace{p(e|q)}_{\frac{1}{Z_2}\exp\left\langle \vec{w}_2, \vec{f}_{E,Q} \right\rangle}$$

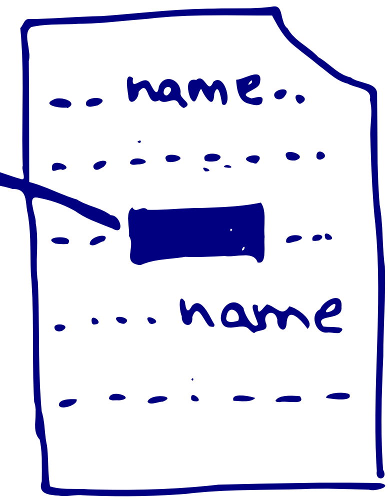
Discriminative probabilistic model based on
Generalized linear models + EM Algorithm
for learning weights w1, w2.

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

Query Expansion with Uncertainties

Taking uncertainty and confidences into account.





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

[Raviv SIGIR16]

Entity Aspects

An entity might be relevant, but:
only some aspects about might make it relevant
=> 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
- some movies were filmed in the UK

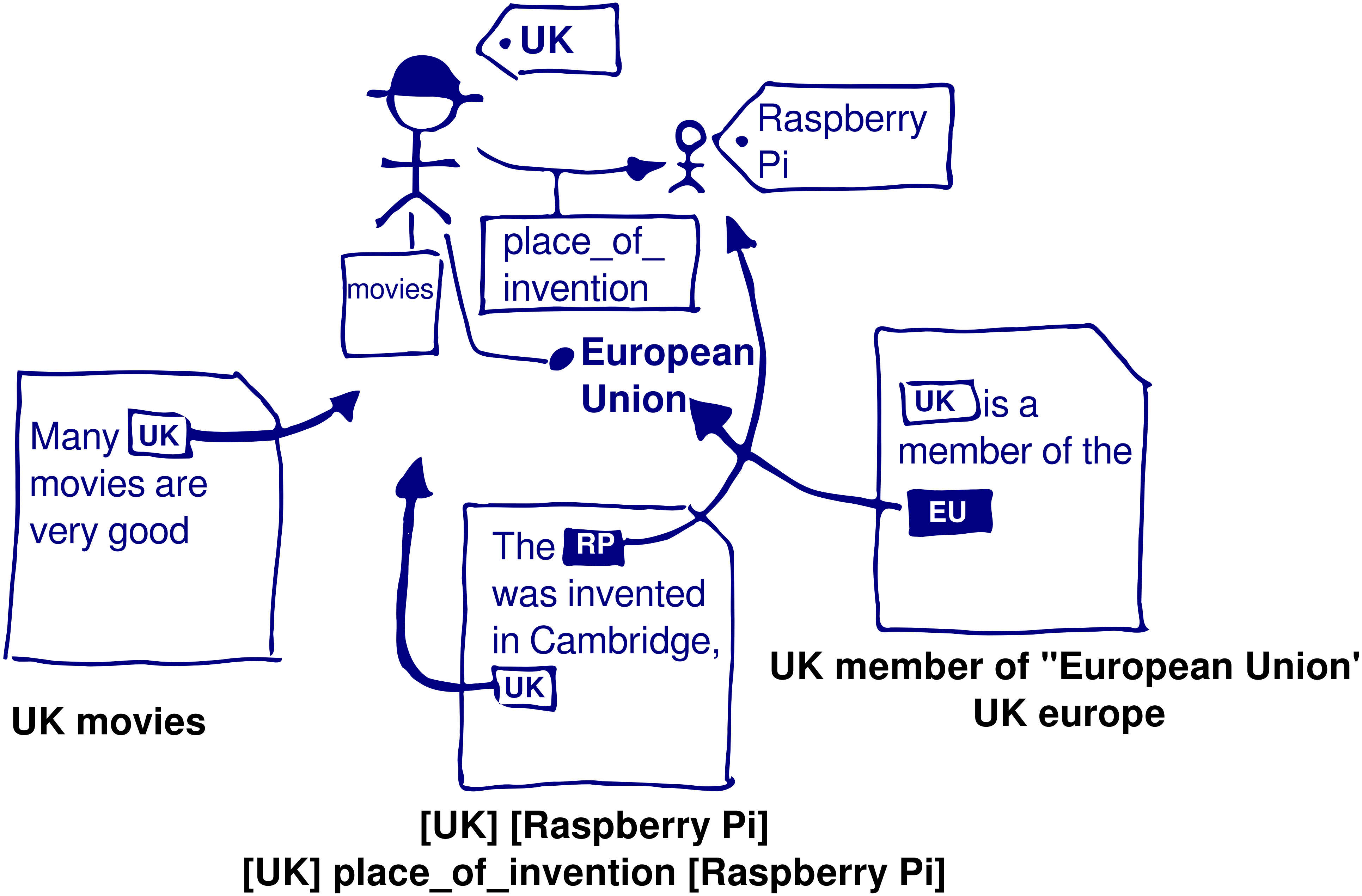
Depending on query, some are relevant, some not.

How to Represent Entity Aspects?

As terms?	UK movies
	brexit
As types?	UK member of "European Union"
As is-a?	UK as a European country
Related entities?	[UK] [Raspberry Pi]
Relations?	[UK] place_of_invention [Raspberry Pi]
Language Model	p(brexit)=0.4
	p(leave)=0.25
	p(immigration)=0.10

[Reinanda SIGIR15, Liu IRJ15, Prasojo CIKM15]

Entity Aspects: Using KG and Text



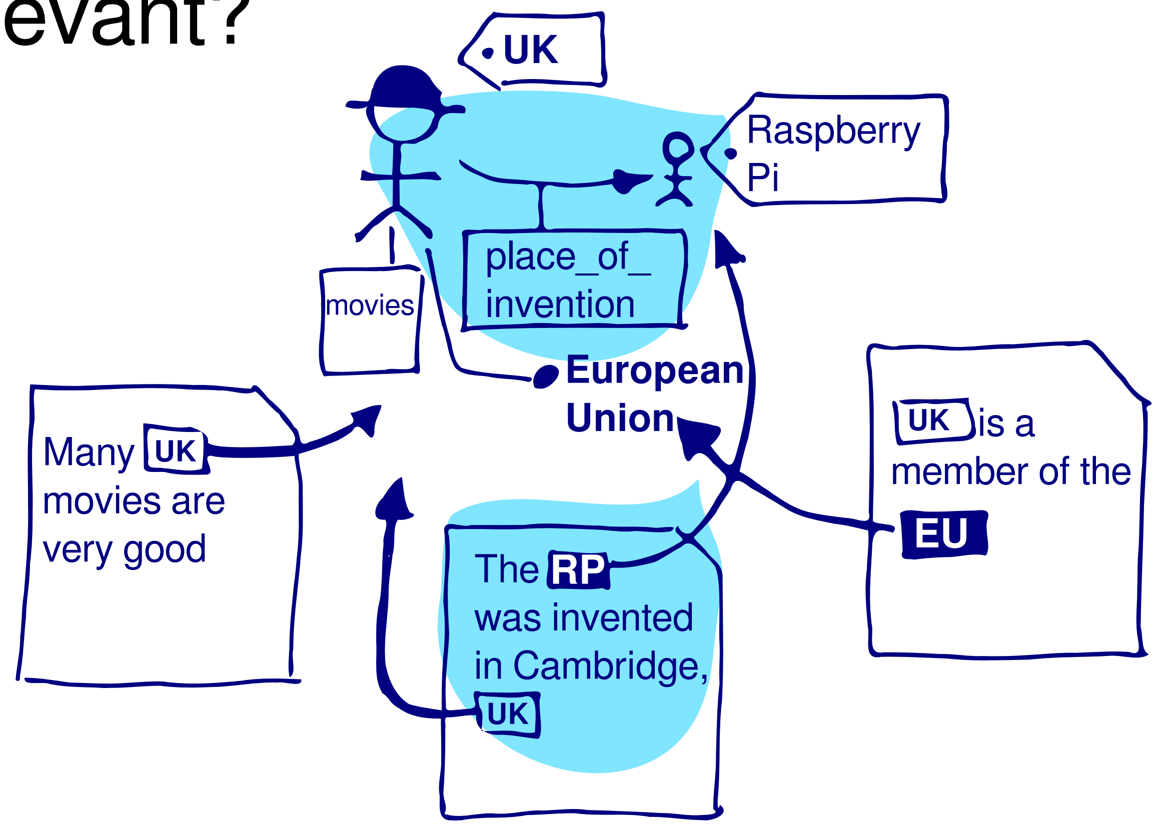
Entity Aspects: Infer Relevance, Match, Extract

1) Relevance:
Which aspects are relevant?

2) Match:
How to match in text?

pseudo relevance feedback

inverse tasks

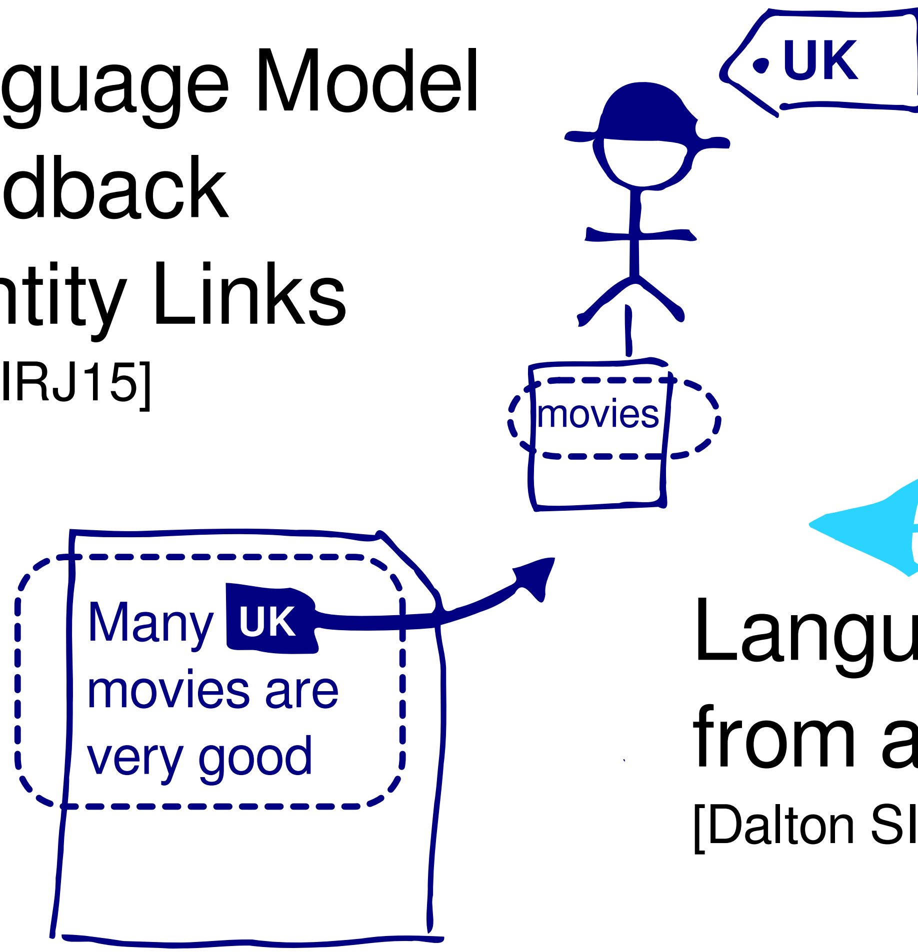


3) Extract:
How to extract new aspects? (KB population)

Entity Aspects as Terms

Passage-Language Model
- Pseudo Feedback
- Surround Entity Links

[Dalton SIGIR14, Liu IRJ15]



UK movies

Language model from article / descr.

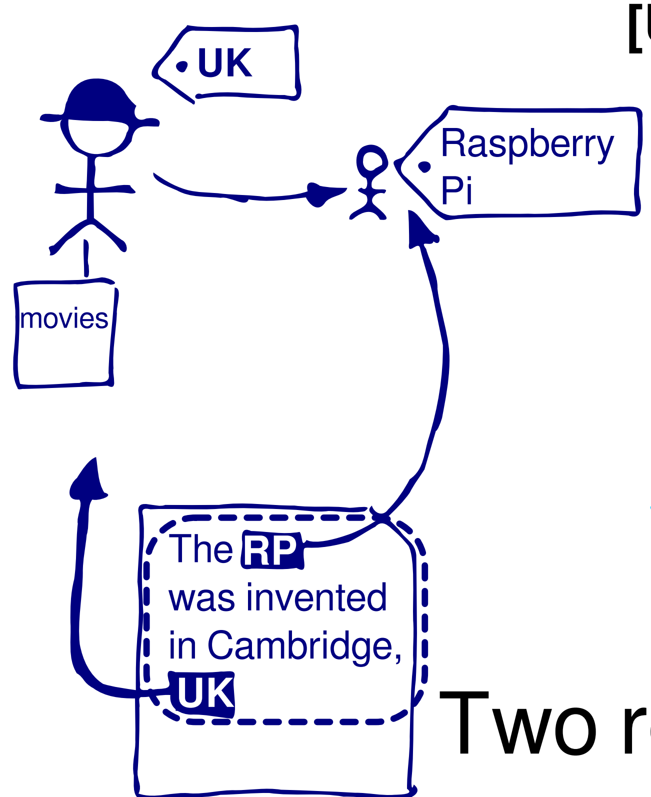
[Dalton SIGIR14, Liu IRJ15]

Extract/Infer relevant Entity Aspects?

- From collocations in pseudo-relevant documents
- From passages surrounding entity links
- Through graph analysis
- What is this frequent among other relevant entities
- Extracting a language model

Entity Aspects through Co-mentioned Entities

[UK] [Raspberry Pi]



Passage with
- link to entity
- matching query terms
=> other enties relevant?

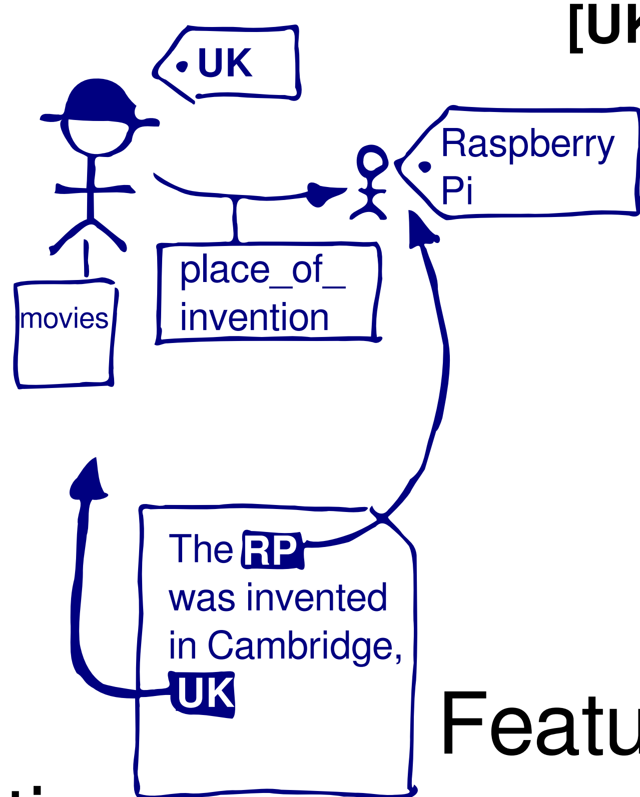
Infer & Extract Aspects

Two relevant Entities which are linked in KG
=> Promote documents that mention both

Match Aspects

Entity Aspects through Relations (Triples)

[UK] place_of_invention [Raspberry Pi]



Relation Extraction:
- Supervised Extraction from Text

[Schuhmacher ECIR16]

Infer & Extract Aspects

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

[Voskarides ACL15]

Match Aspects

Summary (Part 5)

- Query -> Entities -> Documents
- Knowledge graph expansion
- Un-/structured sources of entities:
  Entity Links, Attributes, Article, Type classifier
- Machine learning
- Entity Aspects: Infer relevance, match & extract

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

Outook: Moving Beyond Aggregation of Features

Can we refine the features through a deeper integration of different sources?

Examples:
- Use context of entity links to extract term-models
- Language models from types and link context
- Use terms to find relevantly connected entities
- Factoring in uncertainty from extraction tools