DAT630
Entity Retrieval II.

02/11/2016

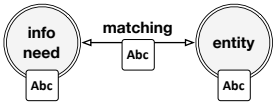
Krisztian Balog | University of Stavanger

Recap

- Entities are meaningful units for organizing information
  - Used for enriching in search engine results
- Knowledge bases store massive amounts of information about entities as RDF triples
- Entities can be represented as documents for retrieval
  - Using document fields can preserve (some of) the underlying structure

So far…

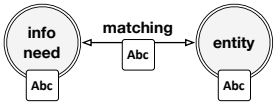
- Term-based retrieval models
- Robust and effective, but ignore *semantics*
  - entity-specific properties (types, relationships, etc.)



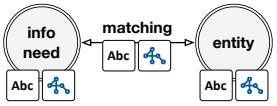
Text-only representation

Incorporating semantics

- working definition:
  semantics = *references to meaningful structures*
- How to capture, represent, and use structure?
  - It concerns all components of the retrieval process!

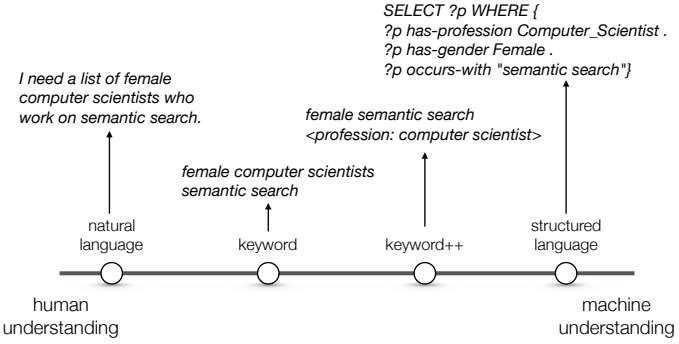


Text-only representation

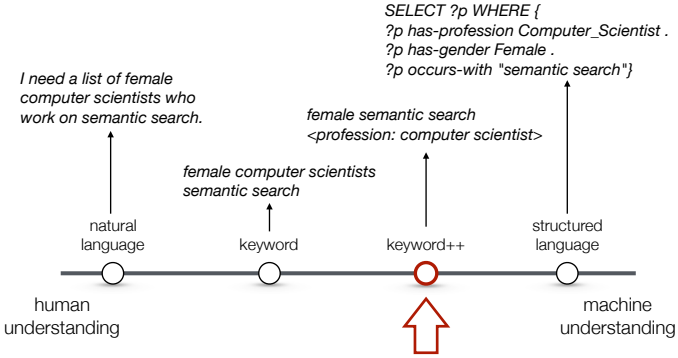


Text+structure representation

Spectrum of queries

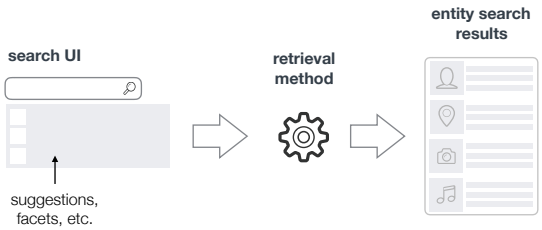


Spectrum of queries

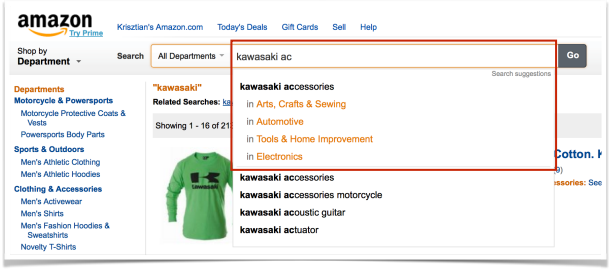


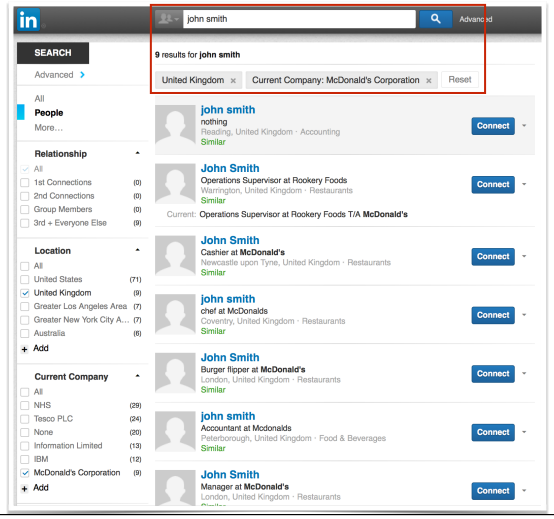
Scenario #1

- User provides keyword++ query



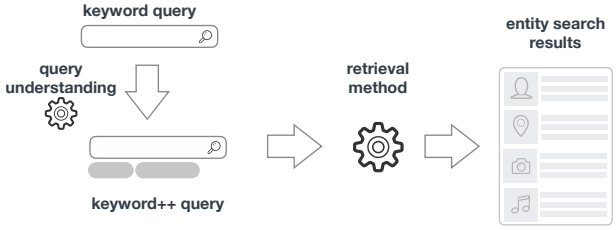
Example keyword++ queries





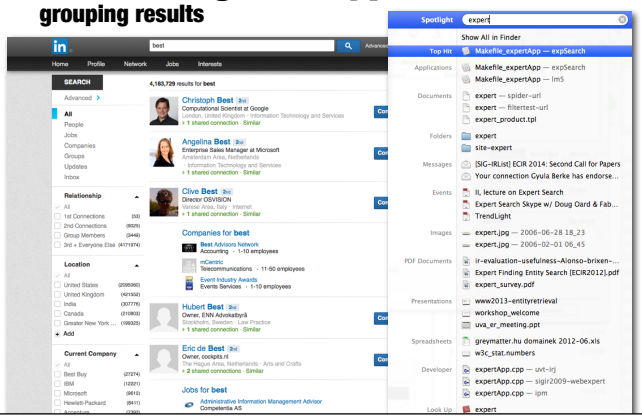
Scenario #2

- *Query understanding* component constructs the keyword++ query (automatically)

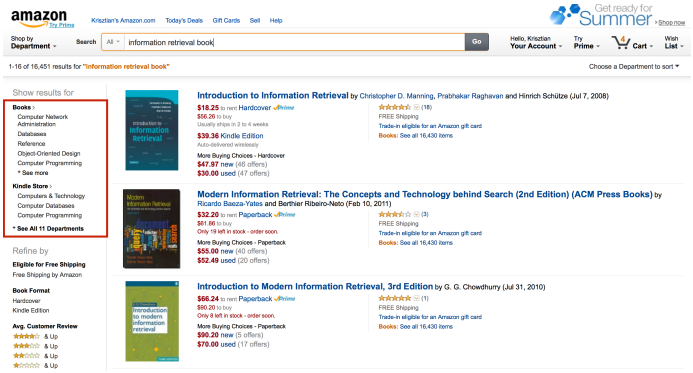


Entity Types

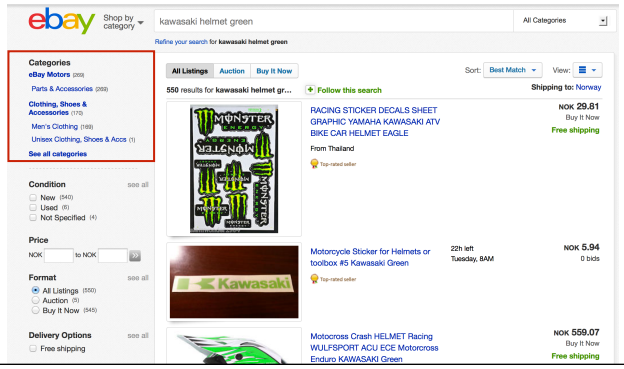
Interacting with types
grouping results



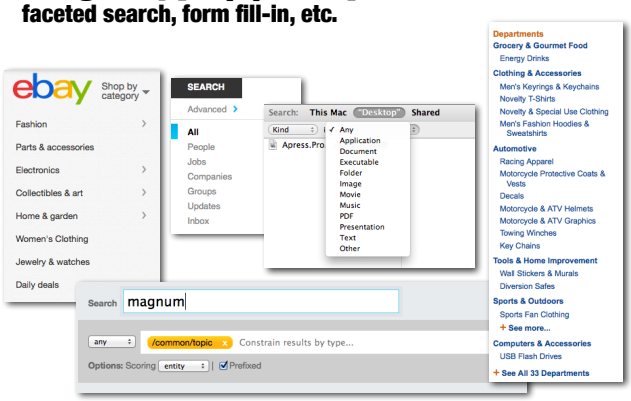
Interacting with types
filtering results



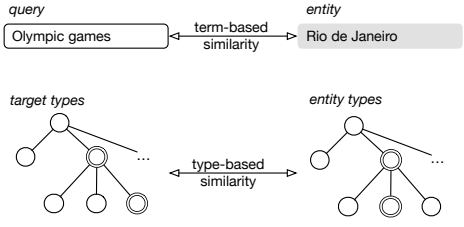
Interacting with types
filtering results



Target type(s) are provided
faceted search, form fill-in, etc.

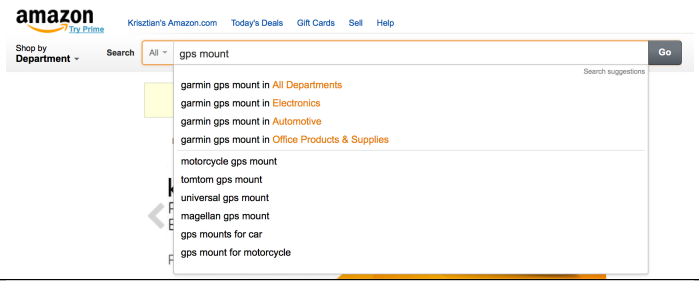


Type-aware ranking



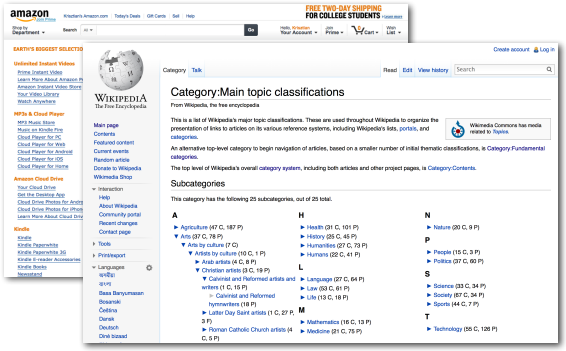
Challenges

- Users are not familiar with the type system



Very many types…

which are typically hierarchically organized



Sense of scale

Table 2: Overview of type taxonomies and their statistics.

Type system	DBpedia	Freebase	Wikipedia categories	YAGO
#types	591	1719	553,571	568,672
#top-level types	58	92	27	61
#leaf-level types	472	1626	392,257	549,775
height	7	2	34	19
#assigning types	390	1626	479,344	314,651
#entities w/ type	3.24M	3.27M	3.61M	2.88M
avg #types/entity	2.79	4.4	28.2	12.2
mode depth	2	2	20	4

In general, categorizing things can be hard

- What is *King Arthur*?
  - Person / Royalty / British royalty
  - Person / Military person
  - Person / Fictional character

DBpedia

Which King Arthur?!

Considerations for type-aware ranking

- Need to be able to handle the imperfections of the type system
  - Inconsistencies
  - Missing assignments
  - Granularity issues
    - Entities labeled with too general or too specific types
- User input is to be treated as a hint, not as a strict filter

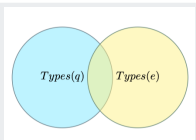
Type-aware retrieval #1

- **Strict filtering** model

$$P(q|e) = \underbrace{P(q_w|e)}_{\text{Term-based similarity}} \cdot \underbrace{\chi[types(q) \cap types(e) \neq \emptyset]}_{\text{Type-based similarity}}$$

w stands for word

1 if the query and entity have some types in common, otherwise **0**



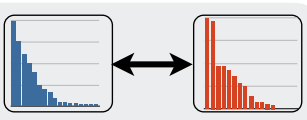
Type-aware retrieval #2

- **Soft filtering** model

$$P(q|e) = \underbrace{P(q_w|e)}_{\text{Term-based similarity}} \cdot \underbrace{P(q_t|e)}_{\text{Type-based similarity}}$$

w stands for word

Compare the type distribution of in the query against that of the entity



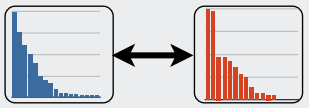
Type-aware retrieval #3

- **Interpolation** model

$$P(q|e) = (1-\lambda)\underbrace{P(q_w|e)}_{\text{Term-based similarity}} + \lambda\underbrace{P(q_t|e)}_{\text{Type-based similarity}}$$

w stands for word

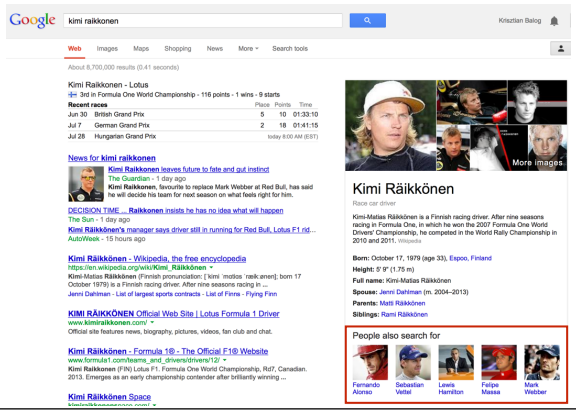
Compare the type distribution of in the query against that of the entity

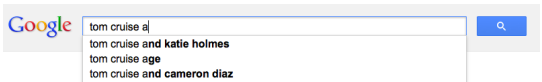


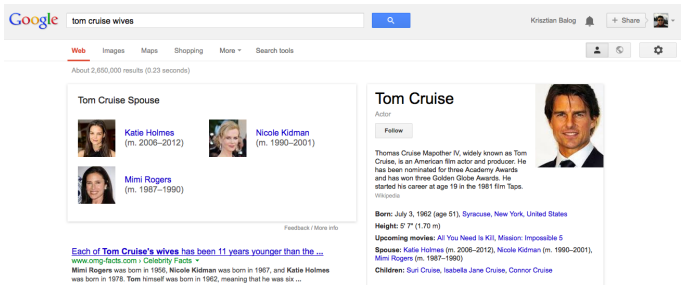
Query types | Entity types

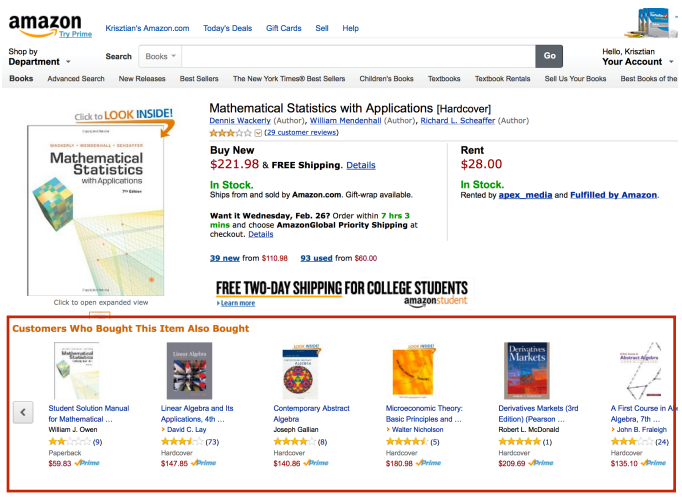
Entity Relationships

Related entities









Searching for arbitrary relations*

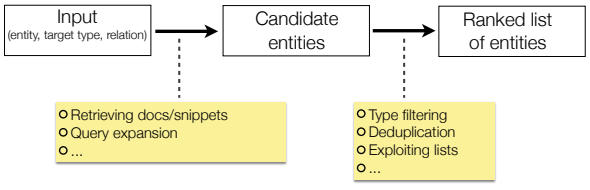
*given an input entity and target type

airlines that currently use Boeing 747 planes
ORG Boeing 747

Members of The Beaux Arts Trio
PER The Beaux Arts Trio

What countries does Eurail operate in?
LOC Eurail

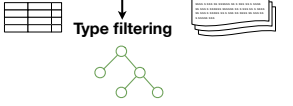
A typical pipeline



Modeling related entity finding

- Ranking entities of a given type (T) that stand in a required relation (R) with an input entity (E)
- Three-component model

$$p(e|E,T,R) \propto \underbrace{p(e|E)}_{\text{Co-occurrence model}} \cdot \underbrace{p(T|e)}_{\text{Type filtering}} \cdot \underbrace{p(R|E,e)}_{\text{Context model}}$$



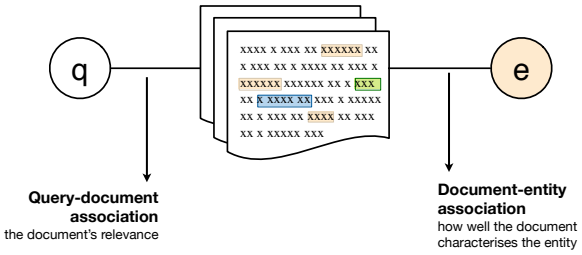
Ranking Entities without Ready-made Descriptions

Scenario

- Entity descriptions are not readily available
- Entity occurrences are annotated
  - manually
  - automatically (i.e., entity linking)

The basic idea

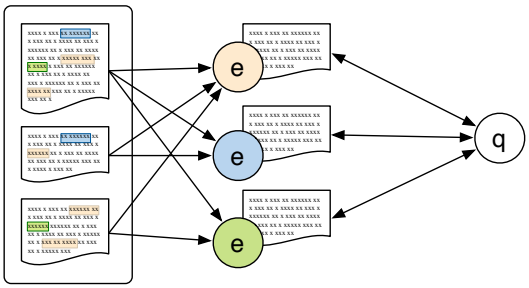
Use documents to go from queries to entities



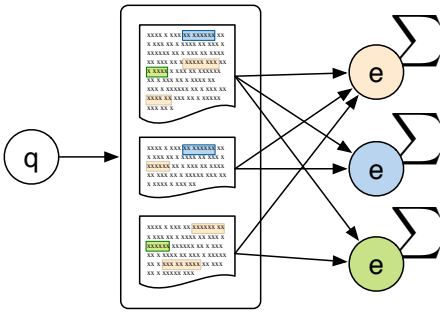
Two principal approaches

- **Profile-based** methods
  - Create a textual profile for entities, then rank them (by adapting document retrieval techniques)
- **Document-based** methods
  - Indirect representation based on mentions identified in documents
  - First ranking documents (or snippets) and then aggregating evidence for associated entities

Profile-based methods



Document-based methods



Many possibilities in terms of modeling

- Generative (probabilistic) models
- Discriminative (probabilistic) models
- Voting models
- Graph-based models

Candidate models ("Model 1")

$$P(q|\theta_e) = \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Smoothing}}^{n(t,q)}$$

Smoothing
With collection-wide background model

$$(1-\lambda)\underbrace{P(t|e)}_{\sum_d P(t|d,e)P(d|e)} + \lambda P(t)$$

$$\sum_d \underbrace{P(t|d,e)}_{\text{Term-candidate co-occurrence}} \underbrace{P(d|e)}_{\text{Document-entity association}}$$

Term-candidate co-occurrence
In a particular document.
In the simplest case: $P(t|d)$

Document-entity association

Document models ("Model 2")

$$P(q|e) = \sum_d \underbrace{P(q|d,e)}_{\text{Document relevance}} \underbrace{P(d|e)}_{\text{Document-entity association}}$$

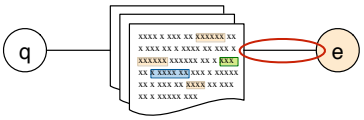
Document relevance
How well document d supports the claim that e is relevant to q

Document-entity association

$$\prod_{t \in q} \underbrace{P(t|d,e)}_{\text{Simplifying assumption (t and e are conditionally independent given d)}}^{n(t,q)}$$

$$P(t|\theta_d)$$

Document-entity associations



- Boolean (or set-based) approach
- Weighted by the confidence in entity linking
- Consider other entities mentioned in the document