

DAT630

Entity Retrieval II.

02/11/2016

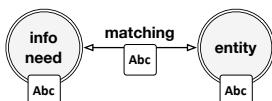
Krisztian Balog | University of Stavanger

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...

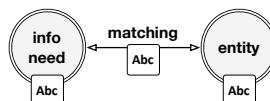
- Term-based retrieval models
- Robust and effective, but ignore *semantics*
 - entity-specific properties (types, relationships, etc.)



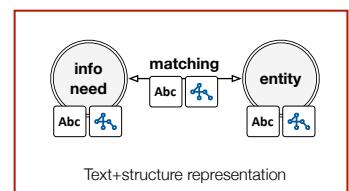
Text-only representation

Incorporating semantics

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

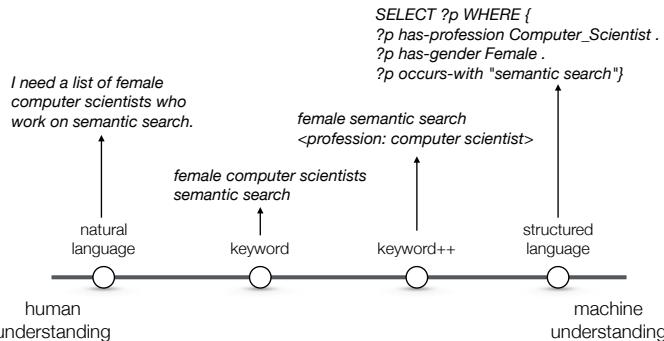


Text-only representation

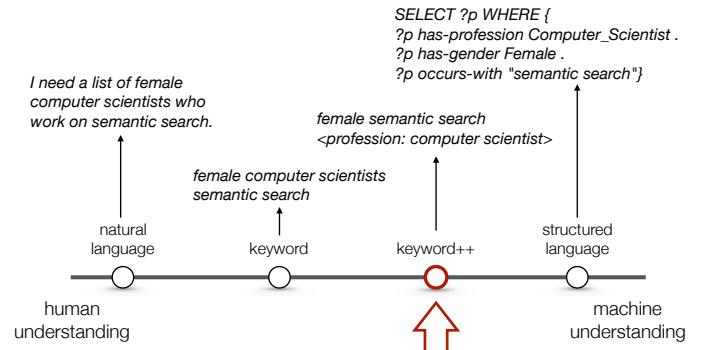


Text+structure representation

Spectrum of queries

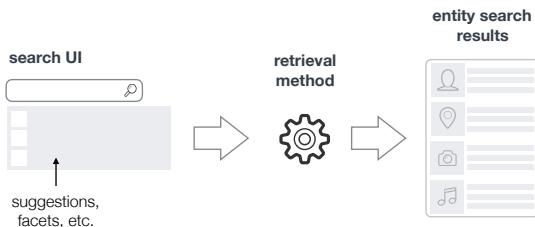


Spectrum of queries

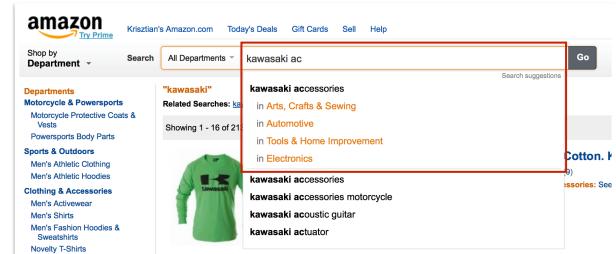


Scenario #1

- User provides keyword++ query



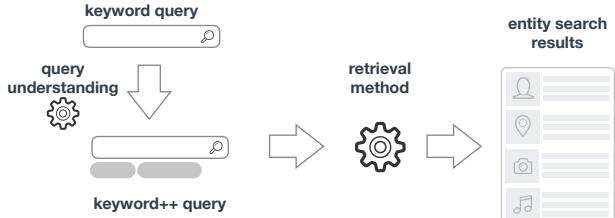
Example keyword++ queries



This screenshot shows the LinkedIn search interface for the query "john smith". It displays 9 results for the user. The results include profiles for "John Smith" from different companies and locations. On the left, there are filters for "Relationship", "Location", and "Current Company". A red box highlights the search bar and the results count.

Scenario #2

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



Entity Types

Interacting with types grouping results

This screenshot shows the LinkedIn search results for the query "best". The results are grouped by entity type, such as "Companies", "People", and "Jobs". Each group contains relevant profiles or job posts. A red box highlights the search bar and the results count.

Interacting with types filtering results

This screenshot shows the Amazon search results for "information retrieval book". The results are filtered by book type, showing various editions and formats. A red box highlights the search bar and the results count.

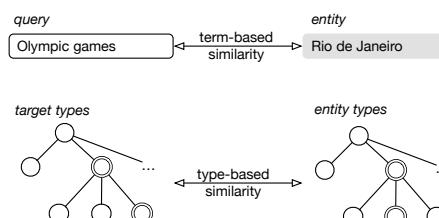
Interacting with types filtering results

This screenshot shows the eBay search results for "kawasaki helmet green". The results are filtered by item type, such as "Categories" (e.g., Parts & Accessories, Clothing, Shoes & Accs), "Condition" (e.g., New, Used), "Price" (e.g., NOK 29.81), "Format" (e.g., All Listings, Auction, Buy It Now), and "Delivery Options" (e.g., Free shipping). A red box highlights the search bar and the results count.

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

This screenshot shows the eBay search interface. The search bar contains "magnum". The "SEARCH" dropdown shows "Advanced" selected. The "Type" facet is expanded, showing options like "Kind", "Any", "Apress.Pro", "Application", "Document", "Executable", "Folder", "Image", "Movie", "Music", "PDF", "Presentation", "Text", and "Other". A red box highlights the search bar and the expanded type facet.

Type-aware ranking



Challenges

- Users are not familiar with the type system

The screenshot shows the Amazon search bar with 'gps mount' entered. Below the search bar, a dropdown menu lists various categories where 'gps mount' can be found, including All Departments, Electronics, Automotive, Office Products & Supplies, and several motorcycle and car-related categories. To the right of the search bar, there's a sidebar with navigation links like 'Shop by Department', 'Search', and 'All - gps mount'. The main search results page is visible below the dropdown.

Very many types... which are typically hierarchically organized

The screenshot shows the Wikipedia category tree interface. At the top, it says 'EARTH'S BIGGEST SELECTED'. Below that is a search bar with 'WIKIPEDIA'. The main content area shows the 'Category:Main topic classifications' page, which lists various categories like Agriculture, Arts, and History. To the right, there's a sidebar with navigation links for 'Category', 'Talk', 'Read', 'Edit', 'View history', and 'Search'. The bottom of the page shows a detailed hierarchical tree structure with counts for each category level.

Sense of scale

Table 2: Overview of type taxonomies and their statistics.

Type system	DBpedia	Freebase	Wikidata 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



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

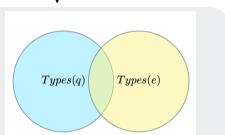
Type-aware retrieval #1

- Strict filtering model

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

w stands for word

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



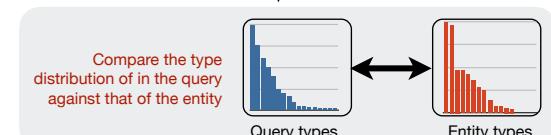
Type-aware retrieval #2

- Soft filtering model

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

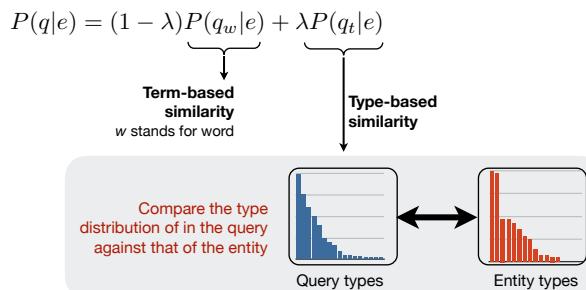
w stands for word

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



Type-aware retrieval #3

- Interpolation model



Entity Relationships

Related entities

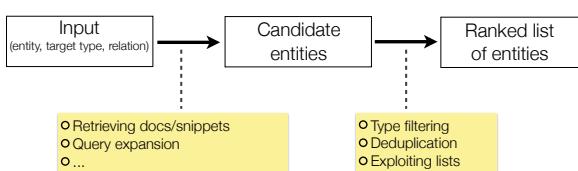
Google search results for "kimi räikkönen":

- Recent races: Shows race results from June 30 to July 7.
- News for Kimi Räikkönen: Includes links to news articles from The Guardian, The Sun, and The Daily Mail.
- Kim Räikkönen - Wikipedia, the free encyclopedia: Provides general information about Kimi Räikkönen, including his racing career and personal life.
- Kim Räikkönen Official Web Site | Lotus Formula 1 Driver: The official website of Kimi Räikkönen, featuring news, biography, pictures, fan club, and contact information.
- Tom Cruise Spouse: Google search results for "tom cruise wives".

Amazon product page for "Mathematical Statistics with Applications":

- Buy New: \$221.98 & FREE Shipping. Details
- Rent: \$28.00 & In Stock. Rented by apex_media and Fulfilled by Amazon.
- Customer reviews: 4.5 stars (23 customer reviews)
- Customers Who Bought This Item Also Bought: A list of related books including "Student Solution Manual for Mathematical Statistics with Applications" and "Linear Algebra and Its Applications".

A typical pipeline



Google search results for "tom cruise a":

- Recent races: Shows race results from June 30 to July 7.
- News for Kimi Räikkönen: Includes links to news articles from The Guardian, The Sun, and The Daily Mail.
- Kim Räikkönen - Wikipedia, the free encyclopedia: Provides general information about Kimi Räikkönen, including his racing career and personal life.
- Kim Räikkönen Official Web Site | Lotus Formula 1 Driver: The official website of Kimi Räikkönen, featuring news, biography, pictures, fan club, and contact information.
- Tom Cruise Spouse: Google search results for "tom cruise wives".

Google search results for "tom cruise wives":

- Recent races: Shows race results from June 30 to July 7.
- News for Kimi Räikkönen: Includes links to news articles from The Guardian, The Sun, and The Daily Mail.
- Kim Räikkönen - Wikipedia, the free encyclopedia: Provides general information about Kimi Räikkönen, including his racing career and personal life.
- Kim Räikkönen Official Web Site | Lotus Formula 1 Driver: The official website of Kimi Räikkönen, featuring news, biography, pictures, fan club, and contact information.
- Tom Cruise Spouse: Google search results for "tom cruise wives".

Searching for arbitrary relations*

*given an input entity and target type

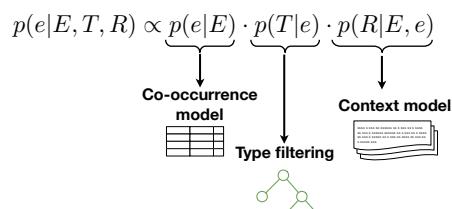
Q airlines that currently use Boeing 747 planes
ORG Boeing 747

Q Members of The Beaux Arts Trio
PER The Beaux Arts Trio

Q What countries does Eurail operate in?
LOC Eurail

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



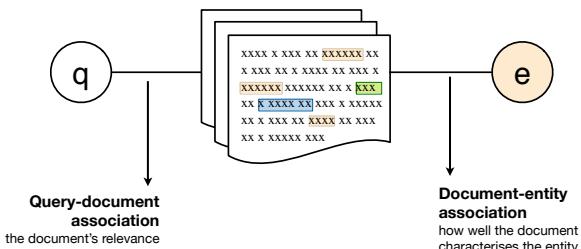
Ranking Entities without Ready-made Descriptions

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

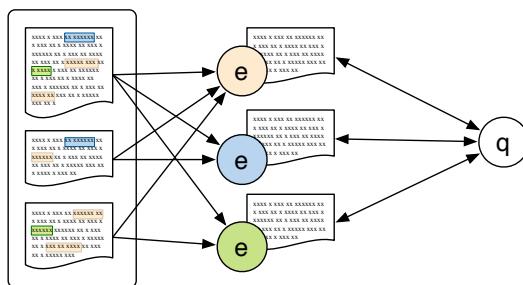
Profile-based methods

- Create a textual profile for entities, then rank them (by adapting document retrieval techniques)

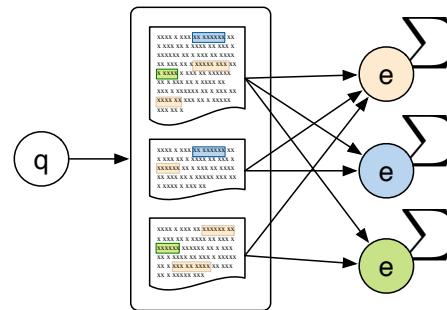
Document-based methods

- 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”)

$$\begin{aligned}
 P(q|\theta_e) &= \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Smoothing}}^{n(t,q)} \\
 &\quad \downarrow \text{With collection-wide background model} \\
 &= (1 - \lambda)P(t|e) + \lambda P(t) \\
 &\quad \downarrow \\
 &= \sum_d \underbrace{P(t|d, e)}_{\text{Term-candidate co-occurrence}} \underbrace{P(d|e)}_{\text{Document-entity association}} \\
 &\quad \downarrow \text{In a particular document.} \\
 &\quad \downarrow \text{In the simplest case: } P(t|d)
 \end{aligned}$$

Document models (“Model 2”)

$$P(q|e) = \sum_d P(q|d, e)P(d|e)$$

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

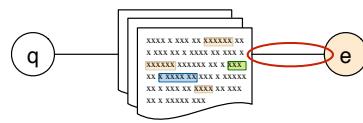
Document-entity association

$$\prod_{t \in q} P(t|d, e)^{n(t, q)}$$

Simplifying assumption (t and e are conditionally independent given d)

$$P(t|\theta_d)$$

Document-entity associations



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