

Facet Annotation Using Reference Knowledge Bases

Riccardo Porrini

Mia-platform
(7Pixel Srl)

(University of Milano-Bicocca)

Matteo Palmonari

Department of Information,
Systems and Communication
University of Milano-Bicocca

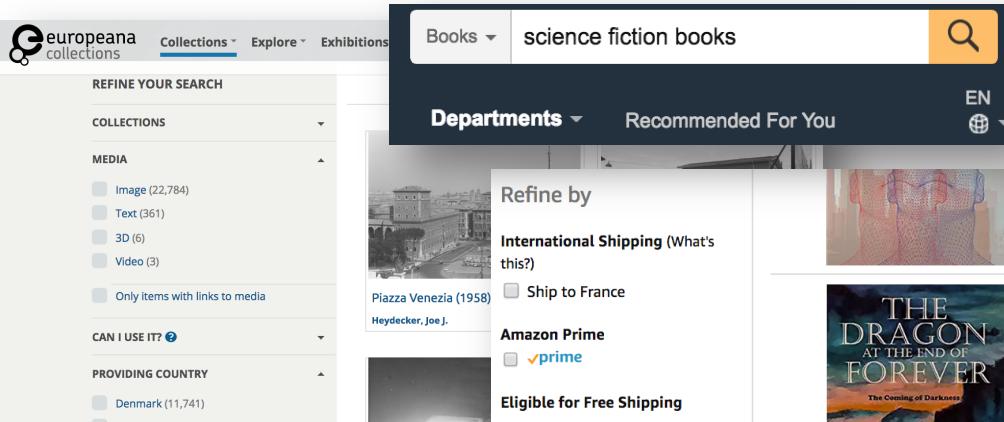
Isabel F. Cruz

Department of Computer Science
University of Illinois at Chicago



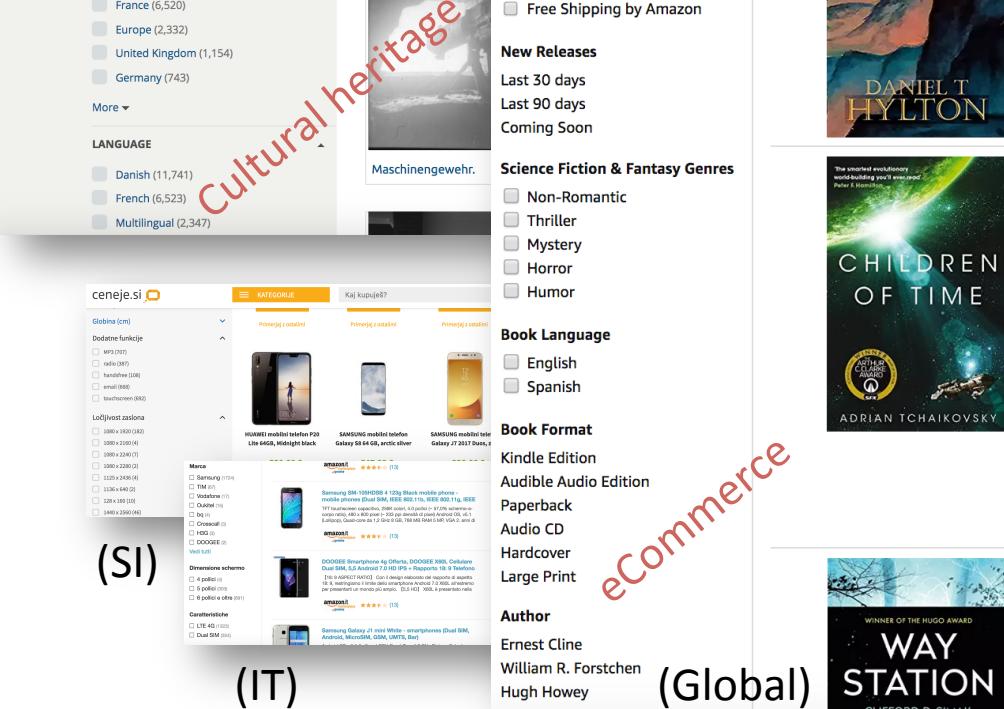
LYON, FRANCE
23 - 27 April 2018

Using and Building Faceted Interfaces



Cultural heritage

The screenshot shows the Europeana collections search interface. It includes a search bar with "science fiction books", a navigation bar with "Books", "Departments", and "Recommended For You", and a sidebar with "REFINE YOUR SEARCH" sections for Collections (Image, Text, 3D, Video), Media (Only items with links to media), Can I use it? (Denmark, France, Europe, United Kingdom, Germany), Providing Country (Denmark, France, Europe, United Kingdom, Germany), and Language (Danish, French, Multilingual). The main content area displays search results for "science fiction books", including images of book covers like "Piazza Venezia (1958)" by Heydecker, Joe J. and "Maschinengewehr.".



eCommerce

The screenshot shows the ceneje.si mobile phone search interface. It includes a search bar with "KATEGORIJE" (Categories) and "Kaj kupujete?" (What are you buying?). The categories section lists "Globus (cm)", "Dodatak funkcije" (functions), "Ločljivost zaslona" (display resolution), "Merce" (merchandise), "Dopravne sheme" (shipping schemes), and "Carakteristike" (characteristics). The price range section shows filters for "Primerjajoči cenovnik" (Comparable price) and "Primejšči cenovnik" (Higher price). The merchandise section shows images of phones like "HUAWEI mobilni telefon P30 Lite 64GB, Midnight black". The characteristics section shows filters for screen size (e.g., 1000 x 2100 (12)), screen resolution (e.g., 1080 x 2160 (4)), and camera (e.g., 1000 x 2000 (2)). The bottom part of the interface shows search results for specific phones like "Samsung Galaxy S10e 64GB, Arctic Silver" and "DOOGEE Y9 Pro".

(SI)

(IT)

(Global)

Faceted interfaces

- Widespread in several domains
- Enhance user experience
 - Competitive advantage for web applications, e.g., in eCommerce

Using and Building Faceted Interfaces

(SI) Europeana collections

The screenshot shows the Europeana collections search interface. It includes a sidebar for refining your search by media type (Image, Text, 3D, Video), providing country, and language. The main search bar contains "science fiction books". Below the search bar are sections for "Departments" and "Recommended For You". A red diagonal watermark labeled "Cultural heritage" is overlaid across the entire interface.

(IT) Mobile Phone Shopping

The screenshot shows a mobile phone shopping interface from ceneje.si. It displays a grid of phones and filters for screen size, weight, and price. A red diagonal watermark labeled "eCommerce" is overlaid across the interface.

(Global) Book Search

The screenshot shows a global book search interface. It features a search bar for "science fiction books", a sidebar for refining by shipping, eligible for free shipping, new releases, and genres (Non-Romantic, Thriller, Mystery, Horror, Humor). The results page shows book covers for "The Dragon at the End of Forever" by Daniel T. Hylton and "Children of Time" by Adrian Tchaikovsky. A blue arrow points from the "Book Language" section to the book covers, indicating how facets like language can be applied to search results.

Faceted interfaces

- Widespread in several domains
- Enhance user experience
 - Competitive advantage for web applications, e.g., in eCommerce

A facet

- Domain, property, values
 - $\langle D, p, \{v_1, \dots, v_n\} \rangle$
 - $\langle Books, book\ language, \{English, Spanish\} \rangle$

Using and Building Faceted Interfaces

The screenshot displays a faceted search interface with three main facets:

- Cell Phone Features**: Includes checkboxes for Camera, LTE, Smartphone, Quad Core Processor, and Dual Camera.
- Avg. Customer Review**: Includes star ratings from 4 to 5 stars and "Up".
- Brand**: Includes checkboxes for various brands like Samsung, ANTYE, MoKo, CIQILY, Goodtrade8, luhan, Schitec, hanende, OSCOO, RAVPower, EpicDealz, Geekercity, DBigness, ECCPP, and FONSTASTIC. A "See more" link is present.

Books facet (highlighted in red dashed box):

- Amazon Prime**: Includes a checked checkbox for prime.
- Eligible for Free Shipping**: Includes a checked checkbox for Free Shipping by Amazon.
- New Releases**: Includes checkboxes for Last 30 days, Last 90 days, and Coming Soon.
- Science Fiction & Fantasy Genres**: Includes checkboxes for Non-Romantic, Thriller, Mystery, Horror, and Humor. A "See more" link is present.
- Book Language**: Includes checkboxes for English and Spanish.
- Book Format**: Includes Kindle Edition, Audible Audio Edition, Paperback, Audio CD, Hardcover, and Large Print.
- Author**: Includes checkboxes for Ernest Cline, William R. Forstchen, and Hugh Howey.

Pet Supplies facet (highlighted in red dashed box):

- Brand**: Includes checkboxes for URPOWER, American Dog, Providence Engraving, Electric Dog Fence™, Electric Dog FenceTM, Crazy Dog T-Shirts, DOGS, DOGNESS, Alpha Dog Series, Dogs Kingdom, Overland Dog Gear, One Dog One Bone, In Dog We Trust, and UltraSonic Dog Whistle. A "See more" link is present.
- Pet Collar Size**: Includes checkboxes for XL, M, L, S, XS, and Adjustable.

Faceted interfaces

- Widespread in several domains
- Enhance user experience
 - Competitive advantage for web applications, e.g., in eCommerce

A facet

- Domain, property, values
 - $\langle D, p, \{v_1, \dots, v_n\} \rangle$
 - $\langle Books, book\ language, \{English, Spanish\} \rangle$

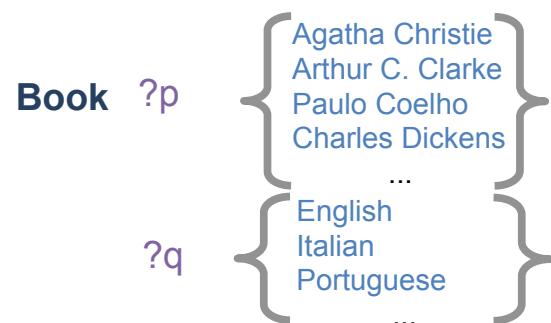
Hard to define well-crafted facets manually

- Many different types of entities require many domain-specific facets
 - E.g., 500+ product types

Facet Extraction & Annotation

Facet extraction

- Clusters of values for different characteristics
 - Offline vs. Online (on search results)



Facet annotation

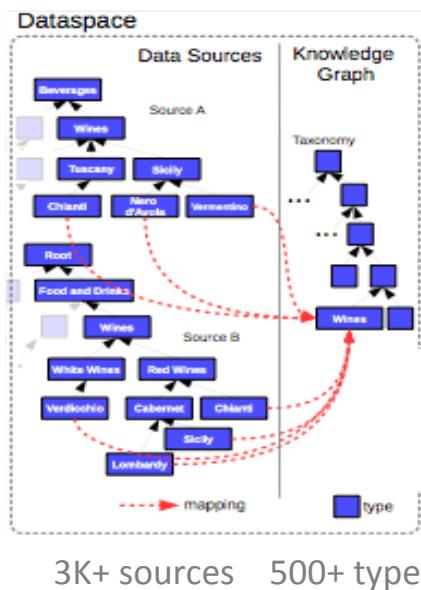
- *Assign a property to the automatically extracted set of values*
- Use a predicate of a KB to capture the semantics of the property
 - Human readable semantics
 - Machine-readable semantics → RDFa markup of faceted interfaces
 - Reduction of terminological entropy across domains



Main goals:

Build faceted interfaces on-the-fly for effective searching
Assist domain experts in creating and maintaining facets

Facet Extraction & Creation in Comparison Shopping Platforms (CSPs)



Automated clustering of facet values
(Porrini&al.CAISE14)

- extracted from 3k+ data sources mapped to a CSP central schema with 500+ leaf types
- clustered by similarity

(used in production in an Italian CSP)

Not Publishing View

Accepted Facets

New group... +

Regione

New facet... +

abruzzo
argentina
basilicata
calabria
campania
emilia romagna
francia
friuli
friuli venezia giulia
lazio

Suggested Facets C

vino bianco
vino rosato
vino rosso
bianco
rosato
rosso
abruzzo
argentina
basilicata

Experts refine the facet with a GUI

- Add/delete values
- Lexical specification of the facet property



Facet Annotation vs. Table Annotation

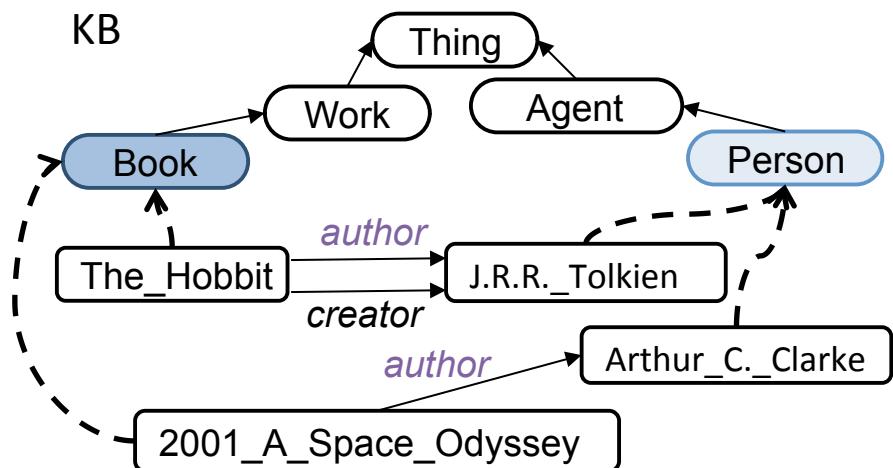
Facet

Book author

{ Agatha Christie
Arthur C. Clarke
Paulo Coelho
Charles Dickens
J. K. Rowling
J. R. R. Tolkien
Anna Sewell
... }



KB



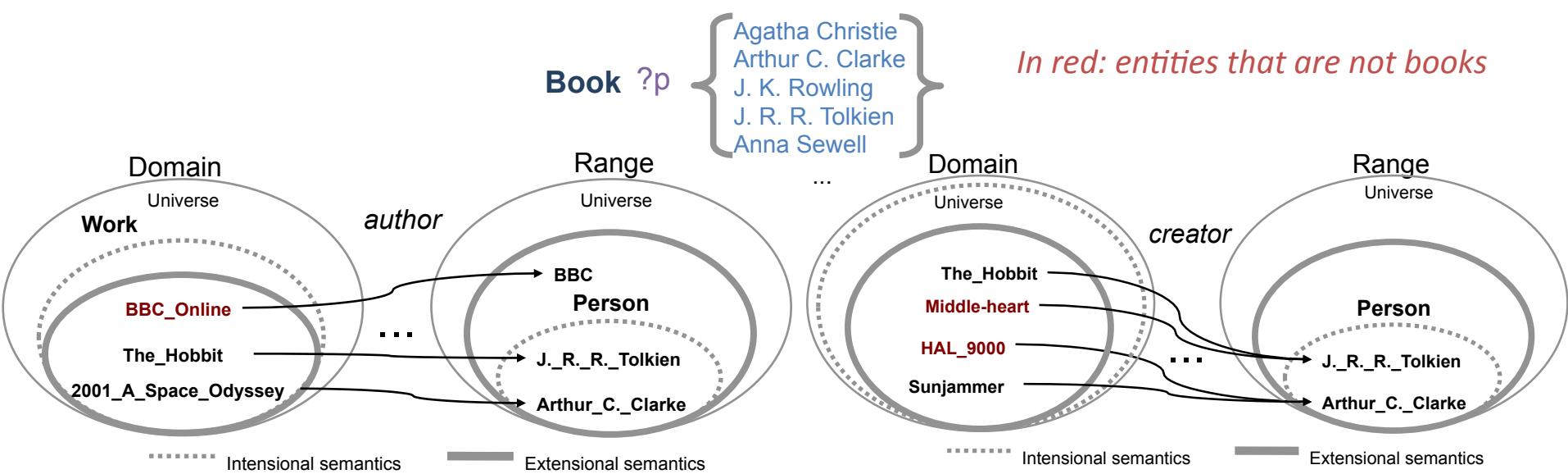
Facet annotation (unbalanced)

Books			
			author
			J. R. R. Tolkien
			Arthur C. Clarke
			Gillian Flynn

Table annotation (balanced)

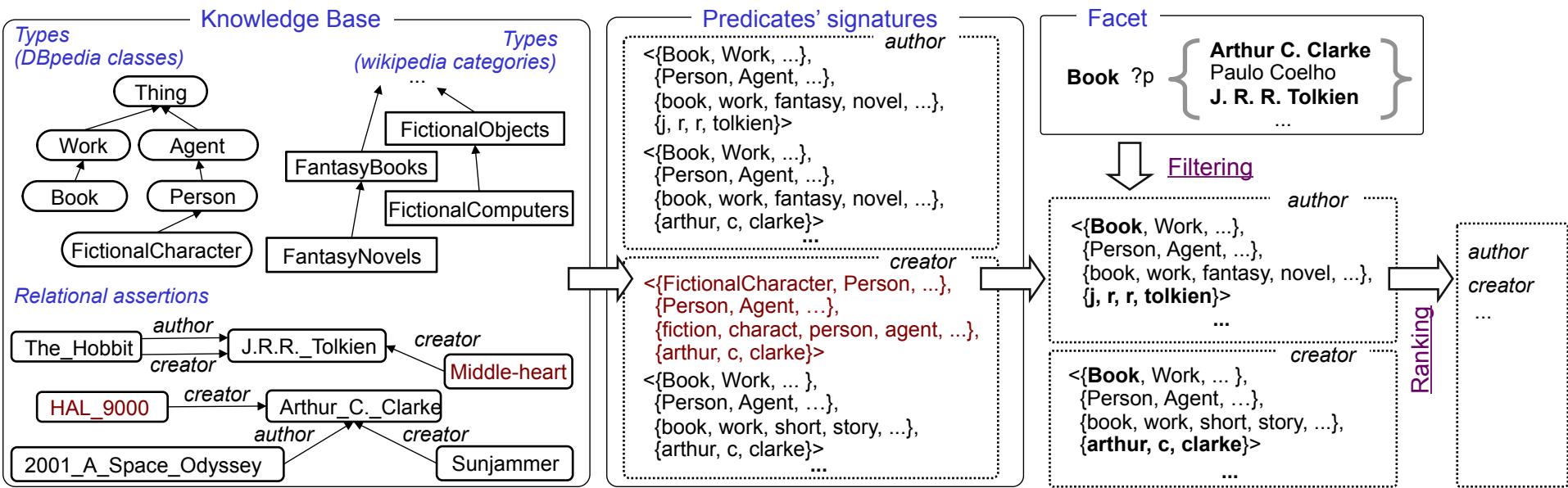
Book	author	Person
book_id	title	genre
1020[...]	The Hobbit	Juvenile fantasy
2013[...]	2001 A Space Odyssey	Science Fiction
2016[...]	Gone Girl	Crime
		Gillian Flynn

Facet Annotation: Challenges



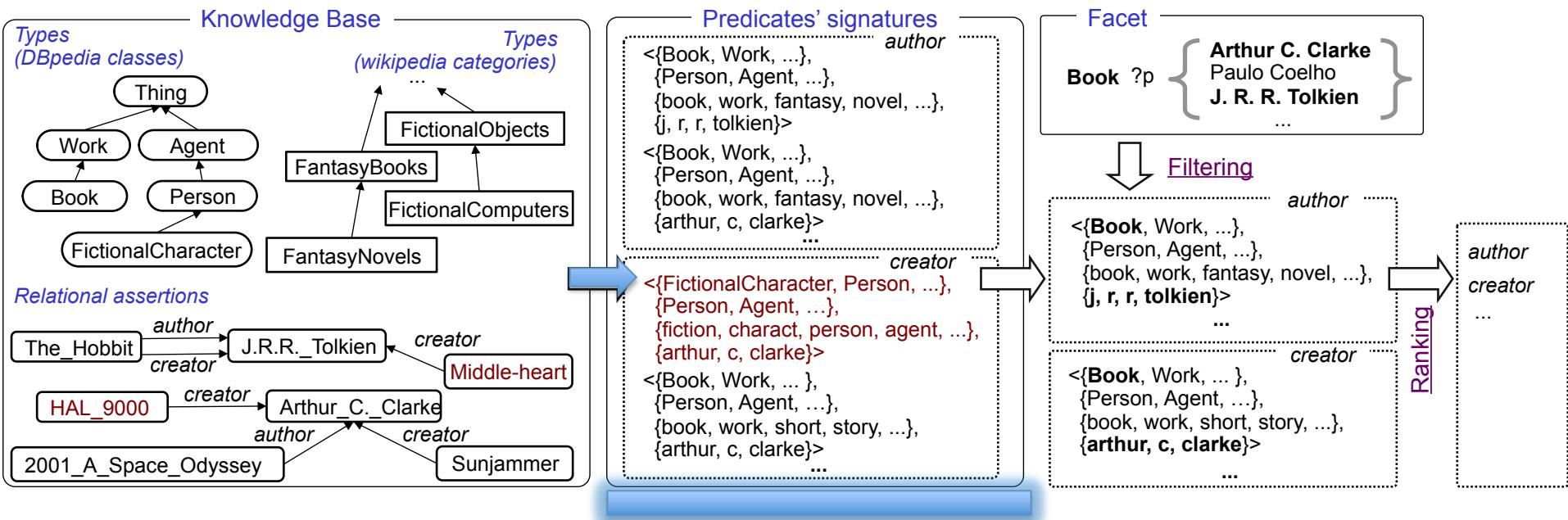
- Challenges:
 - Many predicates, e.g., 53K+ DBpedia properties
 - Weak specification, e.g., no domain/range restrictions
 - Ambiguous values, i.e., same values in different predicates
 - Specificity of the predicate associated with a domain-specific facet

Approach: Overview



- Index the extensions of the KB predicates using **signatures**
- Filter predicates whose signatures match the facet domain and values
- Rank predicates by similarity (three criteria: frequency, coverage, specificity)
- Output: **ranked list of predicates**

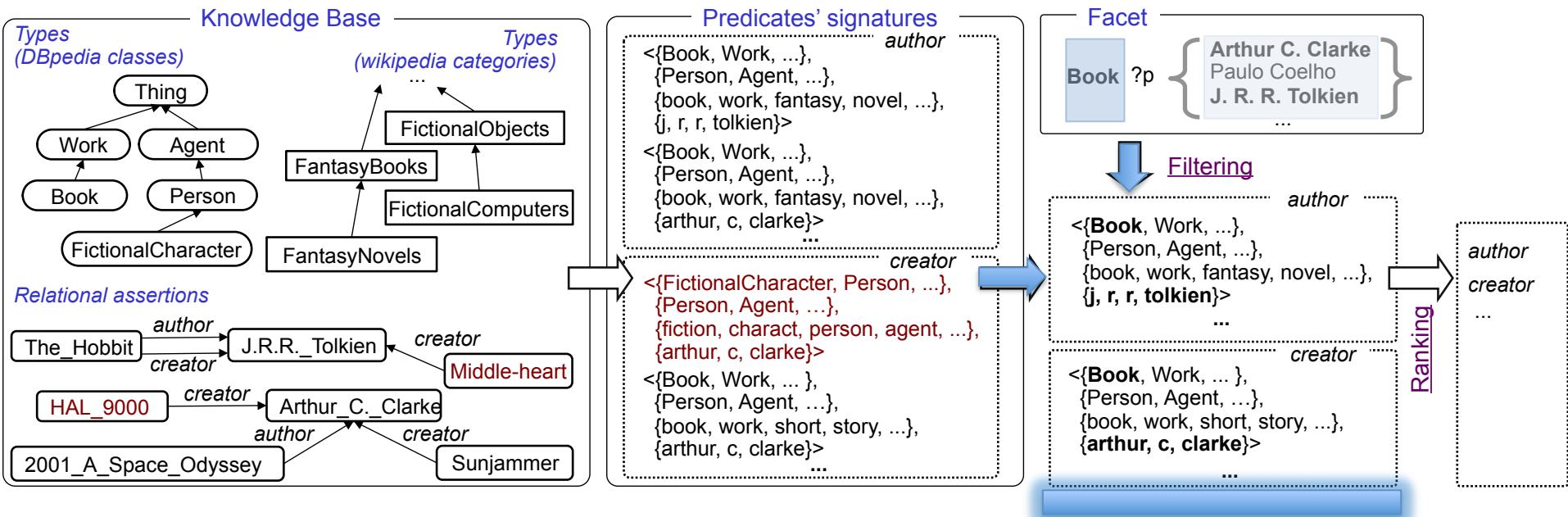
Approach: Signatures



Domain	Range
The_Hobbit	"J.R.R._Tolkien"
Middle-earth	"J.R.R._Tolkien"
HAL_9000	"Arthur_C._Clark"
Sunjammer	"Arthur_C._Clark"

Subject Types' Names	Subject Types' Lexicalizations	Object Types' Names	Objects' Lexicalizations
<{Book, Work, ...}>	{book, work, fantas, novel, ...}	{Person, Agent, ...}	{j, r, r, tolkien}>
<{Thing, FictionalContinents,...}>	{thing, fiction, continent,...,}	{Person, Agent, ...}	{j, r, r, tolkien}>
<{FictionalCharacter, Person, ...}>	{fiction, charact, person, agent, ...}	{Person, Agent, ...}	{arthur, c, clarke}>
<{Book, Work, ...}>	{book, work, short, story, ...}	{Person, Agent, ...}	{arthur, c, clarke}>

Approach: Filtering



D-match

- subject types vs. facet domain (lexicalizations)

R-match

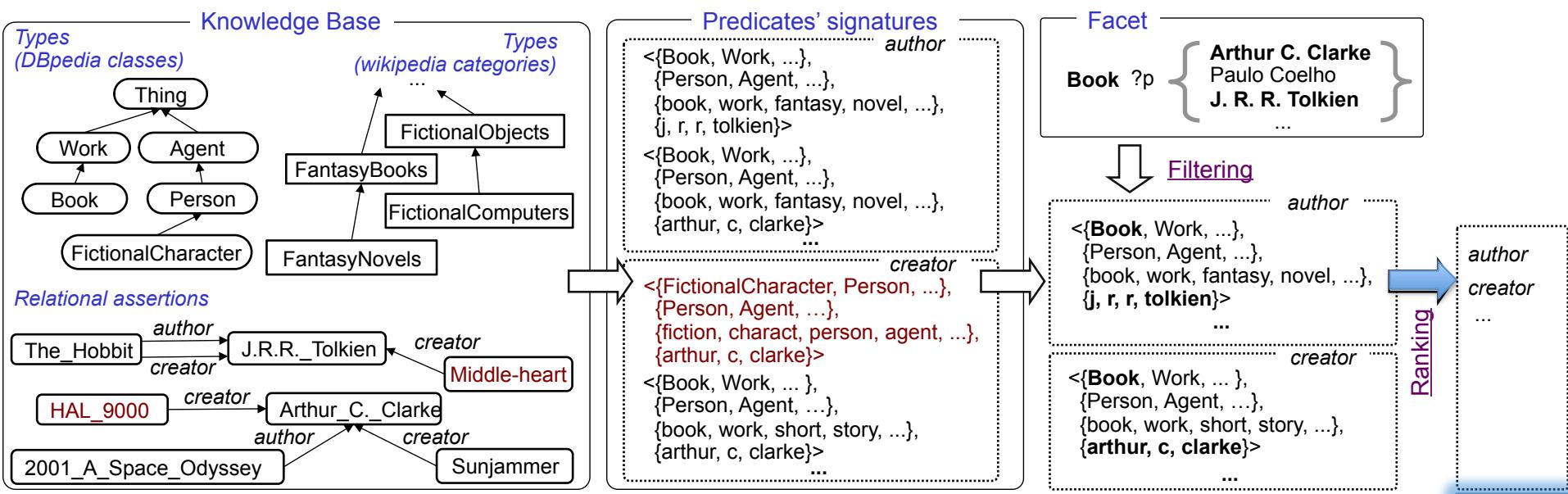
- object vs. at least 1 facet value (lexicalizations)

Signature matches if

- d-match & r-match

Subject Types' Names	Subject Types' Lexicalizations	Object Types' Names	Objects' Lexicalizations
✓ <{Book, Work, ...}>	{book, work, fantas, novel, ...}	{Person, Agent, ...}	{j, r, r, tolkien}>
✗ <{Thing, FictionalContinents,...}>	{thing, fiction, continent,...,}	{Person, Agent, ...}	{j, r, r, tolkien}>
✗ <{FictionalCharacter, Person, ...}>	{fiction, charact, person, agent, ...}	{Person, Agent, ...}	{arthur, c, clarke}>
✓ <{Book, Work, ...}>	{book, work, short, story, ...}	{Person, Agent, ...}	{arthur, c, clarke}>

Approach: Ranking



Three criteria:

- Coverage
- Weighted Frequency
- Specificity

Combination:

- Smoothing for comparability
- Multiplication

Subject Types' Names	Subject Types' Lexicalizations	Object Types' Names	Objects' Lexicalizations
✓ <{Book, Work, ... }>	{book, work, fantas, novel, ...}	{Person, Agent, ...}	{j, r, r, tolkien}>
✗ <{Thing, FictionalContinents,...}>	{thing, fiction, continent,...,}	{Person, Agent, ...}	{j, r, r, tolkien}>
✗ <{FictionalCharacter, Person, ...}>	{fiction, charact, person, agent, ...}	{Person, Agent, ...}	{arthur, c, clarke}>
✓ <{Book, Work, ... }>	{book, work, short, story, ...}	{Person, Agent, ...}	{arthur, c, clarke}>

Ranking: Coverage and Weighted Frequency

- Coverage of a predicate q
 - $\#facet_values_matched_by_q / \#facet_values$
- Weighted Frequency of a predicate q

Scaled down by the number of facet values $V^f \dots$

$$freq(f, q) = \frac{1}{|V^f|} \sum_{\delta_{(s,o)} \in \Delta_q^f} \frac{|L^T(s) \cap L(D^f)|}{|L^T(s) \cup L(D^f)|}$$

Sum of signatures $\delta_{(s,o)}$ matching the facet...

... weighted by the Jaccard similarity between
the lexicalizations of the subject type $L^T(s)$
and of the facet domain $L(D^f)$

Ranking: Specificity

author

creator

Subject Types' Names		Object Types' Names		Subject Types' Names		Object Types' Names		
< {Book, Work, ... }	{...}	{Person, Agent, ...}	{...} >	✓	< {Book, Work, ... }	{...}	{Person, Agent, ...} {...} >	✓
< {Work, ... }	{...}	{Agent, ...}	{...} >	✗	< {Thing, FictionalContinents,...}	{...}	{Person, Agent, ...} {...} >	✗
< {Book, Work, ... }	{...}	{Person, Agent, ...}	{...} >	✓	< {FictionalCharacter, Person, ...}	{...}	{Person, Agent, ...} {...} >	✗
< {Book, Work, ... }	{...}	{Person, Agent, ...}	{...} >	✗	< {Book, Work, ... }	{...}	{Person, Agent, ...} {...} >	✓
< {Book, Work, ... }	{...}	{Person, Agent, ...}	{...} >	✗	< {Book, Work, ... }	{...}	{Company, Agent, ...} {...} >	✗

- Matching signatures vs. non matching signatures
 - “more similar” in *author* than in *creator*

Ranking: Specificity

author

creator

Subject Types' Names	Object Types' Names	Subject Types' Names	Object Types' Names
< {Book, Work, ... } {...}	{Person, Agent, ...} {...}> ✓	< {Book, Work, ... } {...}	{Person, Agent, ...} {...}> ✓
< {Work, ... } * {...}	{Agent, ...} {...}> ✗	< {Thing, FictionalContinents,...} * {...}	{Person, Agent, ...} {...}> ✗
< {Book, Work, ... } {...}	{Person, Agent, ...} {...}> ✓	< {FictionalCharacter, Person, ...} * {...}	{Person, Agent, ...} {...}> ✗
< {Book, Work, ... } * {...}	{Person, Agent, ...} {...}> ✗	< {Book, Work, ... } {...}	{Person, Agent, ...} {...}> ✓
< {Book, Work, ... } * {...}	{Person, Agent, ...} {...}> ✗	< {Book, Work, ... } * {...}	{Company, Agent, ...} {...}> ✗

- Matching signatures vs. non matching signatures
 - “more similar” in *author* than in *creator*
- Compare types in signatures rather than signatures themselves
 - Counting matching vs. non matching signatures is unreliable (noisy, matching signatures << non matching signatures)

Ranking: Specificity

author

creator

Subject Types' Names	Object Types' Names	Subject Types' Names	Object Types' Names
< {Book, Work, ... } {...}	{Person, Agent, ...} {...} > ✓	< {Book, Work, ... } {...}	{Person, Agent, ...} {...} > ✓
< {Work, ... } * {...}	{Agent, ...} {...} > ✗	< {Thing, FictionalContinents,...} * {...}	{Person, Agent, ...} {...} > ✗
< {Book, Work, ... } {...}	{Person, Agent, ...} {...} > ✓	< {FictionalCharacter, Person, ...} * {...}	{Person, Agent, ...} {...} > ✗
< {Book, Work, ... } * {...}	{Person, Agent, ...} {...} > ✗	< {Book, Work, ... } {...}	{Person, Agent, ...} {...} > ✓
< {Book, Work, ... } * {...}	{Person, Agent, ...} {...} > ✗	< {Book, Work, ... } * {...}	{Company, Agent, ...} {...} > ✗

- Matching signatures vs. non matching signatures
 - “more similar” in *author* than in *creator*
- Compare types in signatures rather than signatures themselves
 - Counting matching vs. non matching signatures is unreliable (noisy, matching signatures << non matching signatures)
 - Consider most specific types

Ranking: Specificity

author

creator

Subject Types' Names	Object Types' Names	Subject Types' Names	Object Types' Names		
<{Book, Work, ... }	{...} {Person, Agent, ...}	{...}> ✓	<{Book, Work, ... }	{...} {Person, Agent, ...}	{...}> ✓
<{Work, ... }	{...} {Agent, ... }	{...}> ✗	<{Thing, FictionalContinents,...}	{...} {Person, Agent, ...}	{...}> ✗
<{Book, Work, ... }	{...} {Person, Agent, ...}	{...}> ✓	<{FictionalCharacter, Person, ...}	{...} {Person, Agent, ...}	{...}> ✗
<{Book, Work, ... }	{...} {Person, Agent, ...}	{...}> ✗	<{Book, Work, ... }	{...} {Person, Agent, ...}	{...}> ✓
<{Book, Work, ... }	{...} {Person, Agent, ...}	{...}> ✗	<{Book, Work, ... }	{...} {Company, Agent, ...}	{...}> ✗

Two similarity measures:

- *d-spec (on the domain)*
- *r-spec (on the range)*

Most specific subject types in non matching signatures

Most specific subject types in matching signatures

$$d\text{-spec}(f, q) = \frac{(D[\Delta_q] \cap D[\Delta_q^f])}{(D[\Delta_q] \cup D[\Delta_q^f])}$$

Ranking: Specificity

author

creator

Subject Types' Names	Object Types' Names	Subject Types' Names	Object Types' Names
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✓
<{Work, ... }	{...}	{Agent, ... }	{...}> ✗
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✓
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✗
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✗

Two similarity measures:

- *d-spec (on the domain)*
- *r-spec (on the range)*

$$d\text{-spec}(f, q) = \frac{w\text{-card}(D[\Delta_q] \cap D[\Delta_q^f])}{w\text{-card}(D[\Delta_q] \cup D[\Delta_q^f])}$$

Most specific subject types in non matching signatures

Most specific subject types in matching signatures

$$w\text{-card}(T) = \sum_{t \in T} \frac{\text{depth}(t)}{\max_{t' \in DES[t]} \text{depth}(t')}$$

Each type is weighted by its depth

Ranking: Specificity

author

creator

Subject Types' Names	Object Types' Names	Subject Types' Names	Object Types' Names
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✓
<{Work, ... }	{...}	{Agent, ... }	{...}> ✗
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✓
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✗
<{Book, Work, ... }	{...}	{Person, Agent, ... }	{...}> ✗

Two similarity measures:

- *d-spec (on the domain)*
- *r-spec (on the range)*

$$d\text{-spec}(f, q) = \frac{w\text{-card}(D[\Delta_q] \cap D[\Delta_q^f])}{w\text{-card}(D[\Delta_q] \cup D[\Delta_q^f])}$$

$$r\text{-spec}(f, q) = \frac{w\text{-card}(R[\Delta_q] \cap R[\Delta_q^f])}{w\text{-card}(R[\Delta_q] \cup R[\Delta_q^f])}$$

Most specific subject types in non matching signatures

Most specific subject types in matching signatures

$$w\text{-card}(T) = \sum_{t \in T} \frac{\text{depth}(t)}{\max_{t' \in DES[t]} \text{depth}(t')}$$

Each type is weighted by its depth

Experiments: Set Up

Our approach

DRC

Domain and Range
Comparison

Gold standards (122 facets)

- DBpedia [new!]
 - + properties ($\approx 53K+$)
 - + numbers
 - medium hierarchy
 - Ground truth: graded judgments
(incorrect < correct < accurate)
- YAGO [Limaye&al.PVLDB2010]
 - - properties (= 89)
 - deep hierarchy
 - Ground truth: one correct predicate per facet

Table 1: Gold Standards' statistics.

	#	Facets	Predicates	
		Domains	Accurate	Correct
dbpedia-numbers	8	7	~4	~19
dbpedia-entities	31	13	~7	~7
dbpedia	39	13	~3	~9
yago-explicit	83	17	1	-
yago-ambiguous	83	10	1	-

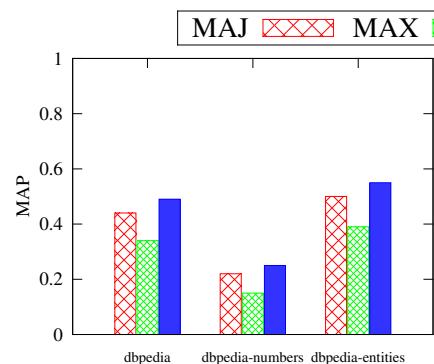
Baselines

- Majority voting (MAG)
- Maximum likelihood (MAX)
 - Adaptation of [Venetis&Al.VLDB2011]

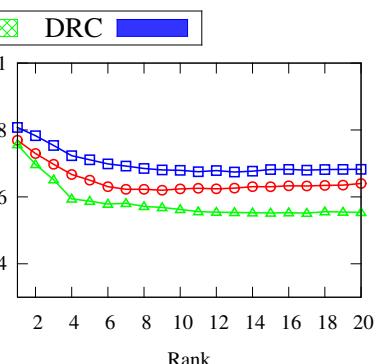
Measures

- Mean Average Precision (MAP) – DBpedia
- Mean Reciprocal Rank (MRR) - YAGO
- Normalized Discounted Cumulative Gain (nDCG) – Dbpedia/YAGO

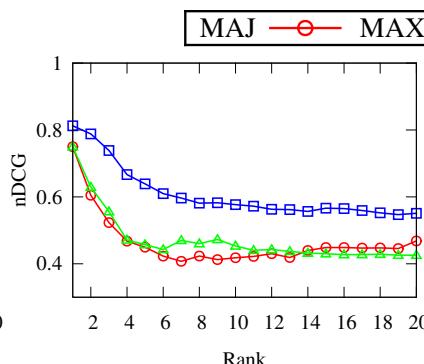
Experiments: DBpedia



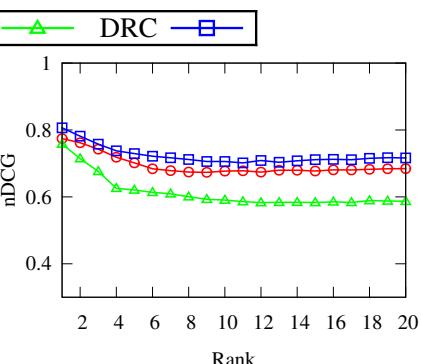
(a) Whole DBpedia dataset.



(b) dbpedia.



(c) dbpedia-numbers.

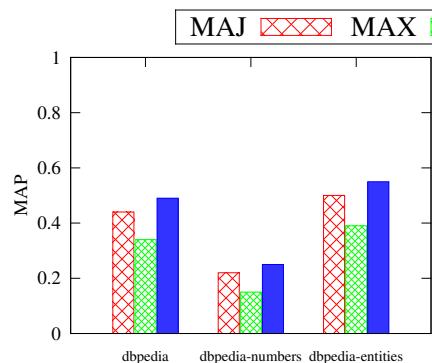


(d) dbpedia-entities.

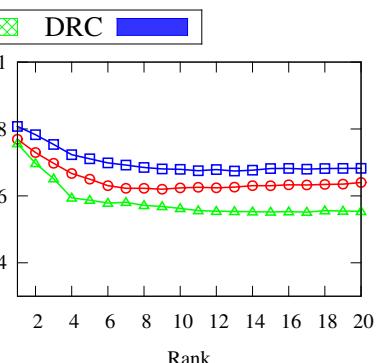
Highlights

- DRC always best results
- DRC particularly better with DBpedia numbers (values are more ambiguous)
- MAJ > MAX (adapted state-of-the-art approach)

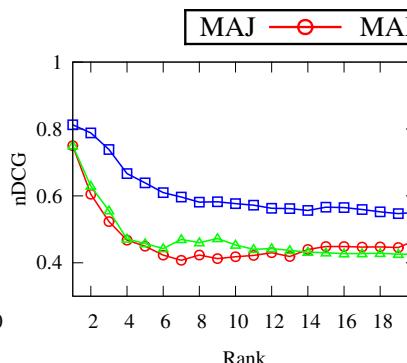
Experiments: DBpedia



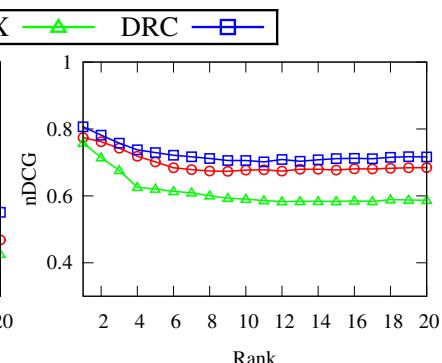
(a) Whole DBpedia dataset.



(b) dbpedia.



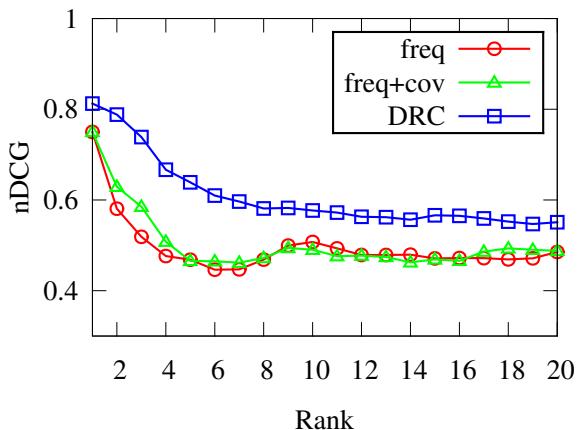
(c) dbpedia-numbers.



(d) dbpedia-entities.

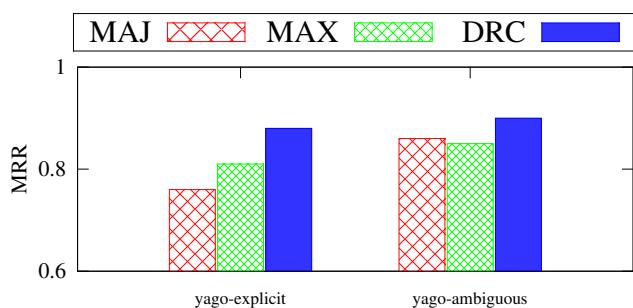
Highlights

- DRC always best results
- DRC particularly better with DBpedia numbers (values are more ambiguous)
- MAJ > MAX (adapted state-of-the-art approach)
- Specificity has impact on the quality of ranking

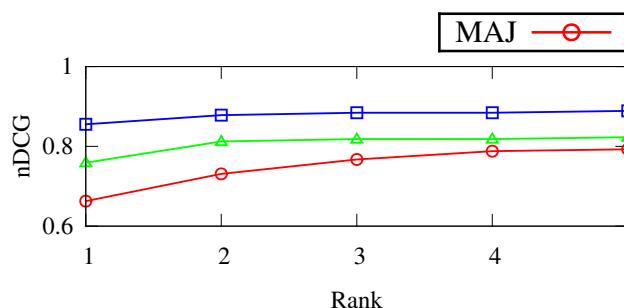


Effect of specificity on DRC

Experiments: YAGO

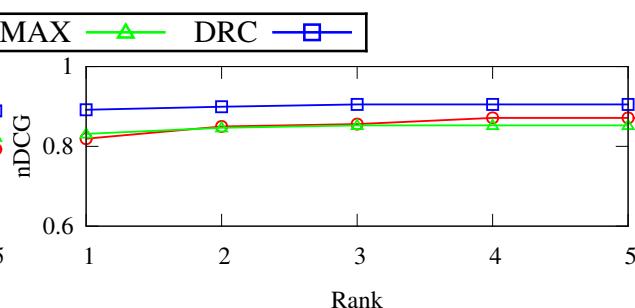


(a) Whole YAGO dataset.



(b) yago-explicit.

Facet domains:
type names in YAGO
(e.g., wikicategory-American-science-fiction-novels)



(c) yago-ambiguous.

Facet domains:
more general and intuitive than type
names in YAGO
(e.g., novels)

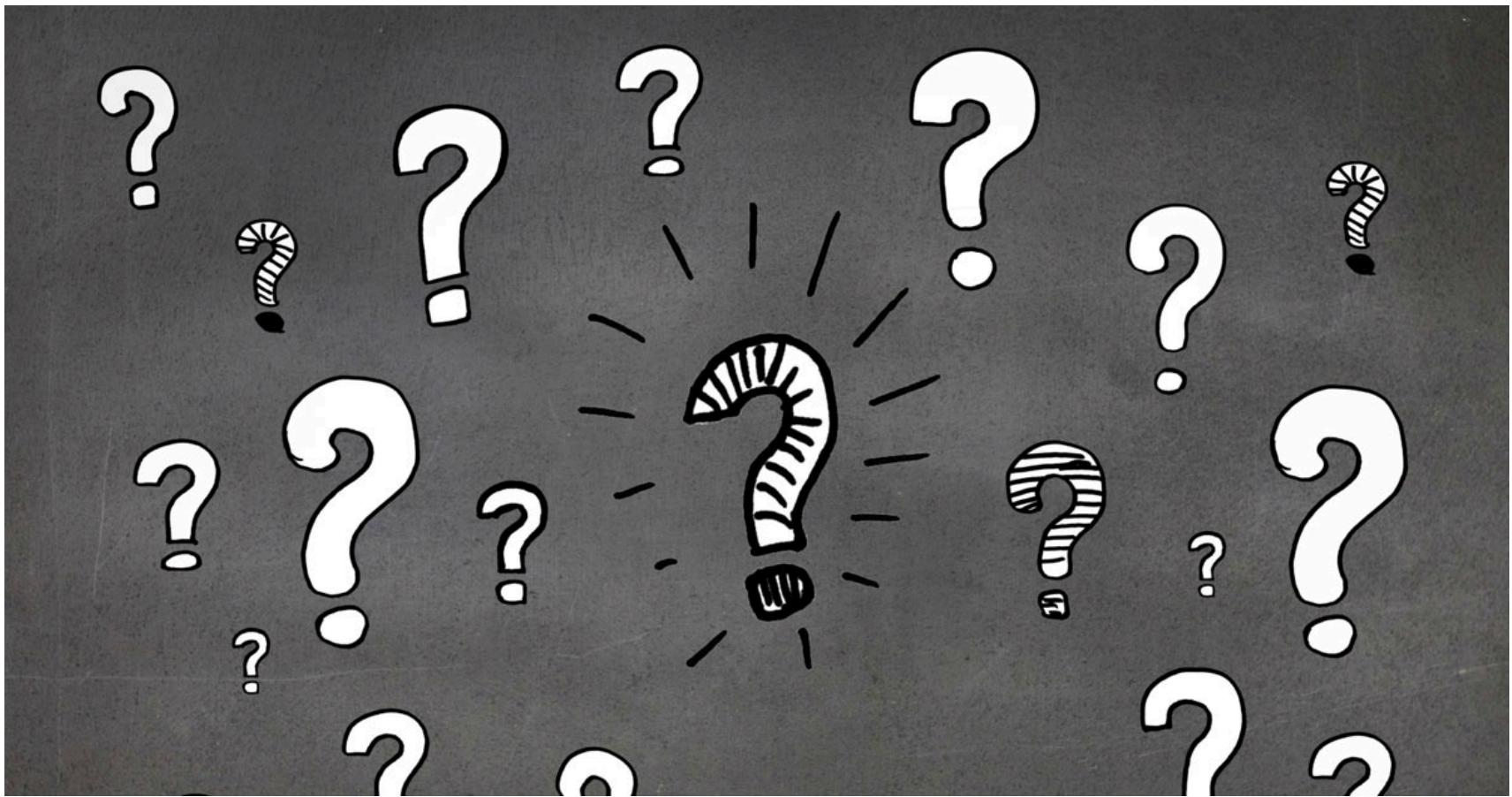
Highlights

- DRC always best results
- DRC and *Specificity* are less effective than with DBpedia (in particular with **yago-ambiguous**)
 - comparison of generic domains vs. very specific YAGO types
 - possible solution: take types of medium depth instead of most specific types

Conclusions & Future Work

- Facet annotation with reference KBs
 - Specificity-driven predicate ranking
 - End-to-end support domain experts in:
 - Creating new facets [Porrini&al.CAISE2014]
 - Labeling and lifting facets so as to transform lexical data into semantic-rich data
- Future work
 - Application to new domains (geospatial healthcare applications)
 - Improvement for KBs with very deep type hierarchy
 - Plug the approach into a table annotation framework



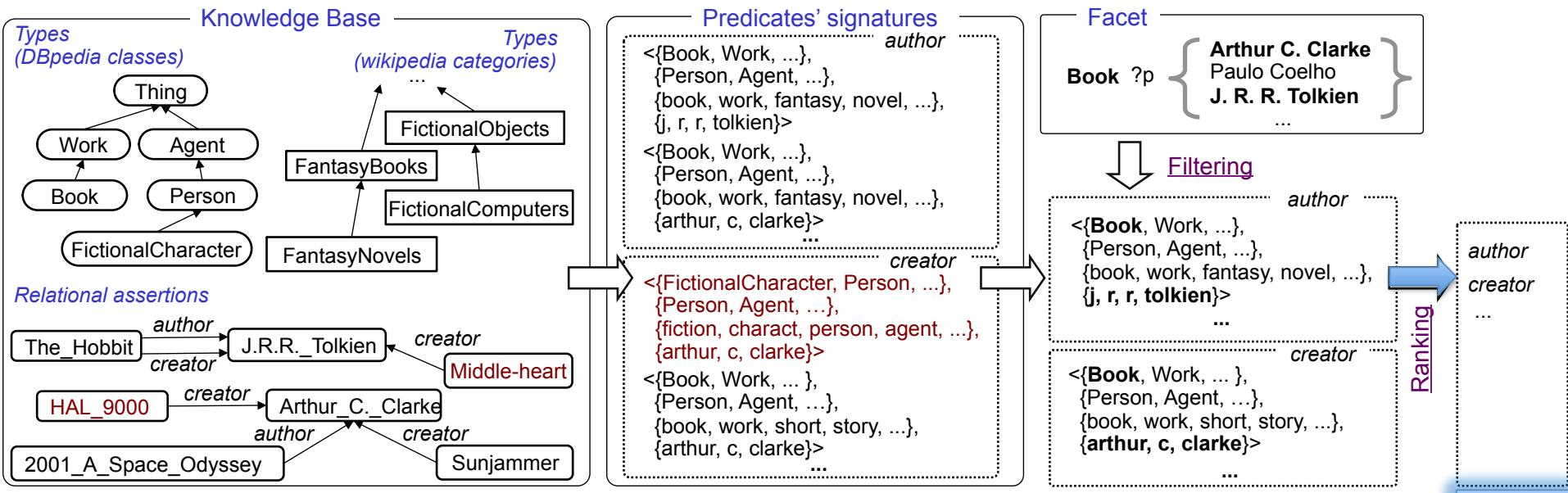


Contact: palmonari@disco.unimib.it

This research was partially supported by

Additional Details

More Details: Combination of Ranking Criteria



Three criteria:

- Coverage, Weighted Frequency, Specificity

Combination:

- Smoothing for comparability
- Multiplication

$$\begin{aligned}
 drc(f, q) = & \left(1 + \ln(d\text{-spec}(f, q) + 1) \right) \cdot \\
 & \left(1 + \ln(r\text{-spec}(f, q) + 1) \right) \cdot \\
 & \left(1 + \ln(cov(f, q) + 1) \right) \cdot \\
 & \left(1 + \ln(freq(f, q) + 1) \right)
 \end{aligned}$$