Collection Ranking & Selection for Federated Entity Search

Krisztian Balog, Robert Neumayer, and Kjetil Nørvåg
Norwegian University of Science and Technology

19th International Symposium on String Processing and Information Retrieval (SPIRE 2012)
Cartagena de Indias, Colombia, October 2012

Motivation

- **Entities are ubiquitous**
  Many information needs revolve around entities (people, products, organisations, places, etc.)
- **Growing amount of structured data**
  Again, organised around entities
- **Entities are often searched by their name**
  If we know (heard of) it, we just ask for it by the name

Top searches of 2011*

Web search	Mobile search
1. iPhone	1. iPhone 5
2. Casey Anthony	2. Powerball
3. Kim Kardashian	3. MLB
4. Katy Perry	4. Scrabble cheat
5. Jennifer Lopez	5. Casey Anthony
6. Lindsay Lohan	6. Hurricane Irene 2011
7. American Idol	7. Kim Kardashian
8. Jennifer Aniston	8. Translator
9. Japan earthquake	9. Amy Winehouse
10. Osama bin Laden	10. May 21, 2011 Rapture

* http://yearinreview.yahoo.com/

Top searches of 2011*

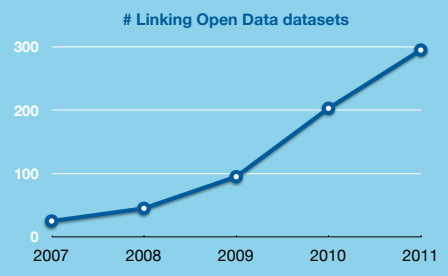
Web search	Mobile search
1. iPhone	1. iPhone 5
9. Japan earthquake	9. Amy Winehouse
10. Osama bin Laden	10. May 21, 2011 Rapture

"users have learned that search engine relevance decreases with longer queries and have grown accustomed to reducing their query (at least initially) to the name of an entity"

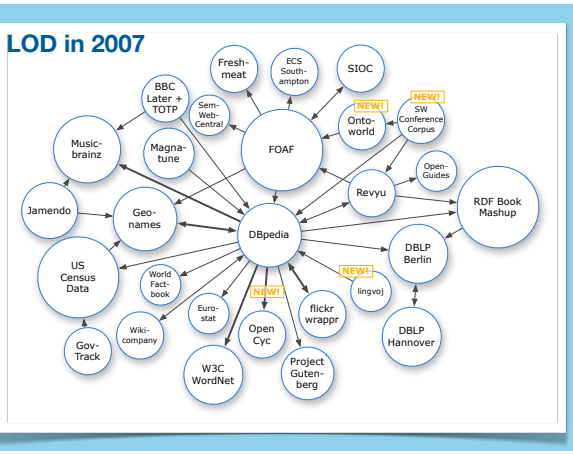
[Blanco et al., 2011]

* http://yearinreview.yahoo.com/

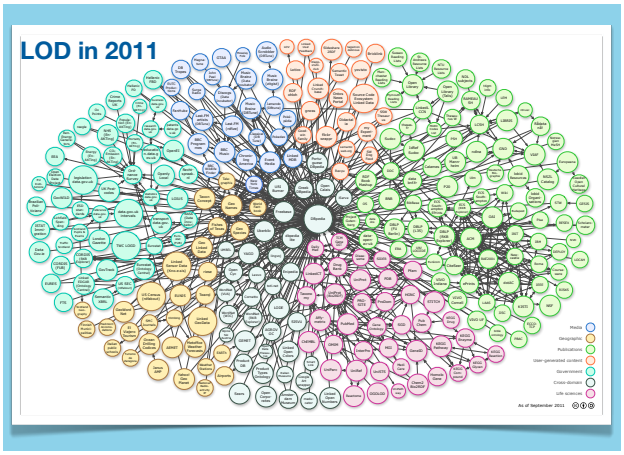
The Web of Data



LOD in 2007



LOD in 2011



Ad-hoc entity search

At the 2010/11 Semantic Search Challenge

- **Task**
  Given a keyword query, targeting a particular entity, provide a ranked list of relevant entities (i.e., URIs)
- **Queries**
  Sampled from web search engine logs (142 in total)
- **Data collection**
  Billion Triple Challenge 2009 (BTC) dataset
- **Relevance judgments**
  On a 3-point scale, collected using crowdsourcing

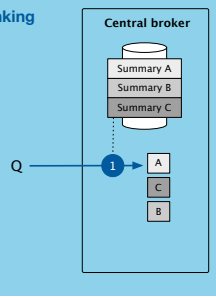
In this talk

- Address the **ad-hoc entity retrieval** task **in a distributed setting**
  - The Web of Data is inherently distributed
  - Some data sources may not be crawleable at all
- Specifically, our focus is on the **collection ranking** and **collection selection** steps

Federated search

A typical broker-based architecture

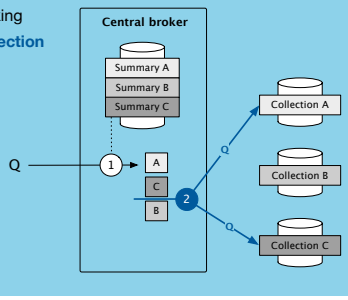
1 Collection ranking



Federated search

A typical broker-based architecture

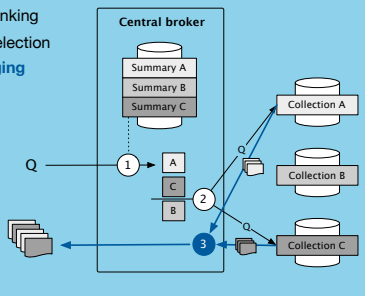
1 Collection ranking
2 Collection selection



Federated search

A typical broker-based architecture

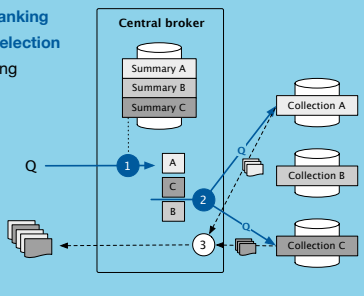
1 Collection ranking
2 Collection selection
3 Result merging



Next: baseline models

for collection ranking and selection

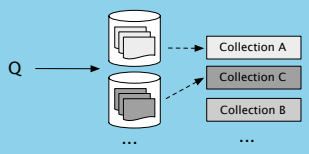
1 Collection ranking
2 Collection selection
3 Result merging



Collection ranking

Collection-centric (CC)

- **Lexicon-based method**
  Treat and score each collection as if it was a single, large document

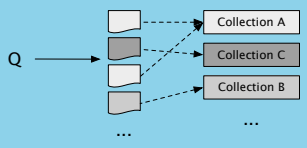


$$P(c|q) \propto P(c) \cdot \prod_{t \in q} P(t|\theta_c)$$

Collection ranking

Entity-centric (EC)

- **Document-surrogate method**
  Model and query individual documents (entities) and aggregate their relevance scores

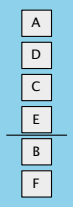


$$P(c|q) \propto \sum_{e \in c, r(e,q) < \gamma} P(e|q)$$

Collection selection

Top-K selection

- **Choosing a fixed cutoff (K) ahead of time**
  K is usually set between 5 and 20



Our method: AENN

"All that an Entity Needs is a Name"

- Exploit that **entities are searched by** their **name**
- The **central broker** maintains a complete **dictionary of entity names** (and corresponding identifiers)
- Utilise this information in the collection selection step to **dynamically adjust the #collections selected**

AENN collection ranking

- **Key observation**
  Different methods—collection-centric (CC) vs. entity-centric (EC)—work best for different queries

- **Idea**
  Combination should give better results than any of the two methods alone

$$\mathrm{AENN}(c, q) = (1 - \lambda) \cdot \mathrm{CC}(c, q) + \lambda \cdot \mathrm{EC}(c, q)$$

AENN collection ranking

Example

CC		EC		AENN	
A	0.35	A	0.65	A	0.5
D	0.3	B	0.3	B	0.2
C	0.2	C	0.05	D	0.15
E	0.15			C	0.125
B	0.1			E	0.075
F	0.05			F	0.025

AENN collection selection

- **Key observation**
  CC has higher recall, while EC has better precision
- **Idea**
  Use the collection rankings generated by EC and/or CC to dynamically adjust the set of collections selected
  - Precision-oriented selection
  - Recall-oriented selection
  - Balanced selection

AENN collection selection

Precision-oriented selection (AENN(p))

- **Only select collections returned by EC**

CC		EC		AENN		→	AENN(p)	
A	0.35	A	0.65	A	0.5		A	0.5
D	0.3	B	0.3	B	0.2		B	0.2
C	0.2	C	0.05	D	0.15		C	0.125
E	0.15			C	0.125			
B	0.1			E	0.075			
F	0.05			F	0.025			

AENN collection selection

Recall-oriented selection (AENN(r))

- **Include collections from CC until all from EC are covered. This defines the cutoff point for AENN**

CC		EC		AENN		→	AENN(r)	
A	0.35	A	0.65	A	0.5		A	0.5
D	0.3	B	0.3	B	0.2		B	0.2
C	0.2	C	0.05	D	0.15		D	0.15
E	0.15			C	0.125		C	0.125
B	0.1			E	0.075		E	0.075
F	0.05			F	0.025			

AENN collection selection

Balanced selection (AENN(b))

- **Include collections from AENN until all from EC are covered**

CC		EC		AENN		→	AENN(b)	
A	0.35	A	0.65	A	0.5		A	0.5
D	0.3	B	0.3	B	0.2		B	0.2
C	0.2	C	0.05	D	0.15		D	0.15
E	0.15			C	0.125		C	0.125
B	0.1			E	0.075			
F	0.05			F	0.025			

AENN collection selection

Comparison of approaches

AENN(p)		AENN(r)		AENN(b)	
A	0.5	A	0.5	A	0.5
B	0.2	B	0.2	B	0.2
C	0.125	D	0.15	D	0.15
		C	0.125	C	0.125
		E	0.075		

Experimental setup

Based on the 2010/11 Semantic Search Challenge

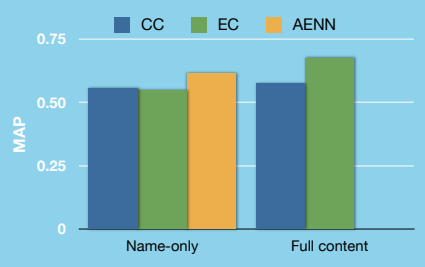
- **Distributed environment**
  Top 100 largest second-level domains from BTC
  - Three sets with different handling of DBpedia
- **Relevance**
  Considered the #relevant entities from each collection
- **Metrics**
  - *Collection ranking:* Standard IR metrics (MAP, MRR, nDCG)
  - *Collection selection:* Analogues of precision and recall, plus the avg. #coll. selected

Test collections

	BTC	BTC \DBpedia	DBpedia
#Entities	68.8M	60.5M	8.8M
#Collections	100	99	100
#Queries	136	116	130
Avg. #rel. entities/query	14.9	4.8	10.1
Avg. #rel. entities/collection	3.4	2.8	9.4

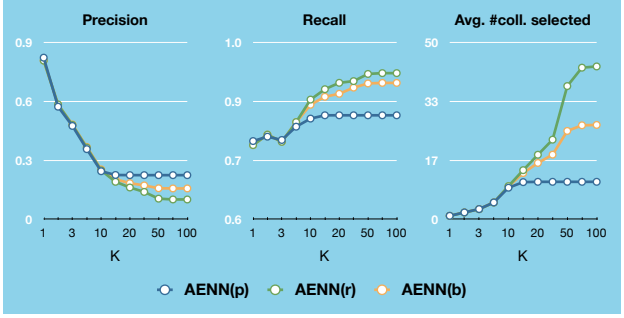
Results

Collection ranking (BTC)



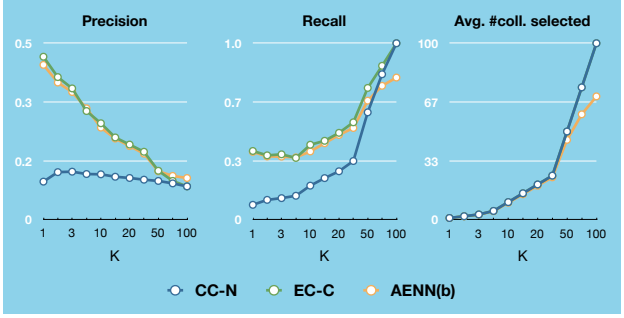
Results

Different collection selection strategies (BTC\DBpedia)



Results

Collection selection (DBpedia)



Summary

- Addressed the task of **ad-hoc entity retrieval** in a **distributed setting**
- Introduced AENN, a **novel collection ranking and selection** method based on a lean name-based entity representation
- Showed experimentally that AENN can **outperform standard baselines** that consider all entity content
- Further, AENN can be geared towards high precision, high recall, or a balanced setting

Questions?

Resources are available at http://bit.ly/OzfYK2

Contact
@krisztianbalog
krisztianbalog.com