Entities, Graphs, and Crowdsourcing for better Web Search

Gianluca Demartini
eXascale Infolab
University of Fribourg, Switzerland

gianlucademartini.net
exascale.info

Gianluca Demartini

demartini@exascale.info

- M.Sc. at University of Udine, Italy
- Ph.D. at University of Hannover, Germany
  - Entity Retrieval
- Worked for UC Berkeley (on Crowdsourcing), Yahoo! Research (Spain), L3S Research Center (Germany)
- Post-doc at the **eXascale Infolab**, Uni Fribourg, Switzerland.
- Lecturer for *Social Computing* in Fribourg
- Tutorial on Entity Search at ECIR 2012, on Crowdsourcing at ESWC 2013 and ISWC 2013
- Research Interests
  - Information Retrieval, Semantic Web, Crowdsourcing

Google tom cruise

Web Images Maps Shopping News More Search tools

6 personal results. 188,000,000 other results.

Tom Cruise - IMDb
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: Top Gun. If you had told 14 year old Franciscan seminary student Thomas Cruise Mapother IV that one day in the not too distant future he ...
Filmography by year - Biography - Rock of Ages - All You Need Is Kill

Tom Cruise - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Tom_Cruise**
Thomas Cruise Mapother IV (/ˈtɒməs ˈkruːz ˈmeɪpɒθər/; born July 3, 1962), widely known as **Tom Cruise**, is an American film actor and producer. He has ...
Tom Cruise filmography - Katie Holmes - Mimi Rogers - List of awards and ...

Official **Tom Cruise**: Oblivion, Movies, Video, Biography, News ...
www.**tomcruise**.com/
Official **Tom Cruise** site: Get the latest Rock of Ages trailer, info & downloads! Watch career movie trailers, videos, and retrospective. Read the **Tom Cruise** ...

TomCruise.com (**TomCruise**) on Twitter
https://twitter.com/**TomCruise**
The latest from **TomCruise**.com (@**TomCruise**). Official http://**TomCruise**.com TeamTC tweets. Does Tom Tweet? Sometimes between family & movies & its ...

Tom Cruise | Facebook
https://www.facebook.com/official**tomcruise**
Tom Cruise. 3883109 likes · 76956 talking about this. Welcome to the Official www.**TomCruise**.com team Facebook page! **Tom Cruise** news, events, pics & video ...

Tom Cruise

Follow

Actor

Thomas Cruise Mapother IV, widely known as Tom Cruise, is an American film actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Taps. Wikipedia

Born: July 3, 1962 (age 50), Syracuse, New York, United States

Height: 5' 7" (1.70 m)

Upcoming movie: All You Need Is Kill

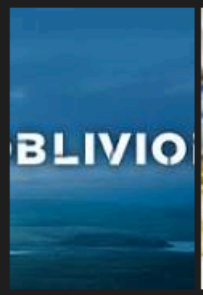
Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)

Children: Suri Cruise, Connor Cruise, Isabella Jane Cruise

Google tom cruise movies

Web Images Maps Shopping News More Search tools

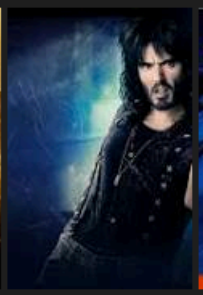
Tom Cruise movies



Oblivion
2013

Jack Reacher
2012

Mission: Impossible – G...
2011

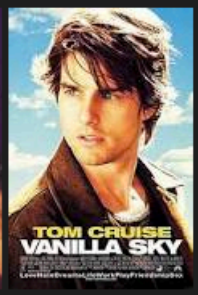

Rock of Ages
2012

Top Gun
1986

Knight and Day
2010

Minority Report
2002

Eyes Wide Shut
1999



Vanilla Sky
2001

Tom Cruise - IMDb
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: **Top Gun**. ... **Movies**. In Theaters; Top 250; US Box Office; Coming Soon; Trailer Gallery; Watch Now on AIV; On DVD ... **Tom Cruise** and Olga Kurylenko at event of **Oblivion** Still of **Tom Cruise** and Mia Sara in Legend Tom ...
Filmography by year - Biography - Rock of Ages - All You Need Is Kill

Tom Cruise filmography - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Tom_Cruise_film**ography
Tom Cruise is an American **film** actor and producer. The following is a ... 1981, Endless Love, Billy, Tom plays a boy who jokingly tells about committing arson.
Filmography - See also - References - External links

Official **Tom Cruise**: **Oblivion**, **Movies**, Video, Biography, News ...
www.**tomcruise**.com/
Official **Tom Cruise** site: Get the latest **Rock of Ages** trailer, info & downloads! Watch career **movie** trailers, videos, and retrospective. Read the **Tom Cruise** ...

Tom Cruise | **Movies** and Biography - Yahoo! **Movies**
movies.yahoo.com/person/**tom-cruise**/
Movies. The biggest star in the world for 20 years, **Tom Cruise** stood atop the ...

Tom Cruise

Actor

Follow

Thomas Cruise Mapother IV, widely known as Tom Cruise, is an American film actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Taps. Wikipedia

Born: July 3, 1962 (age 50), Syracuse, New York, United States

Height: 5' 7" (1.70 m)

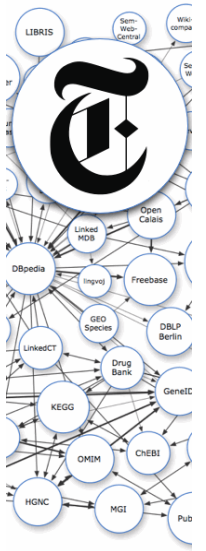
Upcoming movie: All You Need Is Kill

Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)

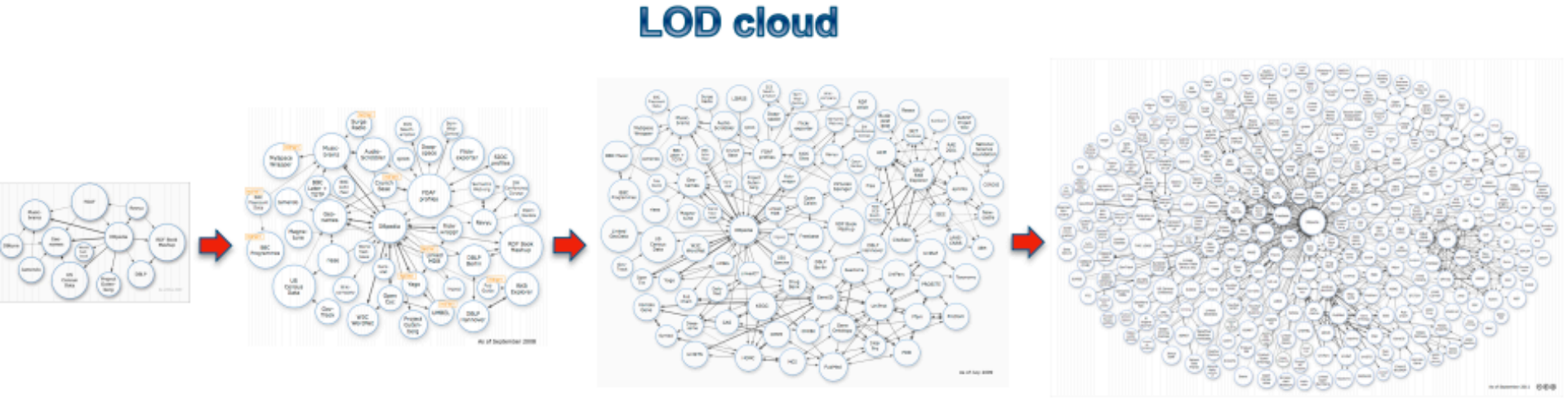
Children: Suri Cruise, Connor Cruise, Isabella Jane Cruise

Web of Data

- Freebase
  - Acquired by Google in July 2010.
  - Knowledge Graph launched in May 2012.
- Schema.org
  - Driven by major search engine companies
  - Machine-readable annotations of Web pages
- Linked Open Data
  - 31 billion triples, Sept. 2011



Linked Open Data



Z. Kaoudi and I. Manolescu, ICDE seminar 2013

I will talk about

- Entity Linking/Disambiguation
  - On the Web using crowdsourcing
  - For Data Integration / For Entity Type Ranking
- Ad-hoc Object Retrieval (Entity Ranking)
  - Using IR and graphs
- Crowdsourced Query Understanding

Disclaimer

- No efficiency evaluation
  - Approaches not distributed
  - But designed to scale out
- No user studies
  - Goal: Obtain high quality data
  - Only TREC-like evaluation on effectiveness

Entity Linking/Disambiguation

APRIL 9, 2012, 1:15 PM MERGERS & ACQUISITIONS

Facebook Buys Instagram for $1 Billion

BY EVELYN M. RUSLI

2:02 p.m. | Updated

Facebook is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.

http://dbpedia.org/resource/Facebook

http://dbpedia.org/resource/Instagram

owl:sameAs

fbase:Instagram

HTML:

```
<p>Facebook is not waiting for its initial public offering to make its first big purchase.</p><p>In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.</p>
```

RDFa enrichment

```
<p><span about="http://dbpedia.org/resource/Facebook"><cite property=”rdfs:label">Facebook</cite> is not waiting for its initial public offering to make its first big purchase.</span></p><p><span about="http://dbpedia.org/resource/Instagram">In its largest acquisition to date, the social network has purchased <cite property=”rdfs:label">Instagram</cite> , the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.</span></p>
```

CNET › News › Mobile

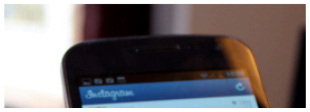
Instagram for Android is now available

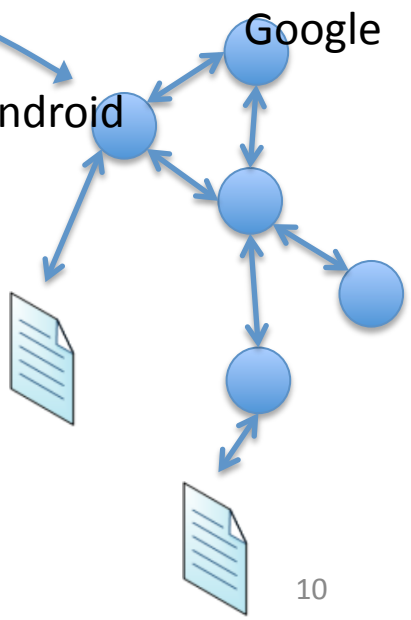
At long last, Instagram finally releases the Android version of its app.

by Jason Cipriani | April 3, 2012 10:07 AM PDT

Follow

Instagram has been around since 2010, available only to iOS devices. Android users have been waiting patiently, with repeated promises of an Android version arriving soon.





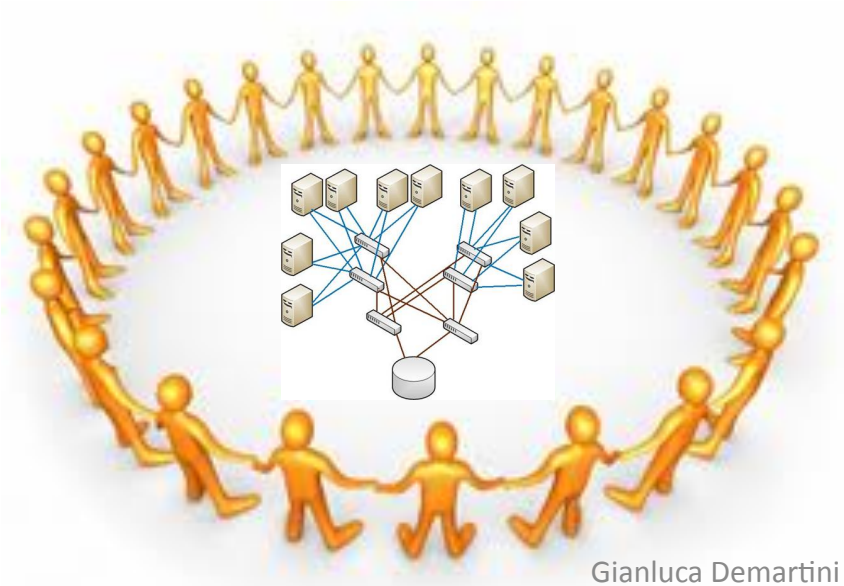
Crowdsourcing

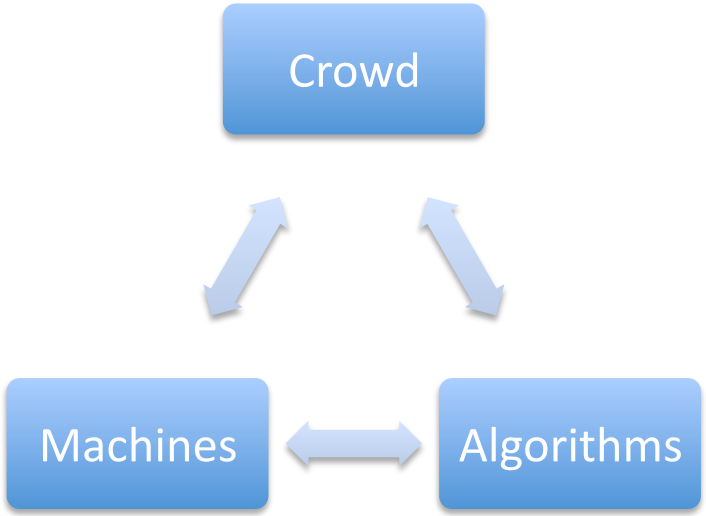
- Exploit human intelligence to solve
  - Tasks simple for humans, complex for machines
  - With a large number of humans (the Crowd)
  - Small problems: micro-tasks (Amazon MTurk)
- Examples
  - Wikipedia, Image tagging
- Incentives
  - Financial, fun, visibility

mechanical turk
Artificial Artificial Intelligence
amazon
beta

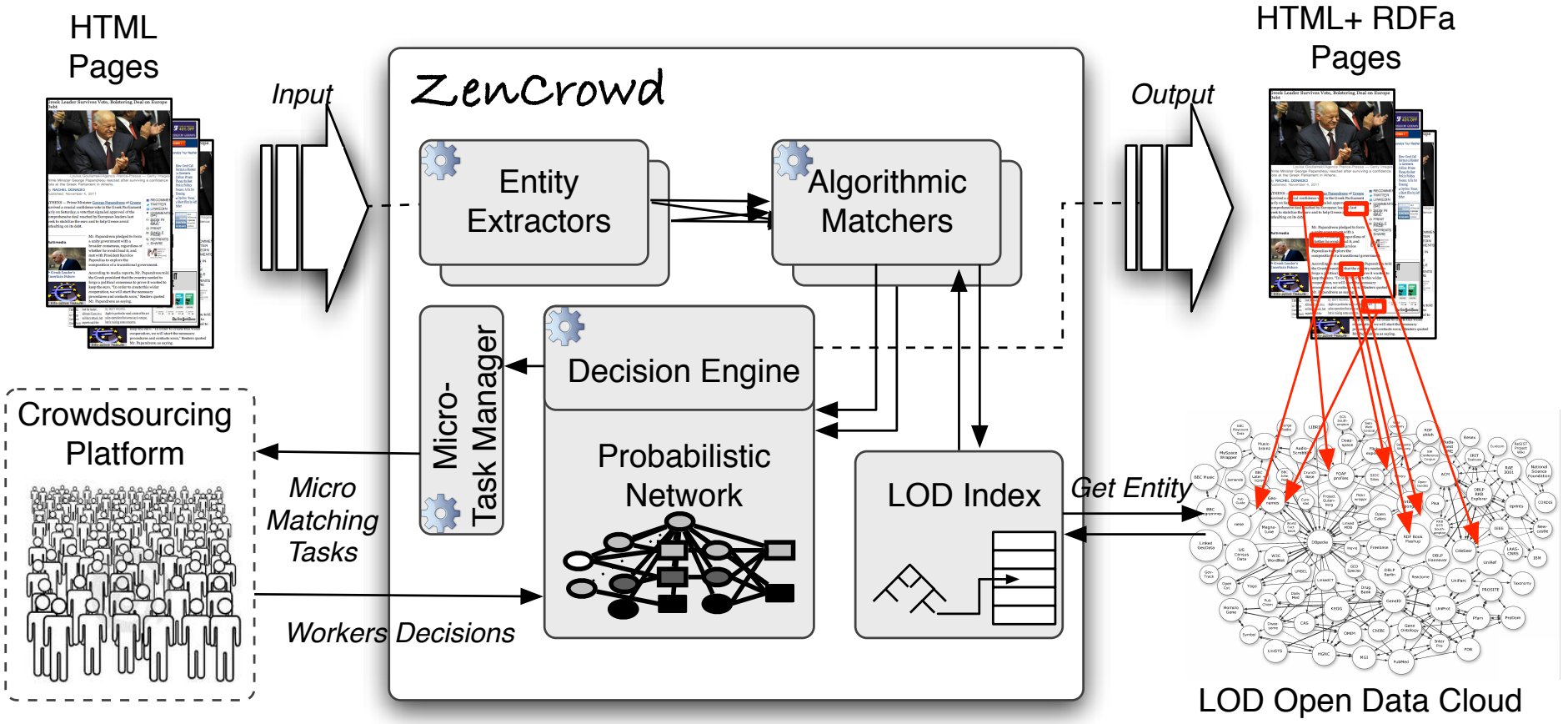
ZenCrowd

- Combine both algorithmic and manual linking
- Automate manual linking via crowdsourcing
- Dynamically assess human workers with a probabilistic reasoning framework





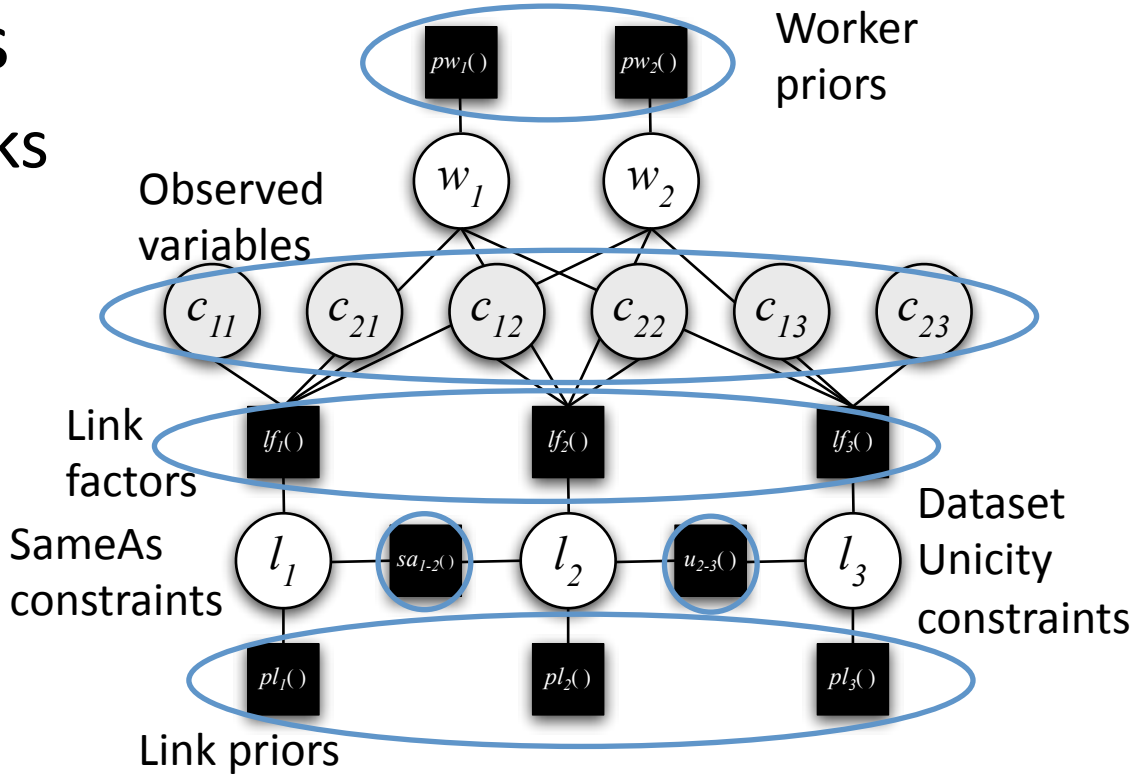
ZenCrowd Architecture



Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In: 21st International Conference on World Wide Web (**WWW 2012**).

Entity Factor Graphs

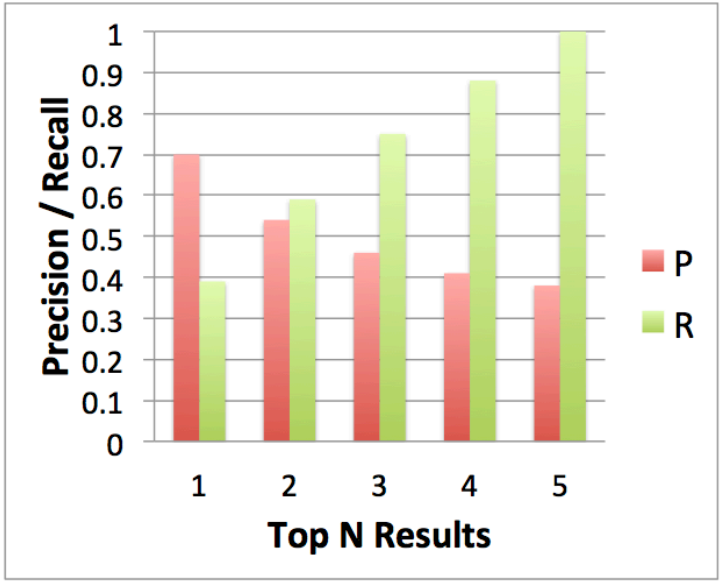
- Graph components
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints
- Probabilistic Inference
  - Select all links with posterior prob >τ



2 workers, 6 clicks, 3 candidate links

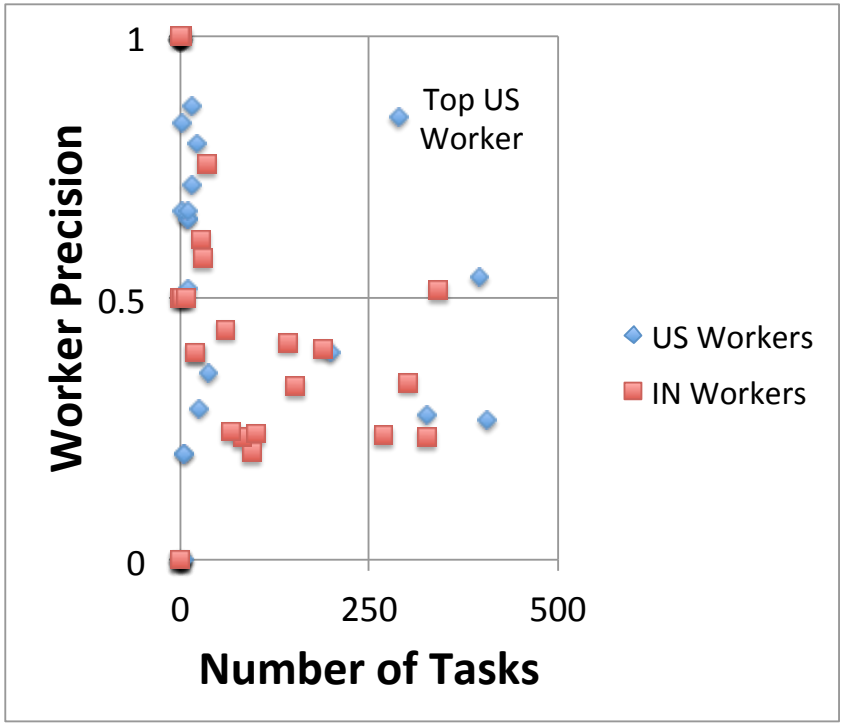
Experimental Evaluation

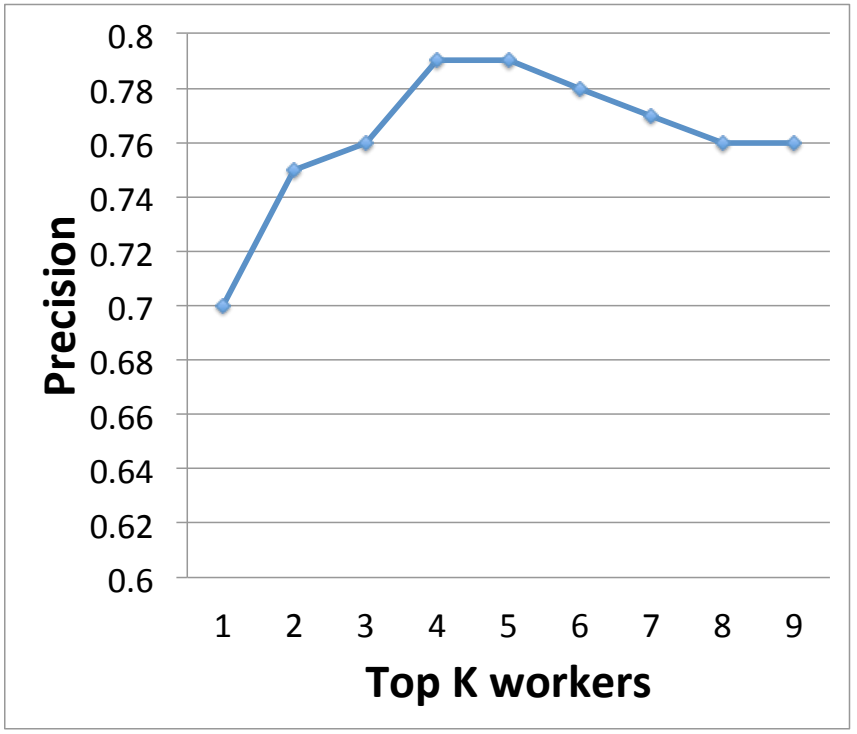
- Datasets
  - 25 news articles from
    - CNN.com (Global news)
    - NYTimes.com (Global news)
    - Washington-post.com (US local news)
    - Timesofindia.indiatimes.com (India news)
    - Swissinfo.com (Switzerland local news)
  - 40M entities (Freebase, DBPedia, Geonames, NYT)



	US Workers			Indian Workers		
	P	R	A	P	R	A
GL News	0.84	0.87	0.90	0.67	0.64	0.78
US News	0.64	0.68	0.78	0.55	0.63	0.71
IN News	0.84	0.82	0.89	0.75	0.77	0.80
SW News	0.72	0.80	0.85	0.61	0.62	0.73
All News	0.80	0.81	0.88	0.64	0.62	0.76

Worker Selection





Lessons Learnt

- Crowdsourcing + Prob reasoning works!
- But
  - Different worker communities perform differently
  - Many low quality workers
  - Completion time may vary (based on reward)
- Need to find the right workers for your task (see WWW13 paper)

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do. In: **WWW2013**

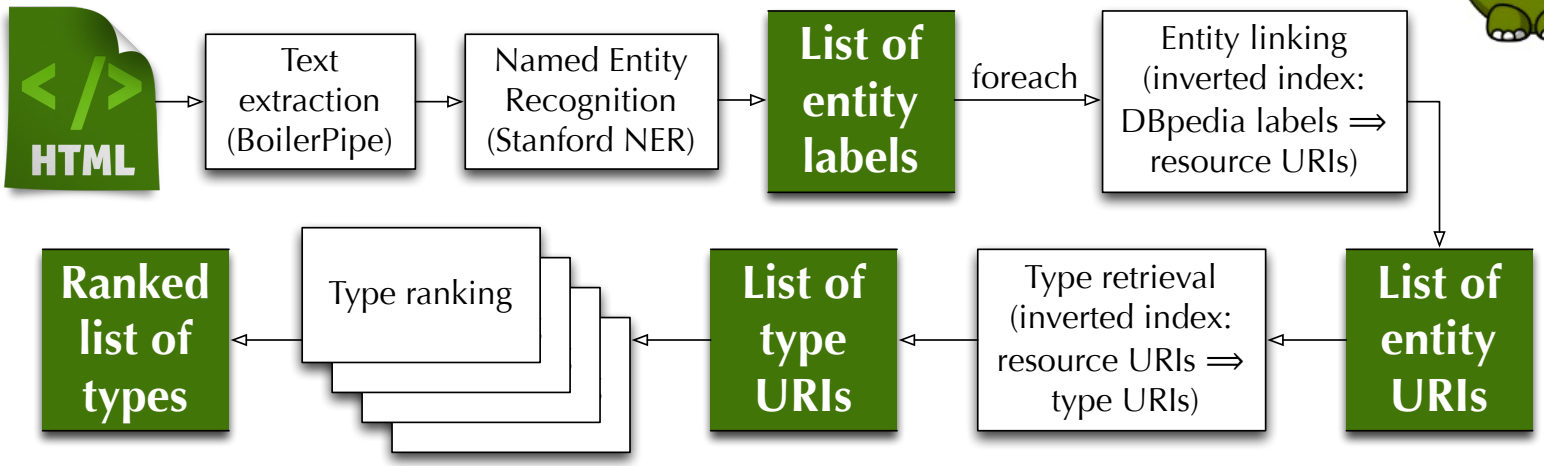
ZenCrowd Summary

- ZenCrowd: Probabilistic reasoning over automatic and crowdsourcing methods for entity linking
- Standard crowdsourcing improves 6% over automatic
- 4% - 35% improvement over standard crowdsourcing
- 14% average improvement over automatic approaches

http://exascale.info/zencrowd/

On top of Entity Linking

- Follow-up works:
  - Also used for instance matching across datasets
  - Ranking Entity *Types*: see TRank at ISWC13, nominated for best paper award



Alberto Tonon, Michele Catasta, Gianluca Demartini, Philippe Cudré-Mauroux, and Karl Aberer. TRank: Ranking Entity Types Using the Web of Data. In: The 12th International Semantic Web Conference (**ISWC 2013**). Sydney, Australia, October 2013.

Three-stage blocking with the Crowd for Data Integration

- 1. Cheap clustering/inverted index selection of candidates
- 2. Expensive similarity measure
- 3. Crowdsource low confidence matches

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Large-Scale Linked Data Integration Using Probabilistic Reasoning and Crowdsourcing. In: **VLDB Journal**, Volume 22, Issue 5 (2013), Page 665-687, Special issue on Structured, Social and Crowd-sourced Data on the Web. October 2013.

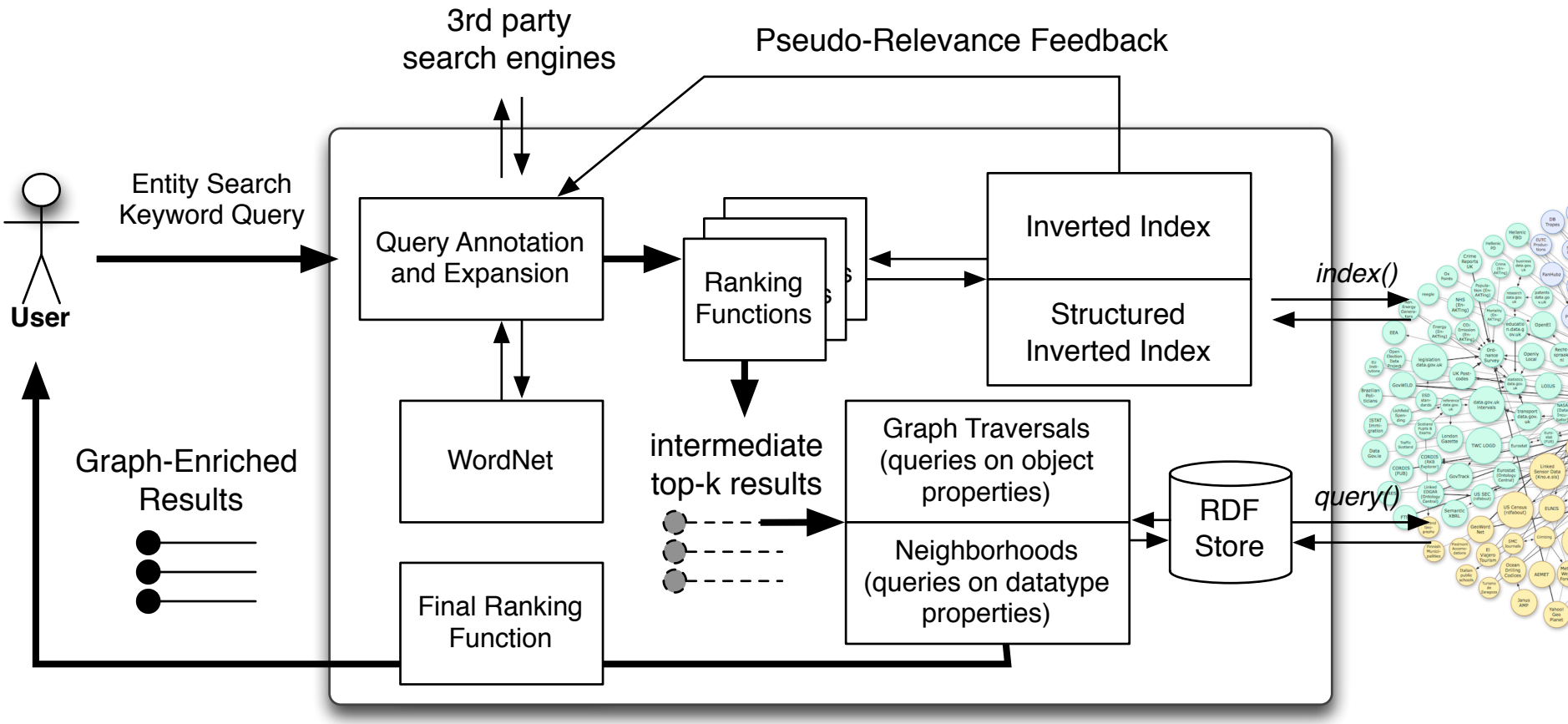
Entity Ranking

Ad-hoc Object Retrieval

- Once entities have been identified…
- We want to rank them as answer to a query

- AOR
  - Given the description of an entity
  - give me back its identifier
  - Input: query *q*, data graph *G*
  - Output: ranked list of URIs from *G*

An Hybrid Approach to AOR



Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining Inverted Indices and Structured Search for Ad-hoc Object Retrieval. In: 35th Annual ACM SIGIR Conference (**SIGIR 2012**).

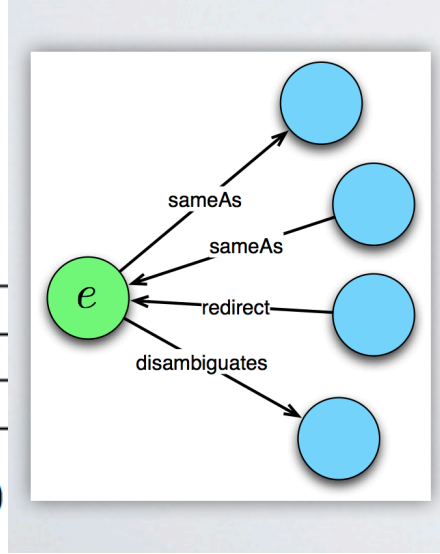
AOR Evaluation

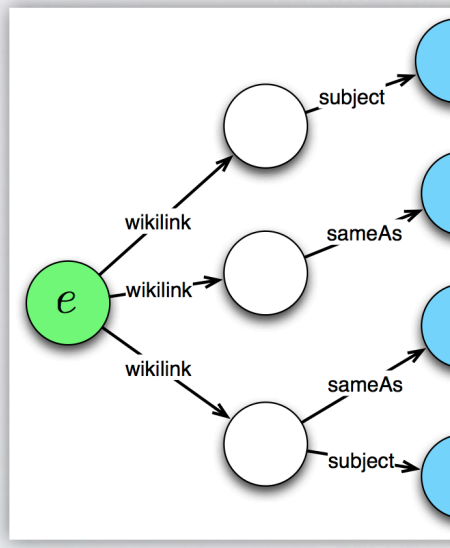
- 1.3 billions RDF triples from LOD cloud
- 92 and 50 queries
- Crowdsourced relevance judgments

- semsearch.yahoo.com

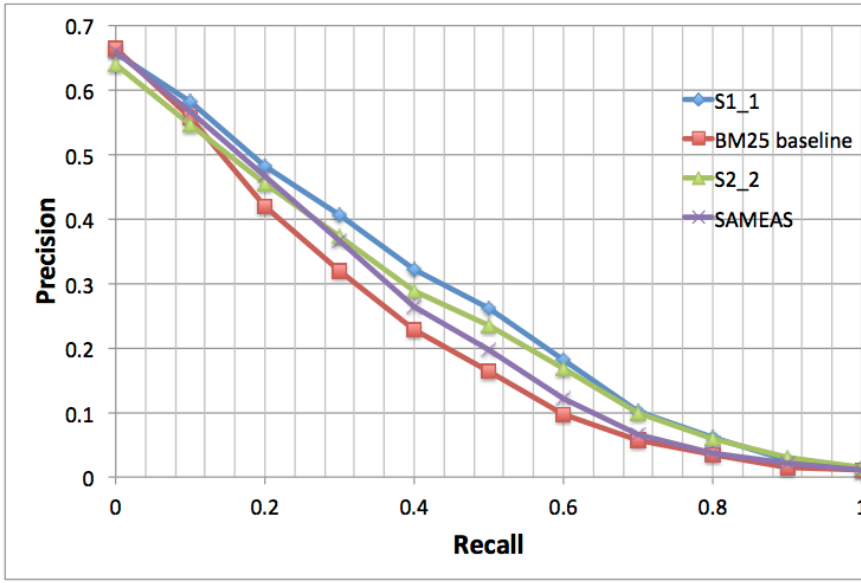
Evaluation Results

	2010 Collection	
	MAP	P10
BM25	0.2070	0.3348
SAMEAS	0.2293* (+11%)	0.363* (+8%)
S1_1	**0.2586*** (+25%)	**0.3848*** (+15%)
S1_2	0.2305* (+11%)	0.3217 (-4%)
S1_3	0.2306* (+11%)	0.3217 (-4%)
S2_1	0.2118 (+2%)	0.3370 (+1%)
S2_2	0.2118 (+2%)	0.3370 (+1%)
S2_3	0.2113 (+2%)	0.3402 (+2%)





Approach	IR time	RDF time	Total time
BM25 Baseline	285	-	285
Extension	580	-	580 (+104%)
Query Autoc.	1447	-	1447 (+408%)
PRF3	2670	-	2670 (+837%)
SAMEAS	285	30	315 (+11%)
S1_1	285	48	333 (+17%)
S1_2	285	84	369 (+29%)
S1_3	285	86	371 (+30%)
S2_1	285	1746	2031 (+613%)
S2_2	285	2192	2477 (+769%)
S2_3	285	105	390 (+37%)



Summary

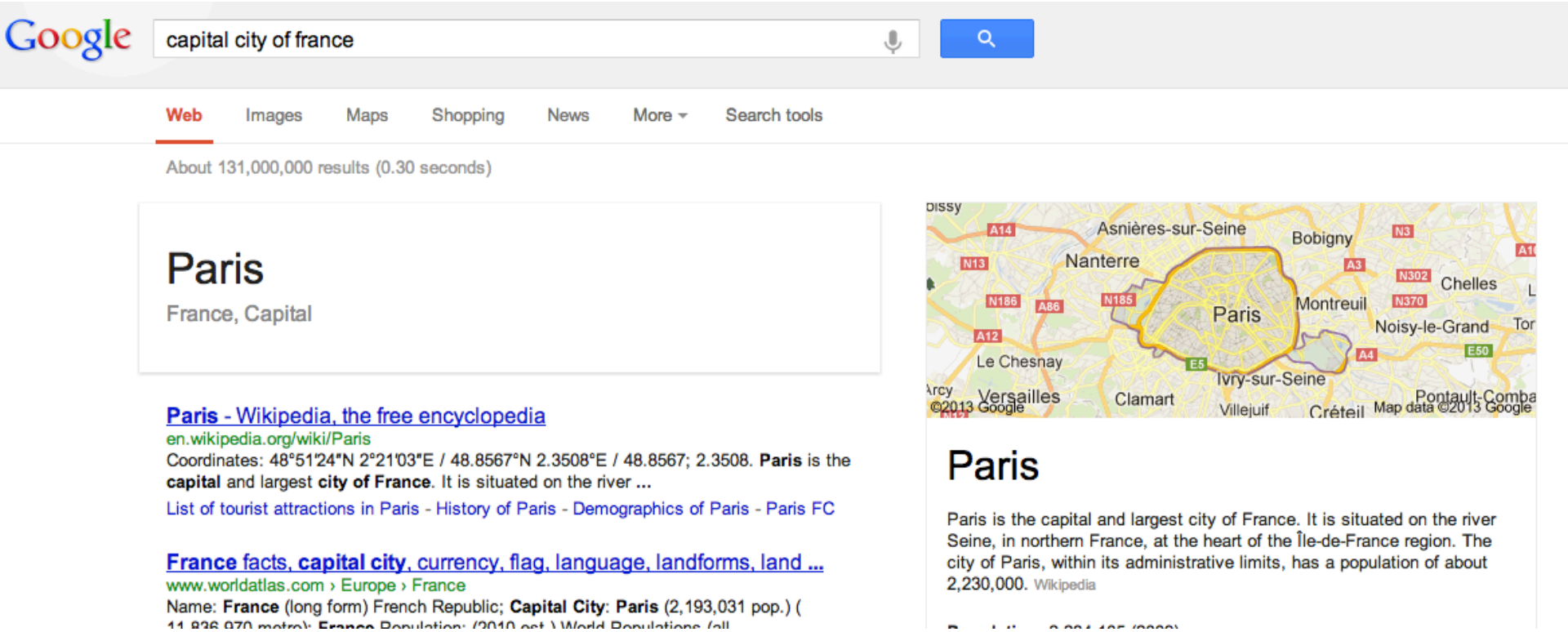
- AOR = *"Given the description of an entity, give me back its identifier"*
- Combining classic IR techniques + structured database storing graph data
- Significantly better results (up to +25% MAP over BM25 baseline).
- Overhead caused from the graph traversal part is limited

http://exascale.info/AOR/

CrowdQ: Crowdsourced Query Understanding

birthdate of mayor of capital city of france

capital city of france



mayor of paris

Google mayor of paris

Web Images Maps Shopping More Search tools

About 98,900,000 results (0.20 seconds)

Bertrand Delanoë

Paris, Mayor

List of **mayors of Paris** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/List_of_**mayors_of_Paris**
Before the French Revolution, the municipality of **Paris** was headed by the provost of the merchants (prévôt des marchands). On 14 July 1789, at the end of the ...

Bertrand Delanoë - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Bertrand_Delanoë
Delanoë has been **mayor of Paris** since 18 March 2001, when control of the city council ... Delanoë won the **mayorship of Paris**, at the head of a coalition of Socialists, ... City Mayors: **Bertrand Delanoe – Mayor of Paris**; ^ "Psychiatric tests for ...
Early life - Career - Personal life - Political career

Letter to the **mayor of Paris**, re: No skyscrapers! | Architecture Here ...
blogs.providencejournal.com/.../letter-to-the-**mayor-of-paris**-re-no-...
Nov 24, 2012 – Anyway, here is my letter: Nov. 24, 2012. Mr. **Bertrand Delanoe Mayor of Paris** Place de l'Hotel de Ville Paris, France 75004. Dear Mr. Delanoe: ...

Bertrand Delanoë

Bertrand Delanoë is a French politician, and has been the mayor of Paris since 2001. He is member of the Socialist Party. Delanoë was born in Tunis, Tunisia to a French-Tunisian father and a French mother. Wikipedia

Born: May 30, 1950 (age 62), Tunis

Party: Socialist Party

People also search for

Martine Aubry

Jacques Chirac

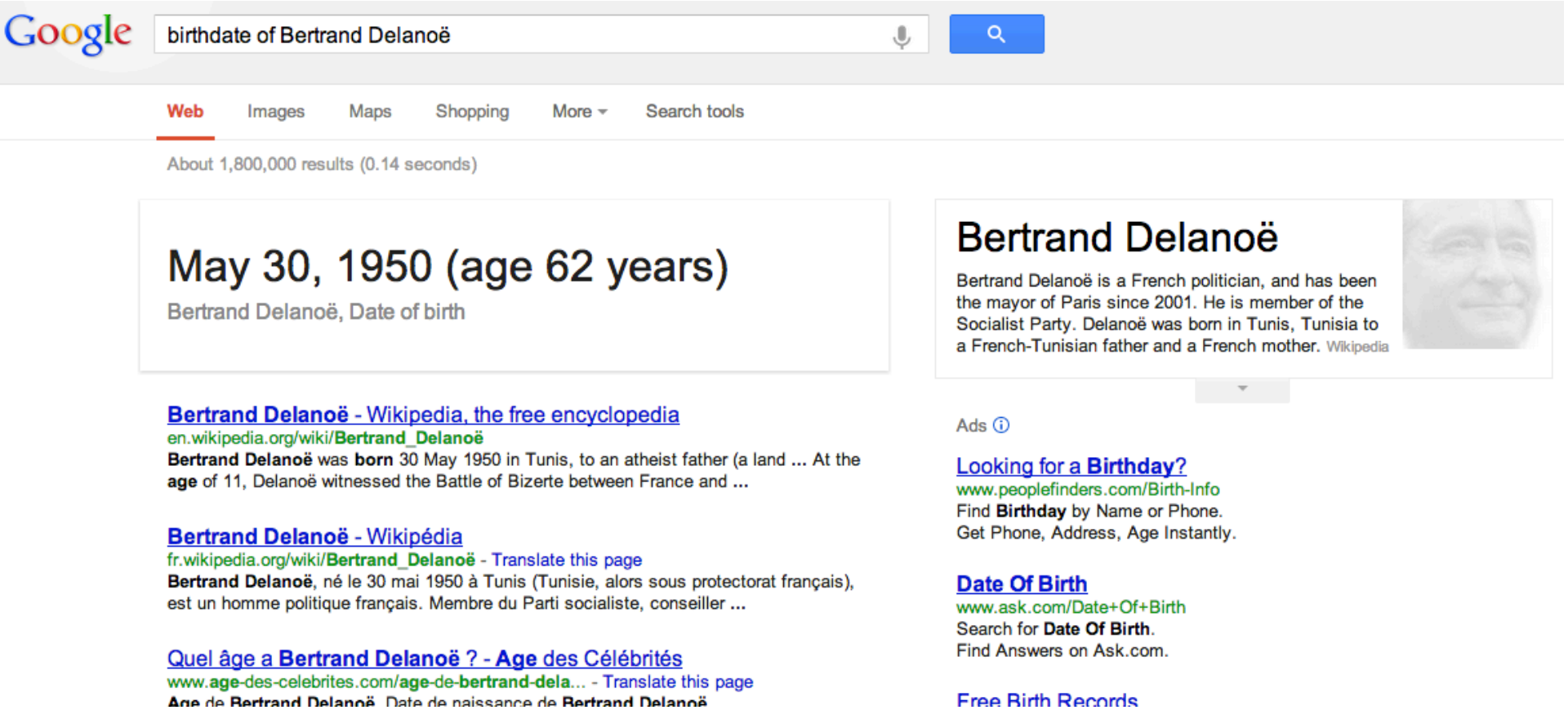
Jean Tiberi

Ségolène Royal

Lionel Jospin

Feedback / More info

birthdate of Bertrand Delanoë



Motivation

- Web Search Engines can answer simple factual queries directly on the result page
- Users with complex information needs are often unsatisfied
- Purely automatic techniques are not enough
- We want to solve it with Crowdsourcing!

CrowdQ

- CrowdQ is the first system that uses crowdsourcing to
  - *Understand* the intended meaning
  - *Build* a structured query template
  - *Answer* the query over Linked Open Data

Gianluca Demartini, Beth Trushkowsky, Tim Kraska, and Michael Franklin. CrowdQ: Crowdsourced Query Understanding. In: 6th Biennial Conference on Innovative Data Systems Research (**CIDR 2013**).

Hybrid Human-Machine Pipeline

Q= birthdate of actors of forrest gump

Query annotation	Noun	Noun	Named entity

Verification	Is forrest gump this entity in the query?
Entity Relations	Which is the relation between: actors and forrest gump → starring
Schema element	Starring → <dbpedia-owl:starring>
Verification	Is the relation between: **Indiana Jones – Harrison Ford** **Back to the Future – Michael J. Fox** of the same type as **Forrest Gump - actors**

Structured query generation

Q= birthdate of actors of fo MOVIE mp

```
SELECT ?y ?x
WHERE { ?y <dbpedia-owl:birthdate> ?x .
        ?z <dbpedia-owl:starring> ?y .
        ?z <rdfs:label> 'Fo MOVIE mp' }
```

Results from BTC09:

```
<http://dbpedia.org/resource/Robin_Wright_Penn> 1966-04-08
<http://dbpedia.org/resource/Tom_Hanks> 1956-07-09
<http://dbpedia.org/resource/Sally_Field> 1946-11-06
<http://dbpedia.org/resource/Gary_Sinise> 1955-03-17
<http://dbpedia.org/resource/Mykelti_Williamson> 1960-03-04
```

Conclusions

- Structured Data make Web Search better
- Exploit the best out of structured and unstructured data (Hybrid AOR)

- Crowd can help in understanding semantics
- Hybrid human-machine systems (ZenCrowd)
- Exploit Human Intelligence at Scale (CrowdQ)

gianlucademartini.net demartini@exascale.info