Entities, Graphs, and Crowdsourcing for better Web Search

Gianluca Demartini
eXascale Infolab
University of Fribourg, Switzerland

gianluca demartini.net
exascale.info
Gianluca Demartini

- 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
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 ˈmæpəθɔːr; 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

Tom Cruise | Facebook
https://www.facebook.com/officialtomcruise
Tom Cruise. 3863109 likes · 76966 talking about this. Welcome to the Official www.TomCruise.com team Facebook page! Tom Cruise news, events, pics & video ...

Tom Cruise
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
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

2007  
2008  
2009  
2011

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
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.

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.
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
ZenCrowd

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

Input

ZenCrowd

Entity Extractors

Algorithmic Matchers

Decision Engine

Probabilistic Network

LOD Index

Output

HTML Pages

HTML+ RDFa Pages

Crowdsourcing Platform

Micro Matching Tasks

Workers Decisions

LOD Open Data Cloud

Entity Factor Graphs

- **Graph components**
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints

- **Probabilistic Inference**
  - Select all links with posterior prob $> \tau$

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)

<table>
<thead>
<tr>
<th></th>
<th>US Workers</th>
<th>Indian Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P  R  A</td>
<td>P  R  A</td>
</tr>
<tr>
<td>GL News</td>
<td>0.84  0.87 0.90</td>
<td>0.67  0.64 0.78</td>
</tr>
<tr>
<td>US News</td>
<td>0.64  0.68 0.78</td>
<td>0.55  0.63 0.71</td>
</tr>
<tr>
<td>IN News</td>
<td>0.84  0.82 0.89</td>
<td>0.75  0.77 0.80</td>
</tr>
<tr>
<td>SW News</td>
<td>0.72  0.80 0.85</td>
<td>0.61  0.62 0.73</td>
</tr>
<tr>
<td>All News</td>
<td>0.80  0.81 0.88</td>
<td>0.64  0.62 0.76</td>
</tr>
</tbody>
</table>
Worker Selection

![Graph showing worker selection](image)

- **Worker Selection**
- **Number of Tasks**
- **Worker Precision**
- **US Workers**
- **IN Workers**

![Graph showing precision vs top K workers](image)

- **Precision**
- **Top K workers**

Gianluca Demartini
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)
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

Three-stage blocking with the Crowd for Data Integration

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

AOR Evaluation

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

• semsearch.yahoo.com
# Evaluation Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>IR time</th>
<th>RDF time</th>
<th>Total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25 Baseline</td>
<td>285</td>
<td>-</td>
<td>285</td>
</tr>
<tr>
<td>Extension</td>
<td>580</td>
<td>-</td>
<td>580 (+104%)</td>
</tr>
<tr>
<td>Query Autoc. PRF3</td>
<td>1447</td>
<td>-</td>
<td>1447 (+408%)</td>
</tr>
<tr>
<td>SAMEAS</td>
<td>285</td>
<td>30</td>
<td>315 (+11%)</td>
</tr>
<tr>
<td>S1.1</td>
<td>285</td>
<td>48</td>
<td>333 (+17%)</td>
</tr>
<tr>
<td>S1.2</td>
<td>285</td>
<td>84</td>
<td>369 (+29%)</td>
</tr>
<tr>
<td>S1.3</td>
<td>285</td>
<td>86</td>
<td>371 (+30%)</td>
</tr>
<tr>
<td>S2.1</td>
<td>285</td>
<td>1746</td>
<td>2031 (+613%)</td>
</tr>
<tr>
<td>S2.2</td>
<td>285</td>
<td>2192</td>
<td>2477 (+769%)</td>
</tr>
<tr>
<td>S2.3</td>
<td>285</td>
<td>105</td>
<td>390 (+37%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approach</th>
<th>MAP</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.2070</td>
<td>0.3348</td>
</tr>
<tr>
<td>SAMEAS</td>
<td>0.2293* (+11%)</td>
<td>0.363* (+8%)</td>
</tr>
<tr>
<td>S1.1</td>
<td>0.2586* (+25%)</td>
<td>0.3848* (+15%)</td>
</tr>
<tr>
<td>S1.2</td>
<td>0.2305* (+11%)</td>
<td>0.3217 (-4%)</td>
</tr>
<tr>
<td>S1.3</td>
<td>0.2306* (+11%)</td>
<td>0.3217 (-4%)</td>
</tr>
<tr>
<td>S2.1</td>
<td>0.2118 (+2%)</td>
<td>0.3370 (+1%)</td>
</tr>
<tr>
<td>S2.2</td>
<td>0.2118 (+2%)</td>
<td>0.3370 (+1%)</td>
</tr>
<tr>
<td>S2.3</td>
<td>0.2113 (+2%)</td>
<td>0.3402 (+2%)</td>
</tr>
</tbody>
</table>
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

[Search results from Google]

France Cities Map, Major Cities of France
www.mapsofworld.com › France › Facts
Dec 21, 2012 – Paris is the capital of France, as well as the largest city in the country with a population of over 2 million, and a metropolitan area population of ...

Nice - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Nice
Giuseppe Garibaldi, born in Nice, strongly opposed the cession to France (arguing ... Two men dominated this period: Jean Médecin, mayor for 33 years from 1928 ... However, it is also the largest city in France that is not a regional capital; the ...

Rouen - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Rouen
In the 14th century urban strife threatened the city: in 1291, the mayor was ... Rouen became the capital city of English power in occupied France and when the duke .... Isaac Oliver (c.1560–1617), French-born English portrait miniature painter ...
mayor of paris
birthdate of Bertrand Delanoë

May 30, 1950 (age 62 years)

Bertrand Delanoë, Date of birth

Bertrand Delanoë - Wikipedia, the free encyclopedia
Bertrand Delanoë was born 30 May 1950 in Tunis, to an atheist father (a land ... At the age of 11, Delanoë witnessed the Battle of Bizerte between France and ... 

Bertrand Delanoë - Wikipédia
Bertrand Delanoë, né le 30 mai 1950 à Tunis (Tunisie, alors sous protectorat français), est un homme politique français. Membre du Parti socialiste, conseiller ...

Quel âge a Bertrand Delanoë ? - Age des Célébrités
Age de Bertrand Delanoë, Date de naissance de 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

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

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 Forrest Gump

SELECT ?y ?x
  ?z <rdfs:label> ‘Forrest Gump’ }
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)

gianluca demartini.net  demartini@exascale.info