Crowdsourcing for Entity-Centric Information Access

Gianluca Demartini
University of Sheffield, UK
gianlucadademartini.net
Entity-Centric Information Access

About 153,000,000 results (0.26 seconds)

In the news

Readers' Poll: The 10 Best Tom Cruise Movies
RollingStone.com - 13 hours ago
Tom Cruise's recent string of bombs coupled with embarrassing revelations about his role in ...

Tom Cruise & Suri: Why He Hasn't Seen Her In Two Years — Report
Hollywood Life - 16 hours ago

Mission Impossible 5 review: Tom Cruise offers us insane fun
Hindustan Times - 3 hours ago

More news for tom cruise

Tom Cruise
Actor · tomcruise.com

Tom Cruise is an American actor and filmmaker. Cruise 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 Endless Love. Wikipedia

Born: July 3, 1962 (age 53), Syracuse, New York, United States
Height: 1.70 m
Full name: Thomas Cruise Mapother IV
Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)
Children: Suri Cruise, Connor Antony Cruise, Isabella Jane Cruise
Tom Cruise is an American actor and producer who made his film debut with a minor role... As of 2015, Cruise has reprised his role as Hunt in four more films in the "..." "Movie Review: Utility Vehicle: 'Days of Thunder': The NASCAR racing footage and Tom Cruise's grin are fine. ... "The Last Samurai Movie Review (2003)". Rock of Ages - Ask the Dust (film) - Losin' It.

Tom Cruise is an American actor and filmmaker. Cruise 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 Endless Love. 

Born: July 3, 1962 (age 53), Syracuse, New York, United States
Height: 1.70 m
Full name: Thomas Cruise Mapother IV
• Entity-seeking queries make up 40-50% of the query volume
  – Jeffrey Pound, Peter Mika, Hugo Zaragoza: Ad-hoc object retrieval in the web of data. WWW 2010: 771-780
  – Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, Ariel Fuxman: Active objects: actions for entity-centric search. WWW 2012: 589-598

• Show a summary of the most likely information-needs
  – Including related entities for navigation
  – Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, Nicolas Torzec: Entity Recommendations in Web Search. ISWC 2013
Web of Data

• Freebase
  – Acquired by Google in July 2010.
  – Knowledge Graph launched in May 2012.
  – Read-only in December 2014 -> WikiData

• Schema.org
  – Driven by major search engine companies
  – Machine-readable annotations of Web pages

• Linked Open Data
  – 31 billion triples, Sept 2011
  – 90 billion triples, Aug 2015 (stats.lod2.eu)
Linked Open Data

Z. Kaoudi and I. Manolescu, ICDE seminar 2013
Today I will talk about

• Crowdsourcing
  – Amazon MTurk as a crowdsourcing platform

• Entity Linking on the Web
  – With the crowd

• Finding the best type for an entity appearing in Web pages
Overview

from http://www.bbc.co.uk/news/magazine-32993891
Crowdsourcing

• *Portmanteau* of "crowd" and "outsourcing," first coined by Jeff Howe in a June 2006 Wired magazine article

• [Merriam-Webster] the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers
Crowdsourcing

- Leverage human intelligence at scale 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

- See my tutorial at ESWC 2013 and ISWC 2013
5-year Analysis of MTurk workload

• Mturk-tracker.com
  – Collects metadata about each visible batch (Title, description, rewards, required qualifications, HITs available, etc), that is, set of similar tasks or HITs
  – Records batch progress (every ~20 minutes)
  – Covers 130M tasks

Country-Specific HITs

Workers from US, India and Canada are the most sought after.
Distribution of *Batch Size*

"Power-law"
Evolution of Batch Sizes

Very large batches start to appear
HIT Pricing

5-cents is the new 1-cent
Requesters and Reward Evolution

Increasing number of New and Distinct Requesters
### Top MTurk Requesters last week

<table>
<thead>
<tr>
<th>Requester name</th>
<th>Hits</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speechpad</td>
<td>41092</td>
<td>$261,239.88</td>
</tr>
<tr>
<td>WorkFusion</td>
<td>2668</td>
<td>$11,552.00</td>
</tr>
<tr>
<td>Jon Breig</td>
<td>115360</td>
<td>$6,000.16</td>
</tr>
<tr>
<td>CastingWords</td>
<td>11037</td>
<td>$5,953.96</td>
</tr>
<tr>
<td>Mark Yatskar</td>
<td>45008</td>
<td>$3,364.97</td>
</tr>
<tr>
<td>VidAngel</td>
<td>181</td>
<td>$2,881.54</td>
</tr>
<tr>
<td>p9r</td>
<td>41754</td>
<td>$2,502.03</td>
</tr>
<tr>
<td>Amazon Requester Inc - browse</td>
<td>35227</td>
<td>$2,113.62</td>
</tr>
<tr>
<td>classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>152</td>
<td>$1,976.00</td>
</tr>
</tbody>
</table>
Distribution of HIT Types

Less Content Access batches

Content Creation being the most popular
Supply Elasticity

Intercept = 2.5
Slope = 0.5%

20% of new work gets completed within an hour
Summary

• HIT reward has increased over time
• **Audio transcription** is the most popular task
• Demand for Indian workers has decreased
• **Surveys** are most popular for US workers
• 1000 new requesters per month join
• 10K new HITs arrive and 7.5K HITs get completed every hour

• Check #mturkdynamics for the main findings
Crowdsourcing for Entity Linking
Facebook Buys Instagram for $1 Billion

By Evelyn 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.

HTML:
<p>Facebook is not waiting for its initial public offering to make its first big purchase.</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.

RDFa enrichment

<spant about="http://dbpedia.org/resource/Facebook">Facebook</spant> is not waiting for its initial public offering to make its first big purchase.</p>

In its largest acquisition to date, the social network has purchased <spant about="http://dbpedia.org/resource/Instagram">Instagram</spant>, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.

http://dbpedia.org/resource/Instagram

http://dbpedia.org/resource/Facebook

owl:sameAs

fbase:Instagram

Google

Android
ZenCrowd

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

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

- Entity Linking with Crowdsourcing and agreement vote (at least 2 out of 5 workers select the same URI)

<table>
<thead>
<tr>
<th>Document Type</th>
<th>US Workers</th>
<th>Indian Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>GL News</td>
<td>0.79</td>
<td>0.85</td>
</tr>
<tr>
<td>US News</td>
<td>0.52</td>
<td>0.61</td>
</tr>
<tr>
<td>IN News</td>
<td>0.62</td>
<td>0.76</td>
</tr>
<tr>
<td>SW News</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td>All News</td>
<td><strong>0.74</strong></td>
<td><strong>0.82</strong></td>
</tr>
</tbody>
</table>

Top-1 precision: 0.70
Experimental Evaluation

• Entity Linking with ZenCrowd
  – Training with first 5 entities + 5% afterwards
  – 3 consecutive bad answers lead to blacklisting

<table>
<thead>
<tr>
<th>Document</th>
<th>US Workers</th>
<th>Indian Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>GL News</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>US News</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>IN News</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>SW News</td>
<td>0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>All News</td>
<td>0.80</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Experimental Evaluation

• 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 WWW2013 and CHI2015 papers)
### Behavioral Patterns of Malicious Workers

<table>
<thead>
<tr>
<th>Category</th>
<th>Instruction</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ineligible Workers (IW)</td>
<td>Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.</td>
<td>‘this is my first task’</td>
</tr>
<tr>
<td>Fast Deceivers (FD)</td>
<td>eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.</td>
<td>‘What’s your task?’ , ‘adasd’, ‘fgfgf gsd ljlkj’</td>
</tr>
<tr>
<td>Rule Breakers (RB)</td>
<td>Identify 5 keywords that represent this task (separated by commas).</td>
<td>‘survey, tasks, history’, ‘previous task yellow’</td>
</tr>
<tr>
<td>Smart Deceivers (SD)</td>
<td>Identify 5 keywords that represent this task (separated by commas).</td>
<td>‘one, two, three, four, five’</td>
</tr>
<tr>
<td>Gold Standard Preys (GSP)</td>
<td>These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions!</td>
<td></td>
</tr>
</tbody>
</table>

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/](http://exascale.info/zencrowd/)
- Follow up-work (VLDBJ, 2013):
  - Also used for instance matching across datasets
  - 3-way blocking with the crowd
Entity Types
...and Why Types?

- “Summarization” of texts

<table>
<thead>
<tr>
<th>Article Title</th>
<th>Entities</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin Laden Relative Pleads Not Guilty in Terrorism Case</td>
<td>Osama Bin Laden</td>
<td>Al-QaedaPropagandists</td>
</tr>
<tr>
<td></td>
<td>Abu Ghaith</td>
<td>Kuwaiti Al-Qaeda members</td>
</tr>
<tr>
<td></td>
<td>Lewis Kaplan</td>
<td>Judge</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>Borough (New York City)</td>
</tr>
</tbody>
</table>

- Contextual entities summaries in Web-pages

- Disambiguation of other entities

- Diversification of search results
Entities May Have Many Types

- American Billionaires
- People from King County
- People from Seattle
- Windows People
- Agent
- Person
- Harvard University People
- American People of Scottish Descent
- Living People
- American Philanthropists
- American Computer Programmers
- Thing
Ranking Algorithms

• Entity centric
• Hierarchy-based
• Context-aware (featuring type-hierarchy)
• Learning to Rank
Hierarchy-Based Approaches (An Example)

- **ANCESTORS**

  \[ \text{Score}(e, t) = \text{number of } t \text{'s ancestors in the type hierarchy contained in } T_e. \]
Context-Aware Ranking Approaches (An Example)

• SAMETYPE

\[ \text{Score}(e, t, c_T) = \text{number of times } t \text{ appears among the types of every other entity in } c_T. \]
Learning to Rank Entity Types

Determine an optimal combination of all our approaches:

- Decision trees
- Linear regression models
- 10-fold cross validation
Datasets

• 128 recent NYTimes articles split to create:
  – *Entity Collection*
  – *Sentence Collection*
  – *Paragraph Collection*
  – *3-Paragraphs Collection*

• Ground-truth obtained by using crowdsourcing
  – 3 workers per entity/context
  – 4 levels of relevance for each type
  – *Overall cost: 190$*
# Effectiveness Evaluation

<table>
<thead>
<tr>
<th>Approach</th>
<th>Entity-only</th>
<th>Sentence</th>
<th>Paragraph</th>
<th>3-Paragraphs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NDCG</td>
<td>MAP</td>
<td>NDCG</td>
<td>MAP</td>
</tr>
<tr>
<td>FREQ</td>
<td>0.6284</td>
<td>0.4659</td>
<td>0.5409</td>
<td>0.3758</td>
</tr>
<tr>
<td>WIKILINK-OUT</td>
<td>0.6874</td>
<td>0.5406</td>
<td>0.6050</td>
<td>0.4521</td>
</tr>
<tr>
<td>WIKILINK-IN</td>
<td>0.6832</td>
<td>0.5342</td>
<td>0.5907</td>
<td>0.4213</td>
</tr>
<tr>
<td>SAMEAS</td>
<td>0.6848</td>
<td>0.5328</td>
<td>0.6049</td>
<td>0.4310</td>
</tr>
<tr>
<td>LABEL</td>
<td>0.6672</td>
<td>0.5067</td>
<td>0.6075</td>
<td>0.4265</td>
</tr>
<tr>
<td>SAMETYPE</td>
<td>-</td>
<td>-</td>
<td>0.6024</td>
<td>0.4452</td>
</tr>
<tr>
<td>PATH</td>
<td>-</td>
<td>-</td>
<td>0.6507</td>
<td>0.4956</td>
</tr>
<tr>
<td>DEPTH</td>
<td>0.7432</td>
<td>0.6128</td>
<td>0.6754</td>
<td>0.5385</td>
</tr>
<tr>
<td>ANCESTORS</td>
<td>0.7424</td>
<td>0.6154</td>
<td>0.6967</td>
<td>0.5637</td>
</tr>
<tr>
<td>ANC.DEPTH</td>
<td>0.7469</td>
<td>0.6236</td>
<td>0.6832</td>
<td>0.5488</td>
</tr>
<tr>
<td>DEC-TREE</td>
<td>0.7614</td>
<td>0.6361</td>
<td>0.7373</td>
<td>0.6079</td>
</tr>
<tr>
<td>LIN-REG</td>
<td>0.7373</td>
<td>0.6079</td>
<td>0.6906</td>
<td>0.5579</td>
</tr>
</tbody>
</table>

Use TRank:
Open Source (Scala)
https://github.com/MEM0R1ES/TRank
Web Service (JSON)
http://trank.exascale.info
Efficiency Evaluation

• Tested efficiency on a CommonCrawl sample of 1TB
  – 1,310,459 HTML pages
  – 23GB compressed

• Map/Reduce on a cluster of 8 machines with 12 cores, 32GB of RAM and 3 SATA disks

• On average, 25 min. processing time (> 100 docs/node x sec)

<table>
<thead>
<tr>
<th>Text Extraction</th>
<th>NER</th>
<th>Entity Linking</th>
<th>Type Retrieval</th>
<th>Type Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.9%</td>
<td>35.6%</td>
<td>29.5%</td>
<td>9.8%</td>
<td>6.2%</td>
</tr>
</tbody>
</table>
Conclusions

• New task: ranking entity types.
  – Useful for: “summarization” of Web-documents, entity summaries, disambiguation.

• Various approaches: entity-centric, context-aware, hierarchy-based, learning to rank.
  – Hierarchy-based and learning to rank are the most effective.

• Hadoop, Map/Reduce, and inverted indices to achieve scalability.
Summary

• Hybrid human-machine systems can
  – Scale over large amounts of data
  – Reach high accuracy by keeping humans in the loop

• Entities are the new entry point to Web content
  – “Things not string”
  – Google Knowledge Graph (but also Bing, Yahoo!, Yandex)

• Users can benefit from entity-centric search, browsing, and exploration of the Web