Human Computation for Entity-Centric Information Access

Gianluca Demartini
gianlucademartini.net
Gianluca Demartini

- B.Sc., M.Sc. at U. of Udine, Italy
- Ph.D. at U. of Hannover, Germany
  - Entity Retrieval
- Worked at the eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
- Senior Lecturer in Data Science at the iSchool, U. of Sheffield

www.gianlucaodemartini.net
Research Interests

• **Entity-centric Information Access** (2005-now)
  – Structured/Unstructured data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  – NER in Scientific Docs (WWW 14), Prepositions (CIKM 14)
  – IR Evaluation (ECIR 16 Best Paper Award, IRJ 2015)

• **Hybrid Human-Machine Systems** (2012-now)
  – ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  – Human Memory based Systems (WWW 14, PVLDB)
  – Hybrid systems overview (COMNET, 2015)

• **Better Crowdsourcing Platforms** (2013-now)
  – Platform Dynamics (WWW 15)
  – Pick-a-Crowd (WWW 13), Malicious Workers (CHI 15)
  – Scale-up Crowdsourcing (HCOMP 14), Scheduling (WWW 16)

Supported by: H2020, EPSRC, U Sheffield, Amazon
Entity-Centric Information Access

Official Tom Cruise: Edge Of Tomorrow, Movies, Bio, News ...
www.tomcruise.com/
OFFICIAL TOM CRUISE SITE: View the latest EDGE OF TOMORROW trailer! Watch career movie trailers, videos, and retrospective. Read the Tom Cruise ...

Tom Cruise - IMDb
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: Top Gun. If you had told fourteen-year-old Franciscan seminary student Thomas Cruise Mapother IV that one day in the not-too-distant future ...

Tom Cruise - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Tom_Cruise
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 ...
Tom Cruise filmography - Mimi Rogers - Katie Holmes - Nicole Kidman

In the news

Scientologist who worked with Tom Cruise condemned to horrific work camp over lesbian kiss
PinkNews - 2 days ago
A former Scientologist, who worked with celebrities like Tom Cruise and John Travolta, has ...

Jerry Bruckheimer confirms Tom Cruise is signed up for Top Gun 2

Tom Cruise
Actor

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: 1.7 m
Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)
Children: Connor Cruise, Isabella Cruise, Suri Cruise, Oli

In the news

Scientologist who worked with Tom Cruise condemned to horrific work camp over lesbian kiss
PinkNews - 2 days ago
A former Scientologist, who worked with celebrities like Tom Cruise and John Travolta, has ...
• 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)
Today I will talk about

- Human Computation
  - Amazon MTurk as a crowdsourcing platform
  - Hybrid Human-Machine Information Systems
- Entity Linking on the Web
  - With the crowd
- Efficiency/effectiveness of Human Computation
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

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
  - 2009-2014

Workers from US, India and Canada are the most sought after.

http://exascale.info/mturk-mrkt/ for interactive visualizations
Distribution of *Batch Size*
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 over 1 month

<table>
<thead>
<tr>
<th>Requester name</th>
<th>hits</th>
<th>reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speechpad</td>
<td>23857</td>
<td>$172,994.63</td>
</tr>
<tr>
<td>Percy Liang</td>
<td>883</td>
<td>$7,320.48</td>
</tr>
<tr>
<td>Princeton Vision</td>
<td>51187</td>
<td>$5,762.44</td>
</tr>
<tr>
<td>Stanford GSB Behavioral Lab</td>
<td>3749</td>
<td>$2,110.70</td>
</tr>
<tr>
<td>Chris Callison-Burch</td>
<td>8157</td>
<td>$2,064.29</td>
</tr>
<tr>
<td>RC.org Mechanical Turk</td>
<td>6591</td>
<td>$2,011.33</td>
</tr>
<tr>
<td>VacationrentalAPI</td>
<td>399</td>
<td>$1,373.50</td>
</tr>
<tr>
<td>Med Expertise</td>
<td>869</td>
<td>$1,303.50</td>
</tr>
<tr>
<td>Bluejay Labs</td>
<td>13613</td>
<td>$1,288.59</td>
</tr>
<tr>
<td>YL Testing</td>
<td>1051</td>
<td>$1,236.83</td>
</tr>
</tbody>
</table>
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 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.
ZenCrowd

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

ZenCrowd

Entity Extractors

Algorithmic Matchers

Decision Engine

Probabilistic Network

LOD Index

Crowdsourcing Platform

HTML Pages

Input

Workers Decisions

Micro-Task Manager

LOD Open Data Cloud

Output

HTML+ RDFa Pages

Entity Factor Graphs

• Graph components
  – Workers, links, clicks
  – Prior probabilities
  – Link Factors
  – Constraints

• Probabilistic Inference
  – Select all links with posterior prob $\tau$
Experimental Evaluation

- Worker Selection
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

• Follow up-work (VLDBJ, 2013):
  – Also used for instance matching across datasets
  – 3-way blocking with the crowd
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)

• Need to make sure high priority tasks are completed fast (see WWW2016 paper)
### Behavioral Patterns of Malicious Workers

<table>
<thead>
<tr>
<th>Role</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 ijikj’</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>

Scheduling HITs

Platform throughput is dominated by large HIT batches
HIT-Bundle

Definition

• Scheduling requires control over the serving process of tasks

• A **HIT-Bundle** is a batch that contains heterogeneous tasks

• All tasks that are generated by the system are published through the HIT-Bundle
Scheduling HITs

• Fair Scheduling
  – Priority of HITs but avoid starvation
  – Assign HITs of the same type (no context switch)

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 Vault (but also Bing, Yahoo!, Yandex)

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