Introduction to Crowdsourcing

Lecture 1

Gianluca Demartini

University of Sheffield
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)
- Lecturer in Data Science at the iSchool, U. of Sheffield, Aug 14
- Tutorial on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013 and ISWC 2013
- Research Interests
  - Information Retrieval, Semantic Web, Human Computation
www.gianluca demartini.net

g.demartini@sheffield.ac.uk
Who are you?
Tentative Menu

• Monday
• Lecture 1 - Introduction to Crowdsourcing
  – An overview of the entire course.
  – Early examples of crowdsourcing (reCAPTCHA, ESP game).
  – Types of incentives: games with a purpose, citizen science, and community based crowdsourcing.
• Lecture 2 - Introduction to Micro-task Crowdsourcing Platforms
  – Key terminology of micro-task crowdsourcing.
  – Popular platforms such as Amazon MTurk and CrowdFlower.
  – How to use such systems a crowd worker
Tentative Menu

• Tuesday
• Lecture 3 – How to Setup a Crowdsourcing Micro-task
  – Dimensions involved in crowdsourcing task design such as pricing, question design, and quality assurance mechanisms (e.g., honeypots).
  – Design and deploy a task during the lecture and see how to collect results back from the crowdsourcing platform.
• Lecture 4 – Micro-task Crowdsourcing Effectiveness
  – Techniques to ensure high quality in crowdsourced tasks (e.g., answer aggregation techniques, push crowdsourcing).
  – Behavior of malicious workers in crowdsourcing platforms.
Tentative Menu

• Wednesday
  • Lecture 5 - Hybrid Human-machine Systems
    – Advanced example uses of crowdsourcing.
    – Systems that combine both the scalability of machines over large amounts of data as well as the quality of human intelligence
  • Lecture 6 - Micro-task Crowdsourcing Scalability
    – In hybrid human-machine systems the latency bottleneck lays on the side of the crowd.
    – Recent research results that proposed techniques to improve the latency of crowdsourcing platforms.
    – Pricing techniques, HIT scheduling
Tentative Menu

• Thursday
• Lecture 7 - Open Research Directions in Micro-task Crowdsourcing
  – In this lecture we will give an overview on open micro-task crowdsourcing research questions.
  – Summarize which communities, conferences, journal, researchers work on crowdsourcing
• Slides:
  – http://www.gianlucademartini.net/crowdsourcing/
Crowdsourcing
How to build social systems at scale?

More people
More machines

Conventional Computation

Big Data

Social Networking

e-infrastructure

online R&D

Dave de Roure
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

• "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers."

[Howe, 2006]
One View of Crowdsourcing

Dimensions of human computation

See also [Quinn & Bederson, 2012]

What is outsourced

- Tasks based on human skills not easily replicable by machines (visual recognition, language understanding, knowledge acquisition, basic human communication etc)

How is the task outsourced

- Explicit vs. implicit participation
- Tasks broken down into smaller units undertaken in parallel by different people
- Coordination required to handle cases with more complex workflows
- Partial or independent answers consolidated and aggregated into complete solution

Who is the crowd

- Open call
- Call may target specific skills and expertise
- Requester typically knows less about the workers than in other work environments

See also [Quinn & Bederson, 2012]
Dimensions of human computation (2)

How are the results validated

- Solutions space closed vs. open
- Performance measurements/ground truth
- Statistical techniques employed to predict accurate solutions
- May take into account confidence values of algorithmically generated solutions

How can the process be optimized

- Incentives and motivators
- Assigning tasks to people based on their skills and performance (as opposed to random assignments)
- Symbiotic combinations of human- and machine-driven computation, including combinations of different forms of crowdsourcing
Aligning incentives is essential

- **Motivation**: driving force that makes humans achieve their goals
- **Incentives**: ‘rewards’ assigned by an external ‘judge’ to a performer for undertaking a specific task
  - Common belief (among economists): incentives can be translated into a sum of money for all practical purposes.
- Incentives can be related to both extrinsic and intrinsic motivations.
- **Extrinsic motivation** if task is considered boring, dangerous, useless, socially undesirable, dislikable by the performer.
- **Intrinsic motivation** is driven by an interest or enjoyment in the task itself.
Citizen Science
Citizen Science

- **WHAT IS OUTSOURCED**
  - Object recognition, labeling, categorization in media content

- **WHO IS THE CROWD**
  - Anyone

- **HOW IS THE TASK OUTSOURCED**
  - Highly parallelizable tasks
  - Every item is handled by multiple annotators
  - Every annotator provides an answer
  - Consolidated answers solve scientific problems
Citizen Journalism and Participatory Sensing
innocentive.com

- Tech Innovation
- via Crowdsourcing
- Competitions
- Monetary Incentive

https://www.innocentive.com/pavilion/NASA
Question Answering Systems
DB specific

- Freebase
The Way Industry Looks At It
Taxonomies

• Doan, Halevy, Ramakrishnan; (Crowdsourcing) CACM 4/11
  – nature of collaboration (implicit vs. explicit)
  – architecture (standalone vs. piggybacked)
  – must recruit users/workers? (yes or no)
  – What do users/workers do?

• Bederson & Quinn; (Human Computation) CHI ’11
  – Motivation (Pay, Altruism, Enjoyment, Reputation)
  – Quality Control (mechanisms)
  – Aggregation (how are results combined?)
  – Human Skill (Visual recognition, language, ...)
  – ...
Participatory Culture - Explicit
Participatory Culture – Implicit

John Murrell: GM SV 9/17/09

...every time we use a Google app or service, we are working on behalf of the search sovereign, creating more content for it to index and monetize or teaching it something potentially useful about our desires, intentions and behavior.
OCR errors: reCAPTCHA
Games with a purpose (GWAP)

See also [von Ahn & Dabbish, 2008]

- Human computation disguised as casual games
- Tasks are divided into parallelizable atomic units (challenges) solved (consensually) by players
- Game models
  - Single vs. multi-player
  - Selection agreement vs. input agreement vs. inversion-problem games
Games with a Purpose

• Tasks leveraging common human skills, appealing to large audiences
  – Selection of domain and task more constrained in games to create typical UX

• Tasks decomposed into smaller units of work to be solved independently

• Complex workflows
  – Creating a casual game experience vs. patterns in microtasks
Games with a Purpose

• Quality assurance
  – Synchronous interaction in games
  – Levels of difficulty and near-real-time feedback in games
  – Many methods applied in both cases (redundancy, votes, statistical techniques)

• Different set of incentives and motivators
Gamification

• A human-based computation technique in which a computational process performs its function by outsourcing certain steps to humans in an entertaining way
How to implement gamification

• **Cosmetic**: adding game-like visual elements or copy (usually visual design or copy-driven)
• **Accessory**: wedging in easy-to-add-on game elements, such as badges or adjacent products (usually marketing-driven)
• **Integrated**: more subtle, deeply integrated elements like % complete (usually interaction-design driven)
• **Basis**: making the entire offering a game (usually product-driven)

http://uxmag.com/design/a-gamification-framework-for-interaction-designers
Transactive Search
Transactive Search

• What if the data to answer your query is not stored on any digital support?
• What if the data is just in people minds?

• Big Data ➔ No Data
Transactive Search

• Search using Transactive (group) Memories
• “Who attended the WWW 2014 conference?”

• **Machines:** Harvest the Web + Data Mining
• **Crowd:** Search twitter, look at event pictures
• **Transactive Memories:** Remember who I met

Transactive Search (2)
Table 3: Effectiveness of machine-based, hybrid, and Transactive Search approaches using Crowdsourcing for ISWC 2013.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors and Tweets</td>
<td>0.3048</td>
<td>0.6906</td>
<td>0.4229</td>
</tr>
<tr>
<td>SVM</td>
<td>0.6632</td>
<td>0.4532</td>
<td>0.5385</td>
</tr>
<tr>
<td>M5P Regression</td>
<td>0.6599</td>
<td>0.4652</td>
<td>0.5457</td>
</tr>
<tr>
<td>Hybrid_uncertain</td>
<td>0.5864</td>
<td>0.4964</td>
<td>0.5377</td>
</tr>
<tr>
<td>Hybrid_unseen</td>
<td>0.4884</td>
<td>0.6043</td>
<td>0.5402</td>
</tr>
<tr>
<td>Hybrid_uncertain_unseen</td>
<td>0.4592</td>
<td>0.6211</td>
<td>0.5280</td>
</tr>
<tr>
<td><strong>Transactive Search</strong></td>
<td><strong>0.9006</strong></td>
<td><strong>0.7136</strong></td>
<td><strong>0.7963</strong></td>
</tr>
</tbody>
</table>

Transactive Search (3)
Discussion

• Sometime data is not on the Web
• The right group of *people can still answer
  – Collaboratively
  – Using Transactive Search
  – Better than machines or anonymous crowds

• Open challenges
  – Incentives
  – Repeatability
  – SNA
Summary

• Crowdsourcing has very many meanings
  – Incentives
  – Explicit/Implicit participation
  – Online/Offline

• This week we will focus on (paid) micro-task crowdsourcing to improve over machine-based systems
Crowdsourcing Incentives

• Paid Crowdsourcing
  – Competition with others (bonus payment for best performance)
  – Surveillance (check before paying)
  – Solidarity (your team will receive a bonus)
  – Accuracy (bonus for correct answers)
  – Agreement with others (bonus for agreeing with the majority)

• Fun (enjoyment)

• Community (belonging, desire to help)
Paid Crowdsourcing Ethics

• People work full-time as crowd workers
• Chinese crowdsourcing platform with 5.5M workers
• Pros
  – Help developing countries
  – Provide cash fast to people == short-term satisfaction
  – Job Flexibility
• Cons
  – No job security
  – No social security
  – Long term satisfaction? Career plans?