Part III
Crowdsourcing and Social Media

Michele @pirroh Catasta
EPFL, Switzerland
Crowdsourcing 1.0

- no social component in MTurk/Crowdflower/etc.
- no notifications / no recommendations
- lack of economical incentives?
What is CS 2.0?

- task prototyping vs custom apps
- Crowdflower partially covers this space, but we need an open source framework for tasks

Join Stanford researchers to form the largest crowdsourcing research project ever

Our goal is to design and develop the next-generation crowdsourcing platform. Want to be a researcher on our team? Join us and sign up by February 16th.
Social Media for Crowdsourcing

- Novel Decentralized Architecture
- The *Push-Crowdsourcing* paradigm
- How do we ensure quality?
Crowdsourcing on Twitter

- large user base
- “notification” system
- assignment problem: “who is the best worker for a certain HIT?”

Diaz-Aviles et al. 
Task Routing

Selecting the Crowd We Need

Task: Extracting Entities from News Articles

<table>
<thead>
<tr>
<th></th>
<th>US News</th>
<th>India News</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>50</td>
<td>75</td>
</tr>
</tbody>
</table>

US Worker
IN Worker

Difallah, Demartini, Cudré-Mauroux
Pick-a-Crowd: Tell me what you like and I’ll tell you what to do. WWW, 2013.
Task Routing

170 Registered

12K Pages

Profile Database

Difallah, Demartini, Cudré-Mauroux
Pick-a-Crowd: Tell me what you like and I'll tell you what to do. WWW, 2013.
Task Routing

Input Example

Title: Actor Identification
Description: Identify Actor from the TV show “How I Met Your Mother”
Contextual:
- Neil Patrick Harris
- Cobie Smulders

Index

Rank workers based on # of relevant “Likes”

Extract Task Content → Task Matching → Relevant Facebook pages
Task Routing

- Workers who like more than 40 pages related to the task category have high accuracy.
ARE ALL TASKS THE SAME?

NO.

Answering Memory Queries using Transactive Search
“A **transactive memory** system is a mechanism through which groups collectively **encode, store, and retrieve knowledge.**”

“[… ] a memory system that is more complex and potentially more effective than that of any of its individual constituents.”

A **transactive search** system discovers and aggregates the information stored in a transactive memory.
INFORMATION NEED
reconstruct the attendees’ list of the 86th Academy Awards (2014)
If only Bradley's arm was longer. Best photo ever. #oscars pic.twitter.com/C9U5N0tGap
THE WINNERS
— Recognizing the year’s best films

BEST PICTURE

American Hustle
Charles Roven, Richard Suckle, Megan Ellison, and Jonathan Gordon, Producers

View Trailer / ►
More Information

Captain Phillips
Scott Rudin, Dana Brunetti and Michael De Luca, Producers

View Trailer / ►
More Information

Nebraska
Albert Berger and Ron Yerxa, Producers

View Trailer / ►
More Information

Philomena
Gabrielle Tana, Steve Coogan and Tracey Seaward, Producers

View Trailer / ►
More Information

Dallas Buyers Club
Robbie Brenner and Rachel Winter, Producers

View Trailer / ►
More Information

Winner
12 Years a Slave
Brad Pitt, Dede Gardner, and Emma Tillinger Koski, Producers

View Trailer / ►
More Information
MISTAKES: not all the nominees participate to the ceremony

PRECISION :-(

MISSING ENTRIES: what about all the people working “behind the scenes”?  

RECALL :-(

FROM THE IDEA...

• for data that is stored in the memories of a group of people, the current query strategies are **suboptimal**

• we need a **new form of human computation**, different from standard crowdsourcing (i.e., no anonymous crowd)

---

**“A taxonomy of Web Search”**
— A. Broder (2002)

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Navigational</strong></td>
<td>The immediate intent is to reach a particular Web site.</td>
</tr>
<tr>
<td><strong>Informational</strong></td>
<td>The intent is to acquire some information assumed to be present on one or more Web pages.</td>
</tr>
<tr>
<td><strong>Transactional</strong></td>
<td>The intent is to perform some Webmediated activity.</td>
</tr>
<tr>
<td><strong>Transactive</strong></td>
<td>The intent is to acquire some information that can be reconstructed <strong>only</strong> by an [ephemeral] social network.</td>
</tr>
</tbody>
</table>
...TO THE TESTING ENVIRONMENT

- We want to **reconstruct the attendees list** of two Semantic Web conferences, ISWC2012 and ISWC2013.

- We were given access to the **ground truth** but, in general, such lists are **not publicly available**.

- Additional data sources: authors list (first author, last author, etc.), mentions in Online Social Networks.
Help us find the participants of ISWC 2013 and ISWC 2012.

We want to test how efficient are "group memories" when it comes to complete a rather trivial task: reconstruct the participant list of a conference.

Each person you add to the list, even if mentioned by other users in the experiment, will receive only ONE email. As such, if this is not the first time you receive a link to this website, please contact Michele Catasta ASAP.

**ISWC2013 participants**

Please insert (one by one) all the names of the people you have met at ISWC2013 during e.g., social events, poster/demo sessions, workshops, paper presentations, etc.

<table>
<thead>
<tr>
<th>Full Name</th>
<th>e-mail (Optional)</th>
<th>Add</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EXPERIMENT ARCHITECTURE

- tailored Web UI + results aggregator

- iterative reconstruction: every time a new person was mentioned, Hippocampus sent her an invitation to contribute to the attendees list
MACHINE LEARNING APPROACHES

• we collected the proceedings information and all the tweets with the conference hashtags

• we trained state-of-the-art classifiers with these features:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>isFirstAuthor</td>
<td>isConference&amp;WorkshopAuthor</td>
</tr>
<tr>
<td>isMiddleAuthor</td>
<td>numberOfPapers</td>
</tr>
<tr>
<td>isLastAuthor</td>
<td>numberOfCoauthors</td>
</tr>
<tr>
<td>isWorkshopAuthor</td>
<td>hasTweeted</td>
</tr>
<tr>
<td>isConferenceAuthor</td>
<td>numberOfTweets</td>
</tr>
</tbody>
</table>

not possible without the ground truth!
ML + CROWDSOURCING APPROACHES

• **Uncertain cases** (precision): we asked the crowd to revise the low-confidence results of the ML classifier. (e.g., people that didn’t attend the conference but tweeted about it)

• **Unseen cases** (recall): we asked the crowd to actively look for attendees not included in the authors list (e.g., organizers mentioned in the Web site)

the crowd has access **only** to public data on the Web!
<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><strong>0.6632</strong></td>
<td>0.4532</td>
<td>0.5385</td>
</tr>
<tr>
<td>M5P Regression</td>
<td>0.6599</td>
<td>0.4652</td>
<td><strong>0.5457</strong></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>Transactive Search</td>
<td><strong>0.9006</strong></td>
<td><strong>0.7136</strong></td>
<td><strong>0.7963</strong></td>
</tr>
</tbody>
</table>

Authors and Tweets: baseline (exhaustive list of authors and twitterers)  
**Machine Learning:** SVM, M5P Regression  
**Machine Learning + Crowdsourcing:** Hybrid (uncertain, unseen, uncertain_unseen)
attendees found over time
Transactive Memory Graph

in green, two isolated “components”
discovered by top-contributors
Result discussion

• for a specific class of queries, our Transactive Search performs up to 46% better than the best alternative approach (i.e., Machine Learning + Crowdsourcing)

• we will explore incentives for Hippocampus, as it is currently two orders of magnitude slower than the alternative approaches

• we reported some initial evidences that, as human memories fade with time, our approach works best with recent events
Transactive Point Queries

what if the information need can be served only by one/few nodes?
What is the name of the delicious cocktail I had during last year’s gala dinner?

This information need can be unlikely satisfied by:
- a Web search (i.e., the conference website does not contain such information)
- a DB query (i.e., the transactions of the restaurant are private)
- a crowdsourcing task (i.e., the anonymous crowd did not participate to the conference)

But (some of) the attendees of the conference could work collectively and come up with an answer
Tapping into Collective Human Memories

• **TransactiveDB**: a decentralized data management system that elicits and processes memories of individuals or groups in order to answer transactive queries.

• **Node**: classical DBMS + transactive operators handling the memories of a particular user (i.e., personal events, contextual data, etc.)

• **Interaction graph**: a subset of the underlying social network connecting different end-users, corresponding to a specific context (e.g., social event, family setting, etc.)
Architecture
Crowdsourcing for Social Media

- Like in many other scientific fields, crowdsourcing is playing a key role in social media research.

- ICWSM2016 proceedings:
  - 37 mentions of “crowdsourcing”
  - 30 mentions of “Mechanical Turk”
  - 6 mentions of “Crowdflower”
• **DATA GATHERING:** “We first employed crowdsourcing to collect Twitter users’ cognitive styles using standard psychometric instruments”

• **SENTIMENT:** “Through a crowdsourcing study, we show that there are marked differences between the overall tweet sentiment and the sentiment expressed towards the subjects mentioned in tweets related to three crises events.”

• **LEXICON:** “We built lexical categories that capture this list of stereotypes by mapping the 2000 most commonly occurring verbs and adjectives in our dataset onto the set of categories through a series of crowdsourcing tasks.”

• **VALIDATION:** ”To calibrate and validate this measure, we turn to crowdsourcing labels on Amazon Mechanical Turk. The results reveal that cosine distance is a strong predictor of similarity.”
CS excels in understanding human nuances

sentiment, sarcasm, jargon, etc.
Timeline SBS A/A Equalized across User Geo

Chart:
- Which tweet is more relevant to the twitter user?
- Which piece of information was most helpful?
- Control Tweet had Monetizable Engagement?
- Test Tweet had Monetizable Engagement?
- Control Tweet had Negative Engagement?
- Test Tweet had Negative Engagement?
- User Type
- User Geo

Segment Bars By:
- None
- Which tweet is more relevant to the twitter user?
- Which piece of information was most helpful?
- Control Tweet had Monetizable Engagement?
- Test Tweet had Monetizable Engagement?
- Control Tweet had Negative Engagement?
- Test Tweet had Negative Engagement?
- User Type
- User Geo

Display:
- Count
- Normalized
- Average value of Which tweet is more relevant to the twitter user?
Annotate entities in Tweets (2010)

megabubbles!
(and extremely expensive...)
Annotate entities in Tweets (today)

NLP pipeline with 80% accuracy in NER on Tweets

low confidence results go to the crowd
• faster
• slash the costs
Virtual Labs (Duncan Watts)

- “big and thin” vs “small and rich”
- SurveyMonkey/Google Forms in a crowdsourcing platform: scale up N of subjects
- what about the need for synchronicity?