Boosting Crowdsourcing

Yuqian Li
Broader Crowdsourcing | Journalism | Project Detail
Broader Crowdsourcing [1]

Bradham's 4 types

- Knowledge discovery and management
  e.g. Peer-to-Patent Community Patent Review

- Broadcast search
  e.g. Innocentive: http://www.innocentive.com/

- Peer-vetted creative production
  e.g. Threadless: http://www.threadless.com

- Distributed human intelligence tasking (HIT)
  e.g. Amazon's Mechanical Turk, Microtask: http://www.microtask.com
Welcome To InnoCentive
Where the World Innovates

Need solutions to real problems?
Click below or call 1-855-CROWDNOW
Learn More ➤

Want to solve important problems
and make a positive impact?
Become A Solver ➤

Are you suffocating under piles of paper forms?
We convert paper forms into Excel spreadsheets

We give you an email address where you can send your forms.

trial@microtask.com
Howe's taxonomy

- Crowd wisdom
  similar to broadcast search
- Crowd creation
  similar to peer-vetted creative production
- Crowd voting
  e.g. prediction markets
- Crowd funding
  e.g. crowd-based microlending institutions, spot.us [http://spot.us]
Jounalism Aspect of Crowdsourcing

- Co-creation, crowdsourcing as journalism
  - meaningful **process** for crowd, unlike CAPCHA, e.g. audience will know more related information about a topic when they help classifying articles/comments

- Crowd wisdom, ask for help
  - It's too hard to get a universal, powerful while easy-to-use tool for reporters. But for a specific task of an reporter, it's easy to find an expert to help him. e.g. TopCoder [http://community.topcoder.com/tc](http://community.topcoder.com/tc), Craigslist [http://raleigh.craigslist.org/cpg](http://raleigh.craigslist.org/cpg)

- Crowd funding + Human Intelligence Tasking
  - We're interested in your story. If we have tim, we can do HITs. If we don't, we donate money to encourage who don't have interests to do so.
A Small Task: Classification

Why choose this? Easy input/output, hard for computer alone, important for journalism

- Filtering by "related or not" among:
  - tens of thousands of comments/tweets
  - thousands of emails in a public mailing-list
  - hundreds of public documents

- Clustering documents by "relation"
  - Clustering by topics, DOC x TOPICS
  - Clustering by pairwise relation, DOC x DOC
Human Boosting

1. Classification Statement
2. Set of Public Docs (emails, comments, etc.)

Classification Result generated by Distributed HIT

Ask for Help
Technical Support

Fund
More Incentive

Share
More Participation

Crowd Funding
Crowd
Social Network
Human Boosting Example

Investigate local gov. economy achievements by analyze job opportunities posted in public mailing lists

Journalist

1. Classify which emails are related to jobs
2. Packed of Emails

TopCoder Craigslist

Classification Result generated by Distributed HIT

Archived Machine/Human Readable Emails

Help me Pack Emails

Crowd

Fund

Share

More Incentive

More Participation

Facebook Twitter

Crowd Funding

spot.us?
Machine Boosting

1. Classification Statement
2. Set of Public Docs (emails, comments, etc.)

Ask for Help
Technical Support
trained classifier
training label
Fund
More Incentive
Share
More Participation

Journalist
Crowd Wisdom
AI, ML
Crowd
Social Network
Crowd Funding
<table>
<thead>
<tr>
<th>Name</th>
<th>Department</th>
<th>Subject</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fletcher Maynard</td>
<td>Duke/inthouse/others</td>
<td>[intilhouse] NO!! BRIDGES LUNCH TOMORROW (11/28/11) - Dear D</td>
<td></td>
</tr>
<tr>
<td>dipys macho</td>
<td>Duke/inthouse/others</td>
<td>[intilhouse] Durham to DC on December 17th or 18th - I would be dri</td>
<td></td>
</tr>
<tr>
<td>Teng-Yee Ng</td>
<td>Duke/inthouse/supply</td>
<td>[intilhouse] MOVING SALE—UPDATED - Dear all, I updated the mov</td>
<td></td>
</tr>
<tr>
<td>Yui</td>
<td>Duke/inthouse/supply</td>
<td>[intilhouse] Bedroom set - Hello, We have a nice full size bedroom</td>
<td></td>
</tr>
<tr>
<td>Tiago Mata (2)</td>
<td>Duke/inthouse/supply</td>
<td>[intilhouse] TOYOTA RAV4 1998 for sale - $3,600 - To UNSUBSCRIBE</td>
<td></td>
</tr>
<tr>
<td>Francisco Duque</td>
<td>Duke/inthouse/supply</td>
<td>[intilhouse] Bedroom in Chapel Hill: One month sublease - To UNSUBSC</td>
<td></td>
</tr>
<tr>
<td>Tiago mata</td>
<td>Duke/inthouse/others</td>
<td>[intilhouse] - To UNSUBSCRIBE from International House List Serv go</td>
<td></td>
</tr>
<tr>
<td>Jiaoni Zhou (2)</td>
<td>Duke/China</td>
<td>Duke/inthouse/others</td>
<td>[intilhouse] Ride from Durham to DC today (Nov 27) - Hi,</td>
</tr>
<tr>
<td>Sambuddha Banerjee</td>
<td>Duke/inthouse/others</td>
<td>[intilhouse] Ride from DC to Durham today - Hi, Is any one traveling from D to C</td>
<td></td>
</tr>
<tr>
<td>Luis Vergara</td>
<td>Duke/inthouse/others</td>
<td>[intilhouse] Hola….Spanish Hour on Monday at 6:30 PM - In Spanish</td>
<td></td>
</tr>
<tr>
<td>Mira Han</td>
<td>Duke/inthouse/supply</td>
<td>[intilhouse] flying pigeon men's bike $90 - To UNSUBSCRIBE from Int</td>
<td></td>
</tr>
</tbody>
</table>
Classify sales/rent/supply in intl@duke.edu mailing list
Interact Boosting

Journalist

Crowd Wisdom

AI, ML

1. Classification Statement
2. Set of Public Docs (emails, comments, etc.)

training label

auto task selection and acceleration

crowd

Fund

More Incentive

Social Network

Share

More Participation

Crowd Funding

Ask for Help

Technical Support

trained classifier
Why Interact?

- Human make errors
  - In previous example, 95% human accuracy, no improvement from 400 to 1000 without interaction
Why Interact?

- **Select Critical Documents**
  - Use classifier to predict which documents are more important to label (human errors, characteristic)

- **Accelerate**
  - Use classifier to predict a new document, use that result as a hint to human

- **Evolve Iteratively**
  - As human give larger amount and more accurate labels, classifiers evolves, and it'll be used to refine/assist/accelerate human labels, and human give even larger amount and more accurate labels, classifiers evolves again...
Boosting Summary

- **Crowd wisdom**
  - ask an expert for preparing a specific problem

- **Crowd funding**
  - Incentive to crowd and from crowd

- **Social Network**
  - More participation

- **Interactive AI/ML module**
  - Better quality
  - Less human work (with hint)
  - Classify more even we don't have so many crowd