Adventures in Crowdsourcing

Panos Ipeirotis
Stern School of Business
New York University

Twitter: @ipeirotis

“A Computer Scientist in a Business School”
http://behind-the-enemy-lines.com
Broad Goal

Integrate machine and human intelligence

Create hybrid “intelligence integration” processes

With **paid** users and with **unpaid** participants
Example Application

Detect Inappropriate Ad Placement
Arizona Suspect’s Online Trail Offers Hints of Alienation
By ERIC LIPTON, CHARLIE SAVAGE and SCOTT SHANE
Published: January 9, 2011

WASHINGTON — His MySpace page included a photograph of a United States history textbook, on top of which he had placed a handgun. He prepared a series of Internet videos in which he posted odd statements about the gold standard, the community college he attended and SWAT teams.

Jared Lee Loughner, in these few public hints, offered a sense of his alienation from society, confusion, anger as well as foreboding that his life could soon come to an end. Friends talked of how he had become reclusive in recent years, and his public postings raised questions, in retrospect at least, about his mental state.

Still, his comments offered little indication as to why, as police allege, he would go to a Safeway supermarket in northwest Tucson on Saturday morning and begin shooting at a popular Democratic congresswoman and more than a dozen others, killing six and wounding 19.

There was evidence of recent trouble, though. Mr. Loughner, 22, was suspended in late September from Pima Community College, where he had been attending classes, because the school became aware of a disturbing YouTube
Anatidaephobia - The Fear That You are Being Watched by a Duck

What Is Anatidaephobia?

Anatidaephobia is defined as a pervasive, irrational fear that one is being watched by a duck. The anatidaephobic individual fears that no matter where they are or what they are doing, a duck watches.

Anatidaephobia is derived from the Greek word “anatidae”, meaning ducks, geese or swans and “phobos” meaning fear.

What Causes Anatidaephobia?

As with all phobias, the person coping with Anatidaephobia has experienced a real-life trauma. For the anatidaephobic individual, this trauma most likely occurred during childhood.

Perhaps the individual was intensely frightened by some species of water fowl. Geese and swans are relatively well known for their aggressive tendencies and perhaps the anatidaephobic person was actually bitten or flapped at. Of course, the Far Side comics did little to minimize the fear of being watched by a duck.

While we may be tempted to smile at the memory of those comics or at the mental image of being watched by a duck, for the anatidaephobic person, that fear is uncontrollable. Whatever the cause, the anatidaephobic person can experience emotional turmoil and anxiety that is completely disruptive to daily functioning.
Detect Inappropriate content

- Ad hoc topics, with no existing training data
  - Hate speech, Violence, Guns & Bombs, Gossip…
- Classification models need to be trained and deployed within days
- Crowdsourcing allows for fast data collection
  - labor is accessible on demand
  - using Mechanical Turk, oDesk, etc
  - but quality may be lower than experts
<table>
<thead>
<tr>
<th>Requester</th>
<th>HIT Expiration Date</th>
<th>Reward</th>
<th>Time Allotted</th>
<th>Hits Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam GONZALES</td>
<td>Dec 13, 2010 (1 week 2 days)</td>
<td>$0.01</td>
<td>30 minutes</td>
<td>39172</td>
</tr>
<tr>
<td>Chris Callison-Burch</td>
<td>Dec 31, 2010 (3 weeks 5 days)</td>
<td>$0.05</td>
<td>15 minutes</td>
<td>14240</td>
</tr>
<tr>
<td>nutella42</td>
<td>Dec 17, 2010 (2 weeks)</td>
<td>$0.08</td>
<td>30 minutes</td>
<td>2446</td>
</tr>
<tr>
<td>Jaime Arquelo</td>
<td>Dec 10, 2010 (7 days)</td>
<td>$0.03</td>
<td>5 minutes</td>
<td>1952</td>
</tr>
<tr>
<td>Andy K</td>
<td>Dec 9, 2010 (6 days 2 hours)</td>
<td>$0.03</td>
<td>60 minutes</td>
<td>1949</td>
</tr>
</tbody>
</table>
Example: Build an “Adult Content” Classifier

- Need a large number of labeled sites for training
- Get people to look at sites and label them as:
  - G (general audience)
  - PG (parental guidance)
  - R (restricted)
  - X (porn)

Cost/Speed Statistics

- Undergrad intern: 200 websites/hr, cost: $15/hr
- Mechanical Turk: 2500 websites/hr, cost: $12/hr
Bad news: Spammers!

Worker ATAMRO447HWJQ

labeled X (porn) sites as G (general audience)
Challenges

- We do not know the true category for the objects
- We do not know the quality of the workers
- We want to label objects with true categories
- We want (need?) to know the quality of the workers
Redundant votes, infer quality

Look at our lazy friend **ATAMRO447HWJQ** together with other 9 workers

- Using redundancy, we can compute error rates for each worker
1. Initialize “correct” label for each object (e.g., use majority vote)
2. Estimate **error rates** for workers (using “correct” labels)
3. Estimate “**correct**” labels (using error rates, weight worker votes according to quality)
4. Go to Step 2 and iterate until convergence

**Expectation Maximization Estimation**

Iterative process to estimate worker error rates

<table>
<thead>
<tr>
<th>Worker</th>
<th>G → G</th>
<th>G → X</th>
<th>X → G</th>
<th>X → X</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAMRO447HWJQ</td>
<td>99.947%</td>
<td>0.053%</td>
<td>99.153%</td>
<td>0.847%</td>
</tr>
</tbody>
</table>

The spammer worker marked *almost all* sites as G.
Challenge: Humans are biased!

Error rates for the CEO, providing “expert” labels

<table>
<thead>
<tr>
<th></th>
<th>G → G</th>
<th>G → P</th>
<th>G → R</th>
<th>G → X</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>20.0%</td>
<td>80.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>P</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>R</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>X</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

We have 85% G sites, 5% P sites, 5% R sites, 5% X sites

- Error rate of spammer (all G) = 0% * 85% + 100% * 15% = 15%
- Error rate of biased worker = 80% * 85% + 100% * 5% = 73%

False positives: Legitimate workers appear to be spammers

(important note: bias is not just a matter of “ordered” classes)
Solution: Fix bias first, compute error rate afterwards

Error Rates for CEO

<table>
<thead>
<tr>
<th>Vote</th>
<th>G</th>
<th>P</th>
<th>R</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>P[G → G]</td>
<td>20.0%</td>
<td>80.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>P[P → G]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>P[R → G]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>P[X → G]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>P[G → P]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>P[P → P]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>P[R → P]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>P[X → P]</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

- When biased worker says G, it is 100% G
- When biased worker says P, it is 100% G
- When biased worker says R, it is 50% P, 50% R
- When biased worker says X, it is 100% X

Small ambiguity for “R-rated” votes but other than that, fine!
Expected Misclassification Cost

- **High cost**: probability spread across classes
- **Low cost**: probability mass concentrated in one class

<table>
<thead>
<tr>
<th>Assigned Label</th>
<th>Corresponding “Soft” Label</th>
<th>Soft Label Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spammer: G</td>
<td>&lt;G: 25%, P: 25%, R: 25%, X: 25%&gt;</td>
<td>0.75</td>
</tr>
<tr>
<td>Good worker: G</td>
<td>&lt;G: 99%, P: 1%, R: 0%, X: 0%&gt;</td>
<td>0.0198</td>
</tr>
</tbody>
</table>

[Assume misclassification cost equal to 1, solution generalizes to arbitrary costs]
Question: How to pay workers?

- **Naïve** solution: Have a quality-score *threshold*
- **Thresholding rewards gives wrong incentives:**
  - Very different outcomes around the threshold, for similar performance
  - Often uncertain about true performance
  - Decent (but still useful) workers get punished
Quality-sensitive Payment

- Set quality goal and price (e.g., $1 for 90%)
  - For workers above goal: Pay full price
  - For others: Payment divided with redundancy needed to reach goal
    - Need 3 workers with 80% accuracy ➔ Payment = $1/3 = $0.33
    - Need 9 workers with 70% accuracy ➔ Payment = $1/9 = $0.11

How to deal with uncertainty?
Instead of blocking: Quality-sensitive Payment

- **Uncertainty hurts:**
  - Small fluctuations in performance may result in drastic payment changes
  - Payment decreases practically equivalent to rejection

- Introduced *uncertainty “penalty”: Pay less* for uncertain estimates (for workers with short working histories)

- **Refund** underpayment when quality estimate more certain
### Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

<table>
<thead>
<tr>
<th>#Tasks</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>Infinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece-rate Payment (cents)</td>
<td>11</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>40</td>
</tr>
</tbody>
</table>

**Fair Payment**
Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

<table>
<thead>
<tr>
<th>#Tasks</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>Infinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece-rate Payment (cents)</td>
<td>11</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>40</td>
</tr>
</tbody>
</table>

Fair Payment: 40

Potential “Bonus”
Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

<table>
<thead>
<tr>
<th>#Tasks</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>Infinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece-rate Payment (cents)</td>
<td>11</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>40</td>
</tr>
</tbody>
</table>

Fair Payment: 40

Potential “Bonus”

Reimbursement

Payment
Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

<table>
<thead>
<tr>
<th># Tasks</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>Infinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece-rate Payment (cents)</td>
<td>11</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>40</td>
</tr>
</tbody>
</table>

Fair Payment: 40
Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

<table>
<thead>
<tr>
<th>#Tasks</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>Infinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece-rate Payment (cents)</td>
<td>11</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>40</td>
</tr>
</tbody>
</table>

Fair Payment: 40
Summary of Experimental Results

- Randomized Controlled Trial on oDesk
  - Thresholding,
  - Quality-based payment (QBP)
  - QBP with reimbursements

- **Retention**: ~150-300% up over thresholding/QBP
  - No significant differences between QBP/thresholding
  - Decrease in pay, same effect as rejection

- **Cost**: 50%-70% reduction, as we pay for performance

- **Work quality**: Stable
Humans Improving Machine Learning

- With just labeling, workers are **passively** labeling the data that we give them.

- Asking instead the workers to search and find **training data**.

- **Vanilla solution**: Use data and build model.
The result? Blissful ignorance…

- Classifier *seems* great: Cross-validation tests show excellent performance

- Alas, classifier fails: The “*unknown unknowns*”

“*Unknown unknowns*” = classifier fails with high confidence
Beat the Machine!

Ask humans to find URLs that

- the classifier will classify incorrectly
- another human will classify correctly

Example:
Find hate speech pages that the machine will classify as benign
Beat the Machine!

Incentive structure:

- $1 if you “beat the machine”
- $0.001 if the machine already knows

Example:
Find hate speech pages that the machine will classify as benign
Error rate for probes significantly higher than error rate on (stratified) random data (10x to 100x higher than base error rate)

**Conclusion**: Humans are good in discovering problematic cases for model testing
Finding People to Beat the Machine

**Question**: Can we find humans that can and are willing to “beat the machine”?

**Example Application**: Improving Automatic Essay Scoring
Audience Discovery?

- How can we automate the process of discovering good users for arbitrary crowdsourcing applications?
Google Knowledge Graph

“Things not Strings”

Kyrgyzstan
Country
Kyrgyzstan, officially the Kyrgyz Republic, is a country located in Central Asia. Landlocked and mountainous, Kyrgyzstan is bordered by Kazakhstan to the north, Uzbekistan to the west, Tajikistan to the southwest and China to the east. Wikipedia

Capital: Bishkek
Currency: Kyrgyzstani som
President: Almazbek Atambayev
National anthem: National Anthem of the Kyrgyz Republic
Official languages: Kyrgyz language, Russian Language
Government: Presidential system, Parliamentary republic, Republic
Still incomplete…

- “Date of birth of Bayes” (…uncertain…)
- “Symptom of strep throat”
- “Side effects of treximet”
- “Who is Cristiano Ronaldo dating”
- “When is Jay Z playing in New York”
- “What is the customer service number for Google”
- …
Key Challenge

“Crowdsource in a **predictable** manner, with **knowledgeable** users, **without** introducing **monetary rewards**”
What is a symptom of Morgellons

- Red eye
- Choreoathetosis
- Skin lesion
- Insomnia
- I don’t know

Correct Answers: 33/67  Correct (%): 49%

Question 1 out of 10
Calibration vs. Collection

- **Calibration** questions (known answer): Evaluating user competence on topic at hand
- **Collection** questions (unknown answer): Asking questions for things we do not know
- *Trust more answers coming from competent users*
Challenges

- Why would anyone come and play this game?
- Why would knowledgeable users come?
- Wouldn’t it be simpler to just pay?
Attracting Visitors: Ad Campaigns

Quiz on disease symptoms
Test how well you can recognize various disease symptoms
www.quizz.us
Treat Quizz as eCommerce Site

- Value of user: \textit{total} information gain contributed
- Information gain is additive: \#questions \times \text{info/question}
Example of Targeting: Medical Quizzes

- Our initial goal was to use medical topics as a evidence that some topics are not crowdsourcable

- Our hypothesis failed: They were the best performing quizzes…

- Users coming from sites such as Mayo Clinic, WebMD, … (i.e., “pronsumers”, not professionals)
Immediate feedback helps

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show if user answer correct</td>
<td>+2.4%</td>
</tr>
<tr>
<td>Show the correct answer</td>
<td>+20.4%</td>
</tr>
<tr>
<td>Score: % of correct answers</td>
<td>+2.3%</td>
</tr>
<tr>
<td>Score: # of correct answers</td>
<td>-2.2%</td>
</tr>
<tr>
<td>Score: Information gain</td>
<td>+4.0%</td>
</tr>
<tr>
<td>Show statistics for performance of other users</td>
<td>+9.8%</td>
</tr>
<tr>
<td>Leaderboard based on percent correct</td>
<td>-4.8%</td>
</tr>
<tr>
<td>Leaderboard based on total correct answers</td>
<td>-1.5%</td>
</tr>
</tbody>
</table>

- Knowing the correct answer 10x more important than knowing whether given answer was correct
- Conjecture: Users also want to learn
### Showing score moderately helpful

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show if user answer correct</td>
<td>+2.4%</td>
</tr>
<tr>
<td>Show the correct answer</td>
<td>+20.4%</td>
</tr>
<tr>
<td><strong>Score: % of correct answers</strong></td>
<td>+2.3%</td>
</tr>
<tr>
<td><strong>Score: # of correct answers</strong></td>
<td>-2.2%</td>
</tr>
<tr>
<td><strong>Score: Information gain</strong></td>
<td>+4.0%</td>
</tr>
<tr>
<td>Show statistics for performance of other users</td>
<td>+9.8%</td>
</tr>
<tr>
<td>Leaderboard based on percent correct</td>
<td>-4.8%</td>
</tr>
<tr>
<td>Leaderboard based on total correct answers</td>
<td>-1.5%</td>
</tr>
</tbody>
</table>

- Be careful what you incentivize 😊
- “Total Correct” incentivizes quantity, not quality
## Competitiveness helps

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show if user answer correct</td>
<td>+2.4%</td>
</tr>
<tr>
<td>Show the correct answer</td>
<td>+20.4%</td>
</tr>
<tr>
<td>Score: % of correct answers</td>
<td>+2.3%</td>
</tr>
<tr>
<td>Score: # of correct answers</td>
<td>-2.2%</td>
</tr>
<tr>
<td>Score: Information gain</td>
<td>+4.0%</td>
</tr>
<tr>
<td><strong>Show statistics for performance of other users</strong></td>
<td>+9.8%</td>
</tr>
<tr>
<td>Leaderboard based on percent correct</td>
<td>-4.8%</td>
</tr>
<tr>
<td>Leaderboard based on total correct answers</td>
<td>-1.5%</td>
</tr>
</tbody>
</table>
Leaderboards are tricky!

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show if user answer correct</td>
<td>+2.4%</td>
</tr>
<tr>
<td>Show the correct answer</td>
<td>+20.4%</td>
</tr>
<tr>
<td>Score: % of correct answers</td>
<td>+2.3%</td>
</tr>
<tr>
<td>Score: # of correct answers</td>
<td>-2.2%</td>
</tr>
<tr>
<td>Score: Information gain</td>
<td>+4.0%</td>
</tr>
<tr>
<td>Show statistics for performance of other users</td>
<td>+9.8%</td>
</tr>
<tr>
<td><strong>Leaderboard based on percent correct</strong></td>
<td><strong>-4.8%</strong></td>
</tr>
<tr>
<td><strong>Leaderboard based on total correct answers</strong></td>
<td><strong>-1.5%</strong></td>
</tr>
</tbody>
</table>

- Initially, strong positive effect
- Over time, effect became strongly negative
- All-time leaderboards considered harmful
Comparison with paid crowdsourcing

Unpaid users vs. hourly (oDesk) vs. piecemeal (MTurk)

% correct

Submitted answers (log)

Source
- MTurk
- oDesk
- Quizz
Citizen Science Applications

- Google gives $10K/month to nonprofits in ad budget

- Climate CoLab experiment
  - Doubled traffic with only $20/day
  - Targets political activist groups (not only climate)

- Additional experiments:
  - Identify users with particular psychological characteristics
  - Engage users with an interest in speech therapy
How can I get rid of users?
Your workers behave like my mice!

An unexpected connection…
Your workers want to use only their motor skills, not their cognitive skills.
The Biology Fundamentals

- Brain functions are biologically expensive (20% of total energy consumption in humans)
- Motor skills are more energy efficient than cognitive skills (e.g., walking)
- Brain tends to delegate easy tasks to part of the neural system that handles motor skills
The Mice Experiment

Cognitive
- Solve maze
- Find pellet

Motor
- Push lever three times
- Pellet drops
How to Train the Mice?

**Confuse** motor skills!

**Reward** cognition!
Punishing Worker’s Motor Skills

- **Punish bad answers** with frustration of motor skills (e.g., add delays between tasks)
  - “Loading image, please wait…”
  - “Image did not load, press here to reload”
  - “404 error. Return the HIT and accept again”

→ Make this **probabilistic** to keep feedback implicit
Experimental Summary

- Spammer workers quickly abandon
- No need to display scores, or ban
- Low quality submissions from ~60% to ~3%
- Half-life of low-quality users from 100+ tasks to less than 5
Thanks!

Q & A?