Crowdsourcing for Analytics

Tim Kraska
<kraska@cs.berkeley.edu>
Machines alone are not enough...

Classification

- full dataset
- training set
- test set

classifier

- accuracy
- prediction
- new entity

Classifica\(\text{on} (\text{full dataset \text{training set test set}}\) classifier

accuracy

g

- new entity

- prediction

- accuracy

- new entity

- prediction

- accuracy

- new entity

- prediction
Machines alone are not enough...

SELECT Image
From Pictures
Where Image contains “Dog”
Adding People to Analytics

Data collection
America's top 10 NASDAQ companies with female CEOs

Data cleaning

Transcription

Creativity/Design/Taste

and more
Structured Data

An entity graph of people, places and things, built by a community that loves open data.

<table>
<thead>
<tr>
<th>Featured Data</th>
<th>Data</th>
<th>Schema</th>
<th>Apps</th>
<th>Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Film</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80 members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>People</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>87 members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Music</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100+ members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Business</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100+ members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Government</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47 members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52 members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Books</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47 members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: July 4th and August 1st statistics show the number of facts and topics.*

**Google Refine**
An open source power tool to fix, discover, experiment, connect and customize your data. [Learn more](#).

**What is Freebase?**
Learn what an entity graph is, what kind of information it contains, and why you should add your data!
[Learn More](#).

**Freebase for Developers**
- powerful queryable API
- JavaScript-based hosting framework
- libraries for other languages
[Learn More](#).
Micro-Task CrowdSourcing

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Get started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you’re satisfied with the results
Microtasking – Virtualized Humans

• Current leader: Amazon Mechanical Turk
• Requestors place Human Intelligence Tasks (HITs)
  – Minimum price: $0.01
  – #of replicas (assignments), expiration, User Interface
  – API-based: “createHit()”, “getAssignments()”, “approveAssignments()”, “forceExpire()”
  – Requestors approve jobs and payment
• Workers (a.k.a. “turkers”) choose jobs, do them, get paid
All HITs

1-10 of 1157 Results

Sort by: HIT Creation Date (newest first) ▼

Huge test
Requester: Mr. Doe
HIT Expiration Date: Aug 27, 2011 (1 day 5 hours)
Reward: $0.01
Time Allotted: 60 seconds
HITs Available: 8424

Validate Brand/Product Information from Product Picture
Requester: Redwood Technologies
HIT Expiration Date: Aug 27, 2011 (11 hours 59 minutes)
Reward: $0.01
Time Allotted: 15 minutes
HITs Available: 6

Copy Brand/Product Information from Product Picture
Requester: Redwood Technologies
HIT Expiration Date: Aug 27, 2011 (11 hours 58 minutes)
Reward: $0.03
Time Allotted: 15 minutes
HITs Available: 8

Huge job test for refactored balancer
Requester: Mr. Doe
HIT Expiration Date: Aug 27, 2011 (1 day 5 hours)
Reward: $0.01
Time Allotted: 60 seconds
HITs Available: 12390
'Are these two pictures of the same person?'

Yes ☐ No ☐

Please ACCEPT the hit before submitting.
Amount per Assignment: $0.01
Amount to Approve Oustanding Assignments: $0.01
Pending Review: 1
Reviewed: 0
Remaining: 0
Total: 1

EXPIRATION DATE
Aug 27 2011, 02:25 PM PDT

» View HIT

Pending Review (1)  Approved  Rejected

Review Submitted Assignments  (showing page 1 of 1)
Select Assignments to approve or reject then click "Submit." When you approve an Assignment, the Worker is paid automatically. You will not be charged for Assignments you reject.

Approve  Reject  Worker ID  Result  Less  Submission Date
All: None  All: None

crowdEqual: no
id: 10202458

You've chosen to:
approve: none selected
reject: none selected
Submit
Samasource.org

1. You send Samasource a project
2. Samasource breaks it down into microwork
3. Work is allocated to our service partners
4. Women, youth, and refugees complete work
5. Samasource compiles work and assures quality
6. Your project gets delivered and helps reduce poverty
Challenges

• Quality
• User Interface Design
• Worker motivation
• Task decomposition
• Leverage worker knowledge/capabilities
• Optimization (time/cost)
• ....
Challenges

• Quality
• Optimization (time/cost)
• User Interface Design
• Worker Motivation
• Task decomposition
• Leverage worker knowledge/capabilities
• ....
How Can You Trust the Crowd?
Quality Techniques

- Approval Rate / Demographic Restrictions
- Gold Sets/Honey Pots
- Redundancy
- Qualification Test
- Verification/Review
- Justification/Automatic Verification
- ...

Quality Techniques

- Approval Rate / Demographic Restrictions
- Gold Sets/Honey Pots
- Redundancy
- Qualification Test
- Verification/Review
- Justification/Automatic Verification
- ...

19
## Approval Rate & Demographic Restrictions

<table>
<thead>
<tr>
<th>Classify text about consumer electronics</th>
<th>View a HIT in this group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Requester:</strong> Buzz Evaluation</td>
<td><strong>HIT Expiration Date:</strong> Sep 7, 2011 (1 week 6 days)</td>
</tr>
<tr>
<td><strong>Time Allotted:</strong> 20 minutes</td>
<td><strong>HITs Available:</strong> 3966</td>
</tr>
</tbody>
</table>

**Description:** Classify text for positive, negative, mixed or neutral tone

**Keywords:** buzz, classify, coding, tag, sentiment, text, analysis, twitter, blog, social

**Qualifications Required:**
- HIT approval rate (%) is not less than 95
- Location is US

### Points:
- + Easy to setup
- + Transparent
- - Easy to defeat
- - Causes a lot of trouble
HIT Group » I recently did **299 HITs for this requester**.... Of the 299 HITs I completed, **11 of them were rejected** without any reason being given. **Prior to this I only had 14 rejections, a .2% rejection rate.** I currently have **8522 submitted HITs, with a 0.3% rejection rate after the rejections from this requester (25 total rejections).** I have attempted to contact the requester and will update if I receive a response. Until then be very wary of doing any work for this requester, as it appears that they **are rejecting about 1 in every 27 HITs being submitted.** posted by ...

fair:2 / 5  fast:4 / 5  pay:2 / 5  comm:0 / 5
Gold Sets / Honey Pots

• Gold derived from
  – Experts
  – Crowd using high quorum
• Interject trap questions
• Block users in trap and invalidate answers
+ Often very effective
+ Cost efficient
- Not always applicable
- Digging gold is hard
Defeating Honey Pots: reCAPTCHA

data

honey

1629 capolats | 1960-73

RICHARD, Redfyll | ansunjon

drôle

Doratil | than Morgicar | Colightly, Bylencor

Type the two words:
Redundancy: Quorum Votes

- Easy to implement
- Hard to defeat
- Increased cost
- Masks cases of ambiguity or diversity, “tail” behaviors
- Does not cover bias
Challenges

• Quality
• User Interface Design
• Worker motivation
• Task decomposition
• Leverage worker knowledge/capabilities
• Optimization (time/cost)
• ....
Use Cases

• Data collection:
  – How do my prices compare to the prices of my competitors
  – Finding job candidates (who is graduating from HPI next year)
  – Find green-tech companies in the Bay Area
  – …
• Data cleaning
  – Verifying customer addresses
  – Duplicate elimination
  – …
• Extending data
  – Labeling (spam/not_spam)
  – …
**CrowdSQL**

**DDL Extensions:**
*Crowdsourced columns*

```sql
CREATE TABLE company ( name STRING PRIMARY KEY, hq_address CROWD STRING);
```

**Crowdsourced tables**

```sql
CREATE CROWD TABLE department ( university STRING, department STRING, phone_no STRING) PRIMARY KEY (university, department);
```

**DML Extensions:**
*CrowdEqual:*

```sql
SELECT * FROM companies WHERE Name ~ "Big Blue"
```

*CROWDORDER operators (currently UDFs):*

```sql
SELECT p FROM picture WHERE subject = "Golden Gate Bridge" ORDER BY CROWDORDER(p, "Which pic shows better %subject");
```
Optimization: Quality

CROWD TABLE professor(\text{name, e-mail})
CROWD TABLE department(\text{name, phone-nb})

SELECT *
FROM professor p, department d
WHERE d.name = p.dep
AND p.name = “Michael J. Carey”
A Bigger(?) Underlying Issue

Closed-World

Open-World
SELECT COUNT(*) FROM IceCreamFlavors

Enter a flavor of ice cream

Requester: trash  Reward: $0.01 per HIT  HITs Available: 51  Duration: 10 minutes
Qualifications Required: HIT approval rate (\%) is greater than 90

Enter a flavor of ice cream

In the textbox below, please enter a flavor of ice cream

answer:

Trashkowsky et al. Getting it All From the Crowd, (in preparation) on arxiv
Estimating Completeness

```
SELECT COUNT(*) FROM US States
```

US States using Mechanical Turk

![Graph showing unique items over time](image)
SELECT COUNT(*) FROM US States

US States using Mechanical Turk

Unique items over time

Crowd set-size estimation

\[ \Phi_{orig} = 0.046 \]
\[ \Phi_{new} = 0.053 \]
Estimating Completeness

SELECT COUNT(*) FROM IceCreamFlavors

• Ice Cream Flavors
  – Estimators don’t converge
  – Very highly skewed (CV = 5.8)
  – Detect that # HITs insufficient (beginning of curve)

Few, short lists of ice cream flavors (e.g. “alumni swirl, apple cobbler crunch, arboretum breeze,...” from Penn State Creamery)
Pay-As-You-Go

• “I don’t believe it is usually possible to estimate the number of species... but only an appropriate lower bound for that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species”
  – Good, 1953

• So instead, can ask: “What’s the benefit of $m$ additional HITs?”

<table>
<thead>
<tr>
<th>m</th>
<th>Actual</th>
<th>Shen</th>
<th>Spline</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>1.79</td>
<td>1.62</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>8.91</td>
<td>8.22</td>
</tr>
<tr>
<td>200</td>
<td>39</td>
<td>35.4</td>
<td>32.9</td>
</tr>
</tbody>
</table>
## Entity Resolution

<table>
<thead>
<tr>
<th>ID</th>
<th>Product Name</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>iPad Two 16GB WiFi White</td>
<td>$490</td>
</tr>
<tr>
<td>$r_2$</td>
<td>iPad 2nd generation 16GB WiFi White</td>
<td>$469</td>
</tr>
<tr>
<td>$r_3$</td>
<td>iPhone 4th generation White 16GB</td>
<td>$545</td>
</tr>
<tr>
<td>$r_4$</td>
<td>Apple iPhone 4 16GB White</td>
<td>$520</td>
</tr>
<tr>
<td>$r_5$</td>
<td>Apple iPhone 3rd generation Black 16GB</td>
<td>$375</td>
</tr>
<tr>
<td>$r_6$</td>
<td>iPhone 4 32GB White</td>
<td>$599</td>
</tr>
<tr>
<td>$r_7$</td>
<td>Apple iPad2 16GB WiFi White</td>
<td>$499</td>
</tr>
<tr>
<td>$r_8$</td>
<td>Apple iPod shuffle 2GB Blue</td>
<td>$49</td>
</tr>
<tr>
<td>$r_9$</td>
<td>Apple iPod shuffle USB Cable</td>
<td>$19</td>
</tr>
</tbody>
</table>
Hybrid Entity Resolution

J. Wang et al. CrowdER: Crowdsourcing Entity Resolution, PVLDB 2012
Human-Tolerant Computing

Adding People into the Analytics Lifecycle:
• Inconsistent answer quality
• Incentives
• Latency & Variance
• Open vs. Closed world
• Hybrid Human/Machine Design

Approaches:
• Statistical methods for error and bias
• Quality-conscious Interface design
• Cost (time, quality)-based optimization
The AMPLab looks into integrating Algorithms, Machines and People for big data analytics

• Crowdsourcing can help with Big Data analytics where machines are not enough

• CrowdDB is a first hybrid Crowd/Cloud data management system following this vision

• Full tutorial: *Crowdsourcing Applications and Platforms: A Data Management Perspective*. VLDB, 2011

• Try it out at mturk.com

Tim Kraska
kraska@cs.berkeley.edu