Crowdsourcing for Relevance Evaluation

Omar Alonso
Microsoft

28 March 2010

Disclaimer

The views and opinions expressed in this tutorial are mine and do not necessarily reflect the official policy or position of Microsoft.
Tutorial Outline

I. Introduction to crowdsourcing
II. Amazon Mechanical Turk
III. Design of experiments

Tutorial objectives

• When to use crowdsourcing for an experiment
• How to use Mechanical Turk
• How to setup experiments
• Apply design guidelines
• Quality control
INTRODUCTION TO CROWDSOURCING

“Honored guests, Madame Chairwoman, distinguished panelists, thank you for inviting me today. I’m a little nervous, as I’m not accustomed to speaking in public, so please forgive me if I begin vomiting.”
Introduction

• What is relevance?
  – Multidimensional
  – Dynamic
  – Complex but systematic and measurable
• How to measure relevance?

Relevance and IR

• Relevance in Information Retrieval
• Frameworks
• Types
  – System or algorithmic
  – Topical
  – Pertinence
  – Situational
  – Motivational
Evaluation

• Relevance is hard to evaluate
  – Highly subjective
  – Expensive to measure
• Click data
• Professional editorial work
• Verticals

“Snow. Snow is relevant.”
You have a new idea

- Novel IR technique
- Don’t have access to click data
- Can’t hire editors
- How to test new ideas?

Crowdsourcing

- Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.
- The application of Open Source principles to fields outside of software.
Crowdsourcing

- Outsource micro-tasks
- Success stories
  - Wikipedia
  - Apache
- Power law
- Attention
- Incentives
- Diversity

Human-based Computation

- Use humans as processors in a distributed system
- Address problems that computers aren’t good
- Games with a purpose
- Examples
  - ESP game
  - Captcha
  - ReCaptcha

Crowdsourcing and relevance evaluation

- For relevance, it combines two main approaches
  - Explicit judgments
  - Automated metrics
- Other features
  - Large scale
  - Inexpensive
  - Diversity

Why is this interesting?

- Easy to prototype and test new experiments
- Cheap and fast
- No need to setup infrastructure
- Introduce experimentation early in the cycle
- In the context of IR, implement and experiment as you go
- For new ideas, this is very helpful
Caveats

- Trust and reliability
- Spam
- Wisdom of the crowd re-visit

Other clarifications

- Adjust expectations
- Crowdsourcing is another data point for your analysis
- Complementary to other experiments
Examples

• A closer look at previous work with crowdsourcing
• Includes experiments using AMT
• Subset of current research
• Wide range of topics
  – NLP, IR, Machine Translation, etc.

NLP

• AMT to collect annotations
• Five tasks: affect recognition, word similarity, textual entailment, event temporal ordering
• High agreement between workers and gold standard
• Bias correction for non-experts

Machine Translation

• Manual evaluation on translation quality is slow and expensive
• High agreement between non-experts and experts
• $0.10 to translate a sentence


Data quality

• Data quality via repeated labeling
• Repeated labeling can improve label quality and model quality
• When labels are noisy, repeated labeling can preferable to a single labeling
• Cost issues with labeling

Quality control on relevance assessments

- INEX 2008 Book track
- Home grown system (no AMT)
- Propose a game for collecting assessments
- CRA Method


Page Hunt

- Learning a mapping from web pages to queries
- Human computation game to elicit data
- Home grown system (no AMT)
- More info: pagehunt.msrlivelabs.com

Snippets

- Study on summary lengths
- Determine preferred result length
- Asked workers to categorize web queries
- Asked workers to evaluate the quality of snippets
- Payment between $0.01 and $0.05 per HIT


Crowdsourcing for Relevance Evaluation

TREC

- Can we get rid of TREC assessors?
- Can we replace TREC-like relevance assessors with Mechanical Turk?


Crowdsourcing for Relevance Evaluation
Experiments

• Selected topic “space program” (011)
• Subset of 29 FBIS documents (14 not relevant, 15 relevant)
• Modified original 4-page instructions from TREC
• Each document judged by 10 workers
• Performed 5 experiments

Results
Results – II

- Workers more accurate than original assessors
- Disagreement in 4 documents
- 40% provided justification for each answer

Worker feedback

- Not relevant documents
  - This report is about the Russian economy, not the space program.
  - The “MIR” in the article refers to a political group, not the Russian space station.
  - This article is about Kashmir, not the space program.

- Relevant documents
  - This is about Japan's space program and even refers to a launch.
  - On the Russian space program, not US, but comments about American interest in the program.
  - The article is relevant, but it seems a non-native English speaker wrote it. For instance the article says the space shuttle will lift off from the “cosmodrome”. NASA doesn't call the launch pad a "cosmodrome."
INEX

- INEX assessment using AMT
- Assessment is done among benchmark participants
- Problem: each topic assessed by 1 or 2 different persons
- Assessor fatigue
- Can we do better with crowdsourcing?


Experiment

- In INEX an assessor highlights (using a tool) relevant passages.
- AMT is form-based so difficult to replicate same interaction
- Solution
  - Perform element-based assessment
  - article, body, sec, and p
- Qualification test on topics
- Binary evaluation, 5 workers, $0.01 per task
- 1 week to complete
Results

• Agreement between INEX and workers

Worker feedback

“Relevant” answers for [Salad Recipes]

Doesn’t mention the word ‘salad’, but the recipe is one that could be considered a salad, or a salad topping, or a sandwich spread.
Egg salad recipe
Egg salad recipe is discussed.
History of salad cream is discussed.
Includes salad recipe
It has information about salad recipes.
Potato Salad
Potato salad recipes are listed.
Recipe for a salad dressing.
Salad Recipes are discussed.
Salad cream is discussed.
Salad info and recipe
The article contains a salad recipe.
The article discusses methods of making potato salad.
This article describes a specific salad. Although it does not list a specific recipe, it does contain information relevant to the search topic.
gives a recipe for tuna salad
relevant for tuna salad recipes
relevant to salad recipes
this is on-topic for salad recipes
Worker feedback - II

“Not relevant” answers for [Salad Recipes]

About gaming not salad recipes.
Article is about Norway.
Article is about Region Codes.
Article is about forests.
Article is about geography.
Document is about forest and trees.
Has nothing to do with salad or recipes.
Not a salad recipe
Not about recipes
Not about salad recipes
There is no recipe, just a comment on how salads fit into meal formats.
There is nothing mentioned about salads.
While dressings should be mentioned with salads, this is an article on one specific type of dressing, no recipe for salads.
article about a swiss tv show
completely off-topic for salad recipes
not a salad recipe
not about salad recipes
totally off base
Another TREC experiment

- A large TREC-8 evaluation on AMT
- All 50 topics
- How to do it?
  - Budget
  - People, queries, documents
  - How to present information for relevance assessment?

Methodology

- Four parameters
  - P (people)
  - T (topics)
  - D (documents)
  - $$
- Data preparation
- Interface design
- Filtering bad workers
- Scheduling
Worker feedback

• Justification
  – Scale may not be appropriate: “some relevance”, “not totally relevant”
  – How people justify not relevant
  – How people justify relevant

• Operational
  – Broken link, site down

• Communication
  – I will post a positive feedback for you at Turker Nation
  – I mean to tag this as ‘relevant’ but clicked ‘submit’ to quickly

Timeline annotation

• Workers annotate timeline on politics, sports, culture

• Bi-partite graph
  – Match a temporal expression to an event
  – Match an event to a temporal expression

• Given a timex (1970s, 1982, etc.) suggest something

• Given an event (Vietnam, World cup, etc.) suggest a timex

Twitter

- Detecting uninteresting content text streams
- Is this tweet interesting to the author and friends only?
- Workers classify tweets
- 5 tweets per HIT, 5 workers, $0.02
- 57% is categorically not interesting

Next steps

- Evidence from a wide range of projects
- Can I *crowdsource* my experiment?
- How do I start?
- What do I need?
AMAZON MECHANICAL TURK
AMT

- Amazon Mechanical Turk (AMT, www.mturk.com)
- Crowdsourcing platform
- On-demand workforce
- "Artificial artificial intelligence": get humans to do hard part
- Named after "The Turk", a fake chess playing machine
- Constructed by Wolfgang von Kempelen in 18th C.

AMT – How it works

- Requesters create "Human Intelligence Tasks" (HITs) via web services API or dashboard
- Workers (sometimes called "Turkers") log in, choose HITs, perform them
- Requesters assess results, pay per HIT satisfactorily completed
- Currently >200,000 workers from 100 countries; millions of HITs completed
The Worker

- Sign up with your Amazon account
- Tabs
  - Account: work approved/rejected
  - HIT: browse and search for work
  - Qualifications: browse and search for qualifications

Example – Relevance evaluation

Judge approximately 45 item search results for relevance

For the following item search results, tell us which of the following items and apply to the search results, give the search term and category to apply:

- Off topic
  - The item is not related to the search query or category at all. A frequent indication of this is when a non-relevant item is prominently displayed in the list or at the bottom of the list.
- Poor
  - The item may be relevant to the search query but the relevance is weak.
- Average
  - The item is relevant to the search query, but the relevance is not strong.
- Good
  - The item is relevant and strong to the search query.
- Excellent
  - The item is relevant and very strong to the search query.

The item image, and where it is in a set. The relevance may be the result of the user's judgment, but be sure to do a set with at least 45 items.
Example – Relevance and ads

Crowdsourcing for Relevance Evaluation

Example – Product search

Crowdsourcing for Relevance Evaluation
Example – Spelling correction

Crowdsourcing for Relevance Evaluation

Sheep Market

- Collection of 10,000 sheep made by workers
- Payment $0.02 to draw a sheep facing left

Crowdsourcing for Relevance Evaluation
Ten Thousand Cents

- Creates a representation of a $100 bill
- Workers painted a part of the bill
- Payment $0.01

Demographics

- Panos Ipeirotis (NYU)
- Survey conducted over 3 weeks
- 1,000 users, payment $0.10 for participating
- 66 countries
  - 46.80% (USA), 34% (India), 19.20% (other)
- Source of income
  - Primary (India)
  - Secondary (USA)
- Complete analysis in Panos blog

behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html
Demographics - II

- Worker population is becoming more international
- Steady increase in the number of male workers
- Younger population
- Average worker earns $2.00/hour
- 18% workers spend more than 15hrs/week on HITs

J. Ross et al. “Who are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk”. CHI 2010

Why do you “turk”? 

- The faces of Mechanical Turk
- Task: upload a picture with a handwritten sign that says “I turk for …”
- Payment
  - $0.05, $0.25, $0.50
- 30 people in total
  - 21 turk for money
  - 9 for fun or boredom

waxy.org/2008/11/the_faces_of_mechanical_turk/
The Requester

- Sign up with your Amazon account
- Amazon payments
- Purchase prepaid HITs
- There is no minimum or up-front fee
- AMT collects a 10% commission
- The minimum commission charge is $0.005 per HIT

Dashboard

- Three tabs
  - Design
  - Publish
  - Manage
- Design
  - HIT Template
- Publish
  - Make work available
- Manage
  - Monitor progress
Dashboard - II

Crowdsourcing for Relevance Evaluation

API

- Amazon Web Services API
- Rich set of services
- Command line tools
- More flexibility than dashboard
Practical discussion

• Dashboard
  – Easy to prototype
  – Setup and launch an experiment in a few minutes

• API
  – Ability to integrate AMT as part of a system
  – Ideal if you want to run experiments regularly
  – Schedule tasks

BREAK
Hands on

- Design two experiments
- Show all details
- Launch and monitor progress
Query classification task

• Ask the user to classify a query
• Show a form that contains a few categories
• Upload a few queries (~20)
• Use 5 workers

Relevance evaluation task

• Relevance assessment task
• Use a few documents from TREC
• Ask user to perform binary evaluation
• Modification: graded evaluation
• Use 5 workers
DESIGN OF EXPERIMENTS

“Whoever designed this place is insane.”
Workflow

- Define and design what to test
- Sample data
- Design the experiment
- Run experiment
- Collect data and analyze results
- Quality control

Crowdsourcing for Relevance Evaluation

Survey design

- One of the most important parts
- Part art, part science
- Instructions are key
- Prepare to iterate
Questionnaire design

• Ask the right questions
• Workers may not be IR experts so don’t assume the same understanding in terms of terminology
• Show examples
• Hire a technical writer

UX design

• Time to apply all those usability concepts
• Generic tips
  – Experiment should be self-contained.
  – Keep it short and simple. Brief and concise.
  – Be very clear with the relevance task.
  – Engage with the worker. Avoid boring stuff.
  – Always ask for feedback (open-ended question) in an input box.
UX design - II

- Presentation
- Document design
- Highlight important concepts
- Colors and fonts
- Need to grab attention
- Localization

Examples - I

- Asking too much, task not clear, “do NOT/reject”
- Worker has to do a lot of stuff
Example - II

- Lot of work for a few cents
- Go here, go there, copy, enter, count ...

Example - III

- Go somewhere else and issue a query
- Report, click, ...
A better example

• All information is available
  – What to do
  – Search result
  – Question to answer

Form and metadata

• Form with a close question (binary relevance) and open-ended question (user feedback)
• Clear title, useful keywords
• Workers need to find your task
TREC assessment example

Payments

- How much is a HIT?
- Delicate balance
  - Too little, no interest
  - Too much, attract spammers
- Heuristics
  - Start with something and wait to see if there is interest or feedback ("I’ll do this for X amount")
  - Payment based on user effort. Example: $0.04 (2 cents to answer a yes/no question, 2 cents if you provide feedback that is not mandatory)
- Bonus
- The anchor effect
Development

• Similar to a UX design and implementation
• Build a mock up and test it with your team
• Incorporate feedback and run a test on AMT with a very small data set
  – Time the experiment
  – Do people understand the task?
• Analyze results
  – Look for spammers
  – Check completion times
• Iterate and modify accordingly

Development – II

• Introduce qualification test
• Adjust passing grade and worker approval rate
• Run experiment with new settings and same data set
• Scale on data
• Scale on workers
Experiment in production

• Lots of tasks on AMT at any moment
• Need to grab attention
• Importance of experiment metadata
• When to schedule
  – Split a large task into batches and have 1 single batch in the system
  – Always review feedback from batch $n$ before uploading $n+1$

Quality control

• Extremely important part of the experiment
• Approach it as “overall” quality – not just for workers
• Bi-directional channel
  – You may think the worker is doing a bad job.
  – The same worker may think you are a lousy requester.
A qualification test

Answer

Properties

# Basic qualification attributes

name= Generic knowledge quiz on topics
description=This qualification tests your general knowledge about a wide range of topics: knowledge, geography, people, places, history, art, current and past events, trec
retrydelayinseconds=3600

# Workers will have 15 minutes to complete this test. 15 minutes = 60 seconds * 15 minutes = 900 testdurationseconds=900
Observations on qualification tests

• Advantages
  – Great tool for controlling quality
  – Adjust passing grade

• Disadvantages
  – Extra cost to design and implement the test
  – May turn off workers
  – Refresh the test on a regular basis

“Plus, we double the accuracy at no extra cost by using our extensive pool of Siamese twins.”
Filtering bad workers

- Approval rate
- Qualification test
  - Problems: slows down the experiment, difficult to “test” relevance
  - Solution: create questions on topics so user gets familiar before starting the assessment
- Still not a guarantee of good outcome
- Interject gold answers in the experiment
- Identify workers that always disagree with the majority

More on quality

- Lots of ways to control quality:
  - Better qualification test
  - More redundant judgments
  - More than 5 workers seems not necessary
- Various methods to aggregate judgments
  - Voting
  - Consensus
  - Averaging
Methods for measuring agreement

• What to look for
  – Agreement, reliability, validity

• Inter-agreement level
  – Agreement between judges
  – Agreement between judges and the gold set

• Gray areas
  – 2 workers say “relevant” and 3 say “not relevant”
  – 2-tier system

Inter-rater reliability

• Lots of research
• Statistics books cover most of the material
• Three categories based on the goals
  – Consensus estimates
  – Consistency estimates
  – Measurement estimates
Statistics

- Cohen’s kappa  
  - Two raters
- Fleiss’ kappa  
  - Any number of raters
- Krippendorff’s alpha

Crowdsourcing for Relevance Evaluation

Was the task difficult?

- Ask turkers to rate the difficulty of a topic
- 50 topics, TREC
- 5 workers, $0.01 per task

<table>
<thead>
<tr>
<th>Exp</th>
<th>Title</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>412</td>
<td>airport security</td>
<td>1.4</td>
</tr>
<tr>
<td>439</td>
<td>inventions, scientific discoveries</td>
<td>1.6</td>
</tr>
<tr>
<td>438</td>
<td>tourism, increase</td>
<td>1.6</td>
</tr>
<tr>
<td>428</td>
<td>declining birth rates</td>
<td>1.6</td>
</tr>
<tr>
<td>424</td>
<td>suicides</td>
<td>1.6</td>
</tr>
<tr>
<td>442</td>
<td>heroic acts</td>
<td>1.8</td>
</tr>
<tr>
<td>436</td>
<td>railway accidents</td>
<td>1.8</td>
</tr>
<tr>
<td>433</td>
<td>Greek, philosophy, stoicism</td>
<td>3.6</td>
</tr>
<tr>
<td>405</td>
<td>cosmic events</td>
<td>3.6</td>
</tr>
<tr>
<td>401</td>
<td>foreign minorities, Germany</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Other quality heuristics

• Justification/feedback as captcha
  – Successfully used at TREC and INEX experiments
  – Should be optional

• Broken URL/incorrect object
  – Leave an outlier in the data set
  – Workers will tell you
  – If somebody answers “excellent” on a graded relevance test for a broken URL => probably a spammer

Dealing with bad workers

• Always pay
• Avoid rejecting workers
• Use bonus as incentive
  – Pay the minimum $0.01 and $0.01 for bonus
  – Better than rejecting a $0.02 task
• You may still be dealing with a sophisticated spammer
  – Block worker for next experiments
Worker feedback

• Real examples of feedback via email after a rejection

• Worker XXX
  I did. If you read these articles most of them have nothing to do with space programs. I’m not an idiot.

• Worker XXX
  As far as I remember there wasn't an explanation about what to do when there is no name in the text. I believe I did write a few comments on that, too. So I think you're being unfair rejecting my HITs.

Exchange with worker

• Worker XXX
  Thank you. I will post positive feedback for you at Turker Nation.

Me: was this a sarcastic comment?

• I took a chance by accepting some of your HITs to see if you were a trustworthy author. My experience with you has been favorable so I will put in a good word for you on that website. This will help you get higher quality applicants in the future, which will provide higher quality work, which might be worth more to you, which hopefully means higher HIT amounts in the future.
Results

• Word of mouth effect
  – Workers trust the requester (pay on time, clear explanation if there is a rejection)
  – Experiments tend to go faster
  – Announcement of forthcoming tasks

Other practical tips

• Sign up as worker and do some HITs
• Eat your own dog food
• Monitor Turker Nation (turkers.proboards.com)
• Discussion forums (aws.amazon.com/mturk/)
• Tweet your experiment
• Establish your fan base
• Address feedback (e.g., poor guidelines, payments, passing grade, etc.)
• Everything counts!
More tips

- Randomize content
- Avoid worker fatigue
  - Judging 100 straight documents on the same subject can be tiring
- Length of the task
- Content presentation

“All I know is that searching for my own name and then clicking on ‘highly relevant’ does wonders for my self-esteem.”
Platform alternatives

- Do I have to use AMT?
- How to build your own crowdsourcing platform
  - Back-end
  - Template language for creating experiments
  - Scheduler
  - Payments?

MapReduce with human computation

- MapReduce meets crowdsourcing
- Commonalities
  - Large task divided into smaller sub-problems
  - Work distributed among worker nodes (turkers)
  - Collect all answers and combine them
- Variabilities
  - Human response time varies
  - Some tasks are not suitable
Challenges and opportunities

• A back-end perspective
• Problems with the current platform
  – Very rudimentary
  – No tools for data analysis
  – No integration with databases
  – Very limited search and browse features
• Opportunities
  – What is the database model for crowdsourcing?
  – MapReduce with crowdsourcing
  – Can you integrate human-computation into a language?
    • crowdsourcing(task,5)

Research questions

• What are the tasks suitable for crowdsourcing?
• What is the best way to perform crowdsourcing?
Conclusions

- Crowdsourcing for relevance evaluation works
- Fast turnaround, easy to experiment, few dollars to test
- But you have to design the experiments carefully
- Usability considerations
- Worker quality
- User feedback extremely useful

Conclusions - II

- Crowdsourcing is here to stay
- Lots of opportunities to improve current platforms
- Integration with current systems
- AMT is a popular platform and others are emerging
- Open research problems
Bibliography

Bibliography - III


Thank You!

For questions about tutorial or crowdsourcing, please email me to: oralonso@gmail.com

Cartoons by Mateo Burtch (buta@mindspring.com)