Design of Experiments for Crowdsourcing Search Evaluation: challenges and opportunities

Omar Alonso
Microsoft
23 July 2010
Disclaimer

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

• Mechanical Turk works
  – Evidence from a wide range of projects
  – Several papers published
  – SIGIR workshop
• Can I **crowdsourcing** my experiment?
  – How do I start?
  – What do I need?
• Challenges and opportunities in relevance evaluation
“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
A methodology

• Data preparation
• UX design
• Quality control
• Scheduling
Questionnaire design

• Instructions are key
• Ask the right questions
• Workers are not IR experts so don’t assume the same understanding in terms of terminology
• Show examples
• Hire a technical writer
• Prepare to iterate
UX design

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

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

Help us describe How-To Videos! Earn $2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the Film & Theater how-to video → 332492 Get a 35mm film look with a depth of field adapter
2. Write a description of the video in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete 25 video descriptions in order to earn the $2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is not necessary to repeat the headline in your entry. It will NOT count toward your word count.
9. Do NOT describe the following: the format, where the video comes from, or how long the video is. This information is IRRELEVANT.
10. Do NOT describe the video in the following manner: “She turns around to face the camera. Then she faces left.” Follow the examples below.

Criteria for REJECTION:

- Entries with obvious and multiple spelling or grammatical errors will be rejected.
- Entries with fewer than 100 words will be automatically rejected.
- Text copied from the web or other places will be rejected. Multiple plagiarized answers will lead to being BLOCKED. You may use a quotation, but the majority of your content must be ORIGINAL.
- Incomplete and blank answers will be rejected. Multiple blank answers will result in being blocked.
- Tasks submitted without descriptions will be rejected.
- Tasks submitted with inaccurate descriptions will be rejected as well.
- Do NOT add any personal opinions. Entries with personal opinions or reviews will be automatically REJECTED.
- If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in rejection.
- Entries that transcribe the video will be REJECTED.
Example - II

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

Search for a topic and collect details about advertisers

Go to www.ezcloud.com. In the Menu on the right side you will find the menu entry "Search". Click on that Menu Entry which will take you to EZCLOUD's Search Page. Or go here. You must use the Search page provided on EZCLOUD's website or your reply will be rejected.

Search for "mustang decal"

1. Copy the url of the search results here

2. Enter the url of the top placed advertiser

3. Count how many different advertisers are shown on the results page. Include all advertisers (don't forget advertisers at the bottom of the page) If results page does not show advertisers enter "no advertisers". We will verify every answer before we approve your reply.
Example - III

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

Report on Total Number of Results for Search

Step 1: Go to http://www.google.com and search for "Hotelscombined singapore"

Step 2: Report on number of results returned (this number is shown on the top right of page)

Step 3: Click on the result ‘Singapore Hotels: Compare Cheap Singapore Accommodation Deals’ at the top of the listing. This will appear below the yellow box. DO NOT click on the sponsored listing in the yellow box.

Link is working

Please provide feedback on this HIT... we'd love to know if the task was clear.
A better example

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

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

Describe your HIT

**Title**: Relevance evaluation for news articles

*Describe the task to workers. Be as specific as possible, e.g., “answer a survey about movies.”*

**Description**: Please help us evaluate relevance for the following document.

*Give more detail about this task. This gives workers a bit more information before they decide to accept it.*

**Keywords**: relevance, news articles, search, TREC, cosmic events, tropical storms, Sch:

Provide keywords that will help workers search for your HITs.

**Task**

Please evaluate the relevance of the following document about cosmic events.

**Description**: What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?

More information: New theories or new interpretations concerning known celestial objects made as a result of new technology are not relevant.

**Qualitative Analysis of Some Methods of Reducing the Asteroid Hazards for the Earth**

[Article by V. V. Ivashkin, V. V. Smirnov, Institute of Applied Mathematics имени М. В. Кельдыш, Russian Academy of Sciences; UDC 629]

[Abstract] The probability of a collision between the Earth and an asteroid such as Amor, Apollo, or Aton is not at all small. The work reported here consists of the results of a preliminary analysis of the use of various methods for preventing such a collision, all of which involve changing the asteroid’s orbit: spacecraft vehicle impact, delivery and attachment of a large-thrust engine to the asteroid, attachment of low-thrust electroreactive engines, attachment of a solar sail, and changing the color (thus, the reflective properties) of the asteroid’s surface. Prevention of the collision is considered to be guaranteed if the flyby distance is at least 1 million kilometers. The asteroid, in the calculations, is taken to be a sphere with a density of 3 g/cm³. Estimates are made for asteroids with radii of 5 m, 50 m, and 500 m and masses of 1.57 x 10⁶ kg and 1.57 x 10⁷ kg. After a comparative analysis of the methods, the researchers chose the impact method as the most effective method and a method that could be performed with existing technology. Numerical
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 MTurk 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 \textit{first}
• Scale on workers
Experiment in production

- Ad-hoc experimentation vs. ongoing metrics
- Lots of tasks on MTurk 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.
Quality control - II

• 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
Methods for measuring agreement

• What to look for
  – Agreement, reliability, validity

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

• Some statistics
  – Cohen’s kappa (2 raters)
  – Fleiss’ kappa (any number of raters)
  – Krippendorff’s alpha

• Gray areas
  – 2 workers say “relevant” and 3 say “not relevant”
  – 2-tier system
And if it doesn’t work ..
Time to re-visit things ...

• Crowdsourcing offers flexibility to design and experiment
• Need to be creative
• Test different things
• Let’s dissect items that look trivial
The standard template

• Assuming a lab setting
  – Show a document
  – Question: “Is this document relevant to the query”?  
• Relevance is hard to evaluate
• Barry & Schamber
  – Depth/scope/specificity
  – Accuracy/validity
  – Clarity
  – Recency
Content quality

• People like to work on things that they like

• TREC ad-hoc vs. INEX
  – TREC experiments took twice to complete
  – INEX (Wikipedia), TREC (LA Times, FBIS)

• Topics
  – INEX: Olympic games, movies, salad recipes, etc.
  – TREC: cosmic events, Schengen agreement, etc.

• Content and judgments according to modern times
  – Airport security docs are pre 9/11
  – Antarctic exploration (global warming)
Content quality - II

• Document length
• Randomize content
• Avoid worker fatigue
  – Judging 100 documents on the same subject can be tiring
Presentation

• People scan documents for relevance cues
• Document design

EU Approves Aid To Boost N. Ireland Peace Efforts

A 36 million pound contribution to peace efforts in the North was approved by the European Commission yesterday.

The cash will go to the International Fund for Ireland (IFI) over the next three years, on top of 72 million pounds the EU (European Union) has provided since 1989.

The IFI was set up by the British and Irish Governments in 1986 to promote reconciliation as well as providing social and economic aid for the North.

But the latest money from Brussels is being billed as a special response to the efforts of the two Governments to end violence.

Commission President Jacques Delors said yesterday's initiative should be seen as an "expression of practical support for the peace process in Northern Ireland" in the wake of the Downing Street Declaration.

The 36 million pounds will be used for projects aimed at bridging the religious divide.

Please rate the above document according to its relevance to Ireland, peace talks as follows. Note that the task is about how relevant the topic the document is.
Presentation - II

Highlighting vs Plain

SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation
Scales and labels

• Binary
  – Yes, No

• Likert
  – Strongly disagree, disagree, neither agree nor disagree, agree, strongly agree

• DCG paper
  – Irrelevant, marginally, fairly, highly

• Other examples
  – Perfect, excellent, good, fair, bad
  – Highly relevant, relevant, related, not relevant
  – 0..10 (0 == irrelevant, 10 == relevant)
  – Not at all, to some extent, very much so, don’t know (David Brent)

• Usability factors
  – Provide clear, concise labels that use plain language
  – Terminology has to be familiar to assessors
Difficulty of the task

- Some topics may be more difficult
- Ask workers
- TREC example

<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>
Document Relevance Evaluation

Please evaluate the relevance of a document to the given topic. A document is relevant if it directly discusses the topic. Each document should be judged on its own merits. That is, a document is still relevant even if it is the thirtieth document you have seen with the same information.

Tips
- Payment based on quality of the work completed. Please follow the instructions and be consistent in your judgments.
- **Bonus** payment if you provide a good justification.
- Please justify your answer, otherwise you may not get paid.
- A document should not be judged as relevant or irrelevant based only on the title of the document. You **must** read the document.

Task

Please evaluate the relevance of the following document about **art, stolen, forged**.

Description: What incidents have there been of stolen or forged art?

More information: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass-produced art reproductions). Pirated software, music, movies, etc. are not relevant.

---

**CHASE ENDS IN ARREST OF 3 AFTER LATEST JEWEL HEIST;**

**CRIME: SANTA ANA ROBBERY FITS A PATTERN OF NEARLY 100 SIMILAR THEFTS IN THE WEST SINCE 1989. THE SUSPECTS MAY BE PART OF A LOS ANGELES COUNTY RING.**

By WENDY PAULSON, TIMES STAFF WRITER

SANTA ANA

A freeway chase from Huntington Beach to Compton ended with the arrests of three men who allegedly robbed a department store jewelry counter at gunpoint, the latest in a series of Southland jewel heists, police said Thursday.

And although Orange County police were tight-lipped about investigations of two similar robberies in the last month, Los Angeles police said the incidents fit a pattern of nearly 100 similar thefts in the western United States since 1989 that may stem from a criminal network in southwest Los Angeles County.

---

Please rate the above document according to its relevance to **art, stolen, forged** as follows. Note that the task is about how relevant to the topic the document is.

- **Relevant.** A relevant document for the topic.
- **Not relevant.** The document is not good because it doesn't contain any relevant information.

Does the topic look difficult? Please rate the difficulty from 1 to 5 (1=easy, 5=very difficult):

- **1** Easy  **2** Somewhat easy  **3** Neither easy nor difficult  **4** Somewhat difficult  **5** Very difficult

Please justify your answer or comment on your selection. Please use your own words. You may get a bonus payment if your comment is useful.
Relevance justification

• Why settle for a label?
• Let workers justify answers
• INEX
  – 22% of assignments with comments
• Must be optional
• Let’s see how people justify
“Relevant” answers

[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.
The recipe is for a dressing for a salad, so the information is somewhat narrow for the topic but is still potentially relevant for a researcher.
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
"Not relevant" answers

[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
Data analysis tools

• Excel
• R
• Databases
• Pivot
Platform alternatives

• Do I have to use MTurk?
• 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
Platform perspective

• Problems
  – 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?
    • $\text{crowdsource}(\text{task}, 5) > 0.80$
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
- Need to question **all aspects** of relevance evaluation
- Be creative
- Lots of opportunities to improve current platforms
- Integration with current systems
- MTurk is a popular platform and others are emerging
- Open research problems
Announcement