

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

Crowdsourcing for Relevance Evaluation

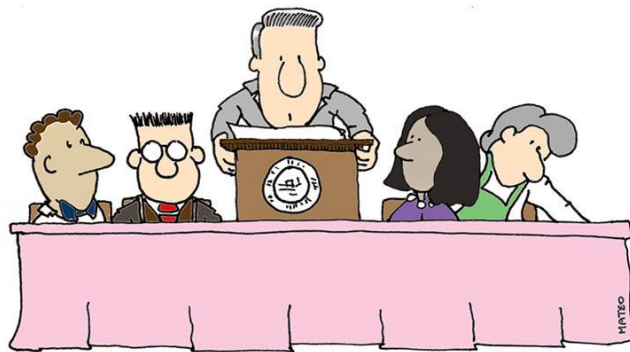
Tutorial objectives

- When to use crowdsourcing for an experiment
- How to use Mechanical Turk
- How to setup experiments
- Apply design guidelines
- Quality control

Crowdsourcing for Relevance Evaluation

INTRODUCTION TO CROWDSOURCING

Crowdsourcing for Relevance Evaluation



"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."

Crowdsourcing for Relevance Evaluation

Introduction

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

Crowdsourcing for Relevance Evaluation

Relevance and IR

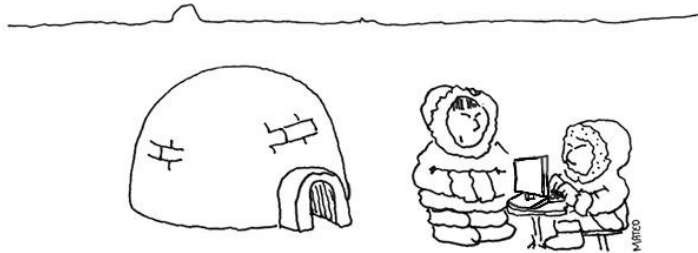
- Relevance in Information Retrieval
- Frameworks
- Types
 - System or algorithmic
 - Topical
 - Pertinence
 - Situational
 - Motivational

Crowdsourcing for Relevance Evaluation

Evaluation

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

Crowdsourcing for Relevance Evaluation



"Snow. Snow is relevant."

Crowdsourcing for Relevance Evaluation

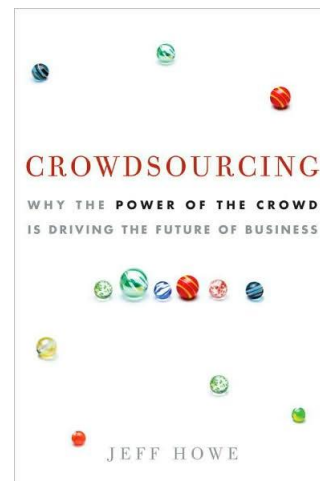
You have a new idea

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

Crowdsourcing for Relevance Evaluation

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 for Relevance Evaluation

Crowdsourcing

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

Crowdsourcing for Relevance Evaluation

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

L. von Ahn. "Games with a purpose". Computer, 39 (6), 92–94, 2006.

Crowdsourcing for Relevance Evaluation

Crowdsourcing and relevance evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

Caveats

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

Crowdsourcing for Relevance Evaluation

Other clarifications

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

Crowdsourcing for Relevance Evaluation

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.

Crowdsourcing for Relevance Evaluation

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

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. "Cheap and Fast But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". EMNLP-2008.

Crowdsourcing for Relevance Evaluation

Machine Translation

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

C. Callison-Burch. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk", EMNLP 2009.

Crowdsourcing for Relevance Evaluation

Data quality

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

V. Sheng, F. Provost, P. Ipeirotis. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers" KDD 2008.

Crowdsourcing for Relevance Evaluation

Quality control on relevance assessments

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

G. Kazai, N. Milic-Frayling, and J. Costello. "Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments", SIGIR 2009.

Crowdsourcing for Relevance Evaluation

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`

H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. "Improving Search Engines Using Human Computation Games", CIKM 2009.

Crowdsourcing for Relevance Evaluation

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

M. Kaiser, M. Hearst, and L. Lowe. "Improving Search Results Quality by Customizing Summary Lengths", ACL/HLT, 2008.

Crowdsourcing for Relevance Evaluation

TREC

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

O. Alonso and S. Mizzaro. "Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment", SIGIR Workshop on the Future of IR Evaluation, 2009.

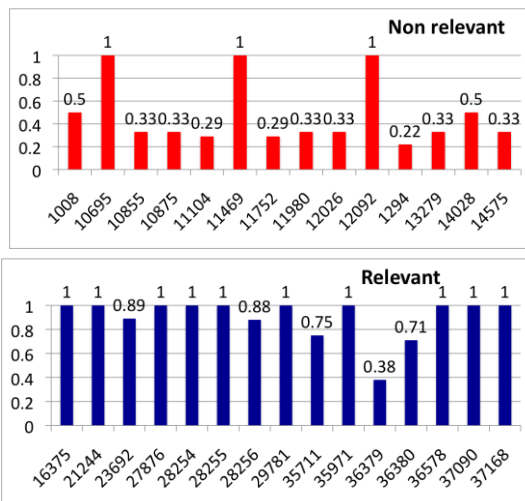
Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

Results



Crowdsourcing for Relevance Evaluation

Results – II

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

Crowdsourcing for Relevance Evaluation

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."

Crowdsourcing for Relevance Evaluation

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?

O. Alonso, R. Schenkel, and M. Theobald . "Crowdsourcing Relevance Assessments for XML Ranked Retrieval", ECIR 2010.

Crowdsourcing for Relevance Evaluation

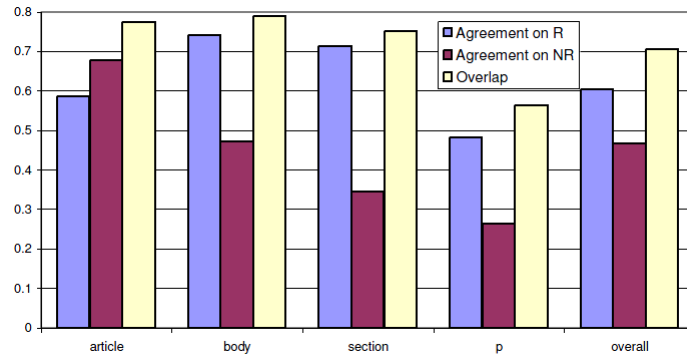
Experiment

- 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

Crowdsourcing for Relevance Evaluation

Results

- Agreement between INEX and workers



Crowdsourcing for Relevance Evaluation

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.
 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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation



“I’m searching on ‘precooked meat product’, but all I’m getting is spam.”

Crowdsourcing for Relevance Evaluation

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?

Crowdsourcing for Relevance Evaluation

Methodology

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

K. Berberich, S. Bedathur, O. Alonso, G. Weikum “A Language Modeling Approach for Temporal Information Needs”. ECIR 2010

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

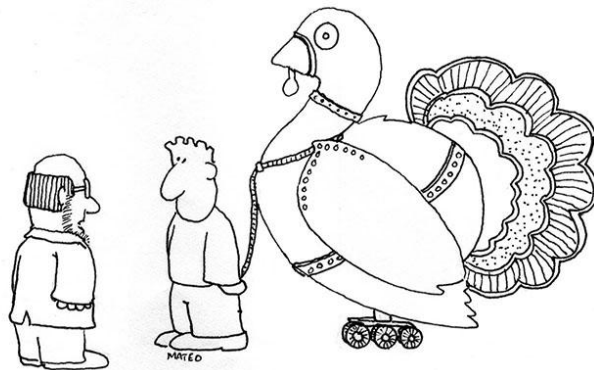
Next steps

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

Crowdsourcing for Relevance Evaluation

AMAZON MECHANICAL TURK

Crowdsourcing for Relevance Evaluation

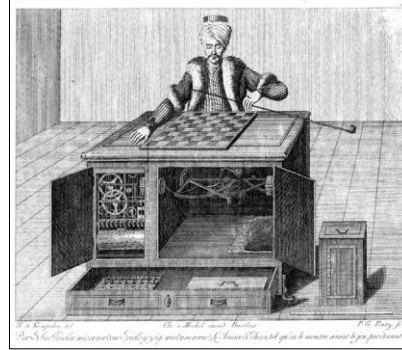


"No, mechanical Turk."

Crowdsourcing for Relevance Evaluation

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.



Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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



Crowdsourcing for Relevance Evaluation

Example – Relevance evaluation

Judge approximately 45 item search results for relevance

Instructions

For the following eBay search results, tell us which of the following labels best apply to the search result, given the search terms and category supplied.

"Off Topic"

An Off Topic result is not related to the search query or category at all. A frequent instance of this are items that contain the search query keywords (usually popular keywords) in the title or subtitle but have no relationship to the item. For example, given the search query "ipod", the following result is considered "Off Topic":



The title, image, and price say that it is a car. iPod-compatibility may be one of the features of the car's stereo, but has nothing to do with a car itself.

"Spam"

A Spam result is one that has no value for the search query or any other eBay search in general. Results that contain a pornographic image are often Spam. For example, for the search query "air conditioner":



Crowdsourcing for Relevance Evaluation

Example – Spelling correction

Evaluate a Spelling Correction for a Product Search Query

Instructions

Imagine that a user is searching for products at an online shopping website. When the user searches for a term, the site suggests a spelling correction, such as "Did you mean: XYZ?". Evaluate whether this spelling correction is **GOOD** or **BAD**. If you aren't sure if the suggestion gives the proper spelling or are not familiar with the search terms, select **I DON'T KNOW**.

When evaluating corrections, ignore capitalization. All search terms and corrections are shown in lower case. A correction can be good even if a space is used instead of a hyphen. For example, "blu ray" and "blu-ray" are both good spelling corrections for "blue ray", even though the trademarked term is "Blu-ray".

Sample search results are provided for context. However, you should base your response on the accuracy of the spelling correction, not the relevance of the results.


Note: We pay bonuses for high-quality responses! You will earn a bonus if your answer is consistent with the majority of respondents. However, if you consistently disagree with the majority, you will be blocked from participating in our future experiments. (An answer is considered to be the majority response when it's selected by five-thirds or more of the workers who complete the test.)

[Instructions](#)


Task

Please evaluate the following spelling correction, using the provided results for context:


User's search query: **eremax** Suggested correction: **erema**




Photographic Print of Coloured X-ray of cancer of the colon from Science Photo Library (Kitchen)
productType: HOME_FURNITURE_AND_DECOR
productGroupID: gl_kitchen
Manufacturer: Science Photo Library
superSaver: false
numberReviews: 0
averageRating: 0.0



Reports on Publications Issued and Registered in the Several Provinces of British India (Paperback)
productType: ABIS_BOOK
productGroupID: gl_book
Author: Home Department, Government of India
superSaver: true
numberReviews: 0
averageRating: 0.0
fastTrack: true
fastTrackIndicate:
ResponseURL: unescapeHtml(\$highlighter.highlight(\$misspelling_diff,\$engine.get('attributes')).get(\$attrkey))
fastTrackQuotedDeliveryCost:
ResponseURL: unescapeHtml(\$highlighter.highlight(\$misspelling_diff,\$engine.get('attributes')).get(\$attrkey))
logPrice: 13.99 GBP



Home esema kit: (Personal Care)
productType: HEALTH_PERSONAL_CARE
productGroupID: gl_drugstore
Manufacturer: Specialist Supplements Ltd.
superSaver: false
numberReviews: 1
averageRating: 4.0



Esema Kit - 3 litre capacity for home and travel (Nec.)
productType: BEAUTY
productGroupID: gl_beauty
Manufacturer: Manifest Health Limited
superSaver: false
numberReviews: 0
averageRating: 0.0

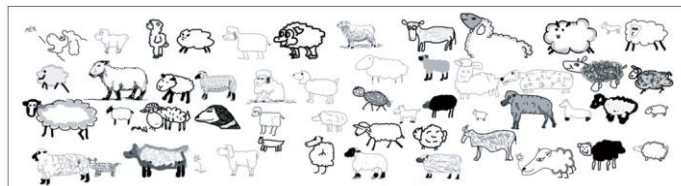
Is the correction of **eremax** to **erema** **GOOD** or **BAD**?

GOOD. Yes, the suggested spelling correction corrects a misspelling.
 BAD. No, the suggested spelling correction is incorrect or unnecessary.
 I DON'T KNOW. Not sure if the suggested spelling correction gives the proper spelling, or not familiar with the search terms.

Crowdsourcing for Relevance Evaluation

Sheep Market

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



www.thesheepmarket.com

Crowdsourcing for Relevance Evaluation

Ten Thousand Cents

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



www.tenthousandcents.com

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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



Crowdsourcing for Relevance Evaluation

Dashboard

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

Edit HIT Template
Specify the properties that are common for all of the HITs created using this template.

1 Enter Properties 2 Design Layout 3 Preview and Finish

Template Name: TREC_Binary_e4 This name is not displayed to workers.

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", instead of "give more detail about this task". This gives workers a lot more information before they decide to view relevance, news articles, search, TREC, airport security, steel production,C

Description Please help us evaluate relevance for the following document.
Give more detail about this task. This gives workers a lot more information before they decide to view relevance, news articles, search, TREC, airport security, steel production,C

Keywords relevance, news articles, search, TREC, airport security, steel production,C
Provide keywords that will help workers search for your HITs.

Working on your HIT

Time allotted per assignment 1 Hours Maximum time a worker has to work on a single task. Be generous so that worker

HIT expires in 10 Days Maximum time your HIT will be available to workers on Mechanical Turk.

Worker must meet the following criteria to work on these HITs:

HIT approval rate (%) greater than or equal to 98 clear

+ Add another criteria. (up to 3)
All criteria must be met for a worker to work on these HITs.

Required for preview [\(what's this?\)](#)
Tip: you can limit your HITs to workers with a certain approval rate. An approval rating of 95% or better is considered

Paying Workers

Reward per assignment \$ 0.04
Tip: Consider how long it will take a worker to complete each task. A \$

Number of assignments per HIT 5
How many unique workers do you want to work on each HIT

Results are automatically approved in 5 Days
After this time, all unreviewed work is approved and workers are paid.

Crowdsourcing for Relevance Evaluation

Dashboard - II

Status		Delete	
Status: Pending Review <div style="display: flex; justify-content: space-around; width: 200px;"> <div style="width: 40%;"><div style="width: 100%; height: 10px; background-color: green;"></div>100% submitted</div> <div style="width: 40%;"><div style="width: 100%; height: 10px; background-color: green;"></div>100% published</div> </div>			
Assignments Completed: 1,035 / 1,035	Average Time per Assignment: 2 Minutes	Average Hourly Rate: \$0.57	
Creation Time: October 15, 2009 9:37 PM PDT	Completion Time: October 17, 2009 6:16 PM PDT		
Settings		Results	
TREC - Graded v2 Description: Please help us evaluate relevance for the following document. Keywords: relevance, news articles, search, TREC, graded relevance, dcg, petroleum exploration, blood-alcohol fatalities Qualification Requirement: HIT approval rate (%) greater than or equal to 98		Assignments pending review: 0 Assignments approved: 1,033 Assignments rejected: 2	
Number of Assignments per HIT: 5 Reward per Assignment: \$0.02 Input File: list2.txt		Cost Summary Estimated Total Reward: \$20.70 Estimated Fees: \$5.175 Estimated Total Cost: \$25.875 <small>These costs are only an estimate until all of the assignments have been submitted and reviewed.</small>	
HIT expires on: EXPIRED Assignment duration: 1 Hours Auto Approval Delay: 3 Days			

Crowdsourcing for Relevance Evaluation

API

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

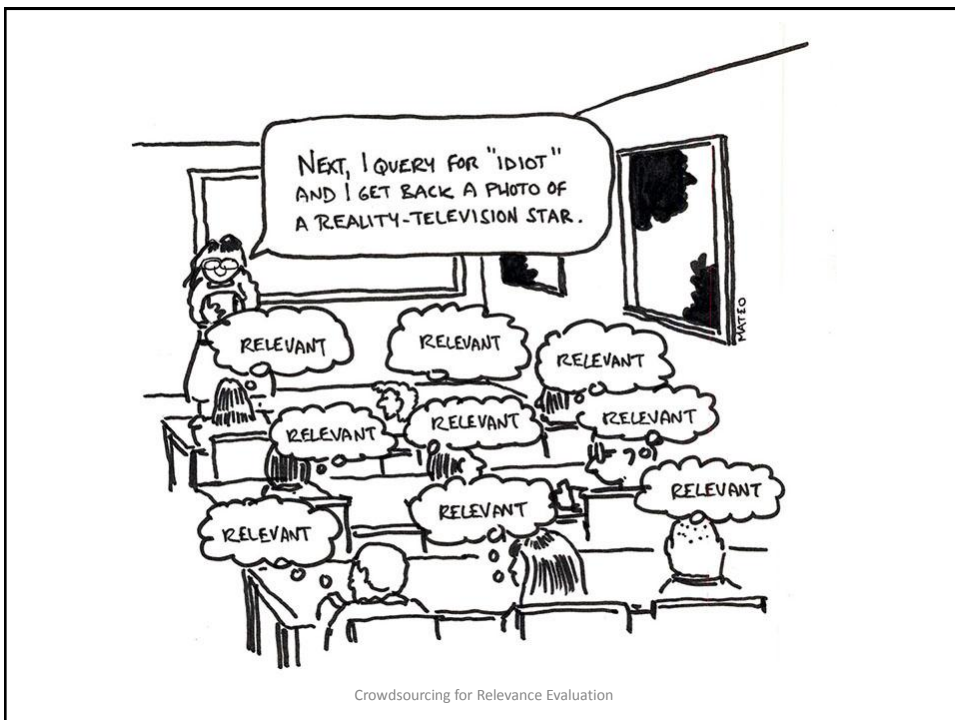
BREAK

Crowdsourcing for Relevance Evaluation

Hands on

- Design two experiments
- Show all details
- Launch and monitor progress

Crowdsourcing for Relevance Evaluation



Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

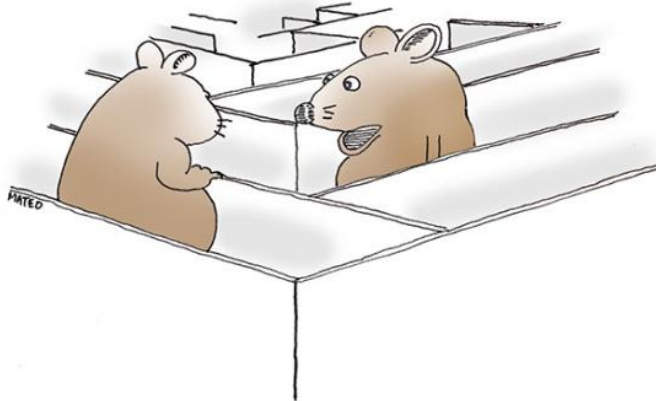
Relevance evaluation task

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

Crowdsourcing for Relevance Evaluation

DESIGN OF EXPERIMENTS

Crowdsourcing for Relevance Evaluation



"Whoever designed this place is insane."

Crowdsourcing for Relevance Evaluation

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



Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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.

Crowdsourcing for Relevance Evaluation

UX design - II

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

Crowdsourcing for Relevance Evaluation

Examples - 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 linked in 1 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 runs around to face the camera. Then she faces left." Follow the examples below:

Current Word Count: 0 Current Character Count: 0 / 2000

Criteria for **REJECTION**:

1. Entries with obvious and multiple spelling or grammatical errors will be rejected.
2. Entries with fewer than 100 words will be automatically rejected.
3. 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**.
4. Incomplete and blank answers will be rejected. Multiple blank answers will result in being **blocked**.
5. Tasks submitted without descriptions will be rejected.
6. Tasks submitted with inaccurate descriptions will be rejected as well.
7. Do **NOT** add any personal opinions. Entries with personal opinions or reviews will be automatically **REJECTED**.
8. 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.
9. Entries that transcribe the video will be **REJECTED**.

Crowdsourcing for Relevance Evaluation

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.ezclout.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 EZCLOUT's Search Page. Or go [here](#). You must use the Search page provided on EZCLOUT'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.

Crowdsourcing for Relevance Evaluation

Example - III

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

Report on Total Number of Results for Search

Step1: Go to <http://www.google.com> and Search for "Hotelscombined singapore"

Step2: 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 listing. 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.

Crowdsourcing for Relevance Evaluation

A better example

- All information is available
 - What to do
 - Search result
 - Question to answer

Task
Please evaluate the relevance of the following document for the query **milton keynes**.

Web Images Videos Shopping News Maps More MSN Hotmail Sign in United States Preferences
bing milton keynes

MILTON KEYNES ALL RESULTS 1-20 of 7,020,000 results advanced Sponsored sites
Milton Keynes Map **Milton Keynes** - Wikipedia, the free encyclopedia **Milton Keynes Hotels**
Milton Keynes Restaurants Milton Keynes, often abbreviated MK, is a large town in Buckinghamshire, in the south east of England, about 45 miles (72 km) north-west of London. It is also the capital of ... Save up to 50% on Hotels and Now Get Our Best Price Guarantee. www.expedia.com

Please rate the above document according to its relevance to **milton keynes** 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.

Crowdsourcing for Relevance Evaluation

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

Describe your HIT	
Title	Relevance evaluation for web pages Describe the task to workers. Be as specific as possible, e.g. "answer a survey about movies", i
Description	Please help us evaluate relevance for the following web page Give more detail about this task. This gives workers a bit more information before they decide
Keywords	relevance, search engines, web, information retrieval Provide keywords that will help workers search for your HITs.


Crowdsourcing for Relevance Evaluation

TREC assessment example

Describe your HIT	
Title	Relevance evaluation for news articles <small>Describe the task to workers. Be as specific as possible, e.g. "answer a survey about movies"</small>
Description	Please help us evaluate relevance for the following document. <small>Give more detail about this task. This gives workers a bit more information before they decide.</small>
Keywords	relevance, news articles, search, TREC, cosmic events, tropical storms, Schi <small>Provide keywords that will help workers search for your HITs.</small>

 **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 imeni M. V. Keldysh, Russian Academy of Sciences, UDC 629]

[Abstract] The probability of a collision between the Earth and an asteroid such as Amor, Apollo, or Aten 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: space 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/cm^3 . Estimates are made for asteroids with radii of 5 m, 50 m, and 500 m and masses of 1.57×10^6 tons and 1.57×10^9 tons. 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

Please rate the above document according to its relevance to **cosmic events** as follows. Note that the task is about how relevant to the topic the document is.

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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 $n+1$

Crowdsourcing for Relevance Evaluation

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.

Crowdsourcing for Relevance Evaluation

A qualification test

```
<QuestionForm xmlns="http://mechanicalturk.amazonaws.com/AWSMechanicalTurkDataSchemas/2005-10-01/QuestionForm.xsd">
  <Overview>
    <Title>Generic knowledge qualification test</Title>
  </Overview>
  <Question>
    <QuestionIdentifier>question1</QuestionIdentifier>
    <QuestionContent>
      <Text>Carbon monoxide poisoning is</Text>
    </QuestionContent>
    <AnswerSpecification>
      <SelectionAnswer>
        <StyleSuggestion>radiobutton</StyleSuggestion>
      <Selections>
        <Selection>
          <SelectionIdentifier>1</SelectionIdentifier>
          <Text>A chemical technique</Text>
        </Selection>
        <Selection>
          <SelectionIdentifier>2</SelectionIdentifier>
          <Text>A green energy treatment</Text>
        </Selection>
        <Selection>
          <SelectionIdentifier>3</SelectionIdentifier>
          <Text>A phenomena associated with sports</Text>
        </Selection>
        <Selection>
          <SelectionIdentifier>4</SelectionIdentifier>
          <Text>None of the above</Text>
        </Selection>
      </Selections>
    </SelectionAnswer>
  </AnswerSpecification>
</Question>
<Question>
  ...
</Question>
</QuestionForm>
```

Crowdsourcing for Relevance Evaluation

A qualification test - II

Answer

```
<?xml version="1.0" encoding="UTF-8"?>
<AnswerKey xmlns="http://mechanicalturk.amazonaws.com/AWSMechanicalTurkDataSchemas/2005-10-01/AnswerKey.xsd">
  <Question>
    <QuestionIdentifier>question4</QuestionIdentifier>
    <AnswerOption>
      <SelectionIdentifier>3</SelectionIdentifier>
      <AnswerScore>10</AnswerScore>
    </AnswerOption>
  </Question>
  <Question>
    ...
  </Question>
</AnswerKey>
```

Properties

```
#
# Basic qualification attributes
#
name= Generic knowledge quiz on topics
description=This qualification tests your general knowledge about a wide range of topics
keywords=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
testdurationinseconds=900
```

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation



"Plus, we double the accuracy at no extra cost by using our extensive pool of Siamese twins."

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Exp	Title	Average
412	airport security	1.4
439	inventions, scientific discoveries	1.6
438	tourism, increase	1.6
428	declining birth rates	1.6
424	suicides	1.6
442	heroic acts	1.8
436	railway accidents	1.8
433	Greek, philosophy, stoicism	3.6
405	cosmic events	3.6
401	foreign minorities, Germany	3.6

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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.

Crowdsourcing for Relevance Evaluation

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.

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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!

Crowdsourcing for Relevance Evaluation

More tips

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

Crowdsourcing for Relevance Evaluation



"All I know is that searching for my own name and then clicking on 'highly relevant' does wonders for my self-esteem."

Crowdsourcing for Relevance Evaluation

Platform alternatives

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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?
 - `crowdsource(task, 5)`

Crowdsourcing for Relevance Evaluation

Research questions

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

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

Crowdsourcing for Relevance Evaluation

Bibliography

- ❑ O. Alonso, D. Rose, and B. Stewart. "Crowdsourcing for relevance evaluation", SIGIR Forum, Vol. 42, No. 2 2008.
- ❑ O. Alonso and S. Mizzaro. "Can we get rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment". SIGIR Workshop on the Future of IR Evaluation, 2009.
- ❑ O. Alonso, R. Schenkel, and M. Theobald. Crowdsourcing Assessments for XML Ranked Retrieval, 32nd ECIR 2010.
- ❑ K. Berberich, S. Bedathur, O. Alonso, and G. Weikum A Language Modeling Approach for Temporal Information Needs, 32nd ECIR 2010
- ❑ N. Bradburn, S. Sudman, and B. Wansink. Asking Questions: The Definitive Guide to Questionnaire Design, Jossey-Bass, 2004.
- ❑ C. Callison-Burch. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk", EMNLP 2009.
- ❑ S. Hacker and L. von Ahn. "Matchin: Eliciting User Preferences with an Online Game", CHI 2009.
- ❑ M. Hearst. "Search User Interfaces", Cambridge University Press, 2009
- ❑ J. Heer and M. Bobstock. "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design", CHI 2010.
- ❑ J. Howe. "Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business". Crown Business, New York, 2008.
- ❑ P. Hsueh, P. Melville, V. Sindhwami. "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria". NAACL HLT Workshop on Active Learning and NLP, 2009.
- ❑ B. Huberman, D. Romero, and F. Wu. "Crowdsourcing, attention and productivity". Journal of Information Science, 2009.

Crowdsourcing for Relevance Evaluation

Bibliography - II

- ❑ M. Kaisser, M. Hearst, and L. Lowe. "Improving Search Results Quality by Customizing Summary Lengths", ACL/HLT, 2008.
- ❑ G. Kazai, N. Milic-Frayling, and J. Costello. "Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments", SIGIR 2009.
- ❑ G. Kazai and N. Milic-Frayling. "On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing". SIGIR Workshop on the Future of IR Evaluation, 2009.
- ❑ D. Kelly. "Methods for evaluating interactive information retrieval systems with users". Foundations and Trends in Information Retrieval, 3(1-2), 1-224, 2009.
- ❑ A. Kittur, E. Chi, and B. Suh. "Crowdsourcing user studies with Mechanical Turk", SIGCHI 2008.
- ❑ K. Krippendorff. "Content Analysis", Sage Publications, 2003
- ❑ G. Little, L. Chilton, M. Goldman, and R. Miller. "TurKit: Tools for Iterative Tasks on Mechanical Turk", KDD-HCOMP 2009.
- ❑ H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. "Improving Search Engines Using Human Computation Games", CIKM 2009.
- ❑ T. Malone, R. Laubacher, and C. Dellarocas. Harnessing Crowds: Mapping the Genome of Collective Intelligence. MIT Press, 2009.
- ❑ W. Mason and D. Watts. "Financial Incentives and the 'Performance of Crowds'", HCOMP Workshop at KDD 2009.
- ❑ S. Mizzaro. Measuring the agreement among relevance judges, MIRA 1999

Crowdsourcing for Relevance Evaluation

Bibliography - III

- ❑ J. Nielsen. "Usability Engineering", Morgan-Kaufman, 1994.
- ❑ J. Ross, L. Irani, M. Six Silberman, A. Zaldivar, and B. Tomlinson. "Who are the Crowdworkers? Shifting Demographics in Mechanical Turk". CHI 2010.
- ❑ F. Scheuren. "What is a Survey" (<http://www.whatisasurvey.info>) 2004.
- ❑ R. Snow, B. OConnor, D. Jurafsky, and A. Y. Ng. "Cheap and Fast But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". EMNLP-2008.
- ❑ V. Sheng, F. Provost, P. Ipeirotis. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers" KDD 2008.
- ❑ S. Weber. "The Success of Open Source", Harvard University Press, 2004.
- ❑ L. von Ahn. Games with a purpose. Computer, 39 (6), 92–94, 2006.
- ❑ L. von Ahn and L. Dabbish. "Designing Games with a purpose". CACM, Vol. 51, No. 8, 2008.

Crowdsourcing for Relevance Evaluation

Thank You!

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

Cartoons by Mateo Burtch (buta@mindspring.com)

Crowdsourcing for Relevance Evaluation