

Design of Experiments for Crowdsourcing Search Evaluation: challenges and opportunities

Omar Alonso

Microsoft

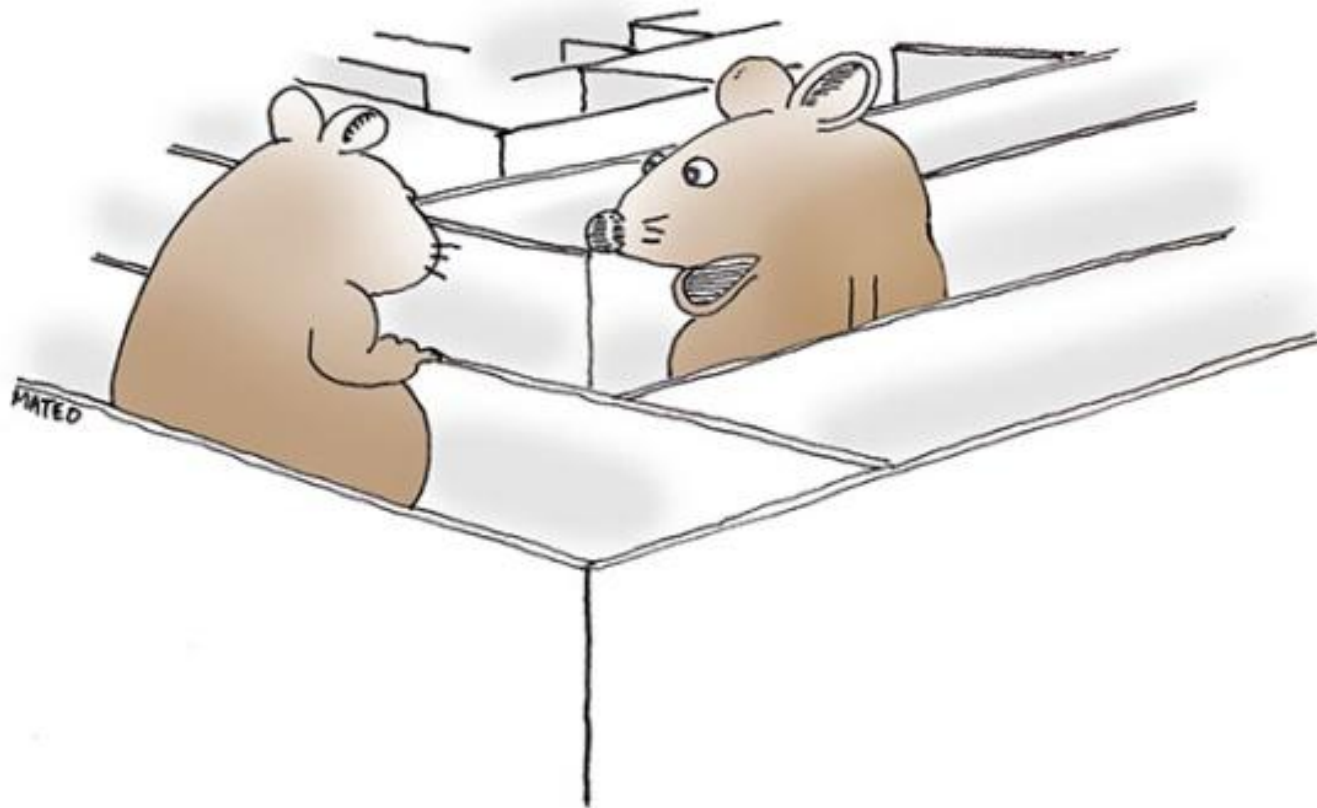
23 July 2010

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 **crowdsource** 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 linked 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.

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**.

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 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

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)

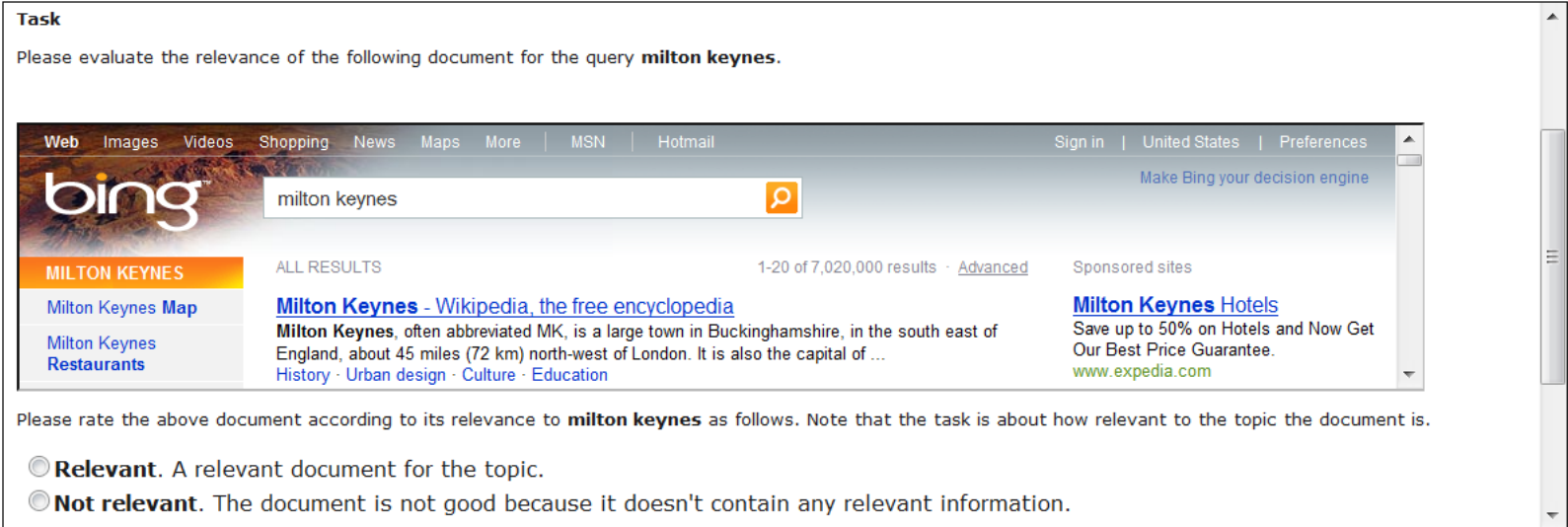
Step 3: Click on the result 'Singapore Hotels: Compare Cheap Singapore Accommodation Deals' at the top of the listing. This will appear below the yellow listing. DO NOT click on the sponsored listing in the yellow box.

Link is working

Please provide Feedback on this HIT... we'd love to know if the task was clear.

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](#)

History · Urban design · Culture · Education

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.

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

Description
Give more detail about this task. This gives workers a bit more information before they decide.

Keywords
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 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 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: 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

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 *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 ..

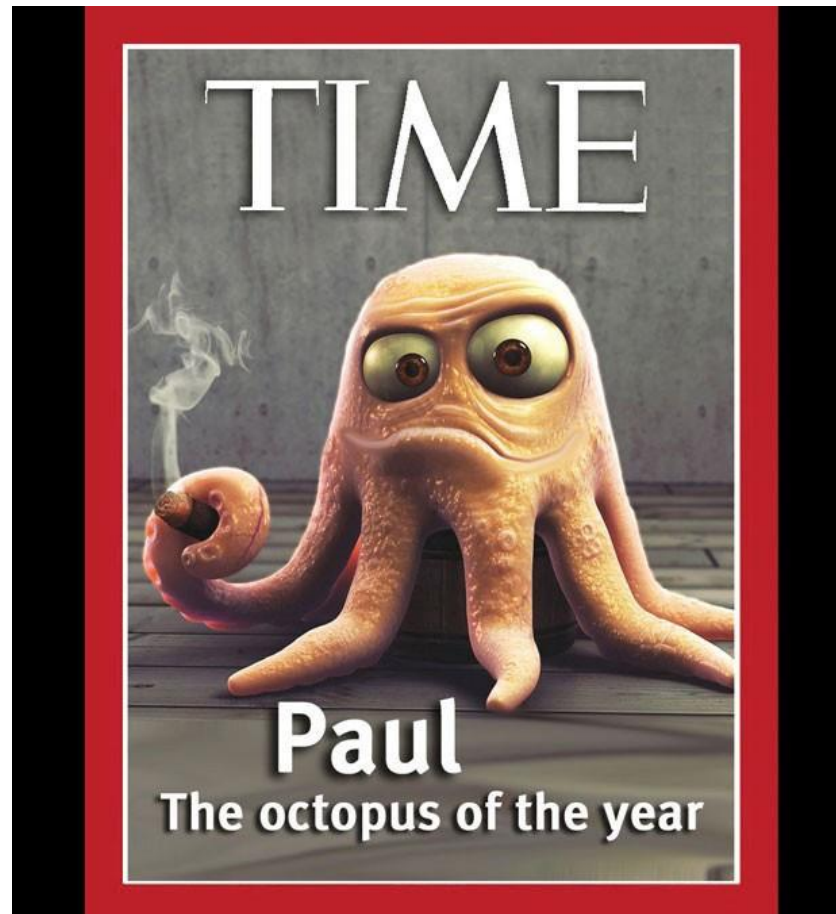


Imagen original del pulpo: <http://weilynn.deviantart.com> By Weilynn

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

PADDOCK, Times Staff Writer NEEDLES Businessman Ron Gaynor would like to dig a huge hole in the Mojave Desert where he can bury low-level radioactive waste. At the same time, biologist Alice Karl has long wanted to protect the desert tortoise from extinction. And so the two have formed an unlikely alliance designed to bring a low-level radioactive waste dump to an isolated spot in the desert that is home to dwindling numbers of the tortoise. Gaynor, senior vice president of US Ecology, a waste management firm, has selected a 100-acre site in the Ward Valley 22 miles west of Needles as the preferred location for California's first low-level radioactive waste facility. To make up for the loss of tortoise territory, the firm proposes to build a barrier two feet high across 7 1/2 miles of the valley. The unusual fence, designed to reclaim a portion of the species' original habitat, would keep the slow-footed reptiles from wandering onto a busy stretch of Interstate 40 that runs through the middle of the valley. Tortoises Come Out Ahead "If we can decrease the effect of the freeway, there is going to be a net benefit for the tortoises in Ward Valley," said Karl, who was recruited by US Ecology to study the animals. The question of where to place the federally required dump underscores a growing conflict between two environmental concerns: preserving the habitats of rare species and finding remote locations to safely store ever-increasing amounts of hazardous waste. Yet, surprisingly, selection of the Ward Valley site has generated few cries of outrage so far. Many residents of Needles have overcome their initial fear of low-level radioactive waste -- material contaminated by radiation, such as protective clothing, used medical supplies and power plant water filters. Needles Mayor Robert C. Prochaska said local people favor putting the dump near their town because of the two dozen permanent jobs it will offer. The Sierra Club, which was among a number of groups consulted by the state government and US Ecology, reluctantly agrees that Ward Valley is the best place to dump the waste. The state, which awarded US Ecology the contract to find a site and operate the dump, insisted that the firm consult local citizens and environmental organizations in an attempt to resolve their concerns, such as protecting the tortoise. As a result, some officials say the involvement of these groups early in the planning process could provide a model for siting other difficult projects in the future. "It's controversial, but it may be a controversy that was resolved because it was honestly approached," said state Health and Welfare Undersecretary Thomas E. Warriner, whose agency

Please rate

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 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.

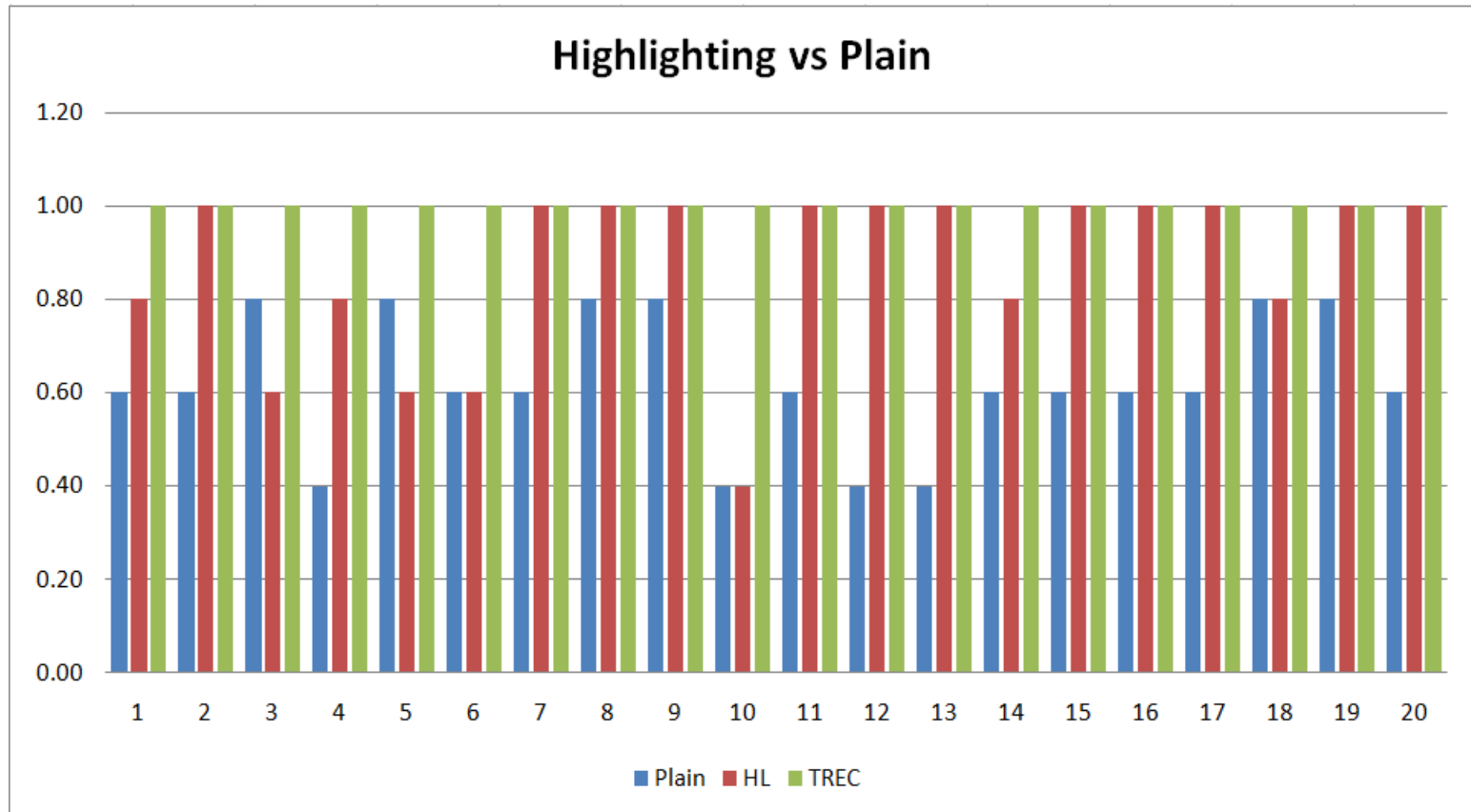
Commission President Jacques Delors said yesterday's initiative should be seen as "an expression of practical support for the **peace** process in Northern **Ireland**" in the wake of the Downing Street Declaration.

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 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.
- I don't know.** I don't know about this topic or the document is difficult to read.

Presentation - II



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

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

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.

Submit

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?
 - `crowdsource(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

Omar Alonso, Gabriella Kazai, and Stefano Mizzaro.

Crowdsourcing for Search Engine Evaluation: Why and How.

To be published by Springer, 2011.