Crowdsourcing for Relevance Evaluation

Daniel E. Rose
Director, Search Relevance
A9.com
What Is A9?

- Search technology subsidiary of Amazon.com
- Product search engine
  - Provides search on Amazon.com and other retail sites
  - Creator of “Search Inside the Book”
- Offers Clickriver advertising marketplace
  - Show targeted service ads in the Amazon network
What A9 Isn’t

• A destination web search service…
• … though we did some experiments in federated web search a while back
  – Developed OpenSearch standard, used today for Firefox search plugins
• … and some work with geographic location
  – Created BlockView, the first human’s-eye view map enhancement
Product Search

- Searching a corpus of retail products
- Data has *more structure* than web
- Search UIs generally support *richer interaction* (sorting, browsing, filtering…)
- Relevance different for every category

Sample queries:
- oakley sport sunglasses
- nintendo ds girls
- official euro 2008 soccer ball
- nylon repair tape
- house
- dell 5150
- ny giant fleece blanket and pillow
- nordic skis
- nine west shoes woman
- nikon 8x40 action binoculars
Evaluating Relevance

- Relevance is notoriously hard to evaluate
- Highly subjective
- Expensive to measure

Human Judgments, Standard Collection

• Predefine test corpora, test queries, and do a one-time relevance assessment.
• Make sets available for re-use in multiple experiments.
Variant 1: Exhaustive Editorial Review

- As found in Cranfield studies (1960s)
- Idea: For a given set of test queries, read every document in corpus and assess its relevance to each query.
- Assessors: Typically, students.
- Problem: Doesn’t scale when corpus is >5k docs
  - or when docs are longer than a paragraph
Variant 2: Pooling

• As found in TREC (1990s)
• Idea: For a given set of test queries, read the top 100 documents retrieved by any participating IR system.
• Assessors: Retired intelligence analysts.

Problems with Standard Collections

- Test sets get stale
- Existing sets may not be appropriate to your domain
- Batch methodology may not be appropriate for measuring your research (e.g. UI)
- Expensive to create new sets
- Breaks down for REALLY big corpora
Variant 3: New Experiment, New Test

- Usually give up on recall, measure DCG
- Either have small group of students, or group of editors
- Problem: Time-consuming and/or expensive

Alternative: Automated Metrics

- Rely on existing user behavior to assess performance
  - e.g. click position in interleaved results
- Problems:
  - Need lots of real users
  - Data not re-usable

Reconciling Two Approaches

Explicit Judgments
- Reusable
- Flexible

Automated Metrics
- Large scale
- Inexpensive

Technique for Evaluating Relevance by Crowdsourcing
Crowdsourcing

“Everyday people using their spare cycles to create content, solve problems, even do corporate R & D.”

Amazon Mechanical Turk (MTurk)

- framework for crowdsourcing
- on-demand workforce
- “artificial artificial intelligence”: get humans to do hard part
- named after faux automaton of 18th C. (really a human)
MTurk: How it works

• Requesters create “Human Intelligence Tasks” (HITs) via web services API.
• 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.
• Currently 21,527 HITs available
Some Sample HITs

- translating text to other languages
- labeling images (x cents each)
- finding “happy hour” times for bars in resort areas

<table>
<thead>
<tr>
<th>Comment and vote on an article, Easy?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester: Product Search</td>
</tr>
<tr>
<td>Time Allotted: 3 hours</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enter 12 pieces of Data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester: Benjamin</td>
</tr>
<tr>
<td>Time Allotted: 11 minutes 40 seconds</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label images of geological formation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester: ndabelle</td>
</tr>
<tr>
<td>Time Allotted: 2 hours</td>
</tr>
</tbody>
</table>

| FREE1 - Get Award or An Info - | HITNAME | |
|---------------------------------|---------|
| Requester: CBS | HIT Expiration Date: Oct 17, 2008 (1 week 4 days) | Reward: $0.01 |
| Time Allotted: 4 hours | HITs Available: 624 |
Hypothetical Relevance Evaluation Task: World Facts

- Want to see whether extracts of CIA World Factbook are relevant to certain locations
- Want test set of (say) 50 queries, and want to judge the first 50 results for each query
  - 2500 query-result pairs
Qualifying Workers

- Requester creates a qualification test
- Ours will be about geography

- Which has a major city named Cairo?
  - Brazil
  - Tunisia
  - Egypt
  - Turkey

- Which is closest to the population of India?
  - 250 million
  - 500 million
  - 750 million
  - 1 billion
Creating HITs

<Question>
    <QuestionIdentifier>question1</QuestionIdentifier>
    <DisplayName>Question 1:</DisplayName>
    <IsRequired>true</IsRequired>
    <QuestionContent>
        <FormattedContent><![CDATA[
            Is the following text relevant to Andorra? Tourism, the mainstay of Andorra's tiny, well-to-do economy, accounts for more than 80% of GDP. An estimated 11.6 million tourists visit annually, attracted by Andorra's duty-free status and by its summer and winter resorts.
        ]]></FormattedContent>
    </QuestionContent>
    <AnswerSpecification>
        <SelectionAnswer>
            <StyleSuggestion>radiobutton</StyleSuggestion>
            <Selections>
                <Selection>
                    <SelectionIdentifier>ir</SelectionIdentifier>
                    <Text>Irrelevant</Text>
                </Selection>
                ...
            </Selections>
        </SelectionAnswer>
    </AnswerSpecification>
</Question>
Creating HITs

[Image of a HITs website interface showing a test question and options for relevance evaluation.]
Cost

- We’ll pay 1 cent per HIT -- a relevance judgment on a query-result pair
- We’ll have 5 workers perform each HIT
- Total cost:
  50 queries x 50 results x 5 workers x $0.01
  = $125
Quality

• Lots of ways to control quality:
  – Better qualification test
  – More redundant judgments
  – Various methods to aggregate judgments
    • Voting
    • Consensus
    • Averaging
  – Methods to filter bad data
    • Look for patterns
    • Look for outliers
Assessing TERC

Advantages:
• Fast Turnaround
• Low Cost
• High Quality
• Flexibility

Limitations:
• Artificiality of Task
• Unknown Population
Conclusions

• TERC is a viable alternative to traditional relevance evaluation methods
• Amazon Mechanical Turk can provide the crowdsourcing framework for TERC.
Questions?

• For more details:

• To learn more about Amazon Mechanical Turk: www.mturk.com