Evaluation with Respect to Usefulness

Some perspectives from industry

Omar Alonso
Microsoft

Bressanone, Italy 4 - 8 February 2013

bing

Disclaimer

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

Outline

- Overview of relevance evaluation
- Crowdsourcing
- Social utility
- Databases & infrastructure

RELEVANCE EVALUATION

Relevance and IR

- What is relevance?
  - Multidimensional
  - Dynamic
  - Complex but systematic and measurable
- Frameworks
- Types
  - System or algorithmic
  - Topical
  - Pertinence
  - Situational
  - Motivational
- How to measure relevance?

IR evaluation

- Relevance is hard to evaluate
  - Highly subjective
  - Expensive to measure
- Two types of IR evaluation
  - Offline: ask users to explicitly evaluate a system
  - Online: see how users interact with a system

Traditional IR relevance evaluation

- The Cranfield tests
- TREC
- Retrieval tasks and metrics
- Approach
  - Design a new retrieval technique
  - Use a test collection
  - Run experiment
  - Collect data and analyze results

Ellen Voorhees and Donna Harman (Eds.) *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

Donna Harman. *Information Retrieval Evaluation.* Morgan & Claypool Publishers, 2011.

bing

Offline evaluation problems

- Expensive
- Slow
- Do users and judges agree on relevance?
- Maintenance of collections and assessments

Online evaluation

- Observable user behavior reflects relevance
- Real users
  - Have a goal
  - They work to satisfy an information need
- Measure performance on real users and queries
- Challenge: how do we know when users are satisfied?

Filip Radlinksi & Yisong Yue. “Practical Online Retrieval Evaluation” ACM SIGIR 2011 Tutorial.

bing

What is online data?

- Links, queries and clicks
- Mouse movement
- User behavior
  - Queries and results: timestamp, IP address
  - Click on results: what order, dwell time
  - Query reformulations

Experimentation

- New ranking mechanism
- Implement logging infrastructure
- Implement re-ranking infrastructure
- Recruit users
- Collect data and analyze results

Comparisons

- Similar to tasting experiment
  - Blind test: Coke or Pepsi?
- Document level
- Ranking level
- Interleaving

Why we need data?

- Relevance metrics
- Machine learning
  - Feature engineering
  - Training sets
- Data science

Paul Bennett, Misha Bilenko and Kevyn Collins-Thompson. “Machine Learning and IR: Recent Successes and New Opportunities”, ECIR 2010 Tutorial.

Other important stuff

- User studies
  - Given a task, observe users
  - Very expensive, lab setting
- Field studies
  - Observe users in-situ
- Eye tracking
  - Heat maps
  - Scanning patterns

Marti Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

Diane Kelly. "Methods for evaluating interactive information retrieval systems with users". *Foundations and Trends in Information Retrieval*, 3(1-2), 1-224. 2009.

bing

Part 1 - Conclusions

- Evaluation is a key part of production systems
- This is a continuous process
- In learning, a common approach is to use:
  - Explicit judgments as ground truth
  - Click data as features
- Everything counts
  - Online
  - Offline
  - User studies

You have a new idea

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

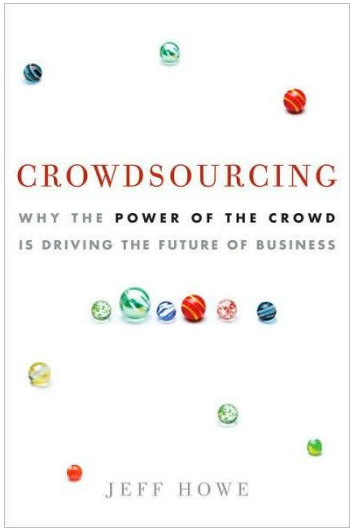
CROWDSOURCING

The rise of crowdsourcing in IR

- Crowdsourcing is hot
- Lots of interest in the research community
  - Articles showing good results
  - Journals special issues (IR, IEEE Internet Computing, etc.)
  - Workshops and tutorials (SIGIR'10, NACL'10, WSDM'11, WWW'11, SIGIR'11/12, VLDB'11, CHI, etc.)
  - HCOMP
  - CrowdConf 2011/2012
- Large companies leveraging crowdsourcing
- Big data
- Start-ups
- Venture capital investment

What is crowdsourcing?

- Take a job traditionally performed by a known agent (often an employee)
- Outsource it to an undefined, generally large group of people via an open call
- New application of principles from open source movement
- Example: Wikipedia



Jeff Howe. *Crowdsourcing. Why the Power of the Crowd is Driving the Future of Business*. Crown Business, 2009.

bing

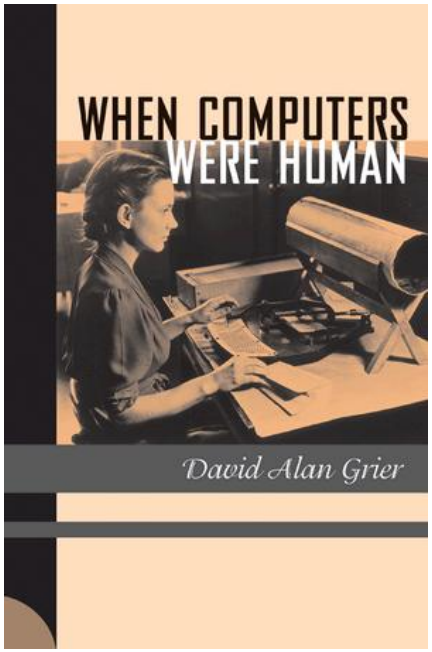
Human-based computation

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

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

Human computation

- Not a new idea
- Computers before computers
- You are a human computer

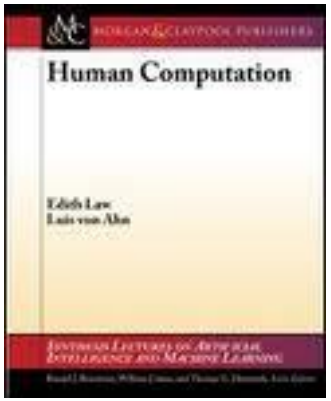


David Alan Grier. *When Computers were Human*. Princeton University Press, 2007.

bing

Some definitions

- Human computation is a computation that is performed by a human
- Human computation system is a system that organizes human efforts to carry out computation
- Crowdsourcing is a tool that a human computation system can use to distribute tasks.

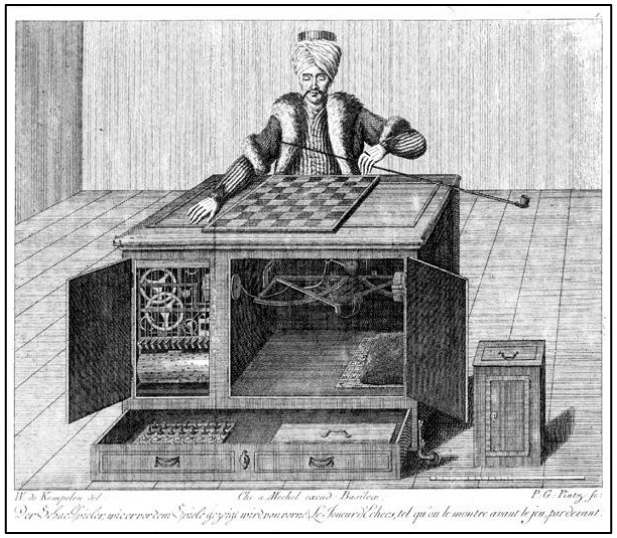


Edith Law and Luis von Ahn. *Human Computation.* Morgan & Claypool Publishers, 2011.

bing

Mechanical Turk

- Amazon Mechanical Turk (AMT, MTurk, www.mturk.com)
- Crowdsourcing platform
- On-demand workforce
- “Artificial artificial intelligence”: get humans to do hard part
- Named after faux automaton of 18th C.



J. Barr and L. Cabrera. “AI gets a Brain”, ACM Queue, May 2006.

bing

Why is this interesting?

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

Caveats and clarifications

- Trust and reliability
- Wisdom of the crowd re-visit
- Adjust expectations
- Crowdsourcing is another data point for your analysis
- Complementary to other experiments

Why now?

- The Web
- Use humans as processors in a distributed system
- Address problems that computers aren't good
- Scale
- Reach

Motivating example: relevance judging

- Relevance of search results is difficult to judge
  - Highly subjective
  - Expensive to measure
- Professional editors commonly used
- Potential benefits of crowdsourcing
  - Scalability (time and cost)
  - Diversity of judgments

Matt Lease and Omar Alonso. “Crowdsourcing for search evaluation and social-algorithmic search”, ACM SIGIR 2012 Tutorial.

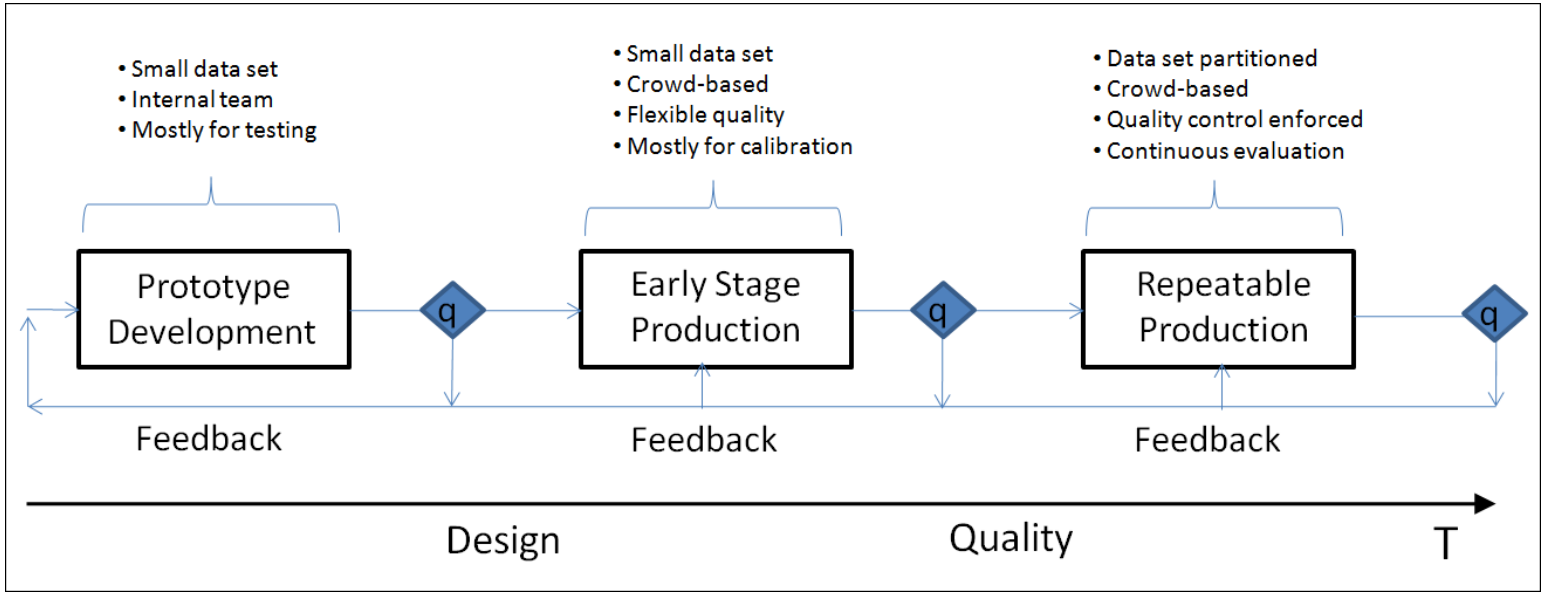
bing

Crowdsourcing and relevance evaluation

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

Development framework

- Incremental approach
- Measure, evaluate, and adjust as you go
- Suitable for repeatable tasks



bing

Asking questions

- Ask the right questions
- Part art, part science
- Instructions are key
- Workers may not be IR experts so don't assume the same understanding in terms of terminology
- Show examples
- Hire a technical writer
  - Engineer writes the specification
  - Writer communicates

N. Bradburn, S. Sudman, and B. Wansink. *Asking Questions: The Definitive Guide to Questionnaire Design*, Jossey-Bass, 2004.

UX design

- Time to apply all those usability concepts
- 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.
- Document presentation & design
- Need to grab attention
- Always ask for feedback (open-ended question) in an input box.
- Localization

Other design principles

- Text alignment
- Legibility
- Reading level: complexity of words and sentences
- Attractiveness (worker's attention & enjoyment)
- Multi-cultural / multi-lingual
- Who is the audience (e.g. target worker community)
  - Special needs communities (e.g. simple color blindness)
- Cognitive load: mental rigor needed to perform task
- Exposure effect

When to assess quality of work

- Beforehand (prior to main task activity)
  - How: “qualification tests” or similar mechanism
  - Purpose: screening, selection, recruiting, training
- During
  - How: assess labels as worker produces them
    - Like random checks on a manufacturing line
  - Purpose: calibrate, reward/penalize, weight
- After
  - How: compute accuracy metrics post-hoc
  - Purpose: filter, calibrate, weight, retain (HR)

How do we measure work quality?

- Compare worker's label vs.
  - Known (correct, trusted) label
  - Other workers' labels
  - Model predictions of workers and labels
- Verify worker's label
  - Yourself
  - Tiered approach (e.g. Find-Fix-Verify)

Comparing to known answers

- AKA: gold, honey pot, verifiable answer, trap
- Assumes you have known answers
- Cost vs. Benefit
  - Producing known answers (experts?)
  - % of work spent re-producing them
- Finer points
  - What if workers recognize the honey pots?

Comparing to other workers

- AKA: consensus, plurality, redundant labeling
- Well-known metrics for measuring agreement
- Cost vs. Benefit: % of work that is redundant
- Finer points
  - Is consensus "truth" or systematic bias of group?
  - What if no one really knows what they're doing?
    - Low-agreement across workers indicates problem is with the task (or a specific example), not the workers

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
  - Percentage agreement
  - Cohen's kappa (2 raters)
  - Fleiss' kappa (any number of raters)
  - Krippendorff's alpha
- With majority vote, what if 2 say relevant, 3 say not?
  - Use expert to break ties
  - Collect more judgments as needed to reduce uncertainty

k coefficient

- Different interpretations of k
- For practical purposes you need to be >= moderate
- Results may vary

k	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Detection theory

- Sensitivity measures
  - High sensitivity: good ability to discriminate
  - Low sensitivity: poor ability

Stimulus Class	"Yes"	"No"
S1	Hits	Misses
S2	False alarms	Correct rejections

Hit rate H = P("yes"|S2)
False alarm rate F = P("yes"|S1)

Part 2 - Conclusions

- Crowdsourcing 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
- Lots of opportunities to improve current platforms

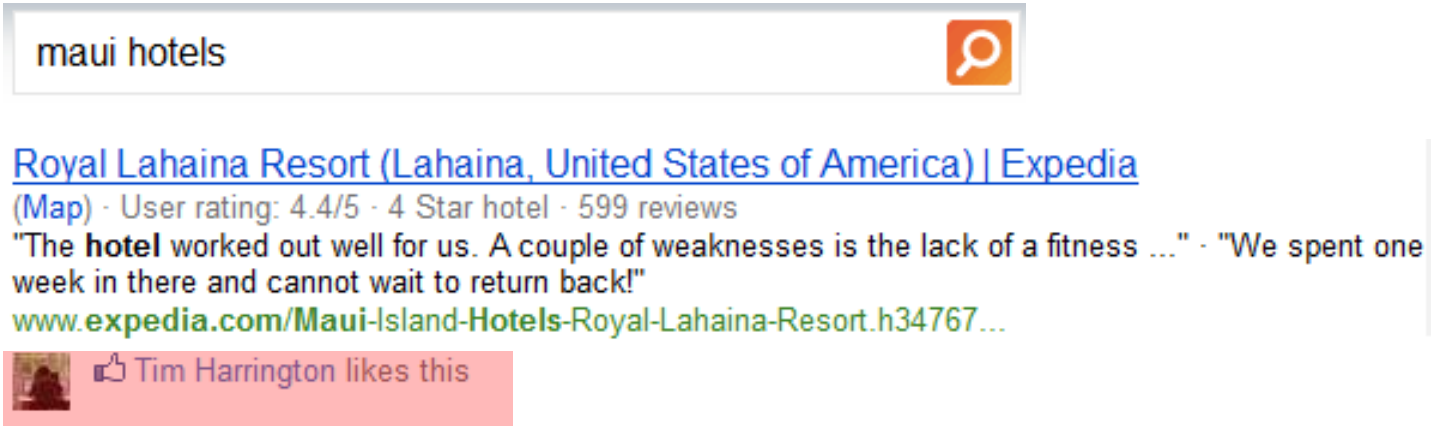
bing

You have a new idea - II

- New source
- How to study relevance in a new domain?

SOCIAL UTILITY

Social features and utility



- Are social features useful?
- Can we measure the utility?

Patrick Pantel, Michael Gamon, Omar Alonso, Kevin Haas. "Social Annotations: Utility and Prediction Modeling". ACM SIGIR 2012, pp. 285-294.

bing

Social features and utility - II

- Social annotations are part of search engines (Bing, Google)
- Benefits
  - Discovery of socially vetted recommendations
  - Personalized search results
  - Connecting to the lives of their friends
  - Result diversity
  - Emotionally connecting with an otherwise static and impersonal search engine
- Problem
  - Very little understanding whether these social features are useful or not
- Some questions
  - Are such endorsements from dearest friends more relevant to the user than from acquaintances or coworkers?
  - Are expert opinions or those from friends who live in the vicinity of the restaurant more valuable?
  - Do annotations on irrelevant results amplify their negative perception?

bing

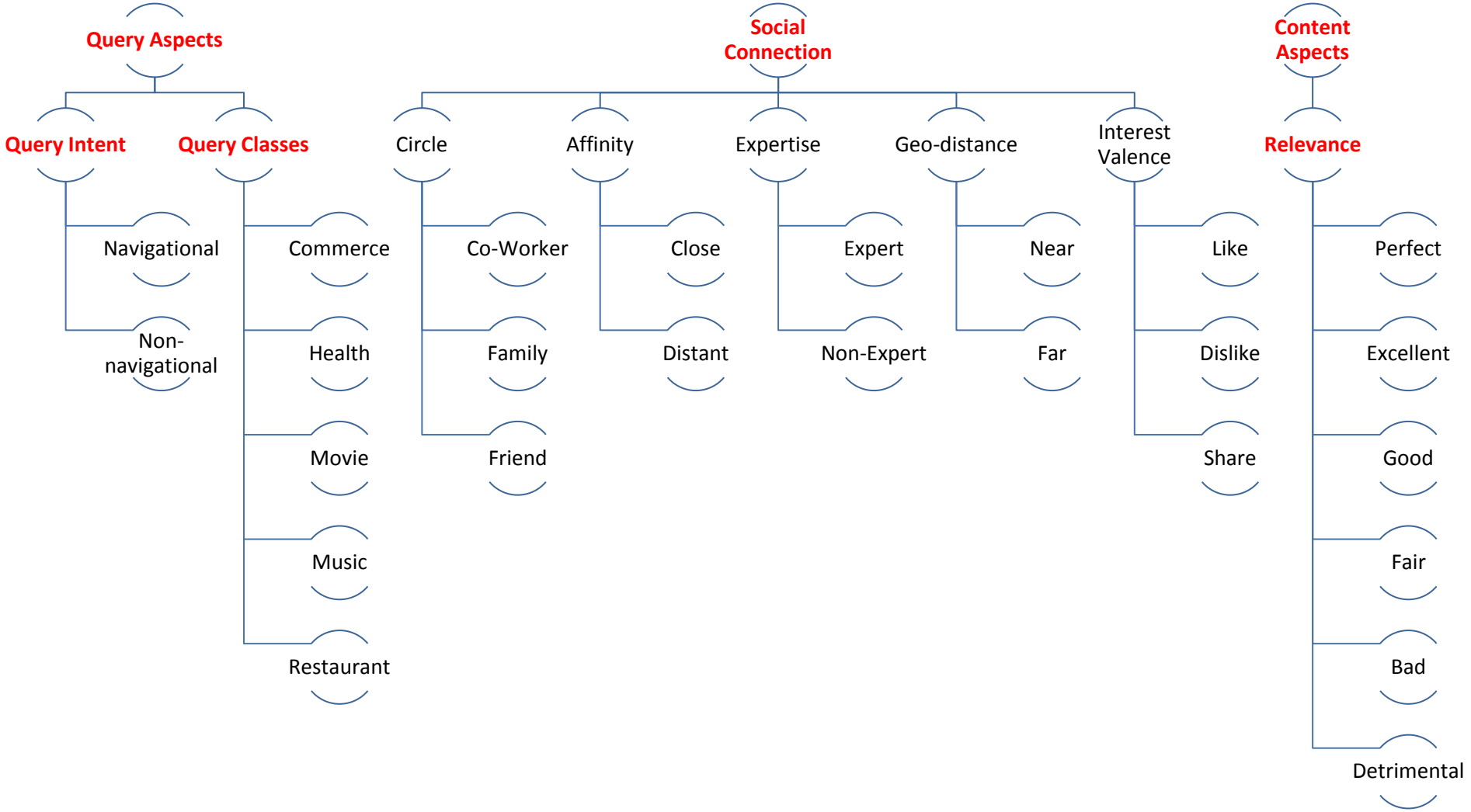
Social features and utility - III

- Social relevance aspects
  - When does a social annotation become relevant?
  - Taxonomy
  - User study
- Can we predict relevance of a social annotation?
  - Feature design and modeling

Aspects of social relevance

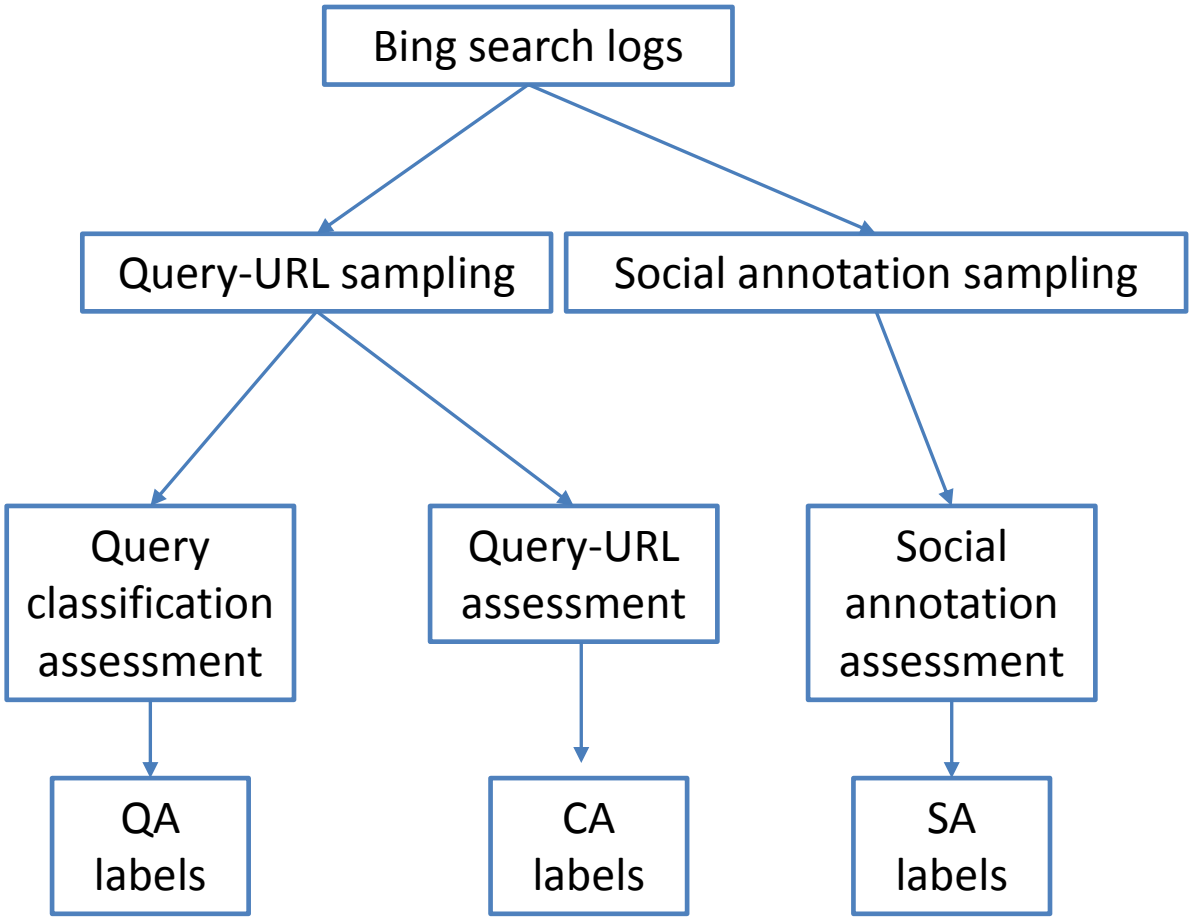
- Social annotation as a tuple *t = {q, u, c, v}*
  - q = query, u = content, c = social network connection, v = interest in the content
  - Interest (like, share, dislike)
  - `t={maui hotels, Expedia hotel page, Tim, like}`
- Taxonomy of Social Relevance aspects
  - Cues that influence the perceived utility of social annotations
  - Query Aspects (`QA`)
    - Query Intent
    - Query Class
  - Social Connection Aspects (`SA`)
    - Circle
    - Affinity
    - Expertise
    - Geographical Distance
    - Interest Valence
  - Content Aspects (`CA`)
    - Graded relevance

Taxonomy of social relevance



Social relevance

- User study
- Approach
  - Query-URL sampling
  - Social annotation sampling
  - Annotation task
    - Scenario
    - Task definition
    - Guidelines
    - Agreement
- 3 weeks spanning 3 months
- 2388 (HEAD) & 1375 (TAIL) queries



bing

Generating social annotations

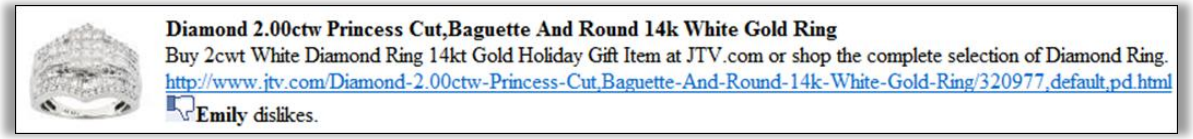
Bing search logs

Original annotation = <Thumbnail + Title + Snippet + URL>

Scenario template

Generator

Simulated social network



bing

What is the value of this social annotation?

- Simulated social network construction
- Scenario template generation
- Disagreements resolution
- Why simulating a network and potential issues

Task

You query Bing for **muniets to midnight** and one of the results is illustrated below. **Jill**, someone in your social network, has liked, disliked or shared this result. Recall that Jill is a **CloseFriend**. Also, assume that **Jill** is a **Expert** who **Dislike** the web page. Local information about **Jill**: **NA**

What is the added value of this social annotation?

Linkin Park - "Leave Out All The Rest
The official music video for "Leave Out All The Rest" from the album Minutes **To Midnight**. Directed by: Joe Hahn.
http://www.youtube.com/watch?v=LBTXNPZPfbE
Jill Dislike.

Please answer the following question:

How relevant is the annotation to the web results?
- **There is significant added value**. The annotation is substantially relevant, useful, or of interest to you.
- **There is some added value**. The annotation is somewhat relevant, useful or of interest to you.
- **No added value**. The annotation is not relevant, useful or of interest to you.
- **Don't know**. I don't have enough information to assess this annotation (please add a comment in the box below).
- **Non English/Service error**. Can't judge because content is non-English or there is a service error (e.g., 404 message, image didn't load, etc.) (please add a comment in the box below).

bing

Social utility

- R(T) - utility of social annotations
- Average utility of each tuple
- Two variants of R(T)
  - Graded relevance utility
  - Binary utility

$$R(T) = \frac{\sum_{t \in T} \frac{\sum_{j \in J(t)} \omega(t,j)}{|J(t)|}}{|T|}$$

$\mathbf{R(T)}$	ω function definition		
$\mathbf{R_{Rel}(T)}$	ω_{Rel}	$= \begin{cases} \texttt{sig-util}: & 1 \\ \texttt{some-util}: & 0.5 \\ \texttt{no-util}: & 0 \end{cases}$	
$\mathbf{R_{Prec}(T)}$	ω_{Prec}	$= \begin{cases} \texttt{sig-util}: & 1 \\ \texttt{some-util}: & 1 \\ \texttt{no-util}: & 0 \end{cases}$	

Social utility results

- No significant difference for HEAD/TAIL queries
- Overall, social annotation is useful
- Social aspects (`SA`) have significant differentiating influence
- The social affinity (`SA-AFF`) shows the most influence followed by expertise (`SA-EXP`) and connection circle (`SA-CIR`),
  - colleagues and friends have equal utility but family members have much higher expected utility
- For interest valence (`SA-INT`)
  - knowing that a connection has liked (`lik`) a link shows more utility than average
  - a share (`shr`) shows a negative utility influence
  - disliking a link (`dis`) has little effect on utility

Social utility results - II

- What is the value of a set of aspects if we know another set of aspects?
  - {`family`} (circle) | {`health`} (query class)
  - Expected gain of knowing `family` given that we know `health`
- Aspect Interplay:
  - `SA-affinity` is more important than `SA-circle`
  - `QA-CLS=movie` knowing `SA-circle=family` -> increases but if `SA-circle=work colleague` -> decreases
  - If affinity is close there is no value in knowing that the circle is a friend
  - If affinity is distant there is value in knowing that circle is work colleague
  - Expertise mostly required for health queries
  - `QA-CLS=health` knowing `SA-expertise=expert`

Feedback analysis

- Inspection of a random sample of comments
- `no-util`: expertise, affinity, interest valence
  - "Because he is a distant friend, neutral and non-expert, his opinion is not going to be useful to me"
- `some-util`: circle, dislike
  - "Chris' dislike might get me to click another link, even though it's what I'm looking for, it could be a bad quality link"
- `sig-util`: expertise, affinity
  - "She is only a colleague, but she is an expert on the result and her opinion matters to me because she knows what she is talking about"

Predicting social relevance

- Can we predict automatically whether a social annotation adds utility to a search result?
- Learning
  - Offline features from the user study
  - Online features available in the search engine

Feature engineering and modeling

- Offline (16)
  - Query class aspects, circle, affinity, expertise, geo-distance, interest-valence, CA
- Online features (150+)
  - Query class (e.g., is-commerce, is-health, etc.)
  - Session metrics (e.g., session duration, page view count, etc.)
  - User metrics (e.g., click count, page view count, etc.)
  - Query metrics (e.g., dwell time, time to 1st click, etc.)
  - Result metrics (e.g. abandonment)
- Model
  - Predict utility using a classification model
  - We use MS internal implementation of MART
  - Model parameters details in paper
- Experimental setup

Prediction is possible

Online features

HEAD

precision +10% (20% recall)

TAIL

precision +15% (25% recall)

Offline features

HEAD

precision +25% (50% recall)

precision +13% (88% recall)

TAIL

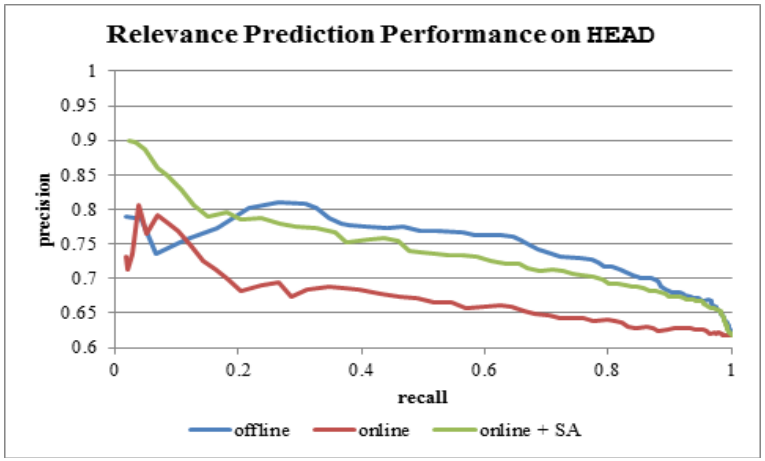
precision +29% (50% recall)

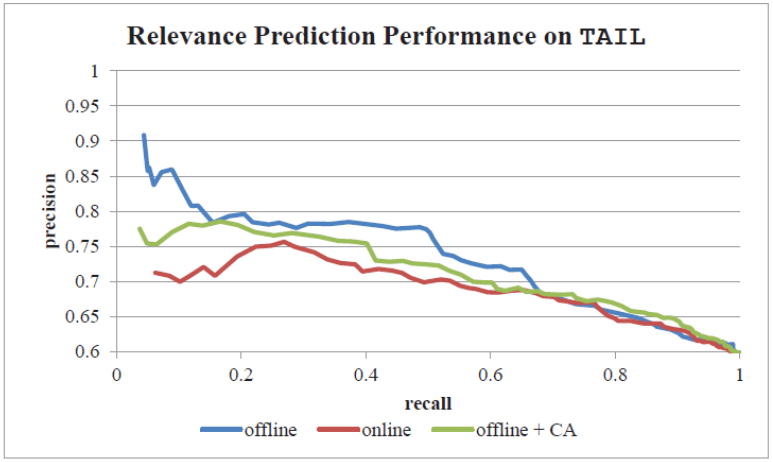
precision +7% (87% recall)

bing

Prediction experiments results

- Analyzed the logs in MART
- For HEAD
  - Social aspects rank higher
  - Circle, affinity and expertise are the most important features
- For TAIL
  - Content aspects rank higher
- Online features are predictive but not as predictive as offline features
- We can increase the performance of online features by adding SA and CA





Part 3 - Conclusions

- Social features are rapidly evolving in search
- Multiple aspects interact to determine relevance of a social annotation on the SERP
- Utility of social features was not well understood
- This work sheds light on this utility for social annotations
  - Aspect taxonomy
  - Utility of each aspect
  - The interplay between aspects
- Social annotation relevance can be predicted to some extent

DATABASES & INFRASTRUCTURE

bing

Datasets

- Millions of queries per day
- Ability to sample and quickly experiment is key
- Data analysis on queries and labels
- Automatic reporting
- Continuous evaluation

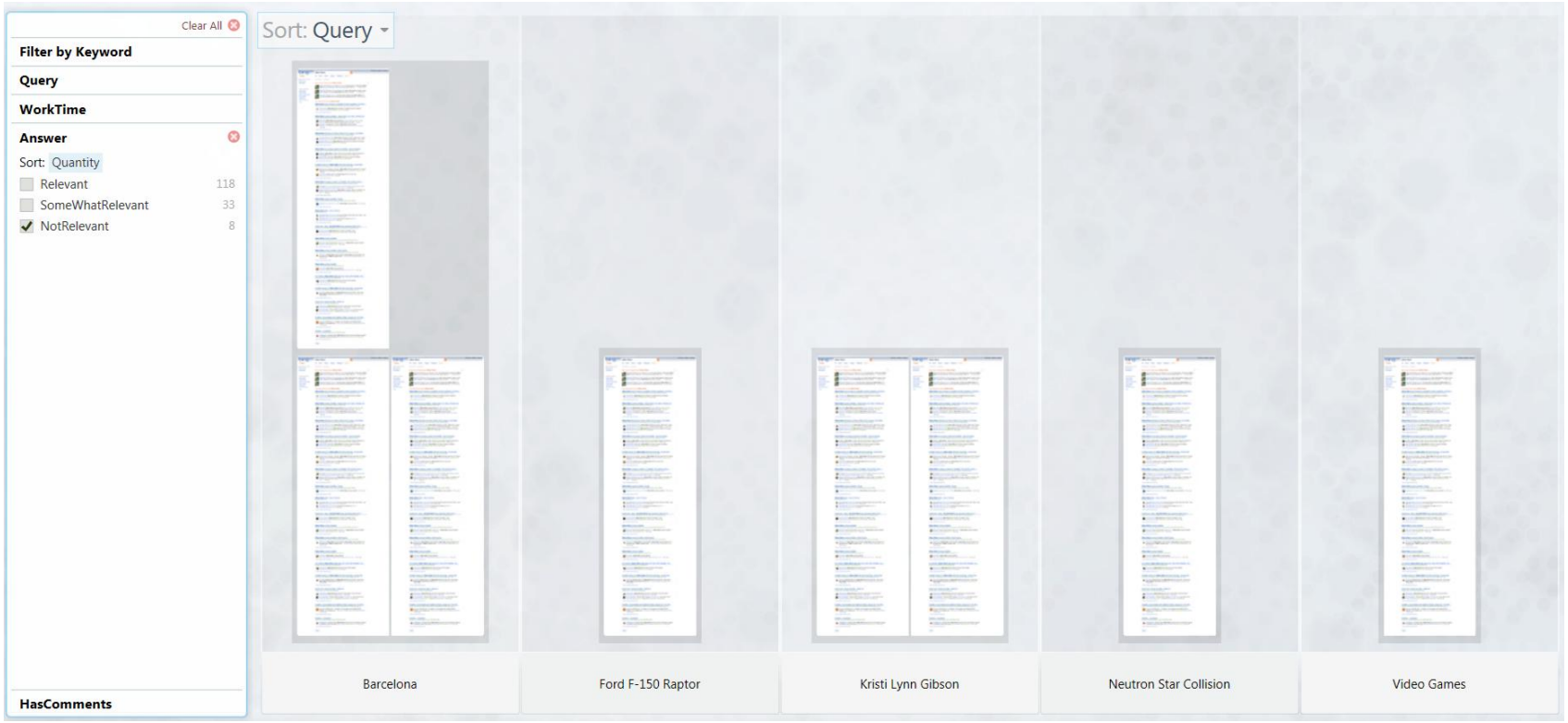
Human computation pipelines

- Or when to mix machines & people
- Lots of experiments needs human labels
- The human in the loop pattern
- Classifiers

Facets for exploratory data analysis

- Production systems contain lots of parameters
- Query sets contain lots of attributes
- Use facets to explore results data sets

PivotViewer



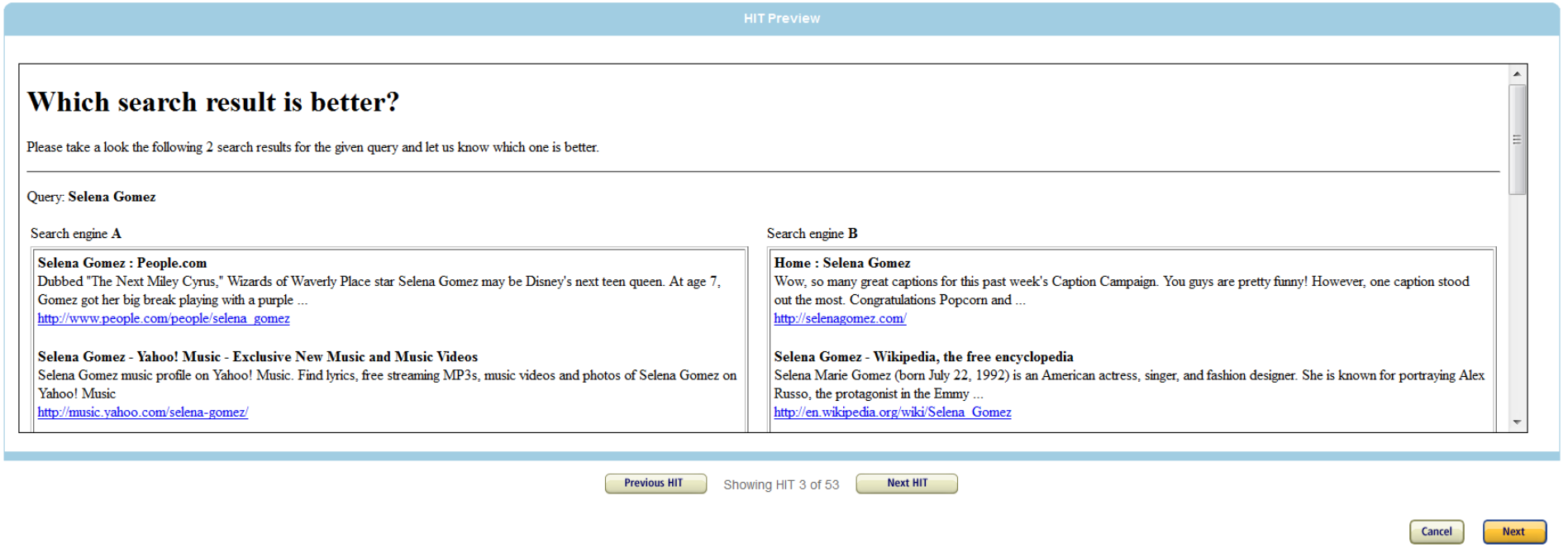
bing

Demo

- Get a query sample
- Using the BingAPI, extract the top 10
  - Plain UX (no brand)
- Split the SERP into 2 lists: top5, top6-10
- Run A-B comparison task
  - Show top5 vs top6-10
  - Which one is better? A, B or the same?
  - Crowdsourcing, 3 workers, a few honey pots
- Facets
  - Query, rank, query length, type, entity, etc.
- Mechanical Turk meets PivotViewer

bing

The task



bing

Inter-rater agreement

- R packages psy and irr
- Example

```
> demo <-
  read.delim(file="C:/Omar/research/Promise-
  2013/test_k.txt", head=TRUE, sep="\t")
> kappam.fleiss(demo)
 Fleiss' Kappa for m Raters

 Subjects = 53
   Raters = 3
    Kappa = 0.508
```

Tutorial summary

- The importance to understand relevance
- Offline and online evaluation
- Crowdsourcing as a cheap mechanism to gather labels
- Relevance criteria
- User studies and taxonomies
- Exploratory data analysis

Additional references

- C. L. Barry and L. Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. Information Processing & Management, 34 (2-3):219–236, May 1998
- P. Borlund. The concept of relevance in IR. Journal of the American Society for Information Science and Technology, 54(10):913–925, May 2003
- S. Mizzaro. Relevance: The whole history. Journal of the American Society for Information Science, 48(9):810–832, 1997.
- T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6):321–343, 1975.
- T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. J. Am. Soc. Inf. Sci., 58(13):1915–1933, 2007.
- T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. J. Am. Soc. Inf. Sci., 58(13):2126–2144, 2007.
- O. Alonso and S. Mizzaro. "Using Crowdsourcing for TREC relevance assessment", Information Processing & Management, 2012.
- M. Bernstein. et al. Soylent: A Word Processor with a Crowd Inside. UIST 2010.
- C. Callison-Burch. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk", EMNLP 2009.
- A. Kittur, E. Chi, and B. Suh. "Crowdsourcing user studies with Mechanical Turk", SIGCHI 2008.
- R. Snow et al. "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 sing Multiple, Noisy Labelers" KDD 2008.

bing