Report on the SIGIR 2009 Workshop on
The Future of IR Evaluation

Shlomo Geva[1] Jaap Kamps[1] Carol Peters[2]
Tetsuya Sakai[3] Andrew Trotman[1] Ellen Voorhees[4,5]

[1] INitiative for the Evaluation of XML Retrieval (INEX)
[2] Cross-Language Evaluation Forum (CLEF)
[3] NII Test Collection for IR Systems (NTCIR)
[4] Text REtrieval Conference (TREC)
[5] Text Analysis Conference (TAC)

Held in Boston, July 23, 2009

Motivation: Is it Time for a Change?

- Evaluation is at the core of information retrieval: virtually all progress owes directly or indirectly to test collections built within the so-called Cranfield paradigm.
- However, in recent years, IR researchers are routinely pursuing tasks outside the traditional paradigm, by taking a broader view on tasks, users, and context.
- There is a fast moving evolution in content from traditional static text to diverse forms of dynamic, collaborative, and multilingual information sources.
- Also industry is embracing "operational" evaluation based on the analysis of endless streams of queries and clicks.

Outline of Workshop and Presentation

- Focus: *The Future of IR Evaluation*
  - ⋆ Jointly organized by the evaluation fora:
    CLEF, INEX, NTCIR, TAC, TREC
- First part:
  - ⋆ Four keynotes to set the stages and frame the problem
  - ⋆ Twenty contributions: Boasters and posters
- Second part (it is a *work*$_{shop}$!):
  - ⋆ Breakout group on 4 themes
  - ⋆ Report out and discussion with a panel

Workshop Setup

- The basic set-up of the workshop was simple. We bring together
  - ⋆ i) those with novel evaluation needs
  - ⋆ and ii) to senior IR evaluation experts
- and develop concrete ideas for IR evaluation in the coming years
- Desired outcomes
  - ⋆ insight into how to make IR evaluation more "realistic,"
  - ⋆ concrete ideas for a retrieval track or task that would not have happened otherwise

Toward More Realistic IR Evaluation

- The questions we expected to address could be succinctly summarized as to make IR evaluation more "realistic."
- There is however no consensus on what then "real" IR is:
  - ⋆ System: from *ranking component* to . . . ?
  - ⋆ Scale: from *megabytes/terabytes* to . . . ?
  - ⋆ Tasks: from *library search/document triage*, to . . . ?
  - ⋆ Results: from *documents* to . . . ?
  - ⋆ Genre: from *English news* to . . . ?
  - ⋆ Users: from *abstracted users* to . . . ?
  - ⋆ Information needs: from *crisp fact finding* to . . . ?
  - ⋆ Usefulness: from *topically relevant* to . . . ?
  - ⋆ Judgments: from *explicit judgments* to . . . ?
  - ⋆ Interactive: from *one-step batch processing* to . . . ?
  - ⋆ Adaptive: from *one-size-fits-all* to . . . ?
  - ⋆ And many, many more...

Part 1: Keynotes

- In the morning we have invited keynotes of senior IR researchers that set the stage, or discuss particular challenges (and propose solutions).
  - ⋆ Stephen Robertson
  - ⋆ Sue Dumais
  - ⋆ Chris Buckley
  - ⋆ Georges Dupret
- I'll try to convey their main points

Richer theories, richer experiments

Stephen Robertson
Microsoft Research Cambridge
and City University
ser@microsoft.com

A caricature

On the one hand we have the Cranfield / TREC tradition of experimental evaluation in IR

- a powerful paradigm for laboratory experimentation, but of limited scope

On the other hand, we have observational studies with real users

- realistic but of limited scale

[please do not take this dichotomy too literally!]

Experiment in IR

The Cranfield method was initially only about "which system is best"

system in this case meaning complete package

- language
- indexing rules and methods
- actual indexing
- searching rules and methods
- actual searching

... etc.

It was not seen as being about theories or models...

Theory and experiment in IR

'Theories and models in IR' (J Doc, 1977):

Cranfield has given us an *experimental* view of what we are trying to do

- that is, something measurable

We are now developing models which address this issue directly

- this measurement is an explicit component of the models

We have pursued this course ever since...

Hypothesis testing

Focus of *all* these models is predicting relevance

(or at least what the model takes to be the basis for relevance)

– with a view to good IR effectiveness

No other hypotheses/predictions sought

... nor other tests made

This is a very limited view of the roles of theory and experiment

Theories and models

So…

- We are all interested in improving our understanding
  - … of both mechanisms and users
- One way to better understanding is better models
- The purpose of models is to make predictions
- But what do we want to predict?
  - useful applications / to inform us about the model

Predictions in IR

1. What predictions would be *useful*?

relevance, yes, of course...

... but also other things

- redundancy/novelty/diversity
- optimal thresholds
- satisfaction
  - ... and other kinds of quality judgement
- clicks
- search termination
- query modification
  - ... and other aspects of user behaviour
- satisfactory termination
- abandonment/unsatisfactory termination
  - ... and other combinations

Predictions in IR

2. What predictions would *inform us about models*?

more difficult: depends on the models

- many models insufficiently ambitious

in general, observables/testables

- calibrated probabilities of relevance
- hard queries
- clicks, termination
- patterns of click behaviour
- query modification

Richer models, richer experiments

Why develop richer models?

- because we want richer understanding of the phenomena
- as well as other useful predictions

Why design richer experiments?

- because we want to believe in our models
- and to enrich them further

A rich theory should have something to say *both* to lab experiments in the Cranfield/TREC tradition, *and* to observational studies

Evaluating IR *In Situ*

Susan Dumais

Microsoft Research

Evaluating Search Systems

- Traditional test collections
  - Fix: Docs, Queries, RelJ (Q-Doc), Metrics
  - Goal: Compare systems, w/ respect to metric
  - NOTE: Search engines do this, but not just this …
- What's missing?
  - Metrics: User model (pr@k, nncg), average performance, all queries equal
  - Queries: Types of queries, history of queries (session and longer)
  - Docs: The "set" of documents – duplicates, site collapsing, diversity, etc.
  - Selection: Nature and dynamics of queries, documents, users
  - Users: Individual differences (location, personalization including re-finding), iteration and interaction
  - Presentation: Snippets, speed, features (spelling correction, query suggestion), the whole page

Kinds of User Data

- User Studies
  - Lab setting, controlled tasks, detailed instrumentation (incl. gaze, video), nuanced interpretation of behavior
- User Panels
  - In-the-wild, user-tasks, reasonable instrumentation, can probe for more detail
- Log Analysis and Experimentation (in the large)
  - In-the-wild, user-tasks, no explicit feedback but lots of implicit indicators
  - The what vs. the why
- Others: field studies, surveys, focus groups, etc.

Sharable Resources?

- User studies / Panel studies
  - Data collection infrastructure and instruments
  - Perhaps data
- Log analysis – Queries, URLs
  - Understanding how user interact with existing systems
    - What they are doing; Where they are failing; etc.
  - Implications for
    - Retrieval models
    - Lexical resources
    - Interactive systems
  - Lemur Query Log Toolbar – developing a community resource !

Sharable Resources?

- Operational systems as an experimental platform
  - Can generate logs, but more importantly …
  - Can also conduct controlled experiments *in situ*
    - A/B testing -- Data vs. the "hippo" [Kohavi, CIKM 2009]
    - Interleave results from different methods [Radlinski & Joachims, AAAI 2006]
  - Can we build a "Living Laboratory"?
    - Web search
      - Search APIs , but ranking experiments somewhat limited
      - UX perhaps more natural
    - Search for other interesting sources
      - Wikipedia, Twitter, Scholarly publications, …
  - Replicability in the face of changing content, users, queries

Closing Thoughts

- Information retrieval systems are developed to help people satisfy their information needs
- Success depends critically on
  - Content and ranking
  - User interface and interaction
- Test collections and data are critical resources
  - Today's TREC-style collections are limited with respect to user activities
  - Can we develop shared user resources to address this?
    - Infrastructure and instruments for capturing user activity
    - Shared toolbars and corresponding user interaction data
    - "Living laboratory" in which to conduct user studies at scale

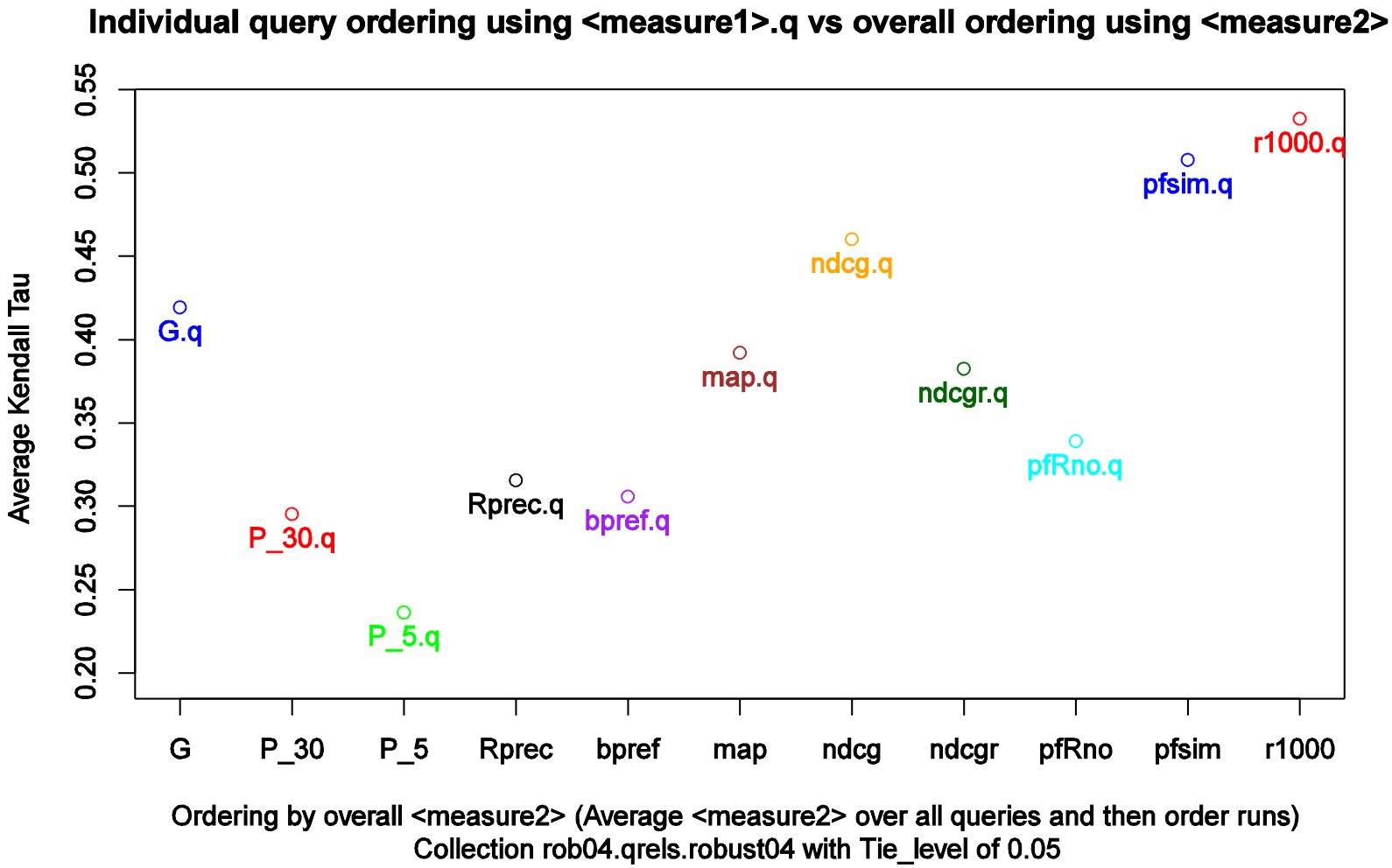
Towards Good Evaluation of Individual Topics

Chris Buckley – Sabir Research

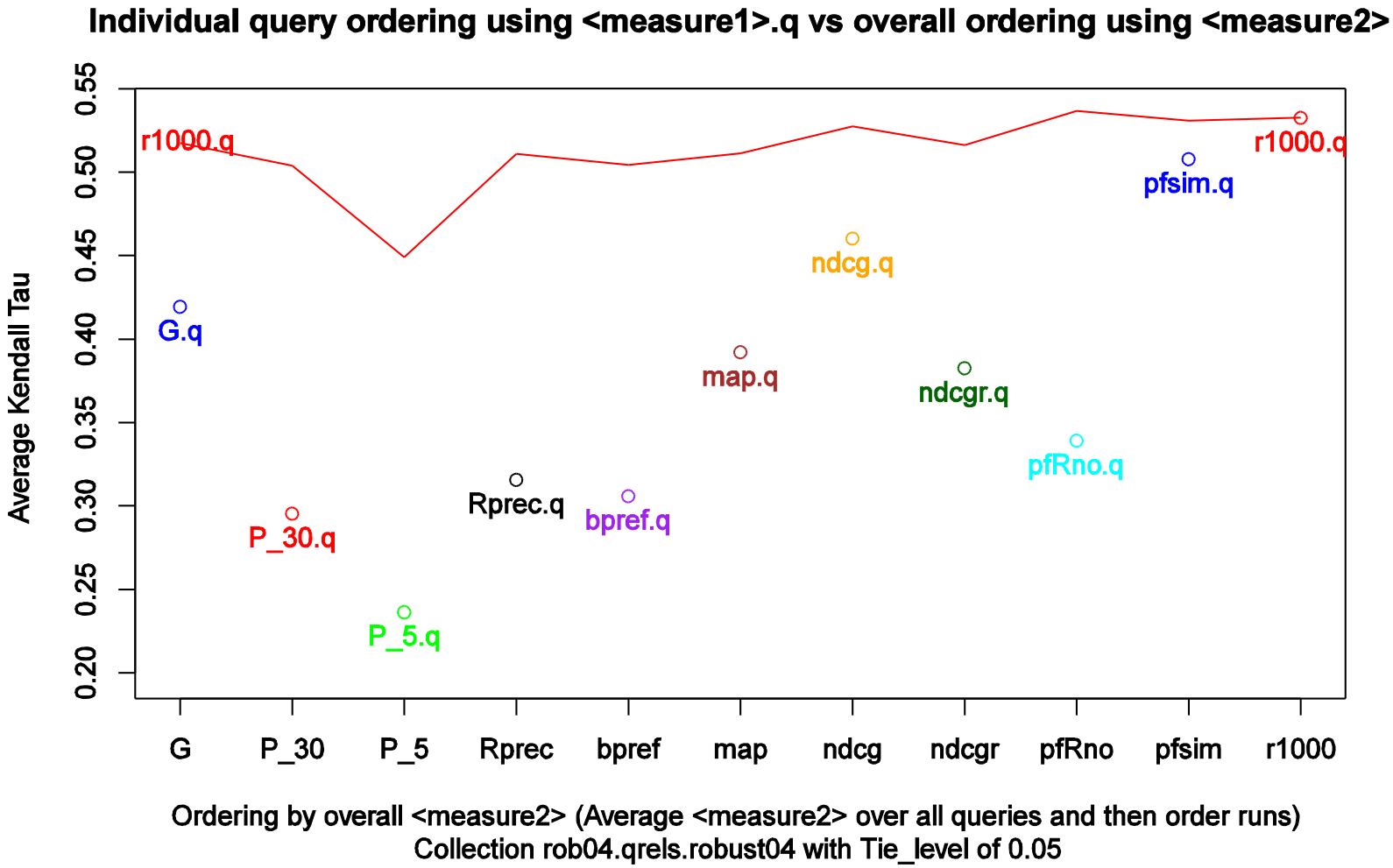
Current Individual Topic Measure Values

- How good are they?
  - Compare ranking of systems on individual topics with the overall ranking of systems. (Kendall Tau)
- Look at what makes a measure better on individual topics
- Initial plots are the Robust04 Track
  - 249 topics
  - All runs are automatic
  - Large number relevance judgments, “Complete”

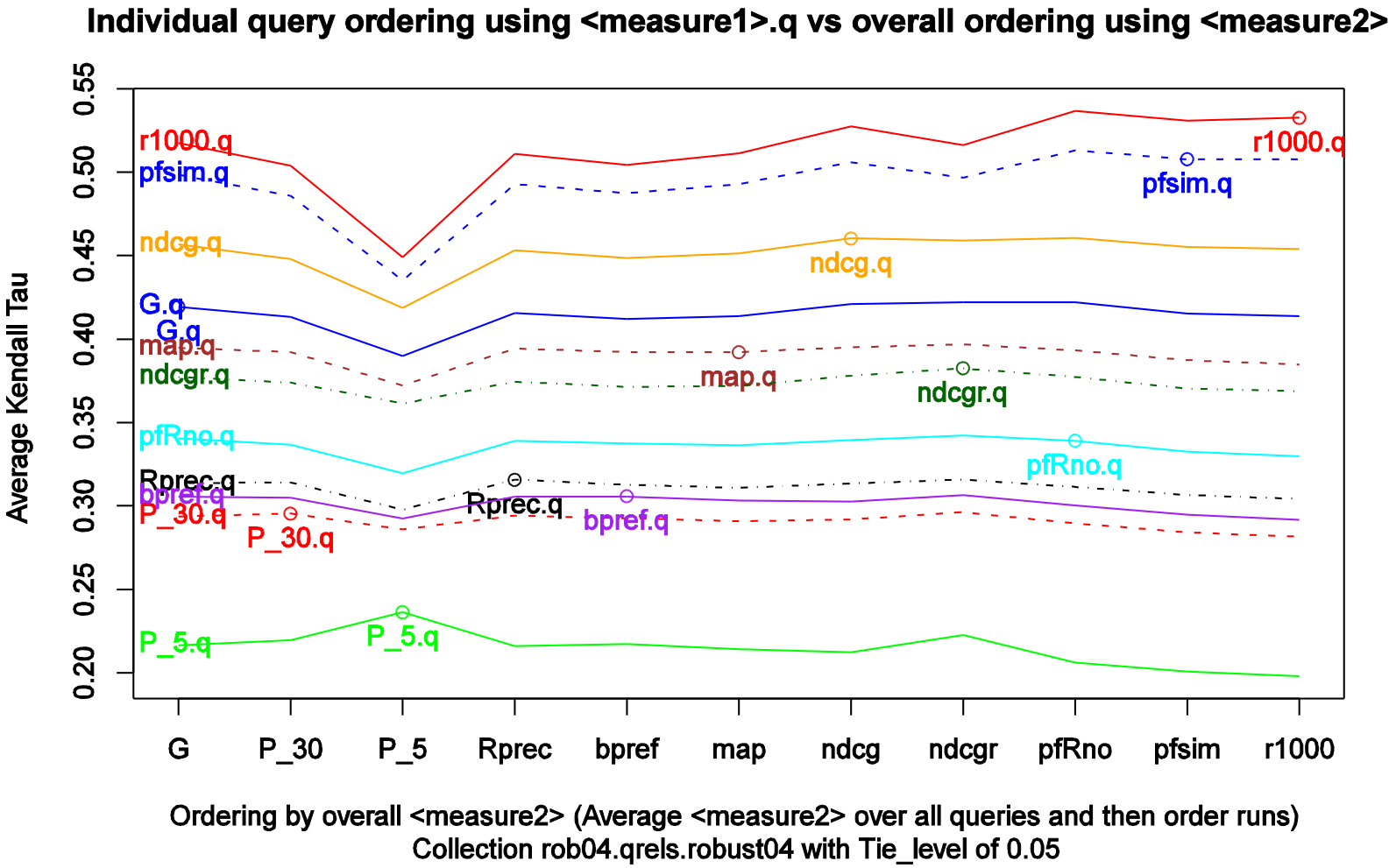
Topics Predicting Overall Rankings (Same Measure)



Topics Predicting Overall Rankings (Recall 1000)



Topics Predicting Overall Rankings (Robust04)



Implications

- Narrow ranges indicates measures are basically the same here, with the exception of P_5
  - Measures do not agree with their own overall average much more than they agree with the other overall measures
- Measures have large differences in predictive power of individual topics
- Measures are ordered by the amount of information used in them
  - Suggests differences show measurement error

Single Topic Evaluation

- Field has neglected, since we want multiple topics to completely compare systems
- Needed for several purposes including failure analysis, error bounds, and understanding
- Current measurement error is high
- Need to use more information in our measures, and more accurate information
  - Must include different user opinions
- Multiple user preference relations a solution

User Models & Metrics

Georges Dupret

August 6, 2009

Summary

1. What are the common assumptions about user behavior implicit or explicit in common metrics?
2. We identify essentially two classes:
   - Assume the user effort is fixed and estimate the session success,
   - Assume the session is successful and estimate the effort.
3. We argue that:
   - Metrics parameters can be estimated thanks to the associated user model,
   - It would be better to fix neither utility nor effort (Pareto frontier),
   - Instead of comparing metrics, we should compare user models.

Mean Average Precision

The average of the precisions at the relevant documents.

$$MAP = \frac{1}{R}\sum_{r=1}^{\infty} \text{precision at } r \times \text{ relevance at } r$$

User Model

- The user decides how many *relevant* documents he needs –say k– and browses sequentially until he finds them [Robertson, 2008].
- [Moffat and Zobel, 2008]: "Every time a relevant document is encountered, the user pauses, asks "Over the documents I have seen so far, on average how satisfied am I" and writes a number on a piece of paper. Finally, when the user has examined every document in the collection –because this is the only way to be sure that all of the relevant ones have been seen– the user computes the average of the values they have written."

Mean Average Precision (cont.)

Relation between the user model and the metric.

1. The level of a user happiness is the precision at k.
   - amount of relevance needed to achieve success is fixed.
   - precision is related to the effort.
2. We don't know the proportion of users who want exactly k documents, hence we assume a uniform distribution.

Utility & Effort

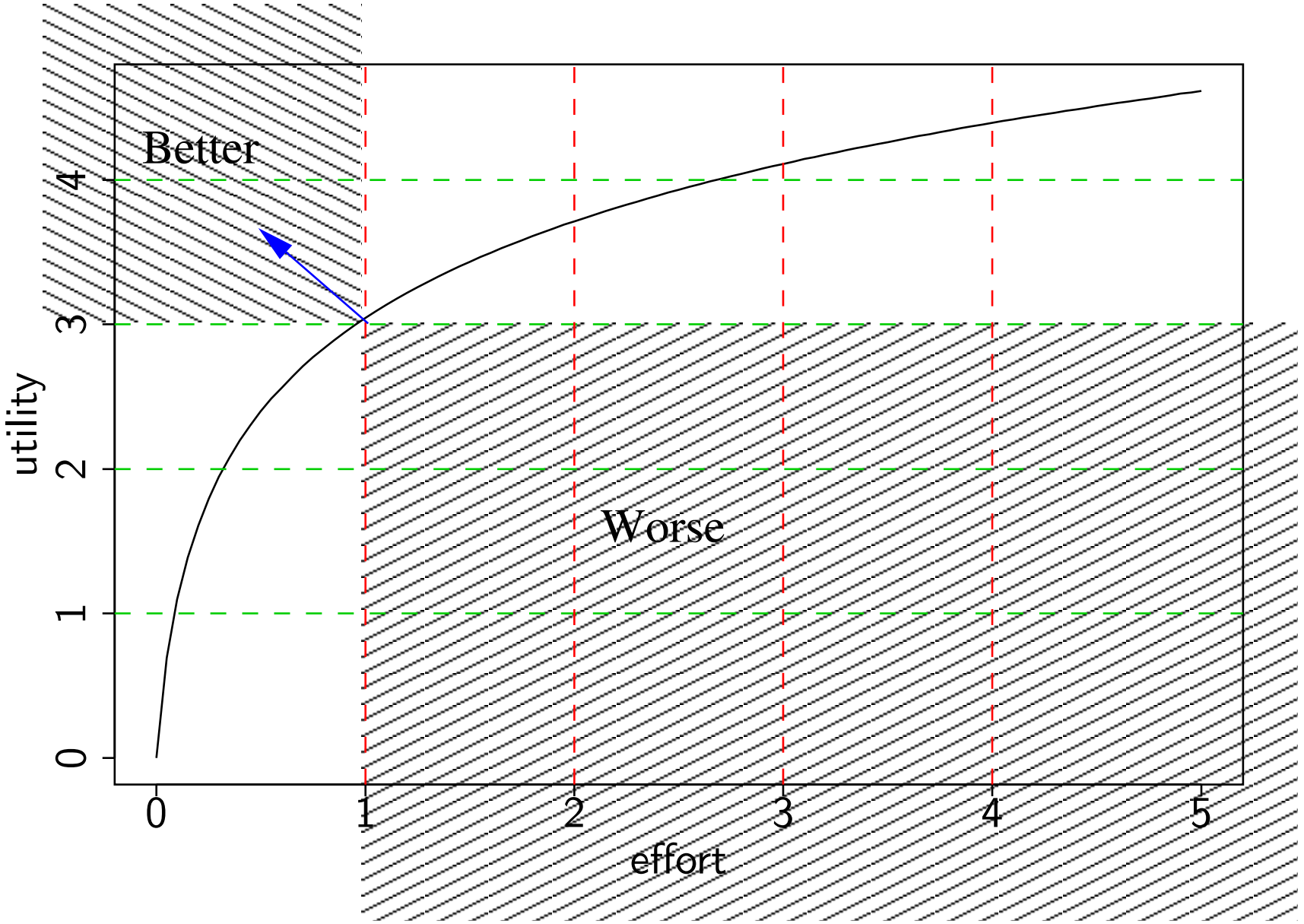
Two classes of metrics:

- DCG fix the effort and marginalize over the utility, MAP fix the utility and marginalize the effort.
- The two metrics are related to the marginalization over the utility / effort

1. User models incorporate both utility and effort to predict session success,
2. A metric derived from such a user model scales naturally: If we know $P(\text{success}, \text{utility}, \text{effort}, \text{session} | \text{ranking function})$ then

$$\text{☺} = \mathbb{E}(\text{success} | \text{utility}, \text{effort}, \text{ranking function})$$

Utility & Effort: Comparing Ranking Function



Utility & Effort: Conclusions

1. We need a metric that includes both effort & utility,
2. This metric needs a realistic user model,
3. The best user model is the one with the best predictive power,
4. The join probability offers a scale free method to compare models

$$P(\text{success}_1 > \text{success}_2, \text{utility}, \text{effort})$$

User Models

- Beware of models... navigational queries are very frequent...
- User choices during a search are limited; We can take advantage of the imposed structure to model user behavior.
  - Example of using the structure: [Piwowarski et al., 2009, Piwowarski et al., 2007],
  - Metric proposal relying on user making choices and decisions: [Fuhr, 2008].

Part 2: Boasters and Posters

- Theme 1: Human in the Loop
  - D.Hawking, P.Thomas, T.Gedeon, T.Rowlands, T.Jones, *New methods for creating testfiles: Tuning enterprise search with C-TEST*
  - N.Belkin, M.Cole, J.Liu, *A Model for Evaluation of Interactive Information Retrieval*
  - C.Paris, N.Colineau, P.Thomas, R.Wilkinson, *Stakeholders and their respective costs-benefits in IR evaluation*
  - M.Smucker, *A Plan for Making Information Retrieval Evaluation Synonymous with Human Performance Prediction*
  - S.Stamou, E.Efthimiadis, *Queries without Clicks: Successful or Failed Searches?*
- Theme 2: Social Data and Evaluation
  - O.Alonso, S.Mizzaro, *Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment*
  - T.Crecelius, R.Schenkel, *Evaluating Network-Aware Retrieval in Social Networks*
  - W.C.Huang, A.Trotman, S.Geva, *A Virtual Evaluation Forum for Cross Language Link Discovery*
  - G.Kazai, N.Milic-Frayling, *On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing*
  - Z.Yue, A.Harplale, D.He, J.Grady, Y.Lin, J.Walker, S.Gopal, Y.Yang, *CiteEval for Evaluating Personalized Social Web Search*

Boasters and Posters (cont'd)

- Theme 3: Improving Cranfield
  - T.Armstrong, J.Zobel, W.Webber, A.Moffat, *Relative Significance is Insufficient: Baselines Matter Too*
  - K.Collins-Thompson, *Accounting for stability of retrieval algorithms using risk-reward curves*
  - A.Hanbury, H.Müller, *Toward Automated Component-Level Evaluation*
  - H.Liu, R.Song, J.-Y.Nie, J.-R.Wen, *Building a Test Collection for Evaluating Search Result Diversity: A Preliminary Study*
  - M.Shokouhi, E.Yilmaz, N.Craswell, S.Robertson, *Are Evaluation Metrics Identical With Binary Judgements?*
- Theme 4: New Domains and Tasks
  - S.Ali, M.Consens, *Enhanced Web Retrieval Task*
  - M.Costa, M.Silva, *Towards Information Retrieval Evaluation over Web Archives*
  - J.Kim, B.Croft, *Building Pseudo-Desktop Collections*
  - N.Lathia, S.Hailes, L.Capra, *Evaluating Collaborative Filtering Over Time*
  - F.Llopis, A.Escapa, A.Ferrandez, S.Navarro, E.Noguera, *How long can you wait for your QA system?*

Part 3: Breakout Sessions

- Four groups on the four themes.
  - ⋆ Most exciting part of the day – but impossible to summarize
  - ⋆ but...

four panelists will give comments #ireval09 #sigir09
1:14 PM Jul 23rd from web

... each step does not have to be user-perceivable #ireval09 #sigir09
12:55 PM Jul 23rd from web

What I wrote down: Ensuring step-by-step progress in terms of user satisfaction/performance ... #ireval09 #sigir09
12:55 PM Jul 23rd from web

Keywords: rich ground truth, longitudinal evaluation #ireval09 #sigir09
12:51 PM Jul 23rd from web

Justin trying to formulate a golden sentence for Improving Cranfield #ireval09 #sigir09
12:48 PM Jul 23rd from web

Please fill out the workshop evaluation at http://tiny.cc/sigir2009ws #ireval09 #sigir09
12:40 PM Jul 23rd from web

Chris: Expand the notion of relevance #ireval09 #sigir09
12:23 PM Jul 23rd from web

Breakout sessions are great. You get to discuss intensively with all these bright people #ireval09 #sigir09
12:20 PM Jul 23rd from web

Still at Improving Cranfield breakout - other groups are laughing, we're not. People writing down their wishes (how to improve Cranfield)
12:13 PM Jul 23rd from web

Improving Cranfield #ireval09 #sigir09
11:46 AM Jul 23rd from web

Part 4: Report out and Discussion

- Four reports
  - ⋆ *Human in the Loop* (Paul Thomas)
  - ⋆ *Social Data and Evaluation* (Ralf Schenkel)
  - ⋆ *Improving Cranfield* (Justin Zobel)
  - ⋆ *New Domains and Tasks* (Mariano Consens)
- Four experts
  - ⋆ Charlie Clarke
  - ⋆ David Evans
  - ⋆ Donna Harman
  - ⋆ Dianne Kelly

Human in the Loop (Paul Thomas)

- Key idea: Evaluate user models, not systems!, by their ability to predict user performance (or satisfaction or behavior or...)
  - ⋆ This solves: better inform UI design, retrieval models, measures
  - ⋆ BUT what should we model exactly? user 'satisfaction'?
  - ⋆ Experimental: Use (extended) test collections as data
  - ⋆ Observational: Could use 'living lab' to collect interaction data plus self-reported satisfaction
  - ⋆ Collaborate with those having data for validation

- Reactions: Dianne: Happy about user-focus but wouldn't this take the user out of the loop?

Social Data and Social Evaluation (Ralf Schenkel)

- Key idea: Use crowd-sourcing (Mechanical Turk) to get to relevance
  - ⋆ This solves: costs (time, volume) of annotation/assessment
  - ⋆ Must compare agreement with traditional approach
  - ⋆ Fit tasks and their distribution to crowd-sourcing with unknown judges (many judges?)
- Reactions:
  - ⋆ Charlie: do it! But sounds like a paper?
  - ⋆ Dianne: lack of control gives problems: what is the population? Motivation to participate? Etc.

Beyond Cranfield (Justin Zobel)

- Key idea: Need rich ground truth, and longitudinal evaluation
  - ⋆ This solves: mismatch between modern search and current 'plain' relevance judgments (context-free, unannotated, etc)
  - ⋆ Compare results between papers and over time (withheld judgments)
  - ⋆ Be open to new methods for gathering user data, e.g. from the community, in an ongoing way, etc.
  - ⋆ Enough queries – more than now – with explicit treatment of ambiguity (temporal, spatial, lexical, referential)
- Reactions:
  - ⋆ Donna: Comparing over time/users/tasks is crucial for progress
  - ⋆ Charlie: Enough out of the box?

New Domains and Tasks (Mariano Consens)

- Key idea: study many different tasks, genres, contexts with direct relation to actual information access problems ('iPhone task'?)
  - ⋆ This solves: more 'realistic' evaluation for given tasks
  - ⋆ Validate techniques across scenario's
  - ⋆ Need different task scenario's and fitting user models
- Reactions:
  - ⋆ Donna: Still no alternative for the 'library search' model
  - ⋆ David: Information Access is more than search; and it is multi-lingual, multi-cultural, etc.

Wrapping Up a Looooooooong Workshop

- Set-up was to discuss concrete practical first steps
  - ⋆ That failed! Majority wanted to discuss fundamentals!
- Piecing things together:
  - ⋆ There is more to IR than system ranking
  - ⋆ We need to connect the system-side to the user-side of IR
  - ⋆ Now is the time: there are powerful ways to gather user data
  - ⋆ Need informal 'user models' underlying tasks, and formal models of information seeking behavior
  - ⋆ Need to evaluate models of users/interaction directly!
- Stephen recalled the 'revolution' of Cranfield, and speculated another 'revolution' may come...

Questions?

- Proceedings and presentations archived at
  http://staff.science.uva.nl/~kamps/ireval/