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

Jaap Kamps[1] **Shlomo Geva**[2] **Carol Peters**[3]
Tetsuya Sakai[4] **Andrew Trotman**[5] **Ellen Voorhees**[6]

[1] University of Amsterdam, The Netherlands & INEX
[2] QUT, Australia & INEX
[3] ISTI-CNR, Italy & CLEF
[4] Microsoft Research Asia, China & NTCIR
[5] University of Otago, New Zealand & INEX
[6] NIST, USA & TREC/TAC

Abstract

On July 23, 2009 the SIGIR Workshop on the Future of IR Evaluation was held as part of SIGIR in Boston. The program consisted of four keynotes, a boaster and poster session with 20 accepted papers, four breakout groups, and a final panel discussion of the breakout group reports. This report outlines the events of the workshop and summarizes the major outcomes.

1 Introduction

Evaluation is at the core of the field of Information Retrieval (IR). One of the greatest achievements of the field is the development of a rigorous methodology to evaluate retrieval effectiveness. This so-called Cranfield approach, as continued by the current evaluation fora, has served us very well: virtually all progress in IR owes directly or indirectly to test collections built within the Cranfield paradigm. However, in recent years, IR researchers are routinely pursuing tasks outside the traditional paradigm. For example, by taking a broader view on tasks, users, and context [28]. 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 sheer endless streams of queries and clicks. The recent MINDS research agenda calls for changes in data and context, in information analysis and organization, and in novel evaluation paradigms [6].

It has been 50 years since the start of the Aslib Cranfield research project that laid the foundations of experimental research in Information Retrieval [7]. The joint chairs of the evaluation fora (CLEF, INEX, NTCIR, and TREC) organized a workshop on *the future of IR evaluation*, aiming to perform a sanity-check on the current IR evaluation fora against

the novel evaluation needs of IR researchers (what are we doing right? and what do we fail to address?) as well as to work out concrete new IR tracks and tasks to take IR evaluation forward.

The questions we expected to address can be succinctly summarized as to make IR evaluation more "realistic." There is however no consensus on what then "real" IR is, or should be, and various directions have been proposed:

- 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...

We envisioned a true *work*shop where all stake-holders, ranging from those with novel evaluation needs to senior IR evaluation experts, are brought together, and develop concrete ideas for IR evaluation in the coming years. The first part of the workshop consisted of four keynotes to set the stages and frame the problem (discussed in Section 2); and a boasters and posters of twenty contributed papers (discussed in Section 3). The second part of the workshop consisted of breakout groups on 4 themes (discussed in Section 4), and a report of the outcome to a panel of experts (discussed in Section 5). The major outcomes of the workshop are discussed in Section 6.

2 Keynotes

The program started with four keynote speakers. Stephen Robertson's keynote was titled "Richer theories, richer experiments" [25]. Stephen took a bird's eye view of IR evaluation in relation to the roles of theory and experiment in the philosophy of science. In particular, there is a (partial) standoff between experimental evaluation in the Cranfield/TREC tradition, which is powerful but of limited scope, and observational studies with real users, which are realistic but of limited scale. IR experimentation has focused almost entirely on evaluating systems with respect to predicting relevance. Without denying the importance of this, it is a rather limited view, and scientifically we should aim at understanding the underlying phenomena. Instead of systems we may endeavor testing other hypotheses, even complete models or theories, and against a broader range of useful predictions than relevance. Think of redundancy/novelty/diversity, optimal thresholds, satisfaction, clicks, satisfactory or unsatisfactory search termination, query modification, etc. The resulting richer models should have something to say about both lab experiments and observational studies.

We have no grand theory of IR, not even the beginning of it. Hence there is a sense of urgency in Stephen's call for research aiming at broadening our understanding the phenomena in IR. But this will be a long road ahead...

Susan Dumais's keynote was titled "Evaluating IR *in situ*" [11]. Sue addressed the limitations of current benchmark test collections to address the scale, diversity, and interaction that characterize information systems today. Fortunately, there are also many ways of collecting user data today, ranging from small-scale user studies and user panels to log analysis and experimentation in the large. There is a need for sharable resources, such as the infrastructure and instruments for capturing user activity, or user interaction data that can be shared by the community. An example is the search logs obtained from the Lemur Query Log Project [21]. An attractive proposal is to set up an operational system by the community [18]. Such a "living laboratory" would not only allow for generating logs but also for conducting controlled experiments with novel search or user interface components.

Nowadays, there are unprecedented ways of capturing user interaction data. Sue's "living laboratory" proposal has a number of advantages that may prove to be invaluable to the field. First, it can greatly facilitate user-centered research in IR, by allowing comparative evaluation across systems and over time. Second, the *in situ* testing of retrieval components holds the promise to build powerful connections between the user-centered research and the system-centered research—Stephen's apparent standoff discussed above.

Chris Buckley's keynote was titled "Towards good evaluation of individual topics" [5]. Test collections are developed to fairly compare systems, and need to average over many topics. Chris analyzed to what extent the resulting score informs us about the per-topic performance of a system. As it turn out the system-ranking on a single topic correlates poorly (0.25-0.50) with the overall system-ranking. The correlation per measure corresponds to the amount of information used: precision at 5 does worst, and recall at 1,000 does best. In fact, measures do not agree more with their own overall average than they agree with the other overall measures. Better per topic evaluation would require richer evaluation information—such as multi-level relevance judgments, partial preference orderings, or multi-user judgments—and novel measures.

The performance at individual topics is of obvious importance, not only to improve our IR measures but also to better understand the effect of topics on performance. After all the user's experience is directly tied to the results for his or her topic.

Georges Dupret's keynote was titled "User models & metrics" [12]. Georges argued that all metrics make (often implicitly) assumptions on user behavior, which can also be evaluated against observations in search logs. If a model better predicts the user behavior on unseen data, then it is arguably more realistic, and hence supports the associated metric. There are two broad families of user models underlying IR metrics. On the one hand, there are effort-based models that fix a certain effort and measure the utility in terms of relevance retrieved. Examples of associated metrics are DCG and other rank-based metrics. On the other hand, there are utility-based models that fix a certain utility and measure the effort needed to obtain it. Examples of associated measures are MAP and other recall-based metrics. Predicting user behavior in search sessions requires user models that neither fix effort and utility, but combine both in a single model.

Georges's focus on the comparison of user models rather than metrics is very attractive, since the fidelity of the user models can be established independently. In fact, such formal user models open up a whole new line of research that naturally incorporate dynamic, interactive

aspects of information seeking behavior.

3 Posters

The program continued with boasters and posters session with twenty accepted papers, roughly falling in four themes. We will discuss the themes and papers in turn.

3.1 Human in the loop

There was a group of papers addressing user centered IR evaluation. Hawking et al. [15] discuss their experience with side-by-side comparison tools for information retrieval evaluation. They also introduce a file format for test files specifying salient features of the queries, the result documents and the ranking. Experiments can use these test files to evaluate and tune enterprise search systems to maximize the actual user satisfaction.

Belkin et al. [4] provide a methodological view to the problem of evaluating interactive IR. The main idea is centered on breaking down the information seeking episode into a sequence of interactions, each with a sequence of information seeking strategies, and propose to evaluate each interaction as well as the overall episode in term of "usefulness."

Paris et al. [24] propose a holistic model of evaluation that considers all participants or stakeholders and both costs and benefits. The four main participant roles are information seeker, information provider, information intermediaries, and system providers. For each of these the respective costs and benefits are listed.

Smucker [27] proposes a community effort (as a TREC track or elsewhere) to collect large amounts of shared user-interaction data, with the goal of enabling researchers to develop models that predict human performance rather than one of the more traditional offline measures such as precision/recall.

Stamou and Efthimiadis [29] describe research aimed at understanding user satisfaction for queries that do not receive any clicks. In a user study, test persons were asked to complete a short questionnaire about each of their searches for one day, and they report on the fraction of queries without clicks as well as the underlying intentions.

3.2 Social data and social evaluation

There was a group of papers addressing social evaluation and evaluating social data search. Alonso and Mizzaro [2] compare the judgments of TREC assessors to the judgments of Mechanical Turk'ers on a set of TREC qrels. The results are promising and show that employing turkers might be a viable solution worth further study.

Crecelius and Schenkel [10] propose to evaluate search and recommendation methods in social tagging networks using community-based relevance judging. The search requests would consist of both a topic and a user, and the judging would be relative to the individual user's position in the network.

Huang et al. [16] propose a virtual evaluation forum for evaluating cross-language link discovery. This task can be evaluated without human judgments, when pages with known links are withhold from the collection and later used as topics. This will also obviate traditional cycles and allow for continual evaluation where runs can be submitted and evaluated at any time against the withheld data. Additional, richer, human assessments can be collected to give fuller evaluations of submitted runs.

Kazai and Milic-Frayling [17] analyze crowdsourcing as a means to obtain the ground truth needed for IR evaluation, based on the experiences during the INEX 2008 Book Track's game model. The diverse backgrounds of the assessors, and the incentives of the crowdsourcing models may influence the trustworthiness and quality of the resulting data. Several indicators of trustworthiness are discussed, such as familiarity with the topic and content, dwell time and changes in dwell time, and agreement amongst judges.

Yue et al. [30] describe the development and properties of a social media test collection, containing a large number of academic bibliographic records with accompanying annotations, plus topics and queries generated and judged by a small set of experts. Experiments on the utility of the annotations for personalized search show promising results.

3.3 Improving Cranfield

A group of papers discussed a range of issues within the current evaluation methodology. Armstrong et al. [3] tabulated effectiveness claims in papers published over the last decade. Whilst many of these papers report significant improvements over a baseline, there is no evidence of an overall gain in absolute retrieval effectiveness: rarely systems outperform the best scores obtained at the original TREC conference. In order to demonstrate verifiable improvements, there is a need for reporting practices that allow for rigorous comparison with prior results. This can be facilitated with a common place where all relevant effectiveness results are brought together, EvaluatIR [13].

Collins-Thompson [8] looks at evaluation using risk-reward curves, addressing not only the average performance but also the stability or variance of the performance. The risk-reward trade-off is used to study the impact of query expansion, which is known to be risky technique. There are several ways in which the risk or variance can be quantified, and we can apply concepts from economics as reasonable starting points.

Hanbury and Müller [14] propose "component-level" evaluation by splitting various processing steps (as required for many IR tasks) into separate components, and to evaluate the impact of each processing step. This will allow for evaluating individual components, and for studying component interactions. Experiments are ongoing in the Grid@CLEF track at CLEF 2008.

Liu et al. [22] develop an IR test collection of ambiguous queries based on Wikipedia disambiguation pages. Queries correspond to disambiguation pages, and the query's ambiguity is measured using average cosine similarity between pairs of disambiguated pages. A preliminary test collection is build by submitted queries to commercial search engines, and then manually judging the results with respect to the different interpretations.

Shokouhi et al. [26] explore how often quite different measures will actually produce different comparative results across a wide range of possible rankings. When comparing over millions of pairs of TREC runs, the metrics are shown to result in very similar comparative results.

3.4 New domains and tasks

The final group of papers discussed new tasks or aspects to evaluate. Ali and Consens [1] discuss a tree-based view of search results, arguing that this may be more appropriate in some cases where the results list has a complex grouped or hierarchical organization. As a

result, we cannot use a ranked-list-of-results-type of evaluation for this scenario. Evaluation may be based on search and click data in transaction logs.

Costa and Silva [9] propose a new evaluation track focused on web archives. This extends earlier web search test collections with a temporal dimension. Web archives have many versions of the same pages that may on the hand help locate relevance URLs, and on the other hand creates the problem of which version(s) of the page to present to a user.

Kim and Croft [19] propose building an artificial test-collection to conduct research into desktop search, a domain where real data is difficult to obtain due to privacy concerns. For people mentioned in the TREC Enterprise Track's W3C collection, a thousand documents with a variety of document types are gathered from the Web. Artificial known-item topics were generated from individual document.

Lathia et al. [20] propose to take the temporal aspect of recommendation systems explicitly into account. The proposal is a test collection that treats the training data as a stream and use at any given time point the earlier historical as the only input. This captures the effects of new users joining the system, and providing more input ratings.

Llopis et al. [23] contend that efficiency is a too-often ignored component of QA system evaluation, and describes and demonstrates an approach for incorporating answer-time using the real-time QA experiment at CLEF 2006. A straightforward but naive method of a linear combination of effectiveness and efficiency is used.

4 Breakout session

Next, the workshop participants were divided into four breakout groups that discussed the four themes:

- *Human in the Loop* (chaired by Paul Thomas);
- *Social Data and Evaluation* (chaired by Ralf Schenkel);
- *Improving Cranfield* (chaired by Justin Zobel); and
- *New Domains and Tasks* (chaired by Mariano Consens).

This was the most exciting part of the day, but impossible to summarize. Figure 1 give an impression of one of the four breakout groups. Fortunately, each of the groups reported on their discussion in the final closing panel, which we will discuss now.

5 Closing Panel

The program continued with the report out of the four breakout groups, and the reactions of a panel consisting of: Charlie Clarke, David Evans, Donna Harman, and Diane Kelly.

5.1 Human in the loop

The breakout group addressed the problem that that the traditional "library consultation" user model is breaking down, and that we want to evaluate our systems with respect to better "user models." However, we don't really know what such a model would look like, and even if we did, we wouldn't know whether it is any good. The proposal, building on the earlier keynote presentations, is to evaluate user models, not systems, by their ability

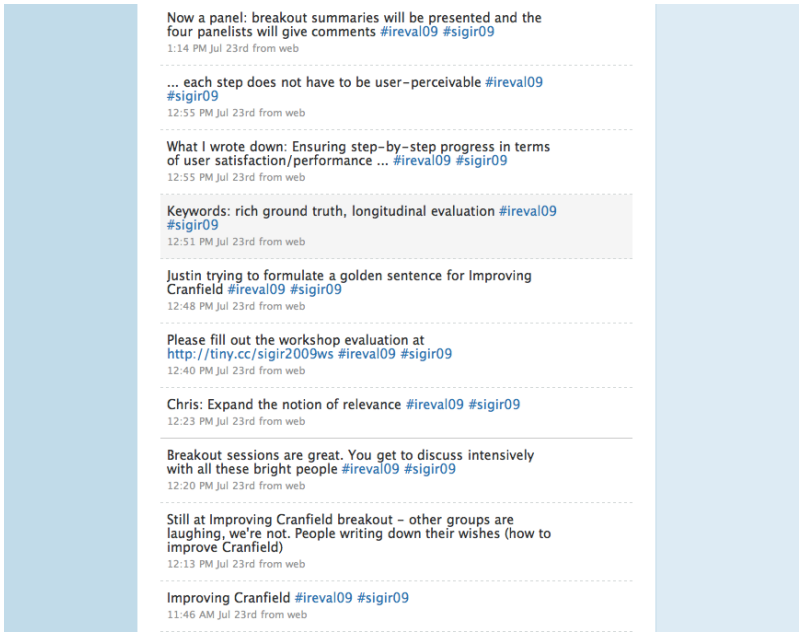

Figure 1: Live report from the "Beyond Cranfield" breakout group (tweets by Dr. T. Sakai).

to predict observable user behavior. The input would be the information needs, details of documents and rankings, etc. The output would be a prediction of user behavior. It is not clear a priori what we should model exactly, user 'satisfaction' or more directly observable behavior. We may be able to use experimental data for this in the form of (extended) test collections. A more ambitious approach would be a "living lab" collecting both interaction data and explicit (self-reported) satisfaction, hence providing some grounded data and the whole interaction path. This should lead to a spiral development where we develop initial models, and evaluate and refine them, etc.

The panel reacted positively to the focus on users and user models, however there was some concern that this could actually result in replacing real users with a formal model—in fact getting rid of users altogether.

5.2 Social data and social evaluation

The breakout groups addressed the problem of getting annotations and assessments in novel ways, in particular using the methodology of crowdsourcing. There is a need for richer assessments, without leading to extravagant expenses, and preferable get them fast. Crowdsourcing methods, such as offered through the Mechanical Turk, offers the workforce needed for this. There are still open questions on the setup of the evaluation task, the quality and scale of the resulting assessments, and the quality of the resulting rankings. The concrete proposal is to systematically compare crowdsourcing judgments to those obtained from traditional assessors, and experiment with various parameters such as the complexity of the assessment task, the grouping or size of results, the number of judges per result, etc.

The panel liked the suggestion, which seems straightforward to set up, but also expressed worries about the lack of control and potential biases: who are these turkers? what population do they represent? and what is their motivation to participate?

5.3 Improving Cranfield

The breakout group address substantial extensions to the current Cranfield paradigm for system measurement. The first part of the proposal is to collect a rich ground truth corresponding to modern information use. This includes differences in user psychology, underspecified queries, and closer user involvement. This is a substantial departure from the current "plain" relevance judgments which are context-free, unannotated, etc. We should be open to new methods for gathering user data, e.g. from the community, in an ongoing way. The second part of the proposal is to facilitate longitudinal evaluation, with rich reporting and recording all runs. This will help demonstrate that, of if, systems improve significantly over time. A fair comparison of results is achieved by comparing against common baselines, and using unseen withheld relevance judgments.

The panel stressed that comparing over time/users/tasks is crucial for progress, but also expressed concerns whether the proposals were radical enough.

5.4 New domains and tasks

The breakout group discussed more "realistic" evaluation scenarios. The key idea is to study many different tasks, genres, and contexts with direct relation to actual information access problems. A broad range of different tasks and scenario's was discussed. For example, think of an "iPhone task" giving access to a variety of different sources. Working on a multitude of tasks will allow us to validate IR techniques across different scenario's, and corresponding user models, aiding to our understanding of what works when, and why.

The panel liked the interest in novel user models that go beyond the "library consultation" scenario, and stressed that information access is more than search, and it is multi-modal, multi-lingual, multi-cultural, etc. and we should study scenario's that do justice to that.

6 Trying to Summarize

The setup of the workshop aimed to focus on concrete first steps for the near future. This failed miserably! The majority of participants wanted to discuss more fundamental aspects of IR evaluation. This gave a very exciting workshop with a lot of "food for thought." Many of the suggestions have potentially far reaching consequences. This also makes it more difficult to summarize the outcome in concrete lessons for the future of IR evaluation.

There seems to be one direction that surfaced throughout the workshop. There is more to IR than the evaluation of systems and their rankings, and important exceptions aside there is a general disconnect between user-centered and system-centered research in IR. Hence, we need to find novel connections between both strands by broadening the scope of system-centered research to provide richer context of the search requests and the judgments, as well as to other useful predictions than relevance. And we have to match these with large-scale user-centered research. The time for this seems exactly right: there are now powerful ways to gather user data. Key element in this is the need of more explicit hypotheses and models of the phenomena we are investigating. We need new informal "user models" underlying tasks, and formal models of information seeking behavior that can be subjected to empirical testing. That is, we need to evaluate models of users/interaction directly.

At the end of his presentation, Stephen Robertson recalled the "revolution" that the Cranfield work caused in IR, and wondered another "IR revolution" may come. We may have seen the seeds of this second revolution being planted at the workshop...

Acknowledgments

We would like to thank ACM and SIGIR for hosting this workshop, in particular Jay Aslam and James Allan for their outstanding support in the organization. We would also like to thank the program committee, consisting of Omar Alonso, Chris Buckley, Charles Clarke, Nick Craswell, Susan Dumais, Georges Dupret, Nicola Ferro, Norbert Fuhr, Donna Harman, David Hawking, Gareth Jones, Noriko Kando, Gabriella Kazai, Mounia Lalmas, Stefano Mizzaro, Iadh Ounis, Stephen Robertson, Ian Ruthven, Anastasios Tombros, Stephen Tomlinson, Justin Zobel, and the six program chairs. Final thanks are due to the paper authors, the panelists, and the participants for a great and lively workshop. The University of Amsterdam is hosting the workshop proceedings and presentations, which are on-line at `http://staff.science.uva.nl/~kamps/ireval/`.

References

[1] S. Ali and M. Consens. Enhanced web retrieval task. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 35–36, 2009.

[2] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

[3] T. Armstrong, J. Zobel, W. Webber, and A. Moffat. Relative significance is insufficient: Baselines matter too. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 25–26, 2009.

[4] N. Belkin, M. Cole, and J. Liu. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 7–8, 2009.

[5] C. Buckley. Towards good evaluation of individual topics. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, page 1, 2009.

[6] J. Callan, J. Allan, C. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai. Meeting of the MINDS: an information retrieval research agenda. *SIGIR Forum*, 41: 25–34, December 2007. `http://doi.acm.org/10.1145/1328964.1328967`.

[7] C. W. Cleverdon. Report on the first stage of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1960.

[8] K. Collins-Thompson. Accounting for stability of retrieval algorithms using risk-reward curves. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 27–28, 2009.

[9] M. Costa and M. Silva. Towards information retrieval evaluation over web archives. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 37–38, 2009.

[10] T. Crecelius and R. Schenkel. Evaluating network-aware retrieval in social networks. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 17–18, 2009.

[11] S. Dumais. Evaluating IR in situ. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, page 2, 2009.

[12] G. Dupret. User models to compare and evaluate web IR metrics. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, page 3, 2009.

[13] EvaluatIR. An online tool for evaluating and comparing ir systems, 2009. `http://evaluatir.org/`.

[14] A. Hanbury and H. Müller. Toward automated component-level evaluation. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 29–30, 2009.

[15] D. Hawking, P. Thomas, T. Gedeon, T. Rowlands, and T. Jones. New methods for creating testfiles: Tuning enterprise search with C-TEST. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 5–6, 2009.

[16] W. C. D. Huang, A. Trotman, and S. Geva. A virtual evaluation forum for cross language link discovery. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 19–20, 2009.

[17] G. Kazai and N. Milic-Frayling. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 21–22, 2009.

[18] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *Computer*, 42:60–66, 2009. `http://dx.doi.org/10.1109/MC.2009.82`.

[19] J. Kim and B. Croft. Building pseudo-desktop collections. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 39–40, 2009.

[20] N. Lathia, S. Hailes, and L. Capra. Evaluating collaborative filtering over time. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 41–42, 2009.

[21] Lemur Query Log Project, 2009. `http://lemurstudy.cs.umass.edu/`.

[22] H. Liu, R. Song, J.-Y. Nie, and J.-R. Wen. Building a test collection for evaluating search result diversity: A preliminary study. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 31–32, 2009.

[23] F. Llopis, A. Escapa, A. Ferrandez, S. Navarro, and E. Noguera. How long can you wait for your QA system? In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 43–44, 2009.

[24] C. Paris, N. Colineau, P. Thomas, and R. Wilkinson. Stakeholders and their respective costs-benefits in IR evaluation. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 9–10, 2009.

[25] S. Robertson. Richer theories, richer experiments. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, page 4, 2009.

[26] M. Shokouhi, E. Yilmaz, N. Craswell, and S. Robertson. Are evaluation metrics identical with binary judgements? In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 33–34, 2009.

[27] M. Smucker. A plan for making information retrieval evaluation synonymous with human performance prediction. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 11–12, 2009.

[28] K. Sparck Jones. What's the value of TREC – is there a gap to jump or a chasm to bridge? *SIGIR Forum*, 40:10–20, 2006. `http://doi.acm.org/10.1145/1147197.1147198`.

[29] S. Stamou and E. Efthimiadis. Queries without clicks: Successful or failed searches? In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 13–14, 2009.

[30] Z. Yue, A. Harplale, D. He, J. Grady, Y. Lin, J. Walker, S. Gopal, and Y. Yang. CiteEval for evaluating personalized social web search. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 23–24, 2009.