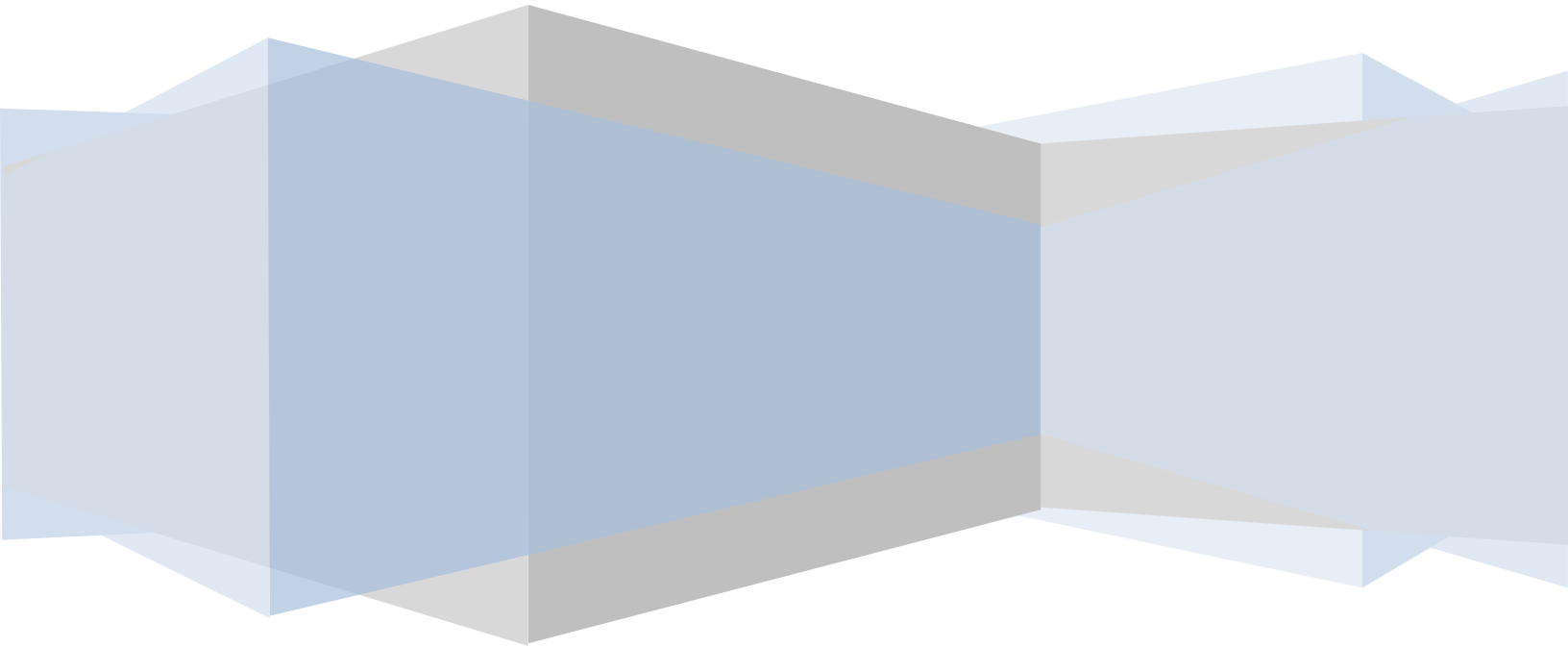
NII Shonan Workshop, Oct 8-12, 2012

Whole-session evaluation of interactive information retrieval systems

Compilation of Homework

Susan Dumais



Participants

Leif Azzopardi, University of Glasgow, UK
Peter Bailey, Microsoft Bing, USA
Nicholas J. Belkin, Rutgers University, USA
Paul Bennett, Microsoft Research, USA
Corrado Boscarino, CWI, The Netherlands
Ben Carterette, University of Delaware, USA
Charles Clarke, University of Waterloo, Canada
Susan T. Dumais, Microsoft Research, USA
Norbert Fuhr, University of Duisburg-Essen, Germany
Viktors Garkavijs, NII, Japan
Kalervo Järvelin, University of Tampere, Finland
Hideo Joho, University of Tsukuba, Japan
Jaap Kamps, University of Amsterdam, The Netherlands
Noriko Kando, National Institute of Informatics (NII), Japan
Evangelos Kanoulas, Google, Switzerland
Diane Kelly, University of North Carolina at Chapel Hill, USA
Gary Marchionini, University of North Carolina-Chapel Hill, USA
Douglas W. Oard, University of Maryland, USA
Jeremy Pickens, Catalyst Repository Systems, USA
Tetsuya Sakai, Microsoft Research, China
Mark Sanderson, RMIT University, Australia
Arjen P. de Vries, Centrum Wiskunde & Informatica, The Netherlands
Max Wilson, University of Nottingham, UK

Leif Azzopardi, Glasgow University, UK

Kelly, D., Dumais, S., and Perderson, J. O. (2009) Evaluation Challenges and Directions for Information-Seeking Support Systems, In Computer, IEEE, p44-50.
This paper points out a number of major challenges in the evaluation of Interactive IR. The main problems identified with current approaches include: (i) user/task models are not adequately captured, (ii) information continually changes over time, (iii) IIR tasks are often very complex and thus hard to model as they evolve, and may not have fixed endpoints, and (iv) IIR often occurs over time and across sessions. While the paper doesn't provide any concrete solutions to these problems, the most promising suggestion is the use of a "living laboratory". The development of such a living lab that is open to researchers would certainly provide a number of ways to evaluate users in the wild - over coming some of the pragmatic problems typically associated with evaluation.

Bookstein, A., (1982) Information Retrieval: A Sequential Learning Process, Journal of the American Soceity for Information Science, 34(5):331-341.
Tague-Sutcliffe, J. (1992) Measuring the Informativeness of a Retrieval Process, In the Proceedings of the 15th ACM SIGIR. p23-36.
These two works suggest that we should focus on the sequence in which users experience, encounter and process information. Bookstein tries to model the retrieval process as a sequence in order to develop a better retrieval system (and is perhaps a pre-cursor to the Interactive Probability Ranking Principle). On the other hand, Tague-Sutclifee tries to measure the informativeness of the process (where informativeness is akin to the novelty and diversity measures being developed). Key in these works is the focus on the order in which the users examine documents.

Smucker, M. D., and Clarke, C., (2012) Time-Based Calibration of effectiveness Measures, In Proceedings of the 35th ACM SIGIR, p95-104.
This paper provides a novel an potentially interesting solution to evaluation across a session. In some respects this work blends developments in HCI with IR. Specifically taking a GOMS like approach by Card and Moran along with Dunlop's work on time, relevance and interaction modelling to produce a "probabilistic GOMS" for IR where the main actions in the search process are assigned a time, and a probability is assigned to these actions. This provides an interesting way to examine and explore a range of potential interactions with the system - as a way to cater for the variety of ways that users interact with systems.

Azzopardi, L. (2009), Usage Based Effectiveness Measures, In Proceedings of 18th ACM CIKM, p631-640.
In terms of evaluating the whole session, I have been particularly interested in developing measures that examine how well a user uses an application. The fundamental idea is that, what should be evaluated is the sequence of interactions and documents that the user examines and inspects during the process (i.e. following on from Bookstein and Tague-Sutcliffee, along with Norman's idea of the user experience is defined by the sequence of interactions.) The experience, whether it be, engagement, utility, fun, etc. at any particular point of time across the experience is monitored, measures and modeled to provide an overview of the user's experience. This stream-based / time centric view is in contrast to rank-based approaches usually used in evaluations and would provide a natural way to measure the whole-session.

Peter Bailey, Microsoft Bing, USA

Jones, R., & Klinkner, K. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs, CIKM 2008.
This is "old", but it's continued to grow on me over time in its framing of the tasks that people conduct in relation to search engines, how to break them down from an analysis perspective into search goals and missions, that may be independent of any "session" (identified through some time-activity/inactivity window) within a search engine log. The data used was from Yahoo search logs, and involved human annotation of 312 searchers behavioral search data, from a 3 day period. A key finding is that search tasks may be interleaved (17% in their data), and 20% are hierarchically organized (that is, that there are multiple tasks comprising a single search mission). Tasks are atomic units of information seeking activity, but may require multiple queries to satisfy. A number of researchers have built on top of this work, investigating many different aspects of more complex user search activity, in areas like task identification and task success, query reformulation, query suggestions, search diversification and more.

Lindley, S., Meek, S., Sellen, A., & Harper, R. It's Simply Integral to What I Do: Enquiries into how the Web is Weaved into Everyday Life, WWW 2012.
This is pretty new, and what I like about this diary/observational study is their identification of different kinds of web activity and how depending on your "mode", you will be seeking and behaving very differently. Five major modes were identified: respite, orienting, opportunistic use, purposeful use and lean-back internet. Ultimately, it's another taxonomy general Web interaction behavior. Understanding the mode of behavior has the potential to condition very different search support interfaces. For example, users involved in orienting can be supported through surfacing common search activities learned from repeat behavior. Whereas opportunistic use might benefit from exploratory search and recommendations support. Historically, I suspect that most commercial search systems have assumed a purposeful mode of activity. A few years ago, Jan Pedersen said to me that it's really helpful to have multiple taxonomies, not just one, as different taxonomies give you "slices" of insight into user understanding and the "triangulation" amongst these help in getting actionable outcomes/modeling; having just one taxonomy is not sufficient typically. I found that this set of analysis has helped to consider search activity in more modal ways, embedded in a larger pattern of information interaction behavior.

Nicholas J. Belkin, Rutgers University, USA

I have two papers which I think are important, plus experience in the TREC Session Track on which I'd like to comment.

One paper is that of Perti Vakkari, cited by Diane in her homework. I find it significant for the same reasons that she cites. Here's the reference again:
Vakkari, P. (2010). Exploratory searching as conceptual exploration. Proceedings of the Fourth Human Computer Information Retrieval Workshop, New Brunswick, NJ, 24-27.

The other work, I'm somewhat embarrassed to say, is by our own group at Rutgers, in which we propose to evaluate whole search sessions according to three degrees of *usefulness*: Usefulness of the system as a whole in support of the task which motivated information seeking; Usefulness of the outcome of the support in each stage of the search session with respect to the searcher's accomplishment of the motivating task; and, Usefulness of the support provided by the system for each Information Seeking Strategy that the searcher engages in during the search session.

Here are two relevant references to this work:
Cole,M., Liu,J., Belkin,N.J., Bierig,R., Gwizdka,J., Liu,C., Zhang,J. & Zhang, X. (2009) Usefulness as the criterion for evaluation of interactive information retrieval. In: Proceedings of the Third Human Computer Information Retrieval Workshop, Washington, DC.

Belkin, N.J. (2010) On the evaluation of interactive information retrieval systems. In: B. Larsen, J.W. Schneider & F. Åström (Eds.) The Janus Faced Scholar. A Festschrift in Honour of Peter Ingwersen (pp. 13-21). Copenhagen: Royal School of Library and Information Science.

With respect to the TREC Session Track, we, and I think all of the other participants, were quite dissatisfied with the only evaluation criterion, and its related measures, that the Track has been able to come up with, using a test collection of search sessions. The criterion that was decided upon was "how much better can a system which takes account of the searcher's behaviors during a search session do in improving the results of the searcher's final query?" Clearly, this criterion cannot address the issue of evaluation of system support for the search session as a whole, yet the participants in the Track could not, and still cannot, identify criteria, measures and methods which could do whole-session evaluation in the context of a test collection.
Our experience in the Session Track, as well as our experiences in the Interactive and HARD TREC Tracks, makes me skeptical of the possibility of test-collection style evaluation of system support for whole search sessions (and even more of sequences of search sessions related to the same motivating goal/tasks).

Paul Bennett, Microsoft Research, USA

Considering the Impact of Types of Interaction on Search Session Evaluation

A variety of work on session analysis and evaluation posits that the final documents, query, or clicks in a search session is a good proxy for the user's information goal [2][3]. However, this fails to distinguish between two common information seeking patterns: (1) users reformulate and issue another query because of poor results or an incomplete answer to their information seeking goal; (2) users successfully find information needed for an aspect of their need and continue to search with a new aspect (the choice of aspect itself is often influenced by the successful result.). The latter of these types of information seeking patterns is much more common in tasks such as exploratory search and comparative shopping. These tasks, which occur quite often in web search engine logs, are typically more complex and may extend across sessions [4], but even within a session there is room for how measures of whole session evaluation can be improved.

For example, consider an information need such as "Find a good college that will admit me near where I live" for a user that lives in Lancaster, Pennsylvania. This user may start a session with the query [us northeast colleges] and click on an article to Forbes.com's *Best Colleges in the Northeast* article. This then may be followed by a series of queries such as [Williams College admissions], [Williams College location], [Princeton University admissions], [us best colleges Pennsylvania], [penn], [penn admissions], [cmu admissions] where nearly all queries provide some relevant information. Certainly only considering the impact on the final query of this session is not a good indicator of the user's overall satisfaction with the search experience. However, at the micro-level, we see that "last item as goal" may be appropriate for clear refinement patterns such as [penn] $\rightarrow$ [penn admissions]. Bennett et al. [1] tried to address these issues by: measuring improvements for ***all*** queries in a session; using dwell time as a proxy for satisfaction to separate clicks on relevant documents from spurious clicks; and "propagating relevance" in a limited fashion by considering documents relevant to later queries to be relevant to earlier queries in cases of micro-patterns of query refinements. However, significant room remains to both improve these measurements and consider other factors (e.g. the number of query words typed vs. the number of documents found).

[1] P.N. Bennett, R.W. White, W. Chu, S.T. Dumais, P. Bailey, F. Borisyuk, and X. Cui (2012). Modeling the Impact of Short- and Long-Term Behavior on Search Personalization. In *Proceedings of SIGIR '12*. 2012.

[2] D. Downey, S. Dumais, D. Liebling and E. Horvitz (2008). Understanding the relationship between searchers' queries and information goals. In *Proceedings of '08.* 2008

[3] E. Kanoulas, B. Carterette, M. Hall, P. Clough, M. Sanderson (2011). Overview of the TREC 2011 Session Track. In *Proceedings of TREC '11.* 2011.

[4] Alex Kotov, Paul N. Bennett, Ryen W. White, Susan Dumais, and Jaime Teevan (2011). Modeling and Analysis of Cross-Session Search Tasks. In *Proceedings of SIGIR '11*. 2011.

Corrado Boscarino, CWI, The Netherlands

We can concile two broad categories of both IR system's design and evaluation, "user-driven" and "system-driven" [Borlund 2003], through a formalisation of the results of studies in search behaviour. We abstracted observed temporal dependency of relevance judgements into a discount model for query expansion [Boscarino 2012] and we tested this retrieval strategy on TREC session track data. In a TREC setting we cast user-driven evaluation into a formal procedure: real users are exchanged for assessors [Kelly 2009], but the outputs are shareable metrics that can be computed on any ranked list.

At the design side we can model user interactions with a search system, as observed in long sessions, also within a more general formal framework and instantiate this formalism with behavioural cues. We developed a model for user interactions based on a probabilistic dynamic epistemic logic [Kooi 2003]. This model accounts for how user interactions induce modifications on the probability space that we use for calculating the distribution of the relevant population by conditioning on the observed events. We tested this approach on 2012 TREC data. Although at the time of writing the track results are not available yet, previous experiments on 2011 data show an improvement of 10% on the RL1 task.

Can we improve on the evaluation side using a similar strategy? More precisely, can we use a combination of formal reasoning and user data to attain the same control on the evaluation process as in TREC tasks, without a need for assessors and instead with real users in the loop?

This point is open for discussion and I can only provide some constraints that models should satisfy and some promising research paths.

Logic models can handle both probabilistic information (example: relevance of a document to a query), non-probabilistic one (example: observation of a click) and higher level information (example: how a system's parameter set change after observing a session). They can formalise users as reasoning agents [Halpern 1995], players in a game [Halpern/Tuttle 1992] or sets of axioms [ten Cate/Shan 2002]. However, their soundness relies on a closed world assumption and they badly scale [Crestani 1995], if large search logs might become available.

Reasoning at the scale of IIR evaluation campaigns requires therefore logics that extend onto large, incomplete, incoherent and changing datasets.
Reasoning onto web data faces similar challenges [Baader 2005] and our community could join its forces towards developing a platform where research groups can plug in their designs of both retrieval algorithms and performance metrics, which can in turn be shared with other groups. Our aim should be eliminating the methodological gap between design and evaluation, and still allowing for comparison of results.

[Borlund 2003]: Pia Borlund, "IIR evaluation model: a framework for evaluation of interactive information retrieval systems"
[Boscarino 2012]: Corrado Boscarino et al., "Adapting Query Expansion to Search Proficiency"
[Kelly 2009]: Diane Kelly, "Methods for evaluating IIR systems"
[Kooi 2003]: Barteld Kooi, "Probabilistic Dynamic Epistemic Logic"
[Halpern 1995]: Joseph Y. Halpern, "Reasoning about Knowledge: a Survey"
[Halpern/Tuttle 1992]: Joseph Y. Halpern and Mark R. Tuttle "Knowledge, Probability, and Adversaries"
[ten Cate/Shan 2002]: Balder ten Cate and Chung-chieh Shan, "Question Answering: from Partitions to Prolog"
[Crestani 1995]: Fabio Crestani et al., "The Troubles with Using a Logical Model of IR on a Large Collection of Documents"
[Baader 2005]: Franz Baader et al., "Description Logic Based Approach to Reasoning about Web Services"

Ben Carterette, University of Delaware, USA

To me, one of the key challenges in session-based evaluation is the construction of portable, reusable test collections that both academics and industrial researchers can use to work on improving retrieval over sessions. The problem is that the traditional Cranfield paradigm of canned topics with short queries and relevance judgments, which I think still has (or can have) a lot of value for ad hoc-type tasks, is inadequate for sessions of interactions. We can have test collections that consist of canned sessions (reformulations of queries), but the problem is that user actions following the first query depend very much on what results are retrieved and ranked, and which the user looks at, for that query. It doesn't seem reasonable to assume that the sequence of queries will be the same regardless of the system that is being tested with the collection, but that is the assumption the Cranfield paradigm requires.

The second key challenge is defining evaluation measures that can work with session test collections. Ideally these measures should take into account that user actions will vary depending on system ranked results. But I'd say that I see this as secondary to the formation of test collections. Having an idea of what we want to measure will lead to the right type of test collection; the specific form of the evaluation measures doesn't matter as much after that.

These challenges were my main motivation for joining the organizing team for the TREC Sessions track, which has the "stealth" goal of learning how to build reusable test collections and evaluation measures in the TREC style (in addition to its advertised goal of learning how to improve retrieval over sessions). In the most recent two years of the track, we have constructed test collections with the type of static sessions I described above. We avoid the problem of dealing with differences between systems by only evaluating the very last query in the session and providing ranked results for the queries prior to that. But this is a compromise; in my opinion it is still far from what a test collection for session evaluation should be.

Charles Clarke, University of Waterloo, Canada

Time-Biased Gain

Over the past year, Mark Smucker and I have been working on a new evaluation framework, called time-biased gain. Time-biased gain unifies and generalizes many traditional effectiveness measures while accommodating aspects of user behavior not captured by these measures. By using time as a basis for calibration against actual user data, time-biased gain can reflect aspects of the search process that directly impact user experience, including document length, near-duplicate documents, and summaries. Unlike traditional measures, which must be arbitrarily normalized for averaging purposes, time-biased gain is reported in meaningful units, such as the total number of relevant documents seen by the user. In work reported at SIGIR 2012, we proposed and validated a closed-form equation for estimating time-biased gain, explored its properties, and compared it to standard approaches. In work reported at CIKM 2012, we used stochastic simulation to numerically approximate time-biased gain, an approach that provides greater flexibility, allowing us to accommodate different types of user behavior and increase the realism of the effectiveness measure. In work reported at HCIR 2012, we extended our stochastic simulation of time-biased gain to model the variation between users. At the workshop, I hope to talk about how the framework can be adapted to whole-session evaluation.

Susan T. Dumais, Microsoft Research, USA

Challenge 1. Sessions are not all alike.
Search sessions are conducted for many different purposes. Some involve simple tasks (e.g., finding a reference for a paper, or an image for a presentation), others more complex (e.g., learning about a new medical condition), and still others are even more involved and may extend over time (e.g., planning a trip or wedding, or purchasing a new car). A better understanding of the different kinds of search goals that are most common in sessions would help focus research and guide the development of models and evaluation measures. Broder's distinction between navigational and informational queries provided a simple taxonomy of web search activities, and I believe that a comparable understanding of common goals in search sessions is an important place to start.

Challenge 2. Sessions are complex.
Search sessions (as observed in log data) are complex, often involving interleaved tasks, at many different levels of granularity, and extending over time. Laboratory experiments typically ask participants to focus on a particular search goal, thus eliminating multi-tasking. The ability to identify activities related to the same task is an important pre-requisite for modeling the progress toward task completion during a session. Further, some tasks cannot be completed during a single session, and are extended over time or over devices. Being able to know when a task is in progress is important in both modeling success of the current session and in supporting task resumption at a later time (Kotov et al. SIGIR 2011). In many ways, tasks (rather than sessions) seem like the appropriate unit of analysis.

Challenge 3. Evaluation methodologies.
The Cranfield style of experiment in which queries, documents and relevance judgments are fixed is not well-suited to interactive information retrieval. The sequence of activities that take place during the course of a session seems critical in determining the relevance of results. It is challenging to accommodate either system or user differences in this kind of highly contextualized environment. One technique that we have explored to address this is to link explicit judgments *in situ* (which are difficult to obtain) and implicit behaviors (which are much more plentiful). Using what we called the "Curious Browser", we asked individuals to judge the relevance of individual results as well as entire search sessions in actual search sessions (Fox et al., 2005). When then developed predictive models to link patterns of implicit activity with explicit judgments. The resulting models can then be used in an open loop to label other sessions. This approach is grounded in observable search behavior that is available in operational search systems or "living laboratories".

Norbert Fuhr, University of Duisburg-Essen, Germany

Markov modeling of search sessions for evaluation, system tuning and user guidance

The interactive PRP [Fuhr 08] characterizes interactive retrieval as a sequence of situations, where, in each situation, the user is confronted with a list of choices. Each choice is described by three parameters, namely the effort for evaluating it, the probability that the user will accept it, and the benefit resulting from acceptance. For estimating these parameters, we have shown in [Tran & Fuhr 12] how we can combine gaze tracking data and query logs for observing cognitive user actions in search sessions. From this data, user effort and acceptance probability for each choice can be estimated immediately. We also can derive Markov models characterizing search behavior. Based on these models, we can compute the expected time for identifying the first/next relevant document for any search situation. From this model, it is also possible to estimate the values of time-based retrieval measures [Smucker & Clark 12], as well as determining the effect of system changes on the resulting overall quality. Besides simulating and tuning systems this way, the approach could also be used for guiding users in order to optimize their search interaction.

The most crucial issue for applying this approach is the situation-specific estimation of the model parameters. Currently, we are working with a small number of situation types, where the parameters are only type-specific. We need to consider not only the (probabilistic) ranking of choice lists produced by the system, but also user-dependent session parameters like e.g. the changes due to query reformulation.

[Fuhr 08] N. Fuhr (2008).
A Probability Ranking Principle for Interactive Information Retrieval. *Information Retrieval 11(3).*
[Tran & Fuhr 12] Vu T. Tran; Norbert Fuhr (2012).
Using Eye-Tracking with Dynamic Areas of Interest for Analyzing Interactive Information Retrieval. In Proc SIGIR 2012, pp. 1165-1166.
[Smucker & Clark 12] Mark D. Smucker; Charles L. A. Clarke (2012). Time-based calibration of effectiveness measures. In Proc SIGIR 2012, pp. 95-104.

Viktors Garkavijs, NII, Japan

Evaluating interactive image search sessions

One of the main objectives of an IR system is the satisfaction of the user information needs. While the systems for text document retrieval deal mostly with natural language, the image retrieval systems may not always rely just on descriptions of objects depicted on the images. Like the adage "a picture is worth a thousand words" says, it may be difficult to formulate a 1-3 words long textual query to retrieve the needed image results. Moreover, the content-based image retrieval (CBIR) methods, capable of detecting the objects, lack the ability to provide annotations for abstract notions, thus, the queries such as "happiness", "memories" or "calmness" are not usable for this type of image retrieval. For these tasks, the systems can rely on user annotations, such as social tags.

The experiments with the IR systems that learn to rank interactively are difficult to repeat, as it is difficult to ensure that exactly the same interactions will take place among different iterations of the experiments. This is true, regardless of whether they are performed by different users or the same user is asked to repeat his or her search session. Repeating the experiment becomes especially difficult, if the system uses devices such as an eye-tracker for interaction. For example, if a system assigns weights to image tags depending on dwell time (or fixation duration in case of eye-tracking), and bases its relevance formula on image tag weights, different dwell time sequences will lead to different ranking results, eventually eliminating from the search result page some of the objects that are required to repeat the interaction sequence. Tuning the relevance formula also results in the same problem, thus making an automatic replay of a search session valid only for the interactions, recorded with that relevance formula and useless for evaluating a new one. Thus, the standard models for IR system evaluation that rely on test collections with relevance assessments are not easily applicable for this kind of systems.

Besides the problems with pure mechanical repeatability, we also face a problem of tracking the changes of mental state of the searcher during the search session. One important difference between text retrieval and image retrieval is that the user can see the whole image at one glance, when comparing to the text search, the user has to at least skim through the document in order to judge about its relevance. A search result page may contain tens of images, however it virtually takes only a few seconds for the user to evaluate all of them, comparing to "ten blue links", which often require to be clicked in order to understand the contents and judge about its relevance. There are evidences about user behavior patterns for web search, such as the "golden triangle", however we don't have such data when it comes to image search. What do we know about user's decisions to click on a specific image thumbnail on a result page? How was it influenced by the other image thumbnails that are on the current page, or were on previous pages? Does the user always click on the most relevant thumbnail on the page? These questions are very difficult to answer, and to our present knowledge there are currently no established evaluation methods for image search sessions.

Kalervo Jarvelin, University of Tampere, Finland

Azzopardi, L.:(2011). The economics in interactive information retrieval. In: Baeza-Yates, R. & al. (Eds.) Proceedings of the ACM SIGIR'11, pp. 15--24.
The paper proposes cost models grounded in cognitive load, and identifies search strategies that minimize the cost of interaction. From the abstract (edited): Searching is inherently an interactive process usually requiring numerous iterations of querying and assessing in order to find the desired amount of relevant information. Essentially, the search process can be viewed as a combination of inputs (queries and assessments) which are used to "produce" output (relevance). The paper adapts microeconomic theory to analyze the dynamics of Interactive Information Retrieval. The search process is taken as an economic problem and the paper simulates sessions on TREC test collections analyzing which combinations of inputs produce relevance. The analysis reveals that the total Cumulative Gain (output) obtained during the course of a search session is functionally related to querying and assessing (inputs. Further analysis using cost models, that are grounded using cognitive load as the cost, reveals which search strategies minimize the cost of interaction for a given level of output.

Baskaya, F. & Keskustalo, H. & Järvelin, K. (2012). Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In: Callan, J., & al. (Eds.), Proceedings of the 35th ACM SIGIR'12, pp. 97—106.
The paper proposes a session simulation model grounded in timing of various actions performed on various devices. It identifies time-wise effective session strategies and shows that traditional rank-based evaluation may hide essential factors that affect users' performance and satisfaction - and even give counter-intuitive results. From the abstract (edited): Real information retrieval consists of sessions, where users search by iterating between various cognitive, perceptual and motor subtasks through an interactive interface. The sessions may follow diverse strategies, which, together with the interface characteristics, affect user effort (cost), experience and session effectiveness. The paper proposes an evaluation approach based on simulation scenarios with explicit subtask costs. The limits of effectiveness of diverse interactive searching strategies in two searching environments (the scenarios) under overall cost constraints are analyzed. This is based on a comprehensive simulation of 20 million sessions in each scenario. Furthermore, the paper also contrasts the proposed evaluation approach with the traditional one, rank based evaluation, and shows how the latter may hide essential factors that affect users' performance and satisfaction - and gives even counter-intuitive results.

Kumpulainen, S. & Järvelin, K. *(2010).* Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels. In: Belkin, N. & al. (Eds.), Proc. of the IIiX 2010, pp. 95--104.
The paper challenges the concept of session: a single work task session may require multiple interleaved and multi-query interactions with several search tools and information systems. It also exemplifies a time-consuming way to produce extremely rich data for interaction analysis.
From the abstract and conclusion (edited): The paper examines empirically task-based information access in Molecular Medicine and analyzes task processes as contexts of information access and interaction, integrated use of resources in information access and the limitations of (simple server-side) log analysis in understanding information access, retrieval sessions in particular. We presented a methodology for task based approach to IR and provide results on three levels. Firstly, the work tasks are analyzed in a real work environment and at three complexity levels. Secondly, we show that interaction between different information channels increases proportionally to the complexity increase. Thirdly, we show that, similarly, the queries concentrate more on resource level in routine tasks, but the prominence of factual and topical queries increases in complex tasks. In task-based information access, interaction logging at any single channel (like a search engine) may give a distorted picture of the searcher's needs and intentions. Therefore, the contribution to system development is that it should not be done in isolation as there is considerable interaction between systems in real world use. Significant benefits may be achieved by taking this into account in system design.

Hideo Joho, University of University of Tsukuba, Japan

I would like to describe our current work on search with constraint. My student and I wanted to see how search constraint affects people's information seeking behavior. We tested three types of constraint: Time (e.g., 15 min), Time + Query (e.g., 10 queries to complete a task), and Time + Click (e.g., 20 clicks to complete a task). We found that they had significant impact on people's attention and behavior and medium effect on search performance. This work led us to think how to measure strategic-ness (how strategic a search is) of search sessions. Many tactics are identified, but hard to measure which worked good and which didn't until you see an outcome of search. However, ideally, we would like to say, a search session was very strategic, but somehow didn't work. Also, we're interested in measuring if people's strategic behavior will improve (or increase) over sessions when searching with constraint.

Fujikawa, K., Joho, H., Nakayama, S. (2012) ”Constraint can affect human perception, behaviour, and performance of search”. In: Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries (ICADL 2012), pp. 39-48, Taipei.

Jaap Kamps, University of Amsterdam, The Netherlands

Papers/Challenges:

(1) Evaluation of Interaction
Stephen E. Robertson, Micheline Hancock-Beaulieu, On the Evaluation of IR Systems, Information Processing and Management 28(4): 457-466 (1992).
http://dx.doi.org/10.1016/0306-4573(92)90004-J
I'd like to slip in this paper which was written when "the ideal test collection" didn't materialize and was in a sense regarded as outdated when it appeared because TREC started happening. However, in recent years IR researcher have looked beyond the classic test collections and are actively pursuing the directions of this paper -- a must read for everyone.

(2) Supporting the whole search task
Nicholas J. Belkin, Charles L. A. Clarke, Ning Gao, Jaap Kamps, and Jussi Karlgren. Report on the SIGIR workshop on “entertain me”:
Supporting complex search tasks. SIGIR Forum, 45(2):51-59, December 2011.
http://www.sigir.org/forum/2011D/workshops/2011d_sigirforum_belkin.pdf

Report on a workshop on supporting complex search tasks. Many of our tasks in our professional and daily lives are complex, yet solving them using standard search technology requires us to slice-and-dice our problem into several queries and sub-queries, and laboriously combine the answers post hoc to solve our tasks. Search with task and person context requires a novel mixture of search and recommendation methods, requiring novel retrieval models and evaluation methods (more than topical relevance). Structured querying and semantic annotation (class labels) become crucial cues, but can be hidden from the searcher. Interactive search requires user-centered evaluation or grounded simulations.

(3) Exploratory/faceted search
Anne Schuth, Maarten Marx: Evaluation Methods for Rankings of Facetvalues for Faceted Search. Multilingual and Multimodal Information Access Evaluation - Second International Conference of the Cross-Language Evaluation Forum, CLEF 2011: 131-136, 2011.
http://dx.doi.org/10.1007/978-3-642-23708-9_15

A perhaps lesser known paper related to the Faceted search task at INEX 2011 and 2012 (Data-Centric and Linked Data Tracks). Submission is not a list of results, but a grouping of a ranked list in facets and facet-values. This small twist generates various new challenges, in obtaining suitable search requests and relevance judgments covering various aspects, and in measures that reflect the quality of a choice of facets against various assumptions.

Noriko Kando, National Institute of Informatics (NII), Japan

Key Challenge #1: Evaluate "How user learned through a whole search session?"

Marcia Bates' paper on berry picking model suggests what are happening in the search sessions and that the search results are the entire stuff that each user learned through a whole session rather than the each search result pages. I believe that this classic paper still provides us the important foundation to address the grand challenge of whole-session evaluation. This view suggested the successful search session is not the continuation of "success of search iteration" which can be evaluated by metrics usable for test-collection style evaluation, i.e., during a search session, a query in a session may produce the retrieved results which are not topically relevant to the user at the time of querying, but may help the user to understand the problem space by trial-and-error or inspire a new unexpected aspect of the problem to worth to explore further, in either case such iteration can contribute to the search session success. The whole-session evaluation is more than the accumulation of the evaluation of each of the search iterations in a session. To address to the problem, evaluating the session 1) by outcome of the session, 2) by understanding the process, or 3) by how users' knowledge changed through search sessions. 1) and 3) shall be done as postmortem evaluation, but 2) has various aspects and variations and some of them may be addressed as online evaluation capable during the search session.

Reference:

Bates, M. J. (1989). Design of browsing and berrypicking techniques for online search interfaces. *Online Review, 13*, 407-424.

Key Challenge #2: Identify each session's interactivity or exploratoritivity, and its online identification.

Some of the search sessions are highly interactive or exploratory, and others can be very straightforward simple look-up type search. If we could have think of the evaluation metrics can cover both types of searches, it would be great. Probably for the initial steps, it would be more practical to indentify the types of search session (or how it is interactive) and apply appropriate metrics for the types of search session. To do so, to define the interactivity or exploravity of the search session, and automatically identify it is inevitable. In addition to this, if the retrieval systems can identify the type of search session online during the search, the systems may provide the better support for the users in the sessions. There are many research IR systems provides interesting functionalities to support users' information seeking or exploratory, but in the user tests, the systems may accepted highly when users have information needs highly interactive or exploratory, but the system may evaluated very low when users have very clear information needs can be solved by one-time query-retrieval iteration. Identifying the degree of interactivity of the search sessions is also usable to provide better user support during search session.

Key Challenge #3: How to bridge the "User-centered" or "behavior studies" to "system-oriented evaluation" or "interaction design to support users' information seeking and/or exploratory"

There are many user-centered evaluation studies of interactive information retrieval (IIR) and the user studies to describe or understand how users behave during search sessions. These studies are detailed and insightful to understand how users behave and think during search sessions, but the scale of the studies are generally small and very labour intensive. The studies are described repeatable, but the results are not reproducible and their data are seldom reused to evaluate other IIR systems. In other hands, IR community worked on the test collection which applicable for large-scale evaluation and reusable for testing other systems in other settings. Although the researchers gradually tried to incorporate the search tasks and users' situations in the test collection design, but these are limited. Bridging two different research communities and utilize the results/insight/approach of the user-centered communities to system-oriented large-scale system evaluation of whole session/ user modeling during search session / interaction design to support users' search sessions (especially) in the highly interactive or exploratory search and information seeking is a Key Challenge. System-side Logs are one of the promising resources. Another approach can be cumulating various users' logs and data of the user-centered studies as an Interaction pool (Joho et al. 2007) with users situation/tasks/background and then utilized them in simulated studies. To do so, how to describe the wide variety of interaction dataset obtained in the various user-centered experiments in a framework, and how to implement the reuse need to be investigated further. There must be more approaches addressing this challenge.

Reference:

Joho, H., Villa, R, and Jose, J. M. (2007) "Interaction Pool: Towards a user-centred test collection". In: Proceedings of the Workshop on Web Information Seeking and Interaction, SIGIR 2007, Amsterdam, Netherlands: ACM.

My Experience towards whole-session evaluation #1: Concept-map

Our group has worked on a project called "Cognitive Research on Exploratory Search (CRES)" since 2008 [5]. The initial purpose of the project was to understand users cognitive aspects and behavior during exploratory search sessions and propose search user interfaces which supporting users exploratory and information seeking, and gradually it has shifted the focus to evaluate Interactive information retrieval sessions. The project started from the small-scale in-depth user studies collecting rich data by eye-tracking, screen-capture, user-side logs, think-aloud and depth interview after the search session while showing the screen capture video with eye-gaze marks[2,4], then analyze users behavior and cognitive aspects qualitatively and quantitatively. CRES also developed some coding schemes and data analysis tools. Based on the detailed analysis of the relationship between qualitative and quantitative analysis, the project gradually shift the focus to the automatically applicable analytical methods to larger number of participants for longer search sessions.

As one of the methods proposed, we have used concept maps to analyse the differences of the users' knowledge structure before and after a search session. The main intention is to evaluate how users learned through a whole session. Although the simple analysis of the number of nodes and links in-common, increased, or disappeared after a search session did not show a strong relationship [1], but qualitative analysis of the changes in the topological features between the pre- and post search concept maps showed promising results to capture some aspect of how user learned through a search session [3,6]. The topological features include depth and width of the maps and density of the networks in the maps. We have developed a tool to analyze these aspects automatically and now under the analysis on the larger dataset to reveal the effectiveness of the approach. This is a kind of evaluation based on the output of a search session and a postmortem evaluation which cannot be used for online evaluation during search, but it can be done automatically (i.e. can be used for larger number of participants in the experiments, and may be able to capture some aspect of how user learned through search session quantitatively)

References

[1] Egusa,Y., Takaku,M., Saito, H., Terai, H., Miwa, M., Kando, N (2010) Using a Concept Map to Evaluate Exploratory Search, Proceedings of the Third Symposium on Information Interaction in Context (IIiX 2010); p.175-184. doi:10.1145/1840784.1840810

[2] Miwa, M., Egusa,Y., Saito, H.,Takaku,M., Terai, H., Kando, N (2011) A method to capture information encountering embedded in exploratory Web searches, Information Research; vol.16; no.3; 87.

[3] Saito, H., Egusa,Y., Takaku,M., Miwa, M., Kando, N (2012) Using Concept Map to Evaluate Learning by Searching, In Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci2012)

[4] Saito, H., Takaku,M., Egusa,Y., Terai, H., Miwa, M., Kando, N (2010) Connecting Qualitative and Quantitative Analysis of Web Search Process: Analysis Using Search Units. In Proceedings of Asian Information Retrieval Society 2010 (AAIRS2010): pp. 173-182 (LNCS 6458), doi:10.1007/978-3-642-17187-1_16

[5] Terai, H., Saito, H., Takaku,M., Egusa,Y., Miwa, M., Kando, N (2008) Differences between Informational and Transactional Tasks in Information Seeking on the Web, Proceedings of the Second Symposium on Information Interaction in Context (IIiX 2008); pp.152-159, doi:10.1145/1414694.1414728

[6] Yoshioka,M., Kando,N., Seki, Y. (2011) Evaluation of Interactive Information Access System using Concept Map, In Proceedings of the 4th International Workshop on Evaluating Information Access (EVIA2011) pp. 20-23

My Experience towards whole-session evaluation #2: Visualize users' interaction process

Another approach that we are currently working is visualize the users interaction process during search sessions. One of the approaches is Linkdepth (http://cres.jpn.org/?LinkDepth) and spatial visualization of search history based on the contents of the pages that users view during search sessions. The qualitative analysis of the search process using these tools indicated that the expert users of the type of task often decompound the tasks into subtasks, and tackled each subtasks one by one during exploratory search or investigation. Linkdepth, especially the density of the actions in the timeline and depth of the links the users viewed, seemed to be affected by the relationship between tasks and the physical characteristics of the Web, but also indicated some aspects of the users familiarity for the tasks and whether the users can manage the search session or she or he has lost the way and need some support to complete the task. Such preliminary analysis implies that users sometimes need higher level or meta-level suggestions like search strategies suitable for the type of tasks rather than simple term suggestions. This is the work undergoing, but the approach can be used for online evaluation during the search session as well.

Evangelos Kanoulas, Google, Switzerland

TREC Interactive Track Proceedings.
The TREC interactive proceedings give an overview of the challenges faced during the interactive evaluation experiments conducted over a number of years by NIST. One of the issues described in these proceedings was the general inability of the framework to allow comparisons between retrieval systems from different participants.

"Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments", by Kalervo Järvelin.
The paper describes an effort to simulate user interaction with a retrieval system (in terms of the original query reformulation) that could allow a Cranfield-like evaluation of interactive IR systems. The modeling of users is simple, but simulating a user population could be a solution to controlling the variability due to users that seem to be an issue in interactive experiments.

"Modeling Expected Utility of Multi-session Information Distillation", Yiming Yang and Abhimanyu Lad.
By using a collection of fixed sessions (i.e. a series of queries) as a test collection the paper develops a measure that (a) probabilistically models the user behaviour in browsing the ranked lists over the entire session, and (b) defines the utility of a user processing the returned results incorporating the novelty component. The development of the measure follows the framework of recent measures such as RBP, ERR, EBU, etc. but extends it to whole-session evaluation.

Diane Kelly, University of North Carolina-Chapel Hill, USA

Bates, M. J. (1989). Design of browsing and berrypicking techniques for online search interfaces. *Online Review, 13*, 407-424.
In this classic paper, Marcia Bates presents an alternative view of what happens during the search process. I believe it is important to keep this model in mind when thinking about how to evaluate search sessions, as it has implications for how we interpret what see. For example, if the user enters a string of seemingly unrelated queries is this good or bad? Does this mean the person is not finding what they need or that they are learning and querying different aspects of the topic? The Berrypicking model characterizes search as an evolving process that unfolds as a person interacts with the system and information. Evolution happens both with respect to the queries entered by users and the documents that are useful at any given point in time during the search. This assumption underlying this model is in contrast with that underlying a measure such as session-based DCG where a penalty, or discount is assessed for relevant results that come later queries in a session.

Jarvelin, K., Price, S. L., Delcambre, L.M.L, & Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query IR sessions. *Proceedings of the 30th European Conference on Information Retrieval (ECIR '08),* 4-15.
Even though I just took a dig at snDCG above, I still think this work is important because at least someone tried something! At the time this paper was published (as far as I know) there really weren't any measures that tried to incorporate multiple queries. Although the underlying assumption might not match all search situations, for some types of tasks such as fact-finding, the assumption might be reasonable.

Vakkari, P. (2010). Exploratory searching as conceptual exploration. *Proceedings of the Fourth Human Computer Information Retrieval Workshop*, New Brunswick, NJ, 24-27.
In this paper, Pertti encourages researchers to consider more process-based evaluation measures that look at what happens during the search, and not just at the end. Specifically, he describes several ways that queries from a session might be analyzed to try to discern when learning is taking place. It is the learning itself, Pertti proposes, that should be used as a yardstick rather than the search output. This call-to-action paper closely captures the assumptions underlying the Berrypicking model. In this paper, Pertti also considers measures that take into consideration search stage, and search across multiple sessions, as an important challenges. I agree.

Gary Marchionini, University of North Carolina-Chapel Hill, USA

1. Classic Papers.

Rather than specific papers, two lines of work are of particular pertinence in my mind. First, the work that Barbara Wildemuth and her collaborators did in the mid-late 1990s is of interest. They had medical students conduct searches in a microbiology database at three different intervals over a 9-month period to understand search tactics as their knowledge of microbiology increased. In addition to single order transition state analysis, they also used an interesting strategy that could be useful today: maximal repeating patterns (MRP). There were several publications in AMIA and other venues in late 90s but the most accessible paper is likely Wildemuth's 2004 JASIST paper: The effects of domain knowledge on search tactic formulation.

A more recent line of work of interest is the studies of search sequence by Jacek Gwizdka and his colleagues at Rutgers. Using a search state transition model adapted from my work in the late 1980s, Jacek analyzed and visualized search transitions (from logs and also from eye-movements) with an aim to provide empirical data on search state transition. The 2010 JASIST paper: Distribution of cognitive load in web search reports on the framework and transaction logs and more recent work includes additional data sources such as eye movement and presents novel visualizations (ASIST poster 2011 illustrates different search strategies).

2. Our work today

With support from a NSF grant, Rob Capra and I have been leading efforts to understand search behavior across multiple sessions and in collaboration. This team included Chirag Shah, whose dissertation presented a framework for synchronous collaborative search based on empirical studies. Our work in the past two years has been focused on looking at the effects of different support tools for asynchronous collaborative search over multiple sessions. As we wrap up this project we are revising my original search process model with theoretical transition probabilities (recognize/accept; define problem, select source, formulate query, execute query, examine results, extract info, reflect/stop) for the collaborative web search environment. Combining transaction logs and think aloud data, we are defining different visualizations of search that progresses across time with two or more participants. More importantly, it is clear that some of the states in the model should be divided into substates and at least one additional state specific to collaborative awareness should be added.

Douglas W. Oard, University of Maryland, USA

Whole-Session Evaluation for Spoken Queries

Background. Sessions involving spoken queries can be quite different from those involving written queries for three reasons: (1) the spoken queries may differ, (2) the system's ability to interpret those queries may differ, and (3) the system's response (to which the user reacts with subsequent queries) may differ. These differences introduce additional issues that should be considered in evaluation design. In this brief note, prepared for the *Shonan Workshop on Whole-Session Evaluation of Interactive Information Retrieval Systems*, I adopt as a context the "Spoken Web," which is an example of what has come to be called a "spoken forum" in which speakers of a low-resource language create user-generated spoken content that other users then search for. The goal of the spoken Web is to facilitate information exchange (e.g., advertising of services to a local community) using inexpensive audio-only cell phones. Additional commentary on some of the issues raised in this can be found in [1]. Other potential applications of spoken queries (e.g., Siri-like personal information services, or hands-free driver-controlled in-car Web search) share many of the same issues.

Evaluation Issues.

1. Spoken queries are different. Web queries are short in part because the typed characters can be interpreted without error. Such is not true for spoken queries, particularly when the potential vocabulary is large, the query is spoken in a low-resource language or dialect, or background noise is present in the speaker's environment. It seems reasonable to expect spoken queries to be longer (perhaps continuing until the system determines that it has enough clues to formulate a reasonable response), somewhat structured (e.g., formulated as fluent questions), or both. Observational studies with situated users (perhaps with Wizard-of-Oz simulation of system behavior to generate representative full-session interactions) will be essential if we are to generate realistic queries that drive technology development in directions that we can reasonably expect would ultimately be useful.
2. Audio-only responses are different. Audio is a far more austere channel than the highly interactive, spatially rich, and high-resolution screens that we typically design for in Web search. If the old hit-enter-and-get-ten-blue-links that we worked with for so many years now seems limiting, try doing a Web search for someone by phone to get a sense for the limitations of the audio channel. Then try limiting yourself to saying things you expect a computer could reasonably generate automatically. You'll get the idea. Many of things we take for granted (like read one past the link you will eventually click on) are far more difficult in audio. Of course, with a little thought we can also imagine some novel things we could do in audio (such as altering intonation to indicate the system's confidence in a result).
3. Test collections will likely be different. Mediaeval 2011 and 2012 have started experimentation with a Spoken Web evaluation scenario [2], but to date only with isolated (and short) queries. Drawing on what we have learned from Cranfield, we may initially want to design simplified "canned" interaction scenarios to help us to learn useful things without modeling rich human-system interactions. Ultimately we might extend this by simulating some types of human responses to system-initiative actions. Because few information retrieval research teams currently also work at the state of the art of speech processing, we may want to distribute some alternative representations for spoken queries (e.g., automatically recognized phoneme or word sequences). Because prior context can be exploited to constrain perplexity (and thus improve recognition), intermediate representations for sessions may be different from those for isolated queries.

References.

[1] Douglas W. Oard. Query by Babbling: A Research Agenda. In *First International Workshop on Information and Knowledge Management for Developing Regions (IKM4DR)*, Maui, HI, November, 2012.

[2] Nitendra Rajput and Florian Metze. Spoken Web Search. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, October, 2011.

Jeremy Pickens, Catalyst Repository Systems, USA

Key Challenge #1: Goals

The first and most important challenge that IIR and session-oriented search face is that of understanding the user's goal. Metrics must match user intent. This is not a surprising position to take; indeed it is the topic of the entire first day of the workshop. I simply wish to enumerate what I see as the four broad reasons a user might engage in an interactive search session. (1) The user is attempting a known-item (lookup, navigational) search, but the system has failed and results are poor. PageRank and/or click-popularity do not work, so the system must rely on more session-embedded signals from the individual user. For example, I know someone named Harry Potter and it is difficult to navigate to that person's web presence in a single round given the popularity of another, fictional Harry Potter. A navigational session would allow the user to interactively override the built-in bias of the search engine. (2) The user has an exploratory search need (cf. Marchionini et al). The goal of such sessions include learning, comparison, synthesis, forecasting, and discovery, and not necessarily document count. (3) The user has a recall-oriented search need, such as during an eDiscovery first or second request. All information that is relevant to a particular legal or regulatory matter must be found and produced. However, concepts such as proportionality are applicable, meaning that while 100% recall is ideal, 90% or even 80% recall might also be acceptable depending on cost/effort metrics. (4) The user is engaged in set-oriented search, wherein a single need is satisfied only after an entire set of information has been found. For example, if the user is trying to build a robotic arm and is doing a parts search for all the pieces necessary to construct that arm, anything short of finding every single piece yields an unsatisfied information need. The robotic arm cannot be built with 80% or even 90% of the pieces.

Another example is trip planning. In order to successfully complete a trip, both flight and hotel need to be found and booked. Not one or the other, but both. While set-oriented search might appear to be a special case of recall-oriented search, with recall set to 100%, a subtle but important difference is that at least (and only?) one instance of each component need be found. A dozen different hotels might satisfy that piece of the need, but a dozen do not need to be found. Finding two hotel possibilities and one flight is better than only finding five hotel possibilities, even though the total "relevant" document count in the latter scenario is greater.

Key Challenge #2: Implicit versus Explicit Interactivity

Should Interactive IR metrics be geared toward implicit or explicit sessions? In other words, if users are engaged with a search system, should that system spent its computational inference resources (and metrics to guide those resources) on trying to detect whether or not the user is attempting interactivity (aka implicitly engaged in session-oriented behavior)? Or should session-based metrics not concern themselves with guessing whether or not a session is in progress, and only begin at the point at which the user has explicitly created an interactive search session? One can ask a similar question about goals. Should the metrics be oriented toward how well we can detect a user's (implicit) goal type? Or should the user explicitly declare the goal, allowing the metrics to primarily be oriented toward how well that goal is supported? Stated another way: The definition of "whole session", of being in a session, takes on a slightly different meaning if that session is implicit or explicit. To what extent does this distinction concern us and the metrics we are attempting to develop?

Key Challenge #3: Monotonicity and Session Length

When user-system interaction is limited to a single round, aka traditional ad hoc search, there is a natural tendency (nay, even a fundamental requirement) that the system always provide the best possible answer or result set for each query that is asked of it. It is also assumed that, minor spelling correction and advances in query rewriting aside, the user is asking the best possible question, so results are geared directly to that question. Results are presented in monotinically-descending order of effectiveness, due to the one-shot nature of the interaction. But does session-oriented search carry with it the same expectation (from both the user and the system) of monotonicity? Should queries and responses only get monotonically better at every round? Or should metrics allow for non-monotonicity, e.g. allowing two steps back now in order to take one step forward in the near future? The question is a leading one, as the theme of this Shonan workshop is "Whole Session Evaluation" I assume that most would agree that it is more important to have globally-best results by the end of the session than locally-best results halfway through a session. If so, then this raises the second part of the key challenge, which is determining how long a session does and/or should last. If the user is willing to take two steps back in order to take three steps forward, then implicit therein is a willingness to engage in a five-step session. But what if five steps back are required in order to take seven steps forward? Is the user willing to engage in a twelve-step session? Or if the user is more interested in backward-to-forward step ratio, rather than in the absolute value of the number of steps, he or she might be willing to both take two back to go three forward, and four back to go six forward. The ratio is the same in both cases. But does that extend indefinitely? Is the user willing to take two thousand steps back in order to go three thousand steps forward? Understanding how long a session either does or should last, and when it does or should terminate, is critical in knowing what to measure and how to measure it.

Tetsuya Sakai, Microsoft Research, China

THREE "IMPORTANT" PAPERS

[1] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In Proceedings of ACM SIGIR 2012, pages 95–104, 2012.
This paper is not about session-based evaluation, but is important in that it proposes to pay attention to users' time instead of ranks, and to consider document lengths. Their current formulation of the Time-Based Gain (TBG) is still rather rank-based in that it revolves around "time to reach rank k,"but it can probably be extended. Also, Smucker and Clarke have an interesting sequel (a CIKM'12 poster) in which they simulate different users in a Monte Carlo fashion within their TBG framework.

[2] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In Proceedings of ACM SIGIR 2012, pages 105–114, 2012.
This paper tackles session-based evaluation, and it considers several basic user actions such as query reformulation. Interestingly, it is also time-oriented. Moreover, the authors advise us not to normalise evaluation metrics, so that we can simulate real user experiences. (Interestingly, TBG does not use normalisation either.)

[3] T. Sakai, M. P. Kato, and Y.-I. Song. Click the search button and be happy: Evaluating direct and immediate information access. In Proceedings of ACM CIKM 2011, pages 621–630, 2011.
This paper is not about session-based evaluation either: it proposes a new method for evaluating a summary. But the method assumes that the user's reading speed is constant, and discounts information units based on their positions within the text. Hence this is also a form of time-based gain discounting. (See also the sequel at AIRS'12.) We are now extending this idea of position-based discounting to seamlessly handle summaries, ranked retrieval (including diversified search), nonlinear traversal (see below) and multi-query sessions.

MY EXPERIENCE IN SESSION-BASED EVALUATION

My colleague Zhicheng Dou and I are now conducting evaluation experiments based on "trailtexts," which represent (concatenations of) texts read by the user. A trailtext could be a summary, a sequence of snippets and documents, or arbitrary fragments of text collected via (say) eyetracking. Potentially, it can even be used for evaluating nonlinear traversal (user reading a document at rank k and then one at rank j(<k)), and multi-query sessions. For conducting click-based session evaluation experiments, we have recently sampled a one-day session data from Bing, under the constraint that every query within a session receives at least one click. From the 19,214,623 sessions thus obtained, we extracted 53,242 (0.277%) "truncated" sessions (i.e. sessions from which all interactions after the first query reformulation have been removed) to conduct nonlinear traversal evaluation. (FYI: 1.439% of the sessions contained at least one nonlinear traversal somewhere in the session.) We have separately extracted 5,610,742 (29.200%) multi-query sessions: some of these sessions contain many queries. We are also in the process of establishing discounting functions over trailtexts, based on the percentage of users who are willing to read at least x characters in total within a session. Potentially, the discounting function can also be designed per user and/or per search task. My hope is that this new evaluation framework will help us evaluate and compare different textual information access modes (e.g. a direct answer vs. an interactive information gathering). While this project currently derives trailtexts based on document relevance assessments or clicks, I believe that more appropriate evaluation will be made possible if we establish methods for clearly defining and deriving information units, as the NTCIR-10 Once Click Access task organisers are currently exploring.
Bottom line: think beyond document ranks and document relevance.

Mark Sanderson, RMIT University, Australia

Hi, I don't have a set of papers to describe rather what I think are the key challenges of session based evaluation. I come from the school of test collection based evaluation. Test collections have for a long time been simple things: documents, topics, qrels, and an evaluation measure. As we all know, there are many evaluation measures available.

The measure simulates the satisfaction a user would have from seeing the document retrieved by a search engine. One of the amazing things about IR research is that it wasn't number until the last few years that anyone bothered to check if the evaluation measures were providing a good simulation of user preferences or satisfaction. It would appear from the work that's been done so far that some measures are clearly better than others.

There isn't yet (to the best of my knowledge) equivalent work for session-based evaluation. Although we have session based test collections, we don't really have much in the way of validation that the way we are evaluating session based searching system is actually an accurate simulation of what users want from a session based search.

Do they want the search to find them lots of different documents? Do they want just one retrieved item? Will they be annoyed to see duplicate documents from earlier searches or will they prefer them? Simple (almost trivial sounding) questions, but it is important that we have a way of answering such question in order to have an accurate evaluation measure that simulates user satisfaction when searching over a session.

Arjen P. de Vries, Centrum Wiskunde & Informatica, The Netherlands

At CWI and Delft we have a few experiences with session-based log file analysis. I try to relate these to two different challenges for both evaluation and system design.

1. Real-life sessions are constructed dynamically, on-the-fly
We have participated in the TREC session track, with the idea that more complex models of the user behaviour should be useful in such a setting. One thing to highlight is that careful analysis of queries in sessions is difficult, because the ground for any observation can vary (and these may thus not be immediately comparable); not just with session length, but also with respect to where in the session we look and how much history is taken into account to what extent. Take for example the following figure (from the SIR workshop paper at ECIR 2011):

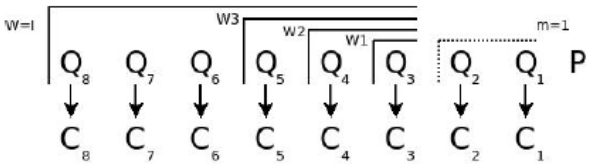


Fig. 1. Descriptive parameters of a search session: length $L = 8$, current observation at step $l = 6$, gap to a purchase $m = 1$ and history window $W = 1, 2, 3, max$.

The main objective of our TREC session work has been that observing behaviour during the session may inform a retrieval system about the user's (expected) search performance, and adapt query expansion techniques accordingly. While we had a few results that indicate we could improve beyond "just" using the last query, applying this in the real life setting would still be harder: a real system has to take a decision about query expansion or not, without knowing whether this is the last query that would be issued.
A recent study (published at IIiX) faces the same issue. We aimed to develop a tool for teachers such that they can measure, in a classroom setting, how well their pupils are completing search tasks (so the teacher would know who needs assistance most, and who are doing fine by their own). We classified search behaviour observed (features derived from the sessions) into children's search roles proposed previously by Alison Druin, and also into binary search success. In this study, we only looked at the classification of full sessions into roles – which is interesting, but again not very useful in practice yet; we need to be able to classify "on-the-go".
As far as I know, the temporal aspect of a session in progress has not been studied in full detail yet – intuitively, there seems to be an opportunity to view the classification of a user's expected search success as an uncertain one, where more evidence makes the measurement more reliable (i.e., a longer session should make us more confident in the classifier's decision). A few proposals exist to model such confidence explicitly for the relevance decision (e.g., the portfolio style models proposed by Jun Wang), but maybe the uncertainty in measurement has to play a more important role in IR.

2. Analysis from logs vs. interactive experiments
We have carried out a few studies where we analysed search logs – especially in an image retrieval setting, where a EU project gave us access to the logs of photo journalists' sessions with-in the Belga image portal (Belga is a news agency). (As an aside, we used LOD data to reduce the sparseness in observed events in the log and help characterize behaviour in higher level patterns than plain query terms; this study has been published as a JASIST paper.) When analyzing, for each Belga query issued, the corresponding result lists by a measure of "coherence", we found that query modifications in session may not behave the way that people usually claim. Basically, the "standard" interpretation (that query term additions correspond to specifications and query term removals to generalisations, accepted

as “natural” in virtually any previous work) does not seem to hold, as users may also issue extra terms with the aim to remove certain interpretations from the top ranked results. We demonstrated the same finding in Bing logs, and published the results at ECIR 2012.However, as we were only looking at logs and interpreting the recorded behaviour, we could not validate that observation directly – it will never be more than just our interpretation of the events observed.
We have made initial attempts to setup an experiment to find complementary evidence in interactive settings, but this turned out harder than expected. Not only does it take quite a lot of time and effort to collect sufficient data (to, e.g., observe sufficient query term removals); also, our own interactive system (that we can control) is inferior to commercial web search engines, and our users notice – and, we have not yet tackled the question how to ask the user what they aimed to do with a modification, without interfering too much with the search.
While these issues are partially explained by the fact that we have insufficient experience with interactive IR experiments in our research team, I do not think that all methodological questions have been resolved elsewhere yet: i.e., how can we validate findings from search logs (which are merely hypotheses) in accompanying follow-up interactive IR experiments?

References:

C. Boscarino, Arjen P. de Vries, V. Hollink, Jacco van Ossenbruggen. Implicit relevance feedback from a multi-step search process: a use of query-logs. *Proceedings of ECIR 2011 Workshop on Information Retrieval Over Query Sessions 2011, Dublin, Ireland*, 2011.

Carsten Eickhoff, Pieter Dekker and Arjen P. de Vries. Supporting Children’s Web Search in School Environments. In *Proceedings of the 4th Conference on Information Interaction in Context (IIiX), Nijmegen, The Netherlands,* 2012

Vera Hollink, Jiyin He, Arjen P. de Vries: Explaining Query Modifications - An Alternative Interpretation of Term Addition and Removal. ECIR 2012: 1-12

Vera Hollink, Theodora Tsikrika, Arjen P. de Vries: Semantic search log analysis: A method and a study on professional image search. JASIST 62(4): 691-713 (2011)

Max Wilson, University of Nottingham, UK

Evaluating depth of learning and sensemaking by analyzing and comparing pre- and post-session written reports.

We often design systems to help people learn, investigate, make sense of, or comprehend information they have found during search. We have many measures of performance and accuracy derived from specific tasks, but it is inherently difficult to evaluate how much a person has learned during a search session. Typical methods involve prescribing what can be learned and evaluating it in a quiz, or by analyzing written reports simply by the breadth and depth of sub-topics described (e.g. Kamerrer et al, 2009). These forms of analysis, that focus on content rather than understanding, can be limited, for example, by participants naively including simplistic facts.

In Wilson & Wilson (2012-ish), we reported on the development of a set of scales derived from Bloom and Engelhart's taxonomy of learning (Bloom and Engelhart, 1956), or rather the revision by Anderson et al (2000) in Figure 1. The three scales are used to determine how deep, according to Bloom's taxonomy levels, the learning is that the participant has achieved. Evaluators read openly written summaries of topics before and after search tasks, and allow them to rate them for their inclusion of: understanding, analysis, and evaluation. The approach has been shown to be more robust to written-summary size, and can be applied to people with both high and low prior knowledge levels.

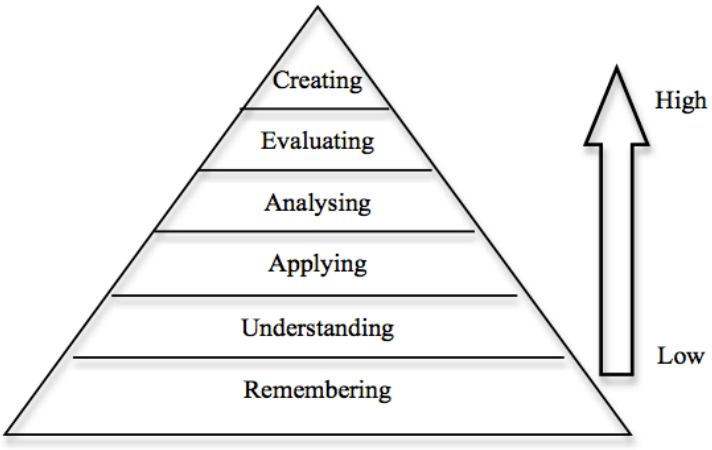


Figure 1: Anderson and Krathwohl's revision of Bloom's Taxonomy of Learning.

Kammerer, Y., Nairn, R., Pirolli, P., & Chi, E. H. (2009). *Signpost from the masses: learning effects in an exploratory social tag search browser.* Paper presented at the Proceedings of the 27th international conference on Human factors in computing systems (CHI'09), Boston, MA, USA.

Wilson, M. J. and Wilson, M. L. (2012) A Comparison of Techniques for Measuring Sensemaking and Learning within Participant-Generated Summaries. In: *Journal of the American Society for Information Science and Technology*, (accepted).

Bloom, B. S., & Engelhart, M. D. (1956). *Taxonomy of educational objectives : the classification of educational goals. Handbook I, Cognitive domain*. London: Longmans.

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., . . . Wittrock, M. (2000). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Version*: Allyn & Bacon.

Collated References (w/ links)

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., . . . Wittrock, M. (2000). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Version*: Allyn & Bacon.

Azzopardi, L. (2009), Usage Based Effectiveness Measures, In Proceedings of 18th ACM CIKM, p631-640.

Azzopardi, L.:(2011). The economics in interactive information retrieval. In: Baeza-Yates, R. & al. (Eds.) Proceedings of the ACM SIGIR'11, pp. 15--24.

Baader, F. Lutz, C., Milicic, M., Sattler, U.and Wolter, F (2005) "Description Logic Based Approach to Reasoning about Web Services", WWW 2005,

Baskaya, F., Keskustalo, H. And Järvelin, K. (2012). Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2012) pp.105-114 **[x 2]**

Bates, M. J. (1989). Design of browsing and berrypicking techniques for online search interfaces. *Online Review, 13*, 407-424. **[x 2]**

Belkin,N.J., Clarke, C.L.A., Gao, N., Kamps, J., Karlgren, J. (2011) Report on the SIGIR workshop on "entertain me": Supporting complex search tasks. SIGIR Forum, 45(2):51-59

Belkin, N.J. (2010) On the evaluation of interactive information retrieval systems. In: B. Larsen, J.W. Schneider & F. Åström (Eds.) The Janus Faced Scholar. A Festschrift in Honour of Peter Ingwersen (pp. 13-21). Copenhagen: Royal School of Library and Information Science.

Bennett, P.N. , White, R.W., Chu, W., Dumais, S.T., Bailey, P. , Borisyuk, F. and X. Cui (2012). Modeling the Impact of Short- and Long-Term Behavior on Search Personalization. In *Proceedings of SIGIR '12*. 2012.

Bloom, B. S. and Engelhart, M. D. (1956). *Taxonomy of educational objectives : the classification of educational goals. Handbook I, Cognitive domain*. London: Longmans.

Bookstein, A. (1982) Information Retrieval: A Sequential Learning Process, Journal of the Amercian Soceity for Information Science, 34(5):331-341.

Borlund, P. (2003) "IIR evaluation model: a framework for evaluation of interactive information retrieval systems", Information Research, Vol. 8 No. 3, April 2003

Boscarino, C. et al (2012) "Adapting Query Expansion to Search Proficiency"

Boscarino, C., de Vries, A. P., Hollink, V. and van Ossenbruggen, J .(2011) Implicit relevance feedback from a multi-step search process: a use of query-logs. *Proceedings of ECIR 2011 Workshop on Information Retrieval Over Query Sessions 2011, Dublin, Ireland*, 2011.

Cole,M., Liu,J., Belkin,N.J., Bierig,R., Gwizdka,J., Liu,C., Zhang,J. and Zhang, X. (2009) Usefulness as the criterion for evaluation of interactive information retrieval. In: Proceedings of the Third Human Computer Information Retrieval Workshop, Washington, DC.

Crestani, F., Ruthven, I., Sanderson, M. and van Rijsbergen, C. J.(1995) "The Troubles with Using a Logical Model of IR on a Large Collection of Documents" In: Proceedings of the Fourth Text Retrieval Conference (TREC-4), 1-3 Nov 1995, Maryland, USA.

Downey, D. Dumais,S., Liebling, D. and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. In *Proceedings of '08.* 2008

Eickhoff, C., Dekker , P. and de Vries, A.P. Supporting Children's Web Search in School Environments. In *Proceedings of the 4th Conference on Information Interaction in Context (IIiX), Nijmegen, The Netherlands,* 2012

Egusa,Y., Takaku,M., Saito, H., Terai, H., Miwa, M. and Kando, N (2010) Using a Concept Map to Evaluate Exploratory Search, Proceedings of the Third Symposium on Information Interaction in Context (IIiX 2010); p.175-184.

Fox, S. Karnawat, K., Mydland, M., Dumais, S. and White, T. (2005). Evaluating implicit measures to improve the search experience. *ACM:TOIS, 23(2),* 147-168.

Fuhr, N. (2008) A Probability Ranking Principle for Interactive Information Retrieval. *Information Retrieval 11(3).*

Fujikawa, K., Joho, H. and Nakayama, S. (2012) "Constraint can affect human perception, behaviour, and performance of search". In: Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries (ICADL 2012), pp. 39-48, Taipei.

Gwizdka, J. (2010) Distribution of cognitive load in web search,Journal of the American Society for Information Science and Technology, Volume 61, Issue 11, pages 2167–2187,

Halpern, J. Y. (1995) "Reasoning about Knowledge: a Survey", Handbook of Logic in Aritifical Intelligence and Logic Programming, Vol. 4 (Eds. By D. Gabbay, C.J. Hogger and J.A. Robinson) pp. 1-34

Halpern, J.Y. and Tuttle, M.R. (1993) "Knowledge, Probability, and Adversaries" Journal of the ACM, Volume 40 Issue 4, Sept. 1993, pp. 917 - 960

Hollink, V., He, J. and de Varies, A.P. (2012) Explaining Query Modifications - An Alternative Interpretation of Term Addition and Removal. ECIR 2012: 1-12

Hollink, V., Tsikrika, T. and de Varies, A.P. (2011) Semantic search log analysis: A method and a study on professional image search. JASIST 62(4): 691-713 (2011)

Järvelin,K. (2009) "Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments". Proceedings of the 18th ACM conference on Information and knowledge management (CIKM 2009) pp. 2053-2056

Järvelin, K., Price, S. L., Delcambre, L.M.L and Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query IR sessions. *Proceedings of the 30th European Conference on Information Retrieval (ECIR '08),* 4-15.

Joho, H., Villa, R, and Jose J. M. (2007) "Interaction Pool: Towards a user-centred test collection". In: Proceedings of the Workshop on Web Information Seeking and Interaction, SIGIR 2007, Amsterdam, Netherlands: ACM.

Jones, R. and Klinkner, K. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs, CIKM 2008.

Kammerer, Y., Nairn, R., Pirolli, P. and Chi, E. H. (2009). *Signpost from the masses: learning effects in an exploratory social tag search browser.* Paper presented at the Proceedings of the 27th international conference on Human factors in computing systems (CHI'09), Boston, MA, USA.

Kanoulas, E., Carterette,B., Hall,M., Clough, P. and Sanderson, M. (2011) Overview of the TREC 2011 Session Track. In *Proceedings of TREC '11.* 2011. **[x 2]**

Kelly, D. (2009) "Methods for evaluating Interactive Information Retreival Systems with Users", oundations and Trends in Information Retrieval, Vol. 3, Nos. 1–2 (2009) 1–224

Kelly, D., Dumais, S., and Perderson, J. O. (2009) Evaluation Challenges and Directions for Information-Seeking Support Systems, In Computer, IEEE, p44-50.

Kooi,B.P. (2003) "Probabilistic Dynamic Epistemic Logic", Journal of Logic, Language and Information 12: 381-408.

Kotov, A.,Paul, B.N., Ryen W., Dumais, S. and Teevan, J. (2011) Modeling and Analysis of Cross-Session Search Tasks. In *Proceedings of SIGIR '11*. 2011. **[x 2]**

Kumpulainen, S. and Järvelin, K. *(2010).* Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels. In: Belkin, N. & al. (Eds.), Proc. of the IIiX 2010, pp. 95--104.

Lindley, S., Meek, S., Sellen, A. and Harper, R. (2012) It's Simply Integral to What I Do: Enquiries into how the Web is Weaved into Everyday Life, WWW 2012.

Miwa, M., Egusa,Y., Saito, H.,Takaku,M., Terai, H. and Kando, N (2011) A method to capture information encountering embedded in exploratory Web searches, Information Research; vol.16; no.3; 87.

Robertson,S.E. and Hancock-Beaulieu, M. (1992) On the Evaluation of IR Systems, Information Processing and Management 28(4): 457-466 (1992).

Saito, H., Egusa,Y., Takaku,M., Miwa, M. and Kando, N (2012) Using Concept Map to Evaluate Learning by Searching, In Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci2012)

Saito, H., Takaku,M., Egusa,Y., Terai, H., Miwa, M., Kando, N (2010) Connecting Qualitative and Quantitative Analysis of Web Search Process: Analysis Using Search Units. In Proceedings of Asian Information Retrieval Society 2010 (AAIRS2010): pp. 173-182 (LNCS 6458)

Sakai,T., Kato,M.P. and Song, Y. –I. (2011) Click the search button and be happy: Evaluating direct and immediate information access. Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM2011) pp. 621-630

Schuth, A. and Marx, M. (2011) Evaluation Methods for Rankings of Facetvalues for Faceted Search. Multilingual and Multimodal Information Access Evaluation - Second International Conference of the Cross-Language Evaluation Forum, CLEF 2011: 131-136, 2011.

Smucker, M.D.; Clarke, C.L.A. (2012) Time-based calibration of effectiveness measures. In Proc SIGIR 2012, pp. 95-104. **[x 4]**

Tague-Sutcliffe, J. (1992) Measuring the Informativeness of a Retrieval Process, In the Proceedings of the 15th ACM SIGIR. p23-36.

ten Cate, B. and Shan, C.C. (2002) "Question Answering: from Partitions to Prolog", Automated Reasoning with Analytic Tableaux and Related Methods (LNCS 2381), pp 251-265

Terai, H., Saito, H., Takaku,M., Egusa,Y., Miwa, M. and Kando, N (2008) Differences between Informational and Transactional Tasks in Information Seeking on the Web, Proceedings of the Second Symposium on Information Interaction in Context (IIiX 2008); pp.152-159

Tran, V.T. and Fuhr, N. (2012) Using Eye-Tracking with Dynamic Areas of Interest for Analyzing Interactive Information Retrieval. In Proc SIGIR 2012, pp. 1165-1166.

Vakkari, P. (2010) Exploratory searching as conceptual exploration. *Proceedings of the Fourth Human Computer Information Retrieval Workshop*, New Brunswick, NJ, 24-27. **[x 2]**

Wildemuth, B. (2004) The effects of domain knowledge on search tactic formulation. Journal of the American Society for Information Science and Technology, Volume 55, Issue 3, pages 246–258

Wilson, M. J. and Wilson, M. L. (2012) A Comparison of Techniques for Measuring Sensemaking and Learning within Participant-Generated Summaries. In: *Journal of the American Society for Information Science and Technology*, (accepted).

Yang, Y. and Lad, A. (2009) "Modeling Expected Utility of Multi-session Information Distillation". In Proceedings of ICTIR 2009 (LNCS 5766) pp. 164-175.