Report from Dagstuhl Seminar 13441

Evaluation Methodologies in Information Retrieval

Edited by

Maristella Agosti[1], Norbert Fuhr[2], Elaine Toms[3], and Pertti Vakkari[4]

1 **University of Padova, IT,** agosti@dei.unipd.it
2 **Universität Duisburg-Essen, DE,** norbert.fuhr@uni-due.de
3 **Sheffield University, GB,** e.toms@sheffield.ac.uk
4 **University of Tampere, FI,** pertti.vakkari@uta.fi

Abstract

This report documents the program and the outcome of Dagstuhl Seminar 13441 "Evaluation Methodologies in Information Retrieval", which brought together 42 participants from 11 countries. The seminar was motivated by the fact that today's information retrieval (IR) applications can hardly be evaluated based on the classic test collection paradigm, thus there is a need for new evaluation approaches. The event started with five introductory talks on evaluation frameworks, user modeling for evaluation, evaluation criteria, measures, evaluation methodology, and new trends in IR evaluation. The seminar participants then formed working groups addressing specific aspects of IR evaluation, such as reliability and validity, task-based IR, learning as search outcome, searching for fun, IR and social media, graph search, domain-specific IR, interaction measures and models, and searcher-aware information access systems.

Seminar 27. October – 01. November, 2013 – www.dagstuhl.de/13441
1998 ACM Subject Classification H.3.4 Systems and Software: Performance evaluation (efficiency and effectiveness), H.5.2 User Interfaces: Benchmarking, Evaluation/methodology
Keywords and phrases evaluation, testbeds, user studies, measures
Digital Object Identifier 10.4230/DagRep.3.10.92
Edited in cooperation with Vu Tran

1 Executive Summary

Maristella Agosti
Norbert Fuhr
Elaine Toms
Pertti Vakkari

Evaluation of information retrieval (IR) systems has a long tradition. However, the test-collection based evaluation paradigm is of limited value for assessing today's IR applications, since it fails to address major aspects of the IR process. Thus there is a need for new evaluation approaches, which was the focus of this seminar.

Before the event, each participant was asked to identify one to five crucial issues in IR evaluation methodology. Pertti Vakkari presented a summary of this homework, pointing out that there are five major themes deemed relevant by the participants: 1) Evaluation frameworks, 2) Whole session evaluation and evaluation over sessions, 3) Evaluation criteria: from relevance to utility, 4) User modeling, and 5) Methodology and metrics.

Evaluation Methodologies in Information Retrieval, *Dagstuhl Reports*, Vol. 3, Issue 10, pp. 92–126
Editors: Maristella Agosti, Norbert Fuhr, Elaine Toms, and Pertti Vakkari
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Based on the evaluation model proposed in Saracevic & Covi [1], the seminar started with four introductory talks covering major areas of IR evaluation: Nick Belkin gave a survey over "Framework(s) for Evaluation (of whole-session) IR", addressing the system components to be evaluated and the context to be considered. In his presentation "Modeling User Behavior for Information Retrieval Evaluation", Charlie Clarke described efforts for improving system-oriented evaluation through explicit models of user behavior. Kal Järvelin talked about "Criteria in User-oriented Information Retrieval Evaluation", characterizing them as different types of experimental variables and distinguishing between output- and (task-)outcome related criteria. "Evaluation Measures in Information Retrieval" by Norbert Fuhr outlined the steps necessary for defining a new metric and the underlying assumptions, calling for empiric foundation and theoretic soundness. Diane Kelly presented problematic issues related to "Methodology in IR Evaluation", such as the relationship between observation variables and criteria, the design of questionnaires, the difference between explanatory and predictive research and the appropriateness of statistical methods when dealing with big data. The round of introductory talks was concluded with Maristella Agosti's presentation "Future in Information Retrieval Evaluation", where she summarized challenges identified in three recent workshops in this area.

For the rest of the week, the participants then formed working groups described in the following.

"From Searching to Learning" focused on the learning as search outcome and the need for systems supporting this process. Learning may occur at two different levels, namely the content level and the search competence level. There is a need for understanding of the learning process, its relationship to the searcher's work task, the role of the system, and the development of appropriate evaluation methods. Approaches may address different aspects of the problem, such as the system, the interaction, the content, the user and the process. For evaluation, the framework from Ingwersen and Jarvelin [2] suggests criteria and measures at the levels of information retrieval, information seeking, the work task and the social-organizational and culture level.

"Social Media" allow users to create and share content, with a strong focus on personal connections. While web search engines are still the primary starting point for many information seeking activities, information access activities are shifting to more personalized services taking into account social data. This trend leads to new IR-related research issues, such as e.g. utility, privacy, the influence of diverse cultural backgrounds, data quality, authority, content ownership, and social recommendations. Traditional assumptions about information seeking will have to be revised, especially since social media may play a role in a broad range of information spaces, ranging form everyday life and popular culture to professional environments like journalism and research literature.

"Graph Search and Beyond" starts from the observation that an increasing amount of information on the Web is structured in terms of entities and relationships, thus forming a graph, which, in turn allows for answering more complex information needs. For handling these, search engines should support incremental structured query input and dynamic structured result set exploration, Thus, in contrast to the classical search engine result page, graph search calls for an incremental query exploration page, where entries represent the answers themselves (in the form of entities, relationships and sub-graphs). The new possibilities of querying and result presentation call for the development of adequate evaluation methods

"Reliability and Validity" is considered as the most central issue in IR evaluation, especially in the current situation where there is increasing discussion in the research community about reproducibility and generalizability of experimental results. Thus, this working group decided to start the preparation of a book on best practices in IR evaluation, which will cover the following aspects: Basic definitions and concepts, reliability and validity in experimentation, reporting out experiments, failure analysis, definition of new measures and methods, guidelines for reviewing experimental papers.

"Domain Specific Information Retrieval" in specific domains like e.g. in cultural heritage, patents and medical collections is not only characterized through the specifics of the content, but also through the typical context(s) in which this information is accessed and used, which requires specific functionalities that go beyond the simple search interaction. Also, context often plays an important role, and thus should be considered by the information system. However, there is a lack of appropriate evaluation methods for considering contexts and new functions.

"Task-Based IR" typically refers to research focusing on the task or goal motivating a person to invoke an IR system, thus calling for systems being able to recognize the nature of the task and to support the accompanying search process. As task types, we can distinguish between motivating tasks, seeking tasks, and search tasks. Task-based IR approaches should be able to model people as well as the process, and to distinguish between the (task-related) outcome and the (system) output.

"Searching for Fun" refers to the interaction with an information system without a specific search objective, like e.g. online window shopping, watching pictures or movies, or reading online. This type of activity requires different evaluation criteria, e.g. with regard to stopping behavior, dwell time and novelty. Also, it is important to distinguish between system criteria and user criteria, where the latter may be subdivided into process criteria and outcome criteria. A major problem in this area is the design of user studies, especially since the starting points (e.g. casual or leisure needs) are difficult to create under experimental conditions. A number of further issues was also identified.

The working group "The Significance of Search, Support for Complex Tasks, and Searcher-aware Information Access Systems" addressed three loosely related challenges. The first topic addresses the definition of IR in the light of the dramatic changes during the last two decades, and the limited impact of our research. The second topic is the development of tools supporting more complex tasks, and their evaluation. Finally, information systems should become more informed about the searcher and the progress in user's task.

"Interaction, Measures and Models" discussed the need for a common framework for user interaction models and associated evaluation measures, especially as a means for achieving a higher degree of reliability in interactive IR experiments. This would allow for evaluating the effect of the interaction and the interface on performance. A possible solution could consist of three components, namely an interaction model, a gain model and a cost model.

Finally, many of the attendees were planning to continue to collaborate on the topics addressed during the seminar since the fruitful discussions were a useful base for future cooperation.

References

1 Tefko Saracevic, Lisa Covi (2000). Challenges for digital library evaluation. In D. H. Kraft (Ed.), Knowledge Innovations: Celebrating Our Heritage, Designing Our Future. *Proceedings of the 63rd Annual Meeting of the American Society for Information Science. Washington, D.C.: American Society for Information Science.* pp. 341–350.

2 Peter Ingwersen, Kalervo Järvelin (2005). The Turn: Integration of Information Seeking and Retrieval. In *Context. Dortrecht, NL: Springer.* ISBN 1-4020-3850-X

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 A Summary of Homework

Pertti Vakkari (University of Tampere, FI)

The major themes in the issues of IR evaluation methodology are presented. They include 1) Evaluation frameworks, 2) Whole session evaluation and evaluation over sessions, 3) Evaluation criteria: from relevance to utility, 4) User modeling, and 5) Methodology and metrics.

3.2 Framework(s) for Evaluation (of whole – session) IR

Nicholas J. Belkin (Rutgers University – New Brunswick, US)

Main reference T. Saracevic,L. Covi, "Challenges for digital library evaluation," in D. H. Kraft (Ed.), Knowledge Innovations: Celebrating Our Heritage, Designing Our Future – Proc. of the 63rd Annual Meeting of the American Society for Information Science, pp. 341–350, American Society for Information Science, 2000.

This presentation uses the structure proposed by Saracevic and Covi (ti00ti) to discuss the constructs and contexts that could specify framework(s) for evaluation of interactive information retrieval (IR). These constructs and contexts are considered from the point of view of the following overall goal for IR systems in general: The goal of (IR) systems is to support people in resolution of the tasks or goals that led them to engage in information seeking in an IR system, through effective interaction with information objects. I propose the following.

An IR system consists of:

- An information resource
- Methods for organizing and representing IOs
- People who have "information problems"
- Methods for representing information problems
- Methods for retrieving and presenting IOs in response to information problems
- Methods for supporting interaction of the people and the other components of the IR system

Most of these elements of the IR system should be evaluated, although some to a greater extent than others. Evaluating information resources is a crucial issue for Web search engines. Evaluating methods for organizing and representing IOs is a classic IR issue. Evaluating people is clearly not our problem, but understanding them is. Methods for evaluating representation of information problems has been ignored, but is increasingly realized as being important. Methods for retrieving and presenting IOs is a classic IR issue, although ideas of presentation have been rather limited.

Saracevic and Covi suggest that systems can be evaluated at Social, Institutional, Individual, Interface, Engineering, Processing and Content levels. It seems likely that the Social level is probably not relevant to IR evaluation; the Institutional level is peripheral; the Individual level is crucial; the Interface level is crucial; the Engineering level is also central;

the Processing level is crucial, but focuses on two aspects: do algorithms work as intended, and do algorithms do what is intended; Content level needs to be evaluated.

If we construe an information seeking episode as a sequence of different kinds of interactions of the information seeker with information objects, with the various IR techniques (i.e. Methods as above) adapting to support the different interactions, then evaluation of their support can be tailored to the goals of each of the different kinds of interaction.

3.3 Modeling User Behavior for Information Retrieval Evaluation

Charles Clarke (University of Waterloo, CA)

Information retrieval systems may be evaluated through user oriented studies or system-oriented tests. User-oriented studies are based on actual user behavior, including laboratory experiments, A/B testing, and the analysis of interaction logs. Unfortunately, these studies can be expensive, requiring substantial time, money, and data. System-oriented tests, often called batch-style or "Cranfield-style" tests, provide a lost-cost and repeatable alternative. Unfortunately, these tests may be criticized for lacking a clear connection with actual user behavior and preferences, and for reporting results in meaningless units.

This presentation describes various efforts to improve system-oriented testing through the addition of explicit models of user behavior. As a specific example, we examine time-biased gain (TBG). TBG provides a unifying framework for information retrieval evaluation, generalizing 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, TBG 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, TBG is reported in meaningful units, such as the total number of relevant documents seen by the user. TBG also provides a method for incorporating user variance into system-oriented tests. The modeling of user variance is critical to understanding the impact of effectiveness differences on the actual user experience. If the variance of a difference is high, the effect on user experience will be low. By incorporating per-query variance, TBG allows for the measurement of the effect size of differences, which allows researchers to understand the extent to which predicted performance improvements matter to real users. The development of TBG is joint work with Mark Smucker appearing in SIGIR 2012, CIKM 2012 and HCIR 2012.

In addition, the presentation provides an overview of a SIGIR 2013 workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013). The workshop brought together researchers interested in improving Cranfield-style evaluation of information retrieval through the modeling of user behavior. After two invited talks and ten short paper presentations, the workshop participants brainstormed research questions of interest and formed breakout groups to explore these questions in greater depth. This presentation summarizes some of the important questions raised by the workshop and briefly outlines some resulting research directions for the improvement of information retrieval evaluation. The organization of the workshop was a joint effort with Luanne Freund, Mark Smucker, and Emine Yilmaz.

3.4 Criteria in User-oriented Information Retrieval Evaluation

Kalervo Järvelin (University of Tampere, FI)

In the presentation, criteria are the dimensions of evaluation. The presentation discusses research designs where the criteria are used as dependent, independent and controlled variables.

The presentation first discussed nested evaluation frameworks for Information Retrieval (IR) – from a specific IR context to increasingly contexts including the information seeking context, the work task context, and the socio-organizational context. Exemplary evaluation criteria were given for each. It was stressed that if the broader contexts are neglected, there is the risk of sub-optimization in IR system development. It was also pointed out that evaluation and theory development in IR go hand in hand because, on the one hand, evaluation requires a model (a theory) of the system being evaluated and some goal to be achieved and, on the other hand, theory grows in instrumental research (like IR) through evaluation. Evaluation requires experimental designs where the evaluation criteria are used as dependent, independent and controlled variables. This allows evaluation / theory development where the effects of some independent variables are tested on one or more dependent variables.

In a nested evaluation framework, the dependent variables (of a narrow framework) may indirectly affect some dependent variable (of a broader framework) that remains outside the evaluation design. The presentation discussed some experimental evaluation designs. It was underlined, that typically in information retrieval research, especially in test-collection based evaluation studies, the evaluation design is specified such that the dependent variable is the search engine's ranking effectiveness (measured through some metric), and the independent variables consist of document representation and topic methods, and of matching methods for comparing the former. However, the ultimate dependent variable is effective information interaction, and it is often believed that the latter is positively correlated with the former. Alternative experimental designs seek to identify context, searcher and system criteria affecting information access and ultimately effective information interaction. The controlled variables may contain some context variables, searcher variables and/or system variables. The independent variables may belong to the same categories. Ingwersen and Järvelin (2005, Chapter 7) discuss these categories of variables.

Technology alone is insufficient in explaining the effectiveness of interactive IR. In order to develop the technology sensibly we need to understand how technology together with users-in-context produces the desired outcomes in information access and the ultimate benefits (Järvelin 2013). Failing to take context and searchers into account in many study designs may be one reason for the views that IR is not a very theoretical field but rather pragmatic. However, it is exactly more theory that is required to manage the complexity of interactions in IR. Experiments for theory building may be based on test collections and simulation, real-user experiments in test collections, or operational systems evaluations. A study design used in Baskaya & al. (2013) was discussed.

In general, user studies are useful for IR systems development when (1) they inform design or (2) guide design. The former may be based on deliberately incorporating systems variables in the study designs. The latter may be based on identifying user or interaction variables that contribute to the dependent variables AND that may be affected by (future) systems variables. However, user studies may also be useful when there is no instrumental (system design) interest. This happens when they focus research on fruitful areas, or help us understand information interaction – in order to later support it.

The ideas presented above are discussed at more depth in the contributions cited below.

References

1 Ingwersen, P. & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval. In *Context. Dortrecht, NL: Springer.* ISBN 1-4020-3850-X

2 Baskaya, F. & Keskustalo, H. & Järvelin, K. (2013). Modeling Behavioral Factors in Interactive Information Retrieval. In: *He, Q. & al. (Eds.). Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM 2013)*, Burlingame, CA, Oct 27 – Nov 1, 2013. New York, NY: ACM Press, pp. 2297–2302. doi>10.1145/2505515.2505660

3 Järvelin, K. (2007). An Analysis of Two Approaches in Information Retrieval: From Frameworks to Study Designs. *Journal of the American Society for Information Science and Technology (JASIST)* 58(7): 971–986.

4 Järvelin, K. (2013). User-Oriented Evaluation in IR. In: Agosti, M. & al. (eds.), *PROMISE Winter School 2012, Heidelberg: Springer, Lecture Notes in Computer Science vol. 7757*, pp. 86–91.

5 Järvelin, K. (2011). Evaluation. In: Ruthven, I. & Kelly, D. (eds.), I*nteractive Information Seeking, Behaviour and Retrieval. London, UK: Facet Publishing*, pp. 113–138. ISBN: 978-1-85604-707-4

6 Järvelin, K. (2009). Explaining User Performance in Information Retrieval: Challenges to IR evaluation. In: *L. Azzopardi & al. (Eds.), Proceedings of the 2nd International Conference on the Theory of Information Retrieval. 2009, Heidelberg: Springer, Lecture Notes in Computer Science vol. 5766*, pp. 289–296.

3.5 Evaluation Measures in Information Retrieval

Norbert Fuhr (Universität Duisburg-Essen, DE)

There is a wide variety of IR measures, but many of them are defined in an ad-hoc way, Basically, the definition of a new metric should consist of the following steps:

- Starting from the chosen criterion, assume a specific user behavior (e.g. stopping after a certain number of relevant documents)
- Define preferences (e.g. the smaller the number of documents seen, the better).
- Define the basic metric obeying the preferences (e.g. precision).
- Furthermore, one can assume a user population, and compute a weighted average of the metric values according to this population (e.g., for average precision, it is assumed that at each relevant document, the same number of users stops).
- Finally, for getting a single result for a set of queries or sessions, an aggregation method has to be chosen (e.g. arithmetic mean).

Many of the current metrics suffer from a number of weaknesses

- The underlying assumptions are not made explicit and/or lack lack empiric foundation (e.g. for mean average precision, Robertson 2008 reconstructed the underlying assumptions; moreover, the assumption of a uniform distribution of users over the possible stopping points seems unrealistic).

- They have theoretic flaws (e.g., reciprocal rank can hardly be seen as interval-scaled, which, however, is a prerequisite for computing mean values).

Based on these observations, one can formulate a number of requirements for the development of new metrics. They should

- allow for more complex user behavior (beyond going through a linear list),
- be able to consider more complex benefits, (like e.g. dependency between documents, or a user searching for fun would like to be entertained all the time),
- have a proper empiric foundation (e.g. with respect to the stopping behavior of a user population),
- be theoretically sound by complying to the fundamentals of measurement theory as well as to basic axioms.

References

1 Luca Busin, Stefano Mizzaro: Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. *ICTIR 2013:8*

2 Stephen Robertson, A new interpretation of average precision, *Proc. SIGIR 2008*, pp. 689–690

3 Warren S. Sarle: Measurement Theory: Frequently asked questions. *Disseminations of the International Statistical Applications Institute, 4th edition*, 1995, Wichita: ACG Press, pp. 61–66. *Revised March 18, 1996.*

3.6 Methodology in IR Evaluation

Diane Kelly (University of North Carolina – Chapel Hill, US)

This talk presents several potentially problematic issues related to evaluation methodology in IR evaluation. The distinction between methodology and methods is made, and questions regarding typical measurement practices, including the convenient practices of associating available and easily obtainable signals (e.g., dwell time) with a number of different constructs (e.g., usefulness, relevance, engagement) without clearly or formally developing a measurement model, and the ad-hoc development of questionnaire items to assess user experience, are raised. A standard psychometric theory is presented, along with a set of steps in which one might engage to create a valid and reliable measure. The talk then examines the differences between explanatory and predictive research and issues related to sample size, power analysis and effect size. A recent study, which questions the appropriateness of some statistical methods to the analysis of big data, is reviewed. The talk closes with some questions to guide workshop discussions about methodology in IR evaluation.

3.7 Future in Information Retrieval Evaluation

Maristella Agosti (University of Padova, IT)

The talk has presented some new challenges in Information Retrieval Evaluation that have been identified thanks to the CULTURA project (http://www.cultura-strep.eu/), the SIGIR 2013 workshop on Exploration, navigation and retrieval of information in cultural heritage (ENRICH 2013, http://www.cultura-strep.eu/events/enrich-2013), and the PROMISE Retreat on Prospects and Opportunities for Information Access Evaluation (Brainstorming workshop held on May 30–31, 2012, Padua, Italy). In fact relevant aspects of the CULTURA project and environment together with ENRICH 2013 and the PROMISE Retreat Report give examples of evaluation challenges that need to be addressed in the next future.

References

1 S. Lawless, M. Agosti, P. Clough, O. Conlan. Exploration, navigation and retrieval of information in cultural heritage: ENRICH 2013. *Proc. of the 36th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2013)*. ACM, New York, USA, page 1136

2 PROMISE Retreat Report on Prospects and Opportunities for Information Access Evaluation. *Brainstorming workshop*, May 30–31, 2012, Padua, Italy

4 Working Groups

4.1 From Searching to Learning

Luanne Freund (University of British Columbia – Vancouver, CA), Jacek Gwizdka (University of Texas – Austin, US), Preben Hansen (Stockholm University, SE), Jiyin He (CWI – Amsterdam, NL), Noriko Kando (National Institute of Informatics – Tokyo, JP), Soo Young Rieh (University of Michigan – Ann Arbor, US)

Search systems to date have been viewed more as tools for the retrieval of content to satisfy information needs, than as environments in which humans interact with information content in order to learn. However, as full-text, information rich search systems become the norm, there is growing recognition of the importance of learning as a search outcome and of the need to provide support for it (Allan et. al., 2012). This is particularly true for environments in which learning is an acknowledged priority, such as collaborative, workplace, and academic search, but learning may also be an important general outcome of search that is not well-served by the drive for ever-greater customization and efficiency. In order to design systems that support learning, we need to investigate when and how learning occurs and develop reliable methods and measures to assess learning through search.

4.1.1 Concepts

Search provides an opportunity for learning on multiple levels, which should be distinguished in order to develop appropriate assessment tools. The primary level relates to learning about the content being searched, which may include acquiring subject knowledge and/or an understanding of the searcher's problem space in relation to the content. At a secondary level, the searcher may also learn about the search system and develop search skills and competencies. Searching may also provide opportunities to learn about oneself and about society through the lens of the content searched. While only a few studies exist that focus on search as learning (Jansen, Booth, & Smith, 2009; Wilson & Wilson, 2013), there is a substantial literature on relevant concepts and frameworks of learning (e.g. Bloom et al, 1956; Kaptelenin & Nardi, 2006; Kintsch, 1998). Across these frameworks, learning is characterized in diverse ways, including learning as knowledge acquisition, learning as sense-making, learning as interpreting, and learning as synthesizing (Säljö 1979; Smith, 2013). Given the breadth of approaches to learning, researchers seeking to assess search as learning need to be explicit about the theoretical framework employed.

4.1.2 Issues

Some of the key issues related to the evaluation of search as learning are:

- How does learning occur through search?
- How does the learning process fit into the searcher's broader Work Task?
- Which system functionalities, components and features of search systems influence learning outcomes?
- What signals are indicators of learning?
- What are appropriate methods and measures?
- Can methods be imported from other fields (learning science, education, cognitive science)?
- Can these incorporated into a methodology to understand the learning process?

4.1.3 Approaches

When approaching the area of search as learning, we may take different viewpoints and look at the problem from different dimensions. Possible approaches could be:

System – How do information access systems including IR systems and tools facilitate learning? When building IR systems, functionalities that support learning should be considered. *Interaction* – How can we design systems that support subject learning? At this dimension, aspects from interactive IR, HCI and Interaction design could be used. How can a system assess the knowledge state of the user? Interaction with content: *Search trails* – predefined trails through content to optimize the searcher learning experience. *Information/resources* – For example, we may prioritize novel content and how to manipulate the quality, quantity of results. The informativeness of the information may be considered. *User* – Several aspects of searchers' knowledge and status are related: how to use the IR system; search strategies and tactics; domain knowledge; task knowledge; socio-cultural background; reading and comprehension ability; ability to conceptualize and integrate; ability to information use *Process* – the learning process need to be acknowledged and taken into account. For example: *Session track research* – that focuses on how systems information should be displayed over the course of a session, depending on the user learning. (e.g.Bates: Berry-picking model). Furthermore, search as learning is just one component of the whole learning process and thus contextual aspects need to be considered.

4.1.4 Measures

Ingwersen and Järvelin provide a conceptual framework for interactive IR evaluation (2005). Based on the framework, learning can occur across the process of search and as an outcome of the search at all four levels. Corresponding measures at each level need to be defined. Table 1 identifies some indicators of learning that could be considered.

Information retrieval Level – Criteria and measures Patterns of query formulation and reformulation, query length, term and terminological variety; number of documents viewed, saved, and downloaded; number of documents assessed and time spent on assessing; pace of interaction, informativeness measures.

Information seeking Level – Criteria and measures Diversity of information seeking strategies; depth, breath, and richness of searchers' understanding of the subject area; searchers' knowledge level and confidence; comprehension test scores (T/F, summary writing), interaction metrics (annotation, notes, writing); searchers' cognitive load, mental workload, affect (happiness, frustration, engagement).

Work Task Level – Criteria and Measures Amount, quality, diversity of the outputs of the searcher's work task, e.g., work report, essay, and decisions made.

Social-organizational and culture Level – Criteria and Measures Success of the organization or social unit, job satisfaction, job promotion, evidence of lifelong learning

4.1.5 Methods

There are a wide range of data collection methods from diverse academic disciplines that could be applied in order to evaluate learning through searching. For example, transaction logs, eye-tracking, think-aloud, observation, self-reports, and interviews could be used. Some methods could be domain- and content-specific, while others are more generic. Methods developed from learning sciences (LS) could provide a useful toolbox. These include pre- and post- tests using instruments such as multiple-choice tests, domain-term lists, concept mapping, essays, comprehension tests, and text understanding measures (sentence verification tasks (SVT), Salmeron et al. 2010).

4.1.6 Conclusion

This summary report aims at providing an initial outline of frameworks and approaches to the evaluation of learning in the context of search in order to better understand how learning takes place and to inform the design of interactive information systems to better support people learning.

References

1 Allan, J., Croft, B., Moffat, A., and Sanderson, M. (2012). Frontiers, challenges, and opportunities for information retrieval: *Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. SIGIR Forum 46:1* (May 2012), pp. 2–32, ACM Press, New York, NY, USA. http://doi.acm.org/10.1145/2215676.2215678

2 Bloom, B. S., Englehard, E., Furst, W., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. New York: McKay.

3 Jansen, B. J., Booth, D., Smith, B. (2009). Using the taxonomy of cognitive learning to model online searching. *Information Processing and Management, 45,* pp. 643–663.

4 Kaptelinin, V. and Nardi, B. (2006). Acting with Technology: Activity Theory and Interaction Design. *Cambridge: MIT Press.*

5 Kintsch, W. (1998) Comprehension: A paradigm for cognition. *New York: Cambridge University Press.*

6 Säljö, R. (1979). Learning in the learner's perspective. I. Some common-sense conceptions, *Reports from the Institute of Education, University of Gothenburg, 76.*

7 Salmeron, L., Gil, L., Braten, I., & Stromso, H. (2010). Comprehension effects of signalling relationships between documents in search engines. *Computers in Human Behavior, 26(3)*, pp. 419–426. doi: 10.1016/j.chb.2009.11.013.

8 Smith, M. K. (2003). Learning theory, in Encyclopedia of Informal Education. http://infed.org/mobi/learning-theory-models-product-and-process/. Retrieved:October 31, 2013.

9 Wilson, M. J. and Wilson, M. L. (2013). A comparison of techniques for measuring sense-making and learning within participant-generated summaries, *Journal of the American Society for Information Science and Technology, 64(2)*, pp .291–306.

4.2 Social Media

Omar Alonso (Microsoft Research – Mountain View, US), Ann Blandford (University College London, GB), Maarten de Rijke (University of Amsterdam, NL), Norbert Fuhr (Universität Duisburg–Essen, DE), Peter Mutschke (GESIS – Köln, DE), Doug Oard (University of Maryland – College Park, US), Max L. Wilson (University of Nottingham, GB)

4.2.1 Introduction

Social media refers to the interaction among people who share different types of information in a particular Internet service. When researchers and practitioners invoke social media, it is usually in the context of social networks like Facebook or Twitter.

All these services have a strong focus on personal connections (e.g., friends, followers) and on user-generated content that is shaped at least in part by those social connections. The use of people to create and enhance content is not new, Wikipedia being but one less "social" example. Having said that, social media has a strong focus on personalization: you are the query.

While search engines like Google and Bing receive millions of queries per day, information dissemination and consumption is also a prominent feature of services with a focus on social characteristics. This phenomenon is changing the landscape of how users access and share information. With users multi-tasking between different information services to get what they are looking for, there is an increased interest to incorporate, to some extent, social data into well-established services. For example, Bing introduced the annotations of Web links with social connections from Facebook, and Google implemented a similar feature using Google+.

Given the huge adoption of social media, what are the implications for the IR community? How do we evaluate the contribution of social data for the next generation of search engines? We need to further investigate user needs, user intent, and the utility of this new source of content and behavioral evidence.

4.2.2 Task categorization

Thinking about tasks and components

We propose the following levels of contexts for characterizing user tasks. We also include components in parentheses.

- IR context: TREC-like, Reputation management (effect prediction), Be the query (contextual search)
- Seeking context: Social utilities, Social load balancing (dynamic routing), Ideation (prognostication detection)
- Socio-organizational & cultural context: A task ecosystem, Buzz exploration (causal reasoning), Event monitoring (interestingness ranking), Groupalization (community detection), Tweet to Powerpoint (contextualization)

Facets

We can further break down the tasks and components into facets to get a different perspective:

- Directionality: Encountering, Monitoring, Influencing, Joining the conversation
- Object: Information, People
- Actor: Human, Machine [on behalf of some human(s)]

4.2.3 Issues

Some of the key issues dealing with social data are:

- Utility: how useful is this data and how can it be used?
- Privacy: how do we explain how the data will be used?
- Differing cultural expectations on privacy
- Controversial content (e.g., adult, racism, etc.) and unsanctioned content (modeling censorship)
- Data quality in an adversarial environment: buying followers (e.g., by celebrities)
- Inferred content (e.g., implicit geo-tagging)
- Informal use of language
- Data cleaning and provenance
- Estimating interestingness (societal, personal, transience)
- Authority detection, personal resolution and the filter bubble
- Content ownership, evolution and curation
- Influence of social recommendations on information seeking behavior
- Feedback, network and virality effects of social media on knowledge dissemination and community building

4.2.4 Opportunities

New assumptions

We need to challenge the traditional assumptions on how users interact with a search engine or IR system. Questioning the established beliefs is essential to understand the potential of social media for the next generation of IR systems. We suggest the following “new” assumptions:

- Information seeking is only one part of the story
- There may be no explicit query; the user and his/her online and offline presence are the query
- The content being searched is neither stable nor bounded
- Emergence can be as important as intent
- Item-based evaluation may concern aboutness (“what”), framing (“how”) as well as reception

- SERP evaluation criteria may include different types of diversity (content, perspective, speed, etc)

Interacting information spaces

Social media does not exist in isolation; people use social media to react to content they find in other media, and other media react in some ways to the activity on social media. Some of these interactions include:

- Social media and the research literature
- Social media and journalism
- Social media and popular culture
- Social media and real life (hybrids of the online and offline worlds)
- Social signals for other IR tasks
- Social media as one lens in an Internet-scale social science "observatory"

Value proposition

Similar to a patient that needs to go to the hospital, we can summarize the main points as follows:

- Know when to go (when to use social media)
- Understand what they say (aggregate and summarize what the "crowd" is saying)
- Learn what they can't tell you (which kind of expertise/knowledge social media can produce)
- Construct strength from adversity (re-construct a story, extract different perspectives)
- Inform their decision making (summarize findings, perceived utility)

4.2.5 Building Bridges

Social media touches much of what we have discussed here at Dagstuhl. Here are some relationships to other discussion groups that formed:

- Task-based retrieval – grounds what we are doing.
- Search as learning – from each other
- Our cultural heritage – is what social media is constructing
- Reliability and validity – are what make our research relevant
- Modeling users – not a single user but lots of users.

4.2.6 Conclusions

This summary report suggest some ways of thinking about social data in the context of IR, and the potential advantage of doing so. We have presented a characterization of tasks and components, identified issues with this type of content, and shed some light on opportunities moving forward.

References

1 O. Alonso, C. Marshall & M. Najork. Are some tweets more interesting than others? #HardQuestion. HCIR, 2013.

2 R. Baeza-Yates. Searching the Future. *SIGIR MF/IR Workshop*, 2005.

3 J. Cook, K. Kenthapadi & N. Mishra. Group chats on Twitter. *WWW*, 225-236, 2013.

4 L. Dantonio, S. Makri, & A. Blandford. Coming across academic social media content serendipitously. *ASIST*, 49(1): 1-10, 2012.

5 M. Efron. Information search and retrieval in microblogs. *JASIST 62(6)*: 996-1008, 2011.

6 S. Erdelez. Investigation of information encountering in the controlled research environment. *IPM 40(6)*, 1013-1025, 2004.

7 J. Hurlock & M. Wilson. Searching Twitter: Separating the Tweet from the Chaff. *ICWSM*, 2011.

8 G. Mishne & M. de Rijke. A study of blog search. *ECIR*, 289-301, 2006

9 P. Mutschke & M. Thamm. Linking social networking sites to scholarly information portals by ScholarLib. *Web Science 2012*, 315-320, 2012.
bibitemPetrov10 S. Petrovic, M. Osborne & V. Lavrenko. Streaming first story detection with application to Twitter. *HLT-NAACL*, 181-189, 2010.

4.3 Graph Search and Beyond

Omar Alonso (Microsoft Research – Mountain View, US), Jaap Kamps (University of Amsterdam, NL)

4.3.1 Motivation

Information on the Web is increasingly structured in terms of entities and relations from large knowledge resources, geo-temporal references and social network structure, resulting in a massive multidimensional graph. This graph essentially unifies both the searcher and the information resources that played a fundamentally different role in traditional IR, and offers major new ways to access relevant information. You are the query.

Graph search affects both query formulation as well as result exploration and discovery. On the one hand, it allows for incrementally expressing complex information needs that triangulate information about multiple entities or entity types, relations between those entities, with various filters on geo-temporal constraints or the sources of information used (or ignored), and taking into account the rich profile and context information of the searcher (and his/her peers, and peers of peers, etc). On the other hand, it allows for more powerful ways to explore the results from various aspects and viewpoints, by slicing and dicing the information using the graph structure, and using the same structure for explaining why results are retrieved or recommended, and by whom.

This new graph based approach introduces great opportunities, but also great challenges, both technical ranging from data quality and data integration to user interface design, as well as ethical challenges in terms of privacy; transparency, bias and control; and avoiding the so-called filter bubbles. The best examples at the time of writing are Facebook Graph Search and related efforts at Bing, Google and other commercial search engines. Similar approaches can be applied to other highly structured data, just to give an example, the hansards or parliamentary proceedings are fully public data with a clear graph structure linking every speech to the respective speaker, their role in parliament and their political party.

4.3.2 Issues

We view the notion of “graph search” as searching information from your personal point of view (you are the query), over a highly structured and curated information space. This goes beyond the traditional two-term queries and ten blue links results that users are familiar, requiring a highly interactive session covering both query formulation and result exploration.

Two Step Interaction

Incremental Structured Query Input: Creating a graph query requires incremental construction of a complex query using a variety of building blocks. Current search engines treat this as a form of query suggestion or query completion, which offers tailored suggestions trying to promote longer queries that cover multiple entity types and relations and various filters. Suggestions and entity types may be based on the user's own activity. This goes beyond prevailing autocompletion techniques, with previews and surrogates from traditional result pages or SERPs (Search Engine Results Page) moving to a more dynamic query suggestion.

Dynamic Structured Result Set Exploration: Results are highly personalized: they are unique for the searcher at a given point in time. The result set is highly structured: rather than just showing the top-10 results from an almost infinite list, a faceted exploration based on your interests is needed. The structure is dynamically derived from the graph structure and the user's point of view, rather than a rigid facet and facet value hierarchy.

When to Use Graph Search?

Rather than a universal solution, the graph search is particularly useful for specific types of information needs and queries. This is also depending on the character of the data available. E.g., Facebook Graph Search emphasizes the social network structure, friends and other persons, locations and location-tagged objects. Social network data is abundantly available (although getting access presents a major barrier) but also notoriously skewed. Rather the searcher personal point of view, it can also be used to show results from the viewpoint of any person in the network. There are many interesting sets of data – both historically or modern – that capture both the persons and related information: think of parliamentary data in public government, or intranet data in organizations.

Query Classification

Graph search also requires a new query classification, beyond the traditional division into navigational, informational, and transactional queries. Is there a new way to characterize queries in this new model? Does the notion of information need change? It is the ultimate form of personalization, with the searcher not only responsible for the query but also determining the (slice of) the data being considered. What shifts in control and transparency are needed to accomplish this?

Graph Search Evaluation

This also presents a range of new evaluation problems. How to evaluate the overall process, given its personalized and interactive nature? How to evaluate the first stage as essentially a form of query autocomplete? And how to evaluate the second stage as to explore and exploit the result set?

4.3.3 Methods

Graph search requires a highly interactive session covering both query formulation and result exploration.

Query Exploration

There is a radical shift towards the control of the searcher, necessitating new tools that help a searcher construct the appropriate graph search query, and actively suggest refinements or filters to better articulate their needs, or explore further aspects. This leads to a far more dynamic interaction than with traditional result lists, or modern hit lists showing summaries of a static set of results.

This suggests a new form of "query autocomplete" that invites and allows users to issue longer queries constructed based on entities, relationships, and templates. In constrast to SERP, we define IQEP as the Incremental Query Exploration Page. IQEP allows the user to explore more the result set as part of the input query. We can think of IQEP as an interactive mechanism that promotes relevant results selected by the user from the traditional SERP to the input box. Figure 1 shows IQEP as a bi-directional channel that moves results from the search list to the input box or viceversa.

There are a range of suitable evaluation methods. The obvious way is by direct evaluation of query suggestion, query recommendations (are they any good?). There is also a range of criteria useful for behavioral observation for in the wild testing: users should issue longer queries, multiple filters, dwell-time, active engagement, structured-query templates. There are query segments where this type of querying is expected to be most useful: torso and tail queries; exploratory scenarios. Traditional head or navigational queries seem less interesting, although these could be part of a more complex underlying information need.

This goes beyond Broder's taxonomy: queries are all navigational, informational, and transactional but they are entity-focused. Queries may aim to return a single or a small set (not unlike traditional Boolean querying over structured data), or there is a need for data analytics on the whole set of results.

Result Exploration

There is a radical shift towards the control of the searcher – small changes in the query can lead to radically different result sets – necessitating active exploration of slices of the data to explore further aspects.

This suggest a new form of search results unique for every user. Similarly to the query exploration mechanism, this interaction encourages users to explore over entities, relationships, and filters. Unlike traditional facetted search options, the result space is highly dynamic, and requires adaptive exploration options tailored to the context and searcher, at every stage of the process.

This is a radical departure from the traditional "ten blue links". The IQEP moves from links to answers, and from answers to suggesting (expressions of) needs. This is an complete shift from the traditional dichotomy between query (the searcher's responsibility) and results (the system's responsibility). Traditional search results have moved to a hit list of result summaries (still a fix set of results, but the shown summaries are tailored to the searcher and her query). These summaries in terms of entities are now answers rather than links to answers. Now these results, or previews of them, are moving into the search box, in the form of structured query suggestions with some sort of preview indicating of the consequences on the result set (often in terms of numbers of results, or entity previews).

There are many options for the evaluation of components: (adaptive) captioning, (adaptive) filters, graph query templates. E.g., captioning should describe (relative to the entity), explain (relative to the user), and be contrastive (relative to the IQEP). There are standard experimental evaluation methods from HCI and UI/UX design. With a running service, evaluation in the wild is very suitable. There are various implicit and explicit criteria: users should explore the result set, usage of multiple filters, dwell-time, active engagement, structured-query templates. Torso and tail queries, and exploratory scenarios are the most suitable query segments.

4.3.4 Conclusions

Graph search gives amazing power, and unleashes the potential of semantically annotated information with many entities, and relations between entities. It brings the control back to the searcher. Graph-based search systems also have the potential to solve part of the old IR problem of conceptual search.

In terms of IR research and required evaluation methods, as discussed in the sections above, there are some open problems. What we need is to work on sharable research data, that exemplifies most of the characteristics we want to study. There is no need to be on Facebook or Twitter, or hand over your personal data. Similar small data sets and systems are available (e.g., so.cl, NYT, Parliamentary data, etc.) It will be hard to share a realistic subset of social network data (unless there are enough volunteers?) but we could work on a simulated set. What would be a concrete task to study on this data? Instead of implementing all features, it is would be useful to select a few components like query suggestion box, filters as facets, and captions to show the potential.

Search engine user interfaces has been very stable in the last 15 years. The input box and the 10 blue links are the still the most optimal way to show search results. Can we do better in terms of user experience? This would be give users a lot of flexibility and options. However, remains to be seen if users would adopt such dynamic interface.

At a high level, graph search seems limited to familiar entity types (e.g. Facebook entities) and templates. How far can this scale? Will this work on truly open domains? Finally, there are a number of ethical issues such as privacy, transparency, bias and control, and filter bubbles.

4.4 Reliability and Validity – A Guide To Best Practices in IR Evaluation

Nicola Ferro (University of Padova, IT), Hideo Joho (University of Tsukuba, JP), Diane Kelly (University of North Carolina – Chapel Hill, US), Dirk Lewandowski (HAW – Hamburg, DE), Christina Lioma (University of Copenhagen, DK), Heather O'Brien (University of British Columbia – Vancouver, CA), Martin Potthast (Bauhaus–Universität Weimar, DE), C.J. Keith van Rijsbergen (University of Cambridge, GB), Paul Thomas (CSIRO – Canberra, AU), Vu Tran (Universität Duisburg–Essen, DE), Arjen P. de Vries (CWI – Amsterdam, NL)

4.4.1 Motivation

Experimental evaluation is one of the backbones of the information retrieval field since its inception. Over the years, it provided both qualitative and quantitative evidence as to which methods, algorithms, and techniques are more effective. Moreover, due to its early and systematic adoption of strong evaluation methodologies, the IR community is often regarded as "leading" in this respect by computer science people but there still many open questions.

Indeed, carrying out thorough experiments is a challenging activity where many "traps" are hidden. For example, there is increasing discussion in the research community about reproducibility and generalizability of our experiments, as it may be difficult to re-use research, methods, measures, data and results.

Moreover, people in IR come from different backgrounds and there is a need to consolidate ideas/expertise from different fields and to establish a common ground around some key concepts (reliability, validity, ...) as well as an understanding of their trade-offs and design decisions.

Finally, a better support for students is needed in order to avoid them to learn best practices in a very fragmented and sometimes inconsistent way, not to say the risk of adopting approaches which have been discarded with the passing of time due to lack of robustness.

Therefore, there is an overall need for a reasoned guide to best practices for IR evaluation which will turn around the two key concepts: reliability is the extent to which a [measure/-method] produces similar results under stated conditions for a stated period of time [inspired by ISO 9126]; and, validity is the extent to which a [measure/method] accurately reflects the phenomena it is intended to reflect.

4.4.2 Goals and Scope

The proposed best practices have the following goals:

- to produce research results with confidence: for communicating with the research and stakeholder community; for assessing their impact, longevity, and generalizability;
- to gain an appreciation of the "trade-offs" and limitations inherent in our studies;
- to encourage good practices for novices and experts;
- to enable/promote repeatability/reproducibility of the experiments.

The proposed best practices have the following scope:

- to understand whether an IR/IIR experiment is valid and reliable, including design, carrying out, analysis, and results presentation;
- to better communicate results
- to make the context explicit: types of methods for IR/IIR evaluation and beyond; experiments as our way to evaluate (lab, insitu, crowdsourcing, log analysis, ...); the kind of context itself (IR context, Seeking context, ...).

The target audience of the proposed best practices are:

- graduate students;
- PhD students;
- researchers (and reviewers).

4.4.3 Structure of the Best Practices

The proposed best practices will be structured as follows:

- Pillar Definitions and Concepts: starting from the definitions of reliability and validity provided above, we will explore and detail them for different methods and measures as well as provide example of factors that demonstrate and/or enhance reliability/validity of measures and methods;

- Reliability and Validity in Experimentation: we will discuss how to ensure reliability and validity in carry out actual experiments, by covering hypotheses, sampling, methods, environments, data analysis, measurement and procedures, as well as pointing at other issues such as ethical/legal issues and privacy and intellectual property rights.
- Reporting Out Experiments: we will discuss how to present experimental results, their limitation, to acknowledge alternative interpretations of the results, to report data analysis as well as verifiable/falsifiable outcomes and we will deal also with archiving and infrastructures for experimental data management (data curation).
- Failure analysis: this is deemed one of the most important activities to actually understand how and why a system behaves differently than expected and why it fails to achieve the desired performances. Unfortunately, this is an extremely demanding activity in terms of time and effort needed to carry it out.
- Definition of new measures/methods: we will cover the steps and the process needed for establishing and motivating a new measure and its trade-offs as well as the checks to ensure its reliability and validity.
- For reviewers: we will look at the previously introduced concept and best practices from the angle of reviewers in order to support them in effectively and fairly reviewing papers reporting experimental data.

References

1 M. H. Birnbaum. Psychological Experiments on the Internet. *Academic Press*, USA, 2000.

2 D. K. Harman and C. Buckley. Overview of the Reliable Information Access Workshop. *Information Retrieval, 12(6)*, pp. 615-641, 2009.

3 D. K. Harman. Information Retrieval Evaluation. *Morgan & Claypool Publishers*, USA, 2011.

4 D. K. Harman and E. M. Voorhees. TREC. Experiment and Evaluation in Information Retrieval. *MIT Press*, USA, 2005.

5 K. Hornbaek: Some Whys and Hows of Experiments in Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction (FnTHCI), 5(4)*, pp. 299-373, 2013.

6 ISO/IEC 25010:2011. Systems and software engineering – *Systems and software Quality Requirements and Evaluation (SQuaRE)* – System and software quality models.

7 K. Järvelin: An analysis of two approaches in information retrieval: From frameworks to study designs. *JASIST 58(7)*, pp. 971-986, 2007.

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

9 F. W. Lancaster. Information Retrieval Systems: Characteristics, Testing and Evaluation. *John Wiley & Sons Inc; 2nd edition*, 1979.

10 H. L. O'Brien, E. G. Toms: The development and evaluation of a survey to measure user engagement. *JASIST 61(1)*, pp. 50-69, 2010.

11 M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)*, 4(4), pp. 247-375, 2010.

12 K. Spärck Jones. Information Retrieval Experiment. *Butterworths*, UK, 1981.

13 P. E. Spector. Summated rating scale construction: An introduction. *Sage Publication*, USA, 1992

14 J. Zobel. Writing for Computer Science. *Springer*, Germany, 2004.

4.5 Domain Specific Information Retrieval

Maristella Agosti (University of Padova, IT), Floriana Esposito (University of Bari, IT), Ragnar Nordlie (Oslo University College, NO), Vivien Petras (HU Berlin, DE), Christa Womser–Hacker (Universität Hildesheim, DE)

The working group domain-specific information retrieval met for one day. After defining domain-specific information retrieval and information systems, the focus of the group was directed at discussing information retrieval and evaluation issues in the domains that were relevant for their use cases (cultural heritage, patent retrieval).

4.5.1 Definition Domain Specific Information System

Domain specific information systems collect, store, preserve, organize, search and display domain specific objects or their (metadata) representations in a digital environment. Good examples of domain of interest are: cultural heritage, patents, and medical collections.

4.5.2 Motivation

For the domains of interest, there may be a challenge to manage collections of documents that the user wants to interact with not only through a query function. This means that some specific features of the domain need to be taken into account when envisaging a system that has to manage the document collection. This also has consequences for information retrieval evaluation.

4.5.3 Summary of Challenges

For the users of some kinds of domain specific information systems to start searching the system – i.e. starting the interaction with an information system through a query – may not always be the optimal mode of access. Domain specific information systems have responded by providing exploratory interaction functionalities like curated digital exhibitions, featured objects, or user-provided stories to present alternative starting options to the user. Other than studying whether users “liked” these features, IR evaluation has not progressed towards a formalized mode of evaluation that would allow comparing the “usefulness” of these features in different applications with respect to the goals of the system (or the user). These efforts have not succeeded in recommendation system improvements based on evaluation. In a domain specific information system, a significant facility will be to provide context for the represented objects, either through links to other objects in the managed collection or through associated text that can be user-provided or producer-provided. Evaluating the quality of this context is a challenge which may be met in different ways: through establishing some measure of semantic similarity between the object and the context, or measuring some degree of user satisfaction and relevance judgment of the context provided. What is needed can only be decided by looking at the outcome of the interactive process with a strong priority on the experts’ involvement. For example Patent people are very conservative and want to continue their work e.g. with Boolean systems. If we stay with Boolean systems no progress will be possible. We – as IR scientists – have to convince them that there are new innovative approaches. Automatically judging the quality of retrieval functions based on observable user behavior could allow for making retrieval evaluation faster, cheaper, and more user centered.

However, the relationship between observable user behavior and retrieval quality is not yet completely investigated.

4.5.4 The No-search / Exploratory Access Evaluation Problem in Cultural Heritage

Cultural heritage user types can possibly be divided into two groups: humanities scholars utilizing cultural heritage information systems for their research and information "tourists" utilizing cultural heritage information systems to get informed or be entertained about or by cultural artefacts. IR evaluation has traditionally been focused on users searching (i.e. more or less targeted querying) a predetermined document pool. For information tourists user types in cultural heritage information systems, search – i.e. starting the interaction with an information system through a query – is probably not the optimal mode of access, because they (a) don't know what content the system provides and (b) often do not have a specific information need in mind that can be translated into a query. Cultural heritage information systems have responded by providing exploratory interaction functionalities like curated digital exhibitions, featured objects, or user-provided stories to present alternative starting options to the user. Other than studying whether users "liked" these features, IR evaluation has not progressed towards a formalized mode of evaluation that would allow comparing the "usefulness" of these features in different applications with respect to the goals of the system (or the user). These efforts have not succeeded in recommendation system improvements based on evaluation. One challenge for IR evaluation in cultural heritage information systems is therefore to develop evaluation scenarios that do not have the conventional query-output (maybe iteration thereof) process in mind, but alternative exploratory options (which might lead to a retrieval-based outcome nevertheless). This would consequently also require the development of new assessment approaches and the creation or adaptation of appropriate measures.

4.5.5 Contextualization and Evaluation of Context in Cultural Heritage

The information retrieval systems for cultural heritage are embedded in a rich context. Documents no longer exist on their own; they are connected to other documents, they are associated with users and they can be mapped onto a variety of ontologies. Retrieval tasks are interactive and are solidly embedded in a user's social and historical context. New challenges in information retrieval will not come from smarter algorithms that better exploit existing information sources, but from new retrieval algorithms that can intelligently use and combine new sources of contextual metadata. Machine learning methods (multirelational learning) could be used: - to automatically create the markup or metadata for existing unstructured documents - to create, merge, update, and maintain ontologies. In a cultural heritage system, a significant facility will be to provide context for the represented objects, either through links to other objects in the database or through associated text, user-provided or producer-provided. Evaluating the quality of this context is a challenge which may be met in different ways: through establishing some measure of semantic similarity between the object and the context, or measuring some degree of user satisfaction and relevance judgment of the context provided. The system's ability to limit or extend the amount of context, encourage the pursuit of context etc should also be evaluated.

4.5.6 Innovative Applications vs. Traditional Values in Patent Retrieval

We do not have fully automatic systems that could perform patent retrieval successfully, yet. Patent retrieval is per se an interactive task sharing human and system intelligence. Since we have more than 10 million patent applications p.a., the need for a solution is very important. On the other hand, we have many system approaches delivering solutions for specific sub-tasks. Here, the need for evaluation steps in. How can proposed solutions be evaluated? What is needed can only be decided by looking at the outcome of the interactive process with a strong priority on the experts' involvement. Patent searchers are very conservative and want to continue their work, e.g. with Boolean systems. If we stay with Boolean systems, no progress will be possible. We – as IR scientists – have to convince them that there are new innovative approaches. How? By showing them that they do better work in shorter time in the interactive scenario. However, control should stay with the experts.

4.5.7 Automatic Observation of User Behavior

Automatically judging the quality of retrieval functions based on observable user behavior could allow for making retrieval evaluation faster, cheaper, and more user centered. However, the relationship between observable user behavior and retrieval quality is not yet completely investigated. A paper studying this relationship for a search engine operating on the arXiv.org e-print archive has shown that none of the eight well known absolute usage metrics (e.g., number of clicks, frequency of query reformulations, abandonment) reliably reflect retrieval quality for the considered sample. Learning techniques have been applied in information retrieval (IR) applications generally for information extraction, relevance feedback, information filtering, text classification and text clustering. Recently, online learning models have been proposed for interactive IR with the aim of providing results of maximum utility to the user. The interaction between human and system takes the following form. The user issues a command (e.g. query) and receives a result in response (ranking). The user then interacts with the results (clicks), thereby providing implicit feedback about the user's utility function. Using online learning models (for example, coactive learning algorithms), the feedback can be inferred from observable user behavior from clicks during-search. In each iteration, a user, drawn from an unknown but fixed distribution, presents a context (e.g., query) to the system and receives a ranking in response. The user is represented by a utility function that determines the actions (e.g. clicks) and therefore the feedback to the learning algorithm. The same utility function also determines the value of the presented ranking. The goal is to learn a ranking function that has high social utility, which is the expected utility over the user distribution.

4.5.8 Support of Exploration and Content Enriching Tools

A domain specific information system has to support the exploration of the managed collection, so it needs to support traditional search-based exploration, but also has to move beyond it. It can support a range of innovative normalization and natural language processing technologies that allow entities and relationships to be extracted from the collection and visualized using a range of specially designed visualizations. A domain specific information system has also to provide for entity oriented search, and allows users to crosswalk from one tool to another, ensuring that their exploration of the collection is flexibly supported. The system has also to provide a comprehensive set of bookmarking, and annotating tools that make it a powerful aid to both extensive and intensive work on content collections.

References

1 M. Agosti, M. Manfioletti, N. Orio, C. Ponchia. Evaluating the Deployment of a Collection of Images in the CULTURA Environment. In: *Proc. of the Int. Conf. on Theory and Practice of Digital Libraries (TPDL 2013)*, Valletta, Malta. LNCS Vol. 8092, Springer, Berlin, 2013, pp. 180–191

2 F. Esposito, C. d'Amato, N. Fanizzi. Fuzzy Clustering for Semantic Knowledge Bases. *Fundamenta Informaticae*, Vol. 99 2/2010, pp. 187–205

3 N. Fanizzi, C. d'Amato, F. Esposito. Inductive Classification of Semantically Annotated Resources Through Reduced Coulomb Energy Networks. *International Journal On Semantic Web And Information Systems*, 2009, Vol. 5(4), pp. 19–38

4 N. Fanizzi, C. d'Amato, F. Esposito. Metric-Based Stochastic Conceptual Clustering For Ontologies. *Information Systems*, 2009, Vol. 34(8), pp. 792–806

5 N. Ferro, R. Berendsen, A. Hanbury, M. Lupu, V. Petras, M. de Rijke, G. Silvello. Context Evaluation. In: *PROMISE Retreat Report. Prospects and Opportunities for Information Access Evaluation.* Brainstorming workshop held on May 30–31, 2012, Padua, Italy

6 J. Jürgens, P. Hansen, C. Womser-Hacker. Going beyond CLEF-IP: The "Reality" for Patent Searchers? In: Catarci, T.; Forner, P.; Hiemstra, D.; Peñas, A.; Santucci, G. (Eds.), *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics Third International Conference of the CLEF Initiative – CLEF 2012*, Rome, Italy, September 2012, pp. 30–35

7 S. Lawless, M. Agosti, P. Clough, O. Conlan. Exploration, navigation and retrieval of information in cultural heritage: ENRICH 2013. In: *Proc. of the 36th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2013).* ACM, New York, USA, p. 1136

8 V. Petras, J. Stiller, M. Gäde. Building for Success (?) - Evaluating Digital Libraries in the Cultural Heritage Domain. In: Cool, C., NG, K. B (Eds.), *Recent Developments in the Design, Construction and Evaluation of Digital Libraries: Case Studies*, ICI Global, 2013

9 M. Sweetnam, M. O. Siochru, M. Agosti, M. Manfioletti, N. Orio, C. Ponchia. Stereotype or Spectrum: Designing for a User Continuum. In: *Proc. of the Workshop of the 36th ACM SIGIR Conference on Exploration, navigation and retrieval of information in cultural heritage (ENRICH 2013)*, Dublin, 2013, pp. 23–30

4.6 Task–Based Information Retrieval

Nick Belkin (Rutgers University – New Brunswick, US), Kalervo Järvelin (University of Tampere, FI), Evangelos Kanoulas (Google Switzerland – Zürich, CH), Birger Larsen (Aalborg University Copenhagen, DK), Thomas Mandl (Universität Hildesheim, DE), Elaine Toms (Sheffield University, GB), Pertti Vakkari (University of Tampere, FI)

4.6.1 Core concepts & definitions

What is meant by task-based analysis of IR?

There are several possible answers to this question. One answer is to focus on the task or goal that motivates a person to engage in information seeking in an IR system. Examples of motivating tasks or goals are work tasks, hobbies, everyday life tasks and leisure time interests. From this point of view, task-based analysis of IR means understanding and responding to motivating tasks and goals, and designing IR systems which can support accomplishment of a variety of such tasks and goals. This answer requires that an IR system be able to recognize different tasks. Another possible answer is understanding the nature of the task or tasks of specific groups, and the design of IR systems which are tailored to support those people as they are engaged in those specific types of tasks. This answer requires the design of many different IR systems.

Task types

We use the following categorization of task for IR purposes:

- Motivating tasks, sometimes called "work tasks", accomplishment of which have led the person to engage in an IR system. These may lead to more than one information seeking or information search session.
- Seeking tasks. These involve deciding where, or with whom to engage in order to obtain information which will be useful in accomplishing the motivating task. This can be from a variety of sources and systems, may include several sessions over time, but may also involve only other people or a combination of both people and systems.
- Search tasks. These are the tasks which a person engages in during an information seeking session, trying to accomplish their intentions while using an IR system. These will involve one or more sequences of behaviors over a search session. Examples of such tasks are formulating a query, learning about a domain, comparing search results, judging usefulness or relevance of search results, etc.

4.6.2 How is task-based IR different from traditional IR, and what does this mean for evaluation of IR system performance?

Traditional IR evaluates performance according the system's response to a single query. However, motivating tasks typically generate several information-seeking intentions, leading to multiple sessions and multiple search tasks within sessions. Task-based IR studies IR as a process, with sequences of behaviors associated with different search tasks during the course of a search session, or over multiple sessions. Evaluation of IR system performance, from the point of view of task-based IR, must be based on some description of what it would take to accomplish the "work task" (i.e. the motivating task context), and of how accomplishment of that task could be measured. Then the IR system's support techniques, and measures for their evaluation, must be justified according to hypotheses about how those techniques will support accomplishment of the motivating task, and of the search tasks, and to how the measures reflect such accomplishment.

4.6.3 Steps toward task-based evaluation of IR system performance

We suggest that the following issues need to be addressed in a coordinated research program, to provide the basis for being able to perform evaluation from the task-based IR point of view

Persons

Minimally, we need to learn more about motivations for engaging in information seeking behavior, about the effects of people's knowledge of task, topic and system, of stage in task accomplishment, and of individual and cultural differences on behaviors and intentions in search sessions. We need to expand upon or perhaps integrate the different classifications of task types that have been proposed by various researchers.

Process of task-based IR

We need a better understanding of the intentions of search tasks, and especially of sequences of search tasks during search sessions of different types. A start toward this goal would be a series of studies, both lab and live, which both observe search behaviors, and elicit search intentions related to the different behaviors.

Outputs and outcomes

Outputs are the products delivered by the system, outcomes are the benefits for the user produced by the system, e.g. task accomplishment. Outputs as well as outcomes are highly task-dependent and may require different measures for evaluating system performance.

4.6.4 Possible actions

There are three overall issues which task-based IR needs to address in order to move toward appropriate evaluation. The first is mapping the territory of task-based IR and identifying thereby areas where knowledge is shallow or nonexistent (white patches). One should also check the borders (or outline) of the map. The second is developing a research program that seeks to systematically analyze the connections of motivating task features, search task features, search behavior features, IR system features, output features and outcome features. And the third is to develop study designs which include variables from larger tasks, consequent search tasks including search processes, outputs and outcomes, and system features. Two kinds of designs are needed: field studies and experimental designs. Due to the complexity of the phenomena investigated, field studies can be used to reveal mechanisms connecting larger tasks with search tasks and consequent search processes and outcomes. These results should provide information for designing evaluation experiments and systems. Experimental designs should identify system features that can most strongly be expected to have a connection to the motivating task features and then examine that connection.

4.7 Searching for fun

Ragnar Nordlie (Oslo University College, NO), Ann Blandford (University College London, GB), Floriana Esposito (University of Bari, IT), Douglas W. Oard (University of Maryland – College Park, US), Vivien Petras (HU Berlin, DE), Max L. Wilson (University of Nottingham, GB), Christa Womser–Hacker (Universität Hildesheim, DE)

4.7.1 Definition

"Searching for fun" might have the double sense of "searching for something that is fun" or "having fun while searching". Our discussion was concerned with the second sense: the activity of interacting with an information system without having a specific search objective in mind. For short, it may be called FII: fun information interaction. This may involve activities such as: online window shopping with nothing to buy, reading online, like reading fiction or the news, watching funny videos or finding funny pictures. It may perhaps also include examples of pursuing more traditional information needs, as in situations where not finding a result is no great concern. Even a traditional search process may waste time, pique interest, be fun.

4.7.2 Motivation

Many information systems are constructed, at least partly, with the objective of inducing the user to interact with the system without a predefined purpose, and to retain this user in interaction; to encourage unexpected discovery, to encourage a certain kind of learning, to support a certain kind of shopping behavior, to expose the user to advertising, or for a number of other reasons. System users frequently engage, also for a number of reasons, in this kind of non-intentional interaction. Evaluation of this kind of system or this kind of system activity calls for different evaluation criteria and measures from those employed for goal-directed information retrieval.

4.7.3 Prior work

Related, but maybe not identical, issues have been discussed at recent workshops, particularly:

- 'Entertain Me' workshop on supporting complex search tasks at SIGIR 2011; followed up by a "contextual suggestion" track at TREC 2012 and 2013
- Searching4Fun workshop at ECIR2012

Despite these efforts, we feel that the evaluation challenges presented by non-outcome-focused interaction have not been exhaustively discussed. This report is a first effort to address these challenges. Discussions How does FII differ from traditional IR evaluation? It changes our assumptions about searching (and browsing, and whatever other activities involve finding things). This changes our criteria, and thus our interpretations of measures. Instances of this include:

- Stopping behavior: Stopping may mean running out of things to find; finding a good result, may be reason to continue (not to stop, as in IR in general);
- Time spent: More time can be good;
- Novelty: Novelty and Repetition might be equally important.

What criteria may be applied to the evaluation of FII? Evaluation of FII may be considered from a system or a user perspective. The system's motivation may be expressed in a simplified manner as "get them in – keep them in – convert entry into experience (learning, shopping, entertainment...)". The user motivation may be to be entertained, to spend time, to make unexpected discoveries...For the user evaluation a number of criteria may identify one or more of these experiences; we discussed, among others: engagement, flow, cognitive load, stimulation, currentness, social engagement, novelty, sensemaking, meaningmaking (contextualization), outcome state, state change, user empowerment. We provisionally concluded that these criteria can be comprised in a process criterion: engagement, and an outcome criterion: state change. A good engagement level for the user involves, for instance, avoiding bad disengagement, avoiding over-engagement. State change may imply changes such as bored to not bored, stressed to not stressed, sad to happy, or changes via a transforming state, such as stressed to relaxed via horrified and surprised. It is difficult to design study conditions for systematic measurements of these criteria. Casual or leisure needs are by nature intrinsic and hard to create under experimental conditions. How can we make participants bored or stressed or sad so that they will naturally entertain themselves, try to relax or be amused? How can we measure engagement when, for instance, a measure like time spent may be either positive or negative depending on the circumstances? Rather than define measures, we identify some open questions regarding FII IR, which may influence the choice of measures, such as

- Is there more to FII than just the distinction between visceral and conscious needs (Taylor, 1962)?
- How does FII relate to things like serendipity?
- Are there gaming measures that are relevant?
- Can we have FII within Serious and Project leisure (Stebbins, 2009)?
- Can we optimize systems for FII behavior?
- Can we detect certain state-change targets (bored to not-bored, stressed to relaxed, etc)?
- How do different demographics differ?
- In what way is the journey more important than the objects found?

What are the current challenges for IR evaluation of FII? We developed the following (incomplete and unordered) list of challenges for FII evaluation:

- Actually studying fun information interaction in action
- Discovering fruitful scenarios/contexts
- Identifying successful FII strategies (If there are strategies for this?)
- Correlating system interactions with study findings
- Determining measures for Fun Information Interaction
- Designing simulated user interaction models that relate to FII
- Create systems that increase engagement
- Identify ways systems can support FII

4.7.4 Actions

Some time has passed since the Searching4Fun workshop, which in itself did not focus primarily on evaluation methods. There might be scope for a new workshop, for instance at IIiX 2014 in Regensburg.

References

1 Ilaria Bordino, Yelena Mejova & Mounia Lalmas (2013). Penguins in Sweaters, or Serendipitous Entity Search on User-Generated Content, CIKM.

2 Mihaly Csikszentmihalyi (1990). Flow: The Psychology of Optimal Experience, Harper Collins.

3 Heather O'Brien (2010). The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. Interacting with Computers, 22(5), pp. 344–352. *Psyclopedia, Effort Recovery Model.* http://www.psych-it.com.au/Psychlopedia/article.asp?id=356

4 Kateriina Saarinen & Pertti Vakkari (2013). A sign of a good book: readers's methods of accessing fiction in the public library. *Journal of Documentation*, 69(5):736–754.

5 Robert Stebbins (2009). Leisure and Consumption: Common Ground, Separate World, Palgrave Macmillan.

6 Robert Taylor (1962). The Process of Asking Questions. *American Documentation*, 13(3):391–396.

4.8 The Significance of Search, Support for Complex Tasks, and Searcher–aware Information Access Systems

Jaap Kamps (University of Amsterdam, NL)

This abstract documents three loosely related challenges. The first challenge is the role and significance of the field in general. There are massive challenges in the way the information available is changing in quantity and in character, and in the ways we create, publish, share, and use information in the always-online world. This urges us to keep 'reinventing search' and redefine the field of information retrieval and its key research problems and research methods. How do these changes affect the core questions we address in the field of IR and what sort of evidence do we need for addressing these questions? How can we factor the larger scope and context into IR evaluation? It is interesting to consider a publication like Salton's "Developments in automatic text retrieval" published in Science in 1991. Salton (1991) is from before the Web happened and discusses all the basic IR aspects: retrieval models, indexing structures, but also hypertext, knowledge resources and semantic search. Articles like Salton (1991) still look surprisingly modern! This raises two question that are perhaps not unrelated: First, why hasn't our research field changed in a dramatic way to suit the revolutionary changes in the information environment. Second, why isn't our field making a larger impact outside our field (Salton published 2 Science articles in 1991) given the dramatic increased role and importance of "search" nowadays.

The second challenge is to work on information access tools that support complex tasks. That is, to build and evaluate information access tools that actively supports a searcher to articulate a whole search task, and to interactively explore the results of every stage of the process. In the prolonged search session, how should we evaluate the overall effectiveness as well as the success at various stages? How can evaluation reflect the different goals of each stage? There is a striking difference in how we ask a person for information, giving context and articulating what we want and why, and how we communicate with current search engines. Current 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. Combining different sources requires opening multiple windows or tabs, and cutting-and-pasting information between them. Current search engines may have reached a local optimum for answering micro information needs with lighting speed. Supporting the overall task opens up new ways to significantly advance our information access tools, by develop tools that are adapted to our overall tasks rather than have searchers adapt their search tactics to the "things that work."

The third challenge is to make information access systems more informed about the searcher. Can we make a retrieval system aware of the searcher's stage in the information seeking process, tailor the results to each stage, and guide the searcher through the overall process? How to evaluate the utility of this (accuracy of the prediction, usefulness of the support, etc)? Can we equate evaluation with observing preferred information interaction patterns? A search session for a non- trivial search task consists of stages with different sub-goals (e.g., problem identification) and specific search tactics (e.g., reading introductory texts, familiarizing with terminology). Making a system aware of a searcher's information seeking stage has the potential to significantly improve the search experience. Searchers are stimulated to actively engage with the material, to get a grasp on the information need and articulate effective queries, to critically evaluate retrieved results, and to construct a comprehensive answer. This may be of particularly great help for those searchers having poor information or media literacy. This is of obvious importance in many situations: e.g., education, medical information, and search for topics "that matter". Some special domains, such as patent search and evidence based practices in medicine, have clearly prescribed a particular information seeking process in great detail. Here building a systems to support (and enforce) this process is of obvious value.

4.9 Interaction, Measures and Models

Gianmaria Silvello (University of Padova, IT), Leif Azzopardi (University of Glasgow, GB), Charlie Clarke (University of Waterloo, CA), Matthias Hagen (Bauhaus–Universität Weimar, DE), Robert Villa (University of Sheffield, GB)

A common framework for user interaction models and a common framework in which to place evaluation measures (i.e., the units of measurement) should be consistent but does not yet exist. Current measures are not comparable as the units used are not clearly defined in terms of real-world outcomes, and vary between measures. Since most measures encode some form of user behaviour as an underlying user interaction model, having measures that use the same unit of measure would enable comparisons between different user interaction models across different systems. As well as making it possible to compare between measures themselves (opposed to viewing them independently in different units).

4.9.1 Motivation

The main goal is to enable assessment of the performances of the system as a whole or specific components in particular. For that we need a repeatable way to say that a system is better than another on a gain base (utility, usefulness, happiness, ...). Ideally, the effect of user attributes that are not salient to the evaluation itself should be minimized (e.g. "what the user had for breakfast"). The measures should be comparable; that is, defined using the same units (i.e. gain, cost, or gain/cost). We would also like to be able to determine the effects of the interface and interaction on the actual performance.

4.9.2 Proposed solution

Integrate the interaction with an IR interface into the measures, e.g. in a TREC- style evaluation, individual IR systems may submit conventional ranked lists. Systems can then be evaluated based on different models of user interfaces or interactions. To extend TREC-style evaluations to accommodate more realistic interfaces, individual systems might submit responses to a variety of user actions, which would then be evaluated across more complex and detailed interfaces and interaction models.

One possible solution would be to decompose measures into components: Interaction model (I) (traditionally: when the user stops) Gain model (G) (traditionally: number of viewed relevant docs) Cost model (C) (traditionally: number of viewed docs with unit costs) An evaluation measure could then be parameterized by the components as M(I,G,C).

An interaction model might be characterized by a sequence of states and for each state some specific interaction with the system taken; potentially depending on the intent and task of the user (e.g., a recall oriented task). The interface of the system could be encoded in the cost model. The gain would model the documents returned to the user (e.g., degree of relevance)

For example, we can deconstruct DCG into the three main components outlined above: the interaction model is "defined" by the discounting function, the gain model is how we sum up the weights of viewed (relevant) documents and the cost model is the number of viewed documents (with a fixed cost for each document). This means that we can fix the gain and the cost models while changing the interaction model still being able to compare measurements.

We could define an idealized interaction between the user and the system (including its interface). Idealized in this case would mean the optimal behavior where users are able to make decisions towards the best possible gains. System comparison based on such idealized interactions seem to be much more reasonable and comparable than based on arbitrary and possibly sub optimal decisions. This approach would also enable us to drop from the models a number of parameters that are difficult to estimate, such as click and query reformulation probabilities.

References

1 Allan, J., Croft, B., Moffat, A., and Sanderson, M. (2012). Frontiers, challenges, and opportunities for information retrieval: *Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. SIGIR Forum 46:1* (May 2012), pp. 2–32, ACM Press, New York, NY, USA . http://doi.acm.org/10.1145/2215676.2215678

2 Azzopardi, L. (2009). Usage based effectiveness measures: monitoring application performance in information retrieval, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, pp. 631–640, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=1646034

3 Azzopardi, L. (2011). The economics in interactive information retrieval. *Proceedings of the 34th international ACM SIGIR International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*, pp. 15–24, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2009916.2009923

4 Azzopardi, L., Kelly, D., and Brennan, K. (2013). How query cost affects search behavior. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, pp. 23–32, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2484049

5 Azzopardi, L., Järvelin, K., Kamps, J. and Smucker, M. D. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction, *SIGIR Forum 44:2 (December 2010)*, pp. 35-47, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=1924475.1924484

6 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'12)*, pp. 105–114, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2348301

7 Belkin, N., Dumais, S., Kando, N. and Sanderson, M. (2012). N*II Shonan meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Held October 2012*. To appear. http://www.nii.ac.jp/shonan/seminar020/

8 Carterette, B. (2011). System effectiveness, user models, and user utility: a conceptual framework for investigation, *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*, pp. 903–912, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2009916.2010037

9 Carterette, B., Kanoulas, E. and Yilmaz, E. (2012). Incorporating variability in user behavior into systems based evaluation. *Proceedings of the 21st ACM international Conference*

on Information and Knowledge Management (CIKM'12), pp. 135–144, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2396761.2396782

10 Sakai, T. and Dou, Z. (2013). Summaries, ranked retrieval and sessions: a unified framework for information access evaluation, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, pp. 473–482, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2484028.2484031

11 Smucker, M. and Clarke, C. (2012). Time-based calibration of effectiveness measures, *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pp. 95–104, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2348283.2348300

12 Smucker, M. and Clarke, C. (2012). Modeling User Variance in Time-Biased Gain, *Proceedings of the Sixth Symposium on Human-Computer Interaction and Information Retrieval (HCIR'12)*, Article 3, 10 pages, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2391224.2391227

Participants

- Maristella Agosti
University of Padova, IT
- Omar Alonso
Microsoft Research – Mountain View, US
- Leif Azzopardi
University of Glasgow, GB
- Nicholas J. Belkin
Rutgers University – New Brunswick, US
- Ann Blandford
University College London, GB
- Charles Clarke
University of Waterloo, CA
- Maarten de Rijke
University of Amsterdam, NL
- Arjen P. de Vries
CWI – Amsterdam, NL
- Floriana Esposito
University of Bari, IT
- Nicola Ferro
University of Padova, IT
- Luanne Freund
University of British Columbia – Vancouver, CA
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Jacek Gwizdka
University of Texas – Austin, US
- Matthias Hagen
Bauhaus-Universität Weimar, DE
- Preben Hansen
Stockholm University, SE
- Jiyin He
CWI – Amsterdam, NL
- Kalervo Järvelin
University of Tampere, FI
- Hideo Joho
University of Tsukuba, JP
- Jaap Kamps
University of Amsterdam, NL
- Noriko Kando
National Institute of Informatics – Tokyo, JP
- Evangelos Kanoulas
Google Switzerland – Zürich, CH
- Diane Kelly
University of North Carolina – Chapel Hill, US
- Birger Larsen
Aalborg Univ. Copenhagen, DK
- Dirk Lewandowski
HAW – Hamburg, DE
- Christina Lioma
University of Copenhagen, DK
- Thomas Mandl
Universität Hildesheim, DE
- Peter Mutschke
GESIS – Köln, DE
- Ragnar Nordlie
Oslo University College, NO
- Heather O'Brien
University of British Columbia – Vancouver, CA
- Doug Oard
University of Maryland – College Park, US
- Vivien Petras
HU Berlin, DE
- Martin Potthast
Bauhaus-Universität Weimar, DE
- Soo Young Rieh
University of Michigan – Ann Arbor, US
- Gianmaria Silvello
University of Padova, IT
- Paul Thomas
CSIRO – Canberra, AU
- Elaine Toms
Sheffield University, GB
- Vu Tran
Universität Duisburg-Essen, DE
- Pertti Vakkari
University of Tampere, FI
- C .J. Keith van Rijsbergen
University of Cambridge, GB
- Robert Villa
University of Sheffield, GB
- Max L. Wilson
University of Nottingham, GB
- Christa Womser-Hacker
Universität Hildesheim, DE