System And User Centered Evaluation Approaches in Interactive Information Retrieval (SAUCE 2016)

Heather L. O'Brien
iSchool
University of British Columbia
Vancouver, British Columbia
Canada
h.obrien@ubc.ca

Nicola Ferro
Department of Information Engineering
University of Padua
Via G. Gradenigo 6/B,
35131 Padua, Italy
ferro@dei.unipd.it

Hideo Joho
Faculty of Library, Information and Media Science, University of Tsukuba
1-2 Kasuga, Tsukuba, Japan
hideo@slis.tsukuba.ac.jp

Dirk Lewandowski
Department of Information
Hamburg University of Applied Sciences, Hamburg, Germany
dirk.lewandowski@haw-hamburg.de

Paul Thomas
CSIRO
GPO Box 664
Canberra, AUSTRALIA
paul.thomas@csiro.au

Keith van Rijsbergen
School of Computing Science
University of Glasgow
Glasgow G12 8QQ
cornelis.vanrijsbergen@glasgow.ac.uk

ABSTRACT

The purpose of this workshop is to bring together academic and industry interactive information retrieval (IIR) researchers with an interest in evaluation methodologies. The workshop articulates contemporary challenges in the investigation of IIR and invites user- and system-oriented researchers to work collaboratively to address these challenges by combining user- and system-centered methodologies in meaningful ways.

General Terms

Measurement, Experimentation, Human Factors

Keywords

Information retrieval; evaluation; system-centered; user-centered

1. RATIONALE, SCOPE AND NOVELTY OF THE WORKSHOP

In 2013, the organizers of this proposed workshop met at the *Evaluation Methodologies in Information Retrieval Seminar* held at the Schloss-Dagstuhl in Germany [1]. This workshop was built around five themes: frameworks for evaluation; evaluating search within and across sessions; evaluation criteria; user modeling; and evaluation methods and metrics. These themes were explored through presentations by leading interactive information retrieval (IIR) researchers and within working groups.

Our working group focused on reliability and validity, cornerstones of effective evaluation. Our group, comprised of system- and user-oriented researchers, spent a great deal of time exploring these terms and their different meanings within our respective frames of reference. We realized that a lack of a shared understanding of these terms was problematic for the advancement of IIR research - particularly in terms of utilizing and accurately assessing the merits of each other's work. This led us to explore possibilities for collaboration that would bridge the gap between user- and system-centred evaluation approaches.

Nowadays, IIR continues to increase in complexity: user tasks and needs are demanding; data and information systems are rapidly evolving and greatly heterogeneous; and the interaction between users and IR systems is more articulated. For example, consider Web search today: highly diversified results are returned from Web pages, news, social media, image and video search, products and more, and all are merged through adaptive strategies driven by current and previous user-systems interactions. As a result, experimental evaluation needs to appropriately model these evolving tasks, needs, data sources and user interactions. An additional challenge pertains to the anticipated outcome of IIR research and application. It is no longer sufficient to focus solely on precision, recall and satisfaction: successful IIR systems must engage, inform, and relate to users, taking into account single session and more long-term use and re-use.

To effectively support the development of next generation IIR systems, it is necessary to bridge system- and user-oriented evaluation methods. Both approaches have advantages and drawbacks: while system-centered methods ensure greater internal validity, they may fail to take into account user and contextual factors that influence IIR; user-oriented methods may better approximate behavior, affect and cognition, but provide less experimental control of independent variables.

The goal of this workshop is to unite system- and user-centered IIR researchers for the purposes of:

- Sharing different user-centered and system-centred research methods, measures, and tools to foster knowledge exchange;
- Exploring the addition of user-centered evaluation strategies to system-oriented studies, and vice versa; and
- Initiating collaborations between user- and system-oriented researchers to further IIR research.

2. DESCRIPTION OF TOPICS

2.1 SYSTEM-CENTERED EVALUATION

A great deal of progress has been made in information retrieval (IR) on the back of so-called "system-centered" evaluations; that is, evaluations which are either abstract from the user or task entirely,

CHIIR'16, March 13–17, 2016, Carrboro, North Carolina, USA
ACM ISBN: 978-1-4503-3751-9/16/03.
DOI: http://dx.doi.org/10.1145/2854946.2886106

or which treat user characteristics as confounds to be controlled. This has enabled dedicated concentration on aspects of information retrieval algorithms and systems. System-oriented evaluation is founded on the Cranfield methodology [14], which makes use of experimental collections consisting of: sets of documents representing domains of interest; topics, which simulate and abstract actual user information needs; and ground-truth, i.e. the "correct"' answers, where relevant documents for each topic are pre-determined. System outputs, in the form of ranked lists of documents in response to a topic, are then scored with respect to ground-truth using a breadth of performance measures [13].

The benefit of the system-oriented approach is the portability and reusability of test collections: it is possible to directly compare systems, or system variants, over exactly the same tasks and with all sources of variation carefully controlled. As such, there are clear advantages in terms of the number of data points available for analysis, the ability to compare findings across different systems, and experimental control when compared to more user-focused techniques. However, since the "user" – indeed the entire context – is represented in a test collection by only a query and set of relevance judgments, there is plenty of reason to be concerned about the external validity and generalizability of the results.

Recent work has taken offline, system-centered evaluation techniques and begun to address these issue, largely by building user behavior models which can be incorporated in system metrics [12]; widening the pools from which topics and judgments are drawn [14]; and considering slightly more variation and richness in the representation of users and their needs, e.g., [4] or [6]).

Online evaluation approaches are another example of the increased interest in approximating user behavior [3; 9]. These approaches infer preferences about what documents are relevant for a given topic directly from user interactions with ranked result lists by considering a click on a document as a proxy for document relevance. There are two main instantiations of this paradigm: A/B testing and interleaving. A/B testing compares two alternative implementations of a system by switching a random sample of system users either to version A or B, and then comparing the clicks in both systems to determine which is "best." An interleaving method presents users with a ranked result list that contains documents from two or more systems and estimates their preferences by interpreting interactions with the interleaved list, e.g. which documents of which system gather more clicks. Online evaluation approaches operate on the basis that more activity, i.e., mouse clicks, is an indication that one system outperforms the other, or is more preferred by users. Yet, there may be other explanations for increased user activity, some of which may not be positive for user outcomes, such as disorientation, uncertainty, or lack of focused attention on the task. Hence, we need user-centered methods, to construct the *why* around the *what* of user behavior.

2.2 USER-CENTRED EVALUATION

User-centered approaches focus on users' affective and cognitive experiences with IIR systems, and the behaviors they exhibit during use or as a result of interacting with information. Thus, the goal in user-centered evaluation is to understand users' motivations, cognitive involvement and processes, and emotional responses to systems and/or search tasks, and how this relates to system performance and other outcomes. In addition to more traditional IR metrics, such as relevance [16] and informativeness [17], emerging work is exploring complex subjective phenomenon, including user engagement [10], learning [5], and serendipity [11].

Interest in subjective user experience, search environments, and outcomes necessitates the inclusion of more social science methods in IIR research [7], including self-reports. Kelly, Harper and Landau [8] deftly summarize and illustrate the challenges (e.g., inflation, demand effects, acquiescence) associated with self-reporting in their study comparing different modes of administering questionnaires during an IR experiment. Nonetheless, questionnaires and other self-report methods, such as focus groups, interviews, and verbal elicitation, are staples of IIR studies [7]. As such, we require self-report measures that are appropriately constructed and robustly tested to ensure they meaningfully contribute to IIR research. This is accomplished, in part, by a solid theoretical foundation upon which to base self-report measures, and also by establishing their validity in relation to objective measures.

There has been much promising work in this regard. For instance, Arapakis et al. [2] linked self-report measures with eye tracking metrics, mouse clicks, and behavioral performance patterns in online news reading. This work, and others of its kind, illustrates attempts to link subjective and objective measures to obtain a more holistic picture of IIR. If mouse clicks can be equated with gaze and self-report measures, there is potential to understand user behavior at a much larger scale, making it possible to evaluate the experience of millions of users in naturalistic search settings (i.e., the Web) [2]. If large- and small-scale methods can inform each other, then it will enhance the robustness and generalizability of both types of methodological approaches.

3. WORKSHOP

This workshop seeks to explore the benefits and drawbacks of user- and system-centered approaches in greater depth. We will acknowledge and discuss various challenges through keynotes, an expert panel, a world café style discussion session, and position papers. Examples of some of the themes to be addressed in the workshop include:

- Scale, with reference to the number of data points collected: What is lost and gained when we investigate IIR with tens, hundreds, or thousands of users (or systems, or tasks)? How might small- and large-scale approaches inform each other?
- The trade-off between internal and external validity, i.e., in the "wild" versus in the laboratory.
- Relevance has been a long-standing measure of interest in IIR. However, there are other valuable outcomes to be measured pertaining to system effectiveness, user experience, and greater societal and political engagement. How might we develop measurement practices to capture IIR beyond relevance and beyond the evaluation of the system itself?
- Temporality, or the ability to examine a single IIR session and repeated or longitudinal system use. Analytic data collection makes it possible to follow user interactions over time, e.g., repeat visits to a website, but user-centered longitudinal studies are less common but nonetheless vital.
- The use of subjective measures, which may be biased, and objective measures, e.g., behavior or physiology), which may require specialized equipment and knowledge to collect and interpret the data. How do we capture and make sense of both the inner world of users and their observed behaviors?
- The collection of measures during user-system interaction (formative or process-based) and post-interaction (summative). What factors of the search process determine search effectiveness [15]? To what degree are we attending validity and reliability of the measures themselves?

3.1 Workshop Outcomes

It is anticipated that the workshop will: increase awareness of evaluation issues from multiple perspectives; facilitate knowledge exchange and spark innovative ideas. A desired outcome of the workshop is increased uptake of user-centered approaches by systems researchers, and vice versa; collaboration between researchers who previously unknown to each other; and the design of new research studies that would begin addressing current evaluation challenges. The main findings of and the lessons learned in the workshop will be summarized in a report in a journal, such as SIGIR Forum, to trigger further research on the topic.

4. WORKSHOP ORGANIZERS

Heather O'Brien is Assistant Professor at the iSchool, University of British Columbia in Vancouver, Canada. Her research focuses on the measurement of subjective user experience, namely the concept of user engagement. She developed a self-report instrument, the User Engagement Scale (UES), and has been concentrating on its utility, reliability, and validity in various information environments.

Nicola Ferro is Associate Professor at the Department of Information Engineering of the University of Padua. His research interests include IR, its experimental evaluation, multilingual information access, and digital libraries. He is the coordinator of the CLEF evaluation initiative of more than 200 research groups world-wide. He is the Chair of ECIR 2016 and has been the coordinator of PROMISE (2010-2013)..

Hideo Joho is Associate Professor at the Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Japan. His research interests include cognitive and affective interactions between search engines and users. He has also been involved in the development of several test collections e.g., GeoCLEF, NTCIR VisEx, NTCIR Temporalia, and NTCIR Lifelog and a Program Co-Chair of NTCIR-9, 10, and 11 (2010-14).

Dirk Lewandowski is Professor of Information Research and Information Retrieval at the Hamburg University of Applied Sciences, Germany. His research areas are Web Information Retrieval, user behavior in Web search and the impact of Web search on knowledge acquisition in society.

Paul Thomas is Senior Research Scientist at CSIRO, Australia. His research includes evaluation techniques for information retrieval systems, especially models of user behavior and how to build offline methods that predict user performance or preference.

Keith van Rijsbergen is Professor Emeritus in the School of Computing Science, University of Glasgow and Honorary Member of the Computer Laboratory, University of Cambridge. His research spans theoretical and experimental aspects of IR, including the specification and implementation of several theoretical models and the design of appropriate logics to model information flow.

5. ACKNOWLEDGMENTS

Our gratitude for the Schloss Dagstuhl *Evaluation Methodologies in Information Retrieval Seminar* where this conversation began.

6. REFERENCES

[1] Agosti, M., Fuhr, N., Toms, E.G. and Vakkari, P. 2013. Evaluation Methodologies in Information Retrieval (Dagstuhl Seminar 13441). *Dagstuhl Report.* 3, 10 (2013), 92-126.

[2] Arapakis, I. Lalmas, M., Cambazoglu, B. B., Marcos, M.-C. and Jose, J.M. 2014. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *J. Assoc. Inform. Sci. Tech.* 65, 10 (March. 2014), 1988-2005.

[3] Chuklin, A., Markov, I. and de Rijke, M. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers, USA.

[4] Dean-Hall, A., Clarke, C. L. A., Kamps, J., Thomas, P. and Voorhees, E. 2014. Overview of the TREC 2014 Contextual Suggestion Track. In *Proceedings of the Text Retrieval Conference*. National Institute of Standards and Technology.

[5] Freund, L., Gwizdka, J., Hansen, P., He, J., Kando, N. and Rieh, S Y. (2014). *Searching as Learning Workshop, Information Interaction in Context* (IIiX) *Conference* (Regensburg, Germany, August 30, 2014).

[6] Gurrin, C., Albatal, R., Joho, H, and Hopfgartner, F. 2015. Lifelog: Pilot Task of NTCIR-12.

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

[8] Kelly, D., Harper, D.J. and Landau, B. 2008. Questionnaire mode effects in interactive information retrieval experiments. *Inform. Process. Manage.* 44, 1 (January. 2008), 122-141.

[9] Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1): 140–181.

[10] Lalmas, M., O'Brien, H. and Yom-Tov, E. 2014. *Measuring User Engagement*. Morgan & Claypool.

[11] McCay-Peet, L. and Toms, E. G. 2011. Measuring the dimensions of serendipity in digital environments. *Information Research: An International Electronic Journal* 16, 3 (September 2011): n3.

[12] Moffat, A., Thomas, P. and Scholer, F. (2013). Users versus models: What observation tells us about effectiveness metrics. In *Proceeding of the 22th International Conference on Information and Knowledge Management*, 659–668. ACM.

[13] Sakai, T. 2014. Metrics, Statistics, Tests. In Ferro, N., ed, *Bridging Between Information Retrieval and Databases PROMISE Winter School 2013, Revised Tutorial Lectures*, 116–163. Lecture Notes in Computer Science (LNCS) 8173, Springer, Heidelberg, Germany.

[14] Sanderson, M. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4(4), 247–375.

[15] Scholer, F., Moffat, A. and Thomas, P. 2013. Choices in batch information retrieval evaluation. In *Proceedings of the Australasian Document Computing Symposium*, 74-81. Brisbane, Australia.

[16] Su, L.T. 1992. Evaluation measures for interactive information retrieval. *Inform. Process. Manage 28*,(4), 503-516.

[17] Tague-Sutcliffe, J. 1992. Measuring the informativeness of a retrieval process. In *Proceedings of the 15th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Copenhagen, Denmark, June 21-24, 1992). ACM, New York, NY, 23-26.