Lessons Learned from the CHiC and SBS Interactive Tracks: A Wishlist for Interactive IR Evaluation

Toine Bogers
Science and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
A.C. Meyers Vænge 15
2300, Copenhagen
toine@hum.aau.dk

Maria Gäde
Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Dorotheenstr. 26
10117, Berlin
maria.gaede@ibi.hu-berlin.de

Mark M. Hall
Department of Computer Science
Edge Hill University
St Helens Road, L39 4QP
United Kingdom
mark.hall@edgehill.ac.uk

Vivien Petras
Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Dorotheenstr. 26
10117, Berlin
vivien.petras@ibi.hu-berlin.de

Mette Skov
E-learning Lab
Department of Communication & Psychology
Aalborg University
Rendsburggade 14
9000, Aalborg
skov@hum.aau.dk

ABSTRACT

Over the course of the past two decades, the Interactive Tracks at TREC and INEX have contributed greatly to our knowledge of how to run an interactive IR evaluation campaign. In this position paper, we add to this body of knowledge by taking stock of our own experiences and challenges in organizing the CHIC and SBS Interactive Tracks from 2013 to 2016 in the form of a list of important properties of any future IIR evaluation campaigns.

KEYWORDS

Interactive IR, evaluation, information seeking

1 INTRODUCTION

Evaluation has always played an important role in IR research. Traditional IR evaluation campaigns have typically focused on measuring system performance in controlled environments, but over the past two decades, IR research has expanded to include a more user-based perspective, which considers the interactive nature of information seeking behavior and the individual contexts surrounding it [13]. However, IIR evaluation campaigns remain relatively ad-hoc and disjointed and still face significant challenges [14].

The first large-scale IIR evaluation campaign was the TREC Interactive Track (1997-2002) [17], followed by the INEX Interactive Track (2004-2010) [18, 20]. These campaigns taught the community important lessons, not the least of which was the difficulty of maintaining a stable evaluation framework, because of the lessons learned after each iteration. Participant recruitment was often difficult, resulting in small convenience samples of students and colleagues. The campaigns also highlighted the influence of GUI design on interaction behavior and engagement with the study. The Cultural Heritage in CLEF (CHiC) Interactive Task took up the mantle of IIR evaluation again in 2013, which continued with the interactive Social Book Search (SBS) task (2014-2016) [8–10, 22].

In this position paper, we take stock of our experiences in the CLEF Interactive Tracks and the lessons learned from previous IIR evaluation campaigns to list important properties for future campaigns. Complex search tasks—the theme of this workshop—are at the heart of information seeking studies. We hope to contribute to the development of an IIR evaluation framework that mirrors the complex nature of information seeking behavior. We make no claims about the completeness of this list, but instead see it as a starting point for iterative refinement and improvement.

2 PROPERTIES OF IIR EVALUATION

2.1 Continuity

The success of any IIR evaluation campaign depends first and foremost on getting a sustainable number of researchers to participate. The ideal IIR evaluation campaign should run continuously over a multi-year period. NewsReel [15] is an example of a campaign that has moved from a rigid campaign schedule to a continuous campaign, running throughout the year.

A continuous campaign has several advantages, including increased researcher participation. Our experience shows that a common barrier to participation is fitting it into a university's schedule, both in terms of teaching obligations as well as being able to use the evaluation campaign as a teaching tool. Clashes and misalignment with these professional obligations was the most commonly stated reason for not participating in the different CLEF Interactive Tracks. A continuous campaign would also aid in participant recruitment as students or other participant groups are not always

CHIIR 2017 Workshop on Supporting Complex Search Tasks, Oslo, Norway.

available throughout the year. System maintenance could easily be accounted for by short, fixed 'downtime' periods.

Workshops could then be organized to highlight interesting research and spark discussion. We suspect that the 12-month TREC or CLEF cycles are too short to achieve meaningful results with IIR studies, to attract enough participants, and account for the longer time it takes to run user studies. Data collected up to a particular point could be presented at workshops in 18-24 month intervals—ideally attached to different conferences to ensure a varied potential audience. Naturally, it should be possible at any given point in time to extract the collected data for a particular research group that is participating, for instance to support its use in teaching.

2.2 Complexity

In real life, searching for information takes place in a broader context. IIR studies have long tried to capture and represent this context through instruments such as simulated work tasks [6]. However, we believe that future IIR evaluation ought to broaden its scope to the entire information seeking process, both individual and collaborative [12]. Searching is only one aspect of information behavior and is often combined with browsing, exploration or interaction with a recommender system. Moreover, information behavior often takes place across and between different devices (desktop vs. smartphone) and modalities (digital vs. paper).

IIR studies should be representative of the real-life variety in users types, domains, and system designs. A comprehensive IIR research program would investigate domain-specific and professional information seeking alongside casual leisure searching, combining specialized systems as well as general web search for resources such as professional documents, books, music, traveling etc. Not all contexts may be studied within the same IIR evaluation campaign, but some could be combined or contrasted. The CHiC and SBS Interactive Tracks aimed to do this through variety in the simulated work tasks that were used, but we believe this was only a first step in the right direction.

Related phenomena such as serendipity should also be incorporated into the IIR evaluation wherever possible, drawing upon the scant lessons that have been learned about evaluating this in a laboratory setting [3]. In general, the complex nature of information behavior should be better represented in IIR evaluation campaigns.

2.3 Flexibility

While the core of an IIR evaluation campaign should be as stable as possible to enable comparisons over time, we wish to advocate for as much flexibility as possible around this stable core to allow for different research questions and approaches within a common IIR framework.

When it comes to individual research contexts and interests of the researchers, experiments often need to be flexible with respect to gathering of background and contextual information. Different research teams may want to focus on different aspects of search behavior or interfaces and may therefore need to ask different pre- and post-task questions of their participants. Ideally, the system would let researchers easily plug in and swap out different questions depending on the research team's interests. Ideally, such question sets could then be stored in the system and used in future experiments by other teams.

Another example is the flexibility in choosing different search interfaces to study the effects of the GUI on information seeking behavior. This was used to great effect in the 2015 and 2016 SBS Interactive Tracks [8, 9] to examine how different interfaces can support the different search stages.

Increasing flexibility is not without its challenges however. Exposing different participant groups to different sets (and sizes) of pre-/post-task questions or interfaces may lead to different levels of fatigue or learning effects, which could influence the integrity of the collected data and any future comparisons. The optimal trade-off between stability and flexibility is far from a solved problem.

2.4 Realism

Another essential property of user-based IR evaluation is realism, which applies to different aspects of the evaluation process. One is the use of a document collection that is realistic in size and content variety. It ensures that participants do not become frustrated, because they cannot find relevant documents, which could affect search behavior as well as engagement.

Ensuring realism also includes participant sampling. Ideally, an IIR evaluation campaign recruits participants that are a realistic representation of the general target population to avoid the introduction of biases [7, p. 241]. Future IIR studies should have a high variety of participants rather than focusing on the academic sector. The influence of the participants' motivation to be part of an IIR study also needs to be taken into account.

Another element is the use of realistic representations of real-world information needs. Typically, simulated work tasks [6] are used to support the search behavior of participants in the form of cover stories that include extra context about the background task. Their formulation based on "the user's own personal information need or in relation to a set of simulated needs" [6, p. 229] is essential for using them to their full potential [4–6].

The question of how best to generate such realistic information needs remains largely unanswered. One possibility is to use search log analysis to identify popular queries and generate a cover story around them. However, such cover stories would always be a best-guess justification. The 2016 SBS Interactive Track [8] was the first edition of the CLEF Interactive Tracks where we successfully used real-world examples of complex information needs taken from the LibraryThing forums as additional, optional work tasks. These tasks were rich in variety and detail, and this sentiment was also expressed by our participants. This does yield two new questions: (1) whether participants would be able to tell the difference between simulated and real-world work tasks, and (2) whether any resulting differences in information seeking behavior would observable and meaningful.

In addition to realism in the information needs, the evaluation data set also should represent a realistic and engaging scenario. The CHiC data set was based on an extract from Europeana and demonstrated a wide range of topics, however the individual items in the

data-set were often sparse in their information and primarily metadata based. In the SBS tasks, the data-set used was based on data aggregated from Amazon, LibraryThing, and the British Library. This provided a realistic and detailed data-set. However, over the duration of the SBS tasks, the data became more and more outdated and users often tried to search for newer books that the data-set did not yet contain. Thus how to have a realistic, engaging, and up-to-date data-set, while at the same time maintaining comparability across iterations of the evaluation is an open question.

Simulated system prototypes have the advantage of full system and content control as well as the possibility to test alternative features and interactions. On the other side, the user experience highly depends on the look and feel of the interfaces and interaction design. Systems need to meet user needs and expectations in order to reduce their influence on behavior and learning effects. Training tasks can be used to create similar requirements for participants as well as familiarity with unknown systems [16].

2.5 Measurability

IIR studies use a variety of quantitative measures, but also qualitative descriptions of information seeking behavior. Both quantitative and qualitative approaches for result representations in this research are rarely comparable—not only because of their different study and audience foci, but also because the reported measures and descriptions are not based on agreed-upon standards. The comparability of results needs to be ensured for long-term-oriented IIR evaluation campaigns.

While the majority of studies gather a variety of performance, interaction, and usability data, no framework for the interpretation of these "numbers" exists [14]. Depending on expected behaviors or user goals, the session duration, number of queries and results viewed can be interpreted in different ways. For example, a long session duration can be a sign of both a good or bad user experience. Binary relevance assessments are rarely applicable when dealing with real users. During an information seeking process, usefulness and satisfaction seem to play a more important role but need to be operationalized for IIR research [1].

Any successful IIR evaluation framework needs to include standardized measures and description approaches, possibly with agreed-upon interpretations—comparable to the common effectiveness measures used in system-based IR evaluation.

Full comparability also requires that IIR experiments can easily be replicated by other researchers. A number of tools have been proposed to enable this replication. Toms et al. (2004) designed and implemented the WiIRE (Web-based Interactive Information Retrieval) system [21], which was used in the TREC-11 Interactive Track, an updated version of which was used at INEX 2007 [23]. Hall et al. (2013) developed the PyIRE system [11], which has been used in the CHiC and SBS Interactive Tracks. Other systems include SCAMP [19] and CIRSE [2]. However, none of these have been used outside of the initial IIR evaluation they were designed for, primarily due to the complexity of setting up the systems and the lack of documentation.

3 WHAT NEXT?

We believe that an essential next step for our wishlist of IIR evaluation properties is to discuss them with the community. The list should then be extended and elaborated upon using other related work with a particular focus on *how* to approach any challenges. Another important step is to investigate how different properties interact and interfere with each other (e.g., flexibility & continuity). A new IIR evaluation campaign and related researcher community could then be designed around these properties to expand our understanding of information (seeking) behavior.

Another aspect of IIR evaluation for future work is the question of real vs simulated users. As the usability and visual design of the evaluation interface have a significant impact on the IIR evaluation process, the question of how these factors should be integrated into user simulation is one that will need further attention.

REFERENCES

[1] Nicholas J Belkin, Michael Cole, and Jingjing Liu. 2009. A Model for Evaluation of Interactive Information Retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 7–8.

[2] Ralf Bierig, Jacek Gwizdka, and M. J. Cole. 2009. A User-centered Experiment and Logging Framework for Interactive Information Retrieval. In *Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and Interpreting User Interactions in Information Search and Retrieval*. 8–11.

[3] Toine Bogers, Rune Rosenborg Rasmussen, and Louis Sebastian Bo Jensen. 2013. Measuring Serendipity in the Lab: The Effects of Priming and Monitoring. In *Proceedings of the iConference 2013*. 703–706.

[4] Pia Borlund. 2003. The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research* 8, 3 (2003).

[5] Pia Borlund. 2016. A Study of the Use of Simulated Work Task Situations in Interactive Information Retrieval Evaluations: A Meta-Evaluation. *Journal of Documentation* 72, 3 (2016), 394–413.

[6] Pia Borlund and Peter Ingwersen. 1997. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation* 53, 3 (1997), 225–250.

[7] Donald O. Case and Lisa M. Given. 2016. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (4th ed.). Emerald Group Publishing, Bingley, UK.

[8] Maria Gäde, Mark Michael Hall, Hugo C. Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skov, Toine Bogers, and David Walsh. 2016. Overview of the SBS 2016 Interactive Track. In *Working Notes of the CLEF 2016 Conference (CEUR Workshop Proceedings)*, Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald (Eds.), Vol. 1609. CEUR-WS.org, 1024–1038.

[9] Maria Gäde, Mark Michael Hall, Hugo C. Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skov, Elaine Toms, and David Walsh. 2015. Overview of the SBS 2015 Interactive Track. In *Working Notes of the CLEF 2015 Conference (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, Gareth J. F. Jones, and Eric SanJuan (Eds.), Vol. 1391. CEUR-WS.org.

[10] Mark Michael Hall, Hugo C. Huurdeman, Marijn Koolen, Mette Skov, and David Walsh. 2014. Overview of the INEX 2014 Interactive Social Book Search Track. In *Working Notes of the CLEF 2014 Conference (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij (Eds.), Vol. 1180. CEUR-WS.org, 480–493.

[11] Mark M Hall, Spyros Katsaris, and Elaine Toms. 2013. A Pluggable Interactive IR Evaluation Work-bench. In *European Workshop on Human-Computer Interaction and Information Retrieval*. 35–38. http://ceur-ws.org/Vol-1033/paper4.pdf

[12] Preben Hansen, Chirag Shah, and Claus-Peter Klas. 2015. *Collaborative Information Seeking*. Springer.

[13] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1–2 (2009), 1–224.

[14] Diane Kelly and Cassidy R Sugimoto. 2013. A Systematic Review of Interactive Information Retrieval Evaluation Studies, 1967–2006. *Journal of the American Society for Information Science and Technology* 64, 4 (2013), 745–770.

[15] Benjamin Kille, Andreas Lommatzsch, Gebrekirstos G. Gebremeskel, Frank Hopfgartner, Martha Larson, Jonas Seiler, Davide Malagoli, András Serény, Torben Brodt, and Arjen P. de Vries. 2016. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-Time Stream-Recommendation Algorithms. In *CLEF '16: Proceedings of the 7th International Conference of the CLEF Association (Lecture Notes in Computer Science)*, Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro

(Eds.), Vol. 9822. Springer, 311–331.

[16] Marijn Koolen, Toine Bogers, Maria Gäde, Mark A. Hall, Hugo C. Huurdeman, Jaap Kamps, Mette Skov, Elaine Toms, and David Walsh. 2015. Overview of the CLEF 2015 Social Book Search Lab. In *CLEF '15: Proceedings of the 6th International Conference of the CLEF Association (Lecture Notes in Computer Science)*, Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth J. F. Jones, Eric SanJuan, Linda Cappellato, and Nicola Ferro (Eds.), Vol. 9283. Springer, 545–564.

[17] Paul Over. 2001. The TREC Interactive Track: An Annotated Bibliography. *Information Processing & Management* 37, 3 (2001), 369–381.

[18] Nils Pharo, Thomas Beckers, Ragnar Nordlie, and Norbert Fuhr. 2011. Overview of the INEX 2010 Interactive Track. In *INEX '10: Proceedings of the Ninth International Workshop of the Initiative for the Evaluation of XML Retrieval*, Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman (Eds.). Springer, Berlin, Heidelberg, 227–235.

[19] Gareth Renaud and Leif Azzopardi. 2012. SCAMP: A Tool for Conducting Interactive Information Retrieval Experiments. In *IIiX '12: Proceedings of the 4th Information Interaction in Context Symposium*. ACM, 286–289.

[20] Anastasios Tombros, Birger Larsen, and Saadia Malik. 2005. The Interactive Track at INEX 2004. In *INEX '04: Proceedings of the Third International Workshop of the Initiative for the Evaluation of XML Retrieval*, Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik (Eds.). Springer, Berlin, Heidelberg, 410–423.

[21] Elaine G Toms, Luanne Freund, and Cara Li. 2004. WiIRE: The Web Interactive Information Retrieval Experimentation System Prototype. *Information Processing & Management* 40, 4 (2004), 655–675.

[22] Elaine G. Toms and Mark M. Hall. 2013. The CHiC Interactive Task (CHiCi) at CLEF 2013. In *Working Notes of the CLEF 2013 Conference (CEUR Workshop Proceedings)*, Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro (Eds.), Vol. 1179. CEUR-WS.org.

[23] Elaine G. Toms, Heather O'Brien, Tayze Mackenzie, Chris Jordan, Luanne Freund, Sandra Toze, Emilie Dawe, and Alexandra Macnutt. 2008. Task Effects on Interactive search: The Query Factor. In *Focused Access to XML Documents*. Springer, 359–372.