Report on the First International Workshop on the Evaluation on Collaborative Information Seeking and Retrieval (ECol'2015)

Leif Azzopardi[1], Jeremy Pickens[2], Tetsuya Sakai[3], Laure Soulier[4], Lynda Tamine[4]
[1] School of Computing Science, University of Glasgow - UK *Leif.Azzopardi@glasgow.ac.uk*
[2] Catalyst Repository Systems - USA *jpickens@catalystsecure.com*
[3] Waseda University - Japan *tetsuyasakai@acm.org*
[4] Université de Toulouse UPS IRIT - France {*soulier,tamine*}*@irit.fr*

November 5, 2015

Abstract

The workshop on the evaluation of collaborative information retrieval and seeking (ECol) was held in conjunction with the 24^{th} Conference on Information and Knowledge Management (CIKM) in Melbourne, Australia. The workshop featured three main elements. First, a keynote on the main dimensions, challenges, and opportunities in collaborative information retrieval and seeking by Chirag Shah. Second, an oral presentation session in which four papers were presented. Third, a discussion based on three seed research questions: (1) In what ways is collaborative search evaluation more challenging than individual interactive information retrieval (IIIR) evaluation? (2) Would it be possible and/or useful to standardise experimental designs and data for collaborative search evaluation? and (3) For evaluating collaborative search, can we leverage ideas from other tasks such as diversified search, subtopic mining and/or e-discovery? The discussion was intense and raised many points and issues, leading to the proposition that a new evaluation track focused on collaborative information retrieval/seeking tasks, would be worthwhile.

1 Introduction

The paradigm of Collaborative Information Seeking/Retrieval (CIS)/(CIR) refers to methodologies and technologies that support collective-knowledge sharing within a team in order to solve a shared complex problem [5, 6]. One main challenge in CIS/CIR is to satisfy the mutual beneficial goals of both individual users and the collaborative group while maintaining a reasonable level of cognitive effort [11]. Evaluating CIS/CIR is problematic because there is a variety of confounding factors such as the multi-user context, the exploratory aspect of the search through multi-session search activities, the multiplicity of relevance factors, the individual vs. collective value of relevance, and the search interfaces supporting the collaborative interactions. In previous work, CIS/CIR evaluation relies on simulation-based

protocols [4, 16], user or log-based studies [9, 12, 15] and metrics leveraging the individual vs. collaborative dimensions of search effectiveness [11]. The most common evaluation strategy is to undertake both qualitative and quantitative evaluation according to the search tasks characteristics and evaluation objectives such as cognitive effort, usability, and individual vs. collective effectiveness. However, it is unclear how to best evaluate CIS/CIR. For "non-collaborative" information retrieval and seeking tasks substantial research advances in evaluation have been made, generally through international evaluation campaigns such as TREC, CLEF and NTCIR as they focus the attention of the community on particular tasks. To date and to the best of our knowledge there has not been a major evaluation campaign launched for collaborative information seeking and retrieval tasks. Yet, collaborative searches are common practice. We believe that there is an important need to investigate the evaluation challenge in CIS/CIR with the hope of building standardized evaluation frameworks that would foster the research area. Accordingly, the ECol 2015 workshop [1] on the evaluation of collaborative information retrieval and seeking[1] set out to discuss and explore the following questions (without being exhaustive):

- What are the relevant factors for evaluating CIS/CIR scenarios, and what are the appropriate evaluation protocols?
- What are the properties of current evaluation metrics in CIS/CIR, and to what extent these properties are adequate to measure the benefice vs. the cost of collaboration?
- How to leverage between individual relevance and collective relevance within the evaluation of a CIS/CIR setting, and how to measure the satisfaction of mutual beneficial goals surrounding individual and collective relevance?
- Should the collaborative interfaces be evaluated similarly to usual search interfaces?
- How to design standard evaluation framework for CIS/CIR?

After the introduction to the workshop running through the above questions, we started with a table round in which each participant introduced him or herself. The workshop was composed of students, industrialists and academics with different levels of experience. This was followed an engaging and interactive keynote talk by Chirag Shah which helped to share the understanding of the concepts, dimensions, and challenges ahead in the area. After the break, four paper were presented within two interactive and open sessions. Finally, all the participants took part in an in-depth discussion around three seed research questions:

1. In what ways is collaborative search evaluation more challenging than IIR evaluation? Any possible solutions?
2. Would it be possible/useful to standardise experimental designs/data for collaborative search evaluation? Why or why not? Can we do any better?
3. For evaluating collaborative search, can we leverage ideas from other tasks such as diversified search/subtopic mining?

Below is a summary of the events of the day, along with preliminary conclusions, actions and next steps.

[1] urlhttp://irit.fr/ECol2015/

2 Keynote

Chirag Shah, Associate Professor at the University of Rutgers, kicked off the discussions with his keynote entitled, "Social and Collaborative information seeking: state of the union"[2] [10]. During this talk, Chirag presented an overview of social and collaborative information seeking frameworks, particularly from the search iteration side in which user-to-user (social) interaction dimensions are prevalent. The state-of-the-art of collaborative information seeking approaches have been reviewed and differentiated: ethnographic, passive support, and active support. The evaluation issue has been addressed from both the methodologies and measurements perspectives. The talk also highlighted the ideal domain-applications of collaborative information seeking (legal, education, health, intelligence analysis, etc.) and pointed out the remaining challenges and relevant opportunities in the research area. Chirag then gave an overview of the main scientific events dealing with the topic and a roadmap of (i) what we currently know: usefulness of collaboration, collaboration cost, issues related to awareness, privacy, control, transparency, communication as well as coordination, and (ii) what should be mainly investigated in the future: fundamentals of social and collaborative information seeking, the impact on society, and the computational models that can be designed and evaluated. The keynote session was highly interactive leading to a fruitful discussion, but also raising many issues and challenges.

3 Paper Presentations

Below, we briefly summarize each of the four paper presentations.

Yamamoto et al. [18] presented a user study for evaluating the impact of roles on the search effectiveness and query formulations in a collaborative search setting. The authors particularly studied the asymmetric roles of gatherer and surveyor and highlighted the differences between these roles in terms of the search precision and the similarity between the issued queries. They wanted to understand the relationships between the roles and search behaviors to draw insights into developing algorithms such as query suggestions or document rankings such that they are adaptive to the roles and behaviors.

Wang [17] focused on collaborative information retrieval within peer review platforms, Specifically, she described how the open science movement which has resulted in the post-publication peer-review (PPPR) model creates a new type of scientific collaboration, in which authors, reviewers, and registered users, share information and shape scientific findings. She first gave an overview of the collaboration manifestations within such platforms and then highlight the research opportunities that would enhance the collaborative aspect of the peer review process. She argued that a new research initiative is needed to first understand collaborative information behaviors in this context and then to determine how to evaluate the systems designed to support the process.

Knight [7] presented a different kind of CIS/CIR tasks within an learning environment. Specifically, groups of students were given the task of writing a report on the "best-supported claims" around the risks of a substance. The success of task is based on what information is found by the group and how well it is synthesized. Thus the interaction with each other and the system will dictate how well they perform on the task. To evaluate their performance, he

[2] The keynote slides are available at ?????.

presented a number of different possible evaluation metrics for measuring the effectiveness of the collaborative learning task, that could be used. However, he argues that the addition of social or/and collaborative elements to task and system design adds an additional layer of complexity to such evaluation, and questioned whether the existing simple metrics typically used were adequate.

Knight and Mitsui [13] then presented their work on the impact and effect of time on learning and collaboration, and how it is an important and relevant dimension for evaluating a collaborative learning task. The authors review the temporal behavioural features used for the study of CIR search tasks. The authors also point out on the lack of time-oriented evaluation metrics.

4 Discussion

The discussion was split into three parts based on the interests of the group, these were: (i) going beyond individual interactive information retrieval (IIIR), (ii) suggestions for robust experimentation and evaluation, and (iii) where to next for CIS/CIR evaluation?

First, participants discussed possible challenges that are specific to collaborative information retrieval and seeking (CIR/CIS), as opposed to individual interactive information retrieval (IIIR). The following are the main points raised. In IIIR user experiments, we hire participants independently. In contrast, in CIR/CIS user experiments, we need to determine not only how to sample participants but also how to couple them. The success of the CIR/CIS approach probably depends on how people are coupled. Thus, in addition to system, task and user effects, synergy effect plays a significant part. Combining multiple users implies that the variance is going to be very high. IIIR is about finding relevant information, or possibly finding out first about the user's own information need through interaction. In contrast, CIS also involves finding about the user's partner (e.g. surveyor finding out about his gatherer and vice versa). Also, it may involve consensus making within the collaboration group if what constitutes relevant information or what comprises a satisfactory solution to a given problem (e.g. find a "good" bar for a post-workshop party). Another way to put it is that in CIS, we are interested not only in the final outcome but also in the entire process. Because of the collaborative nature of CIR/CIS, finding realistic and suitable search tasks for experimentation is generally harder than IIIR.

A more philosophical point raised was where do we draw the line when thinking of collaborative vs. individual search and retrieval vs? Is whether collaboration is working with a system, or whether it involves working with others or your future self. But what if we are collaborating with artificially intelligent agents. Would these be considered collaborative, too? When creating CIS/CIR experiments, how should the group be formed? Should roles be imposed, or should we let participants adopt or assume particular roles in the collaborative process? Then in order to determine the overall performance and determine whether the collaborative was indeed beneficial, do we also need to measure the individual differences of searchers beforehand? if so, how do we normalise the differences? This lead to the point that how do we delineate the performance of the system from the performance and interaction of the users? Should we measure the performance in terms of gain per time [14] or consider how economical the collaborative interactions are [2]?

In TREC style experiments most of the context and interaction is removed or ignored to make it simple, tractable and reproducible. However, in CIS/CIR the multiple participants

(even 2 collaborators plus search system) and the multiple roles that can be assumed during the process add even greater variance than in the individual setting [8]. Furthermore, during the CIS/CIR process the information gathered often needs to be externalized and shared with others in the group so that they can all make sense of it, and to help ensure that the space is explored efficiently. This adds other complexities, i.e. CIR/CIS has higher task dependence, goes beyond searching, and incorporates sense making and learning. It was not clear whether it was possible or sensible whether these could be broken down into discrete measurable units.

The next set of issues discussed by participants focused on practical strategies for robust experimentation for CIR/CIS. The following ideas were shared. It was felt that experiments should be task-driven, and task accomplishment needs to be defined clearly (e.g. task accomplishment example: user learns how to conduct a t-test with given data, and understands the underlying relevant concepts in statistics). A range of tasks suitable for CIS experiments that were noted include: learning about a controversial topic (different opinions in different documents), travel planning, finding a suitable collaborator for a given research topic, asymmetric collaborations such as a doctor and a patient trying to find information on a disease, and recall-oriented tasks such as e-discovery. It was felt that for each of these tasks an reasonably clear objective could be defined. Once task accomplishment has been clarified, we wondered whether the goal of CIS should be to accomplish it more efficiently than IIIR, or whether other aspects needed to be considered?

Another issue that arose was when designing experiments was regarding the motivation and incentivisation of task. It was felt that an artificial task could seriously affect the outcomes reports, and so it is important to carefully construct tasks that provide a reason for the collaboration and to motivate the participants to perform the task appropriately and diligently (much like when designing simulated work task for individuals the task need to be realistic, believable and relevant [3]). This led to an interesting suggestion, whether participants are involved in adversial collaborations, where another participant is working against, rather than with the person, whether it be competing goals, lack of motivation on behalf of one participant, different interpretations of the task, or just distracting. This again raised the question of what is being measured, the human-computer interaction or the human-human-computer interaction.

In term of how to assign roles, it might be useful to let the participants decide their roles (e.g. surveyor/gatherer) for themselves during the search sessions (perhaps via finding out about what the participant him/herself is good at, and what the partner is good at). However, this itself adds in the complexity of forming teams, and again are we trying to measure how well people work together or how well they can use the system, and does that depend on the roles that adopt?

In terms of statistics, while IIIR has a lot of previous work from with which we can conduct power analysis and sample size design (e.g. determining the right number of participants for a new experiment), CIS lacks such data. Consequently, more data needs to be accumulated, from which we can estimate variance estimates for sample size design.

Building a test collection for CIS: collecting real collaborative search logs is difficult. However, if we can collect IIIR search logs and extract sessions for similar information needs, it may be possible to artificially match them up and simulate collaboration between these different users. In some CIS tasks, not only relevance but also diversity is important. It might be possible to incorporate ideas from diversity-related tasks or even from other tasks.

At the end of the discussions session, participants talked about the future plans for

evaluation of CIS/CIR. One possibility that was raised was to propose an ECol track to TREC or another evaluation forum. There could be a main collaborative ad-hoc search task, plus subtasks for finding the best collaborator from a given pool of people or even finding the optimal collaborative group size. Of course, an appropriate task would need to be selected - or perhaps one of the existing track tasks could be used. However, it was clear from the day that there was a multitude of issues that needed to be addressed in terms of evaluation, so it was felt that a second ECol workshop would be of benefit, where one or two tasks and experimental designs could be mooted in more detailed (to be held in conjunction with future SIGIRs, ECIRs or CIKMs).

5 Conclusions

The first edition of the ECol workshop provided a comprehensive overview of current research work on collaborative information retrieval and seeking, perceived as an emerging and important topic in the short-term. The discussion gave raise to the importance of designing a track for evaluating a collaborative search setting in order to foster the research in the area.

Concrete actions were planned: (i) meetings within the coming ECIR and SIGIR conference editions in order to discuss around the evaluation track design and (ii) a second edition of the Ecol workshop in conjunction with the SIGIR'2017 conference in Tokyo.

6 Acknowledgments

We thank ELIAS fundation for their grant #5907. We also thank the CIKM organizers, Chirag Shah, the authors and the participants for their contributions and involvement in the discussions as well as our programme committee members for their timely reviews.

References

[1] *ECol '15: Proceedings of the 2015 Workshop on Evaluation on Collaborative Information Retrieval and Seeking*, New York, NY, USA, 2015. ACM.

[2] L. Azzopardi. Modelling interaction with economic models of search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 3–12, 2014.

[3] P. Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, 8(3), 2003.

[4] C. Foley and A. F. Smeaton. Synchronous collaborative information retrieval: Techniques and evaluation. In *Springer ECIR*, pages 42–53, 2009.

[5] J. Foster. Collaborative information seeking and retrieval. *Annual Review of Information Science and Technology*, 40(1), 2006.

[6] P. Hansen and K. Järvelin. Collaborative information retrieval in an information-intensive domain. *Inf. Process. Manage.*, 41(5):1101–1119, 2005.

[7] S. Knight. Learning indicators in scis tasks. In *Proceedings of the first workshop on the evaluation of collaborative information retrieval and evaluation (ECol), in conjuction of the conference on information knowledge and management (CIKM)*, 2015.

[8] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled evaluation over query variations: Users are as diverse as systems. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1759–1762. ACM, 2015.

[9] J. Pickens, G. Golovchinsky, C. Shah, P. Qvarfordt, and M. Back. Algorithmic mediation for collaborative exploratory search. In *ACM SIGIR*, pages 315–322, 2008.

[10] C. Shah. Social and collaborative information seeking: State of the union. In *Proceedings of the 2015 Workshop on Evaluation on Collaborative Information Retrieval and Seeking*, ECol '15, pages 1–1, New York, NY, USA, 2015. ACM.

[11] C. Shah and R. González-Ibáñez. Evaluating the synergic effect of collaboration in information seeking. In *ACM SIGIR*, pages 913–922, 2011.

[12] C. Shah, J. Pickens, and G. Golovchinsky. Role-based results redistribution for collaborative information retrieval. *Inf. Process. Manage.*, 46(6):773–781, 2010.

[13] K. Simon and M. Matthew. Temporal analysis in epistemic scis tasks. In *Proceedings of the first workshop on the evaluation of collaborative information retrieval and evaluation (ECol), in conjuction of the conference on information knowledge and management (CIKM)*, 2015.

[14] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 95–104, 2012.

[15] L. Soulier, C. Shah, and L. Tamine. User-driven system-mediated collaborative information retrieval. In *SIGIR*, pages 485–494, 2014.

[16] L. Soulier, L. Tamine, and W. Bahsoun. On domain expertise-based roles in collaborative information retrieval. *Inf. Process. Manage.*, 2014, to appear.

[17] P. Wang. Collaborative interaction behavior in the era of open science movement. In *Proceedings of the first workshop on the evaluation of collaborative information retrieval and evaluation (ECol), in conjuction of the conference on information knowledge and management (CIKM)*, 2015.

[18] T. Yamamoto, M. Yamamoto, and K. Tanaka. Analyzing effect of roles on search performance and query formulation in collaborative search. In *Proceedings of the first workshop on the evaluation of collaborative information retrieval and evaluation (ECol), in conjuction of the conference on information knowledge and management (CIKM)*, 2015.