Report on the CHIIR 2018 Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2018)

Gareth J. F. Jones
Dublin City University, Ireland
gareth.jones@dcu.ie

Nicholas J. Belkin
Rutgers University, USA
belkin@rutgers.edu

Séamus Lawless
Trinity College Dublin, Ireland
seamus.lawless@scss.tcd.ie

Gabriella Pasi
University of Milano-Bicocca, Italy
pasi@disco.unimib.it

Abstract

The Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2018) was held in conjunction with the ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR 2018) in New Brunswick, USA. The purpose of WEPIR 2018 was to bring together researchers from different backgrounds, interested in advancing the evaluation of personalisation in information retrieval. The workshop focused on developing a common understanding of the challenges, requirements and practical limitations of evaluation of personalisation in information retrieval.

1 Introduction

One of the key goals of information retrieval research is to advance the development of search applications which enable individual searchers to satisfy their information needs more effectively and efficiently. Given that the information need underlying their engagement with a search application relates to their personal needs for information, the search process should seek to retrieve information most likely to be useful to them personally.

Identifying information useful to individual searchers potentially requires search applications to make use of all available information relating to the searcher to be incorporated into the search process. There are many ways in which personal information might be modeled and represented within some form of user model, and how this model might be used within the information retrieval process. In order to determine how best to implement a personalised information application, an evaluation strategy is required.

WEPIR 2018 sought to develop a shared understanding of the challenges involved in evaluating personalised information retrieval (PIR) in interactive search settings, and to identify issues and topics for consideration in evaluation of personalisation in information retrieval based on both user-centered studies and laboratory-based algorithmic research.

The workshop began with a keynote address by Ben Carterette, this was followed by presentation of five peer reviewed contributed papers, a panel session and then a breakout session exploring themes relevant to the topics of the workshop identified in the earlier discussions, and a final reporting back session from the breakouts.

2 Background

A number of current and previous initiatives and workshops have focused on topics relevant to WEPIR. While each of these has aspects relevant to the WEPIR, none of them directly addresses the focus or encompasses the scope of this workshop.

The key relevant activity exploring this topic from the perspective of the user is the Interactive Track at the TREC conferences, which ran for twelve years [4], and is of relevance to this workshop for several reasons. One is that it developed methods for evaluating various aspects of system performance over entire search sessions, a crucial aspect of evaluation of personalisation. Another is that one of the main findings of this track was the difficulty, perhaps impossibility, of applying the general TREC/Cranfield evaluation model to the dynamic situation of interactive information retrieval, again, a key aspect of the personalisation situation.

More recently the TREC Session Track held from 2010 to 2014, sought to provide test collections and evaluation measures for studying information retrieval over user sessions with multiple stages of query reformulation rather than one-time queries. This track introduced modified evaluation metrics for session based search [9], but had the limitation that the information need was assumed to remain static for a query across the session.

The 2012 NII-Shonan Seminar on Whole-Session Evaluation of Interactive Information Retrieval Systems [2], and the 2013 Dagstuhl Seminar on Evaluation Methodologies in Information Retrieval [1], each addressed evaluation issues relevant to this workshop, including evaluation measures for entire search sessions, and user modeling for evaluation, but stopped short of the problem of evaluation of personalization of information retrieval.

The recent interest in conversational information retrieval is also related to the topic of WEPIR. The International Workshop on Conversational Approaches to Information Retrieval, held at the 2107 ACM SIGIR conference in Tokyo (CAIR 2017) addressed some personalization issues, including system adaptation and clarification dialogues, but discussion of evaluation of such techniques was minimal.

Introduced at CLEF 2017, the Personalised Information Retrieval (PIR-CLEF) task is seeking to develop a framework for the repeatable evaluation of user models and search algorithms for personalised information retrieval (PIR) [11]. The PIR-CLEF 2017 task introduced a Pilot Task that provided data gathered during a single search session by ten users; these data are related to various activities undertaken during their search session by each participant, including details of relevant documents as marked by the searchers [12]. The Pilot task was the preliminary edition of a Lab dedicated to the theme of personalised search that is being included as a full task at CLEF 2018.

Unlike the Information Retrieval (IR) research community, the User Modeling research community has traditionally not had a significant focus on comparative evaluation or shared evaluation tasks. However, this situation is changing with the emergence of the EvalUMAP workshop series exploring the evaluation of user modeling, adaptation and personalization' which began at the UMAP 2016 conference [3], and is currently being held on an annual basis.

3 Keynote

WEPIR 2018 began with an invited keynote presentation **Offline Evaluation of Personalization Using Logged Data** by Ben Carterette, University of Delaware & Spotify.

In this presentation Ben outlined and reflected on some of his experiences examining evaluation of applications in his work including examples of his activities at Spotify.He explained that evaluation of personalisation in applications, it is very challenging. While batch experimentation using a test collection is fast, it has high start-up costs, often requires very strong assumptions about users and their needs in context, and can introduce biases if the data has not been collected very carefully. User studies are slow and have high variance, making them difficult to generalize and certainly not possible to use for iterative development. Online experimentation using A/B tests, pioneered and refined by companies such as Google and Microsoft, can be fast but is limited in other ways, in particular that it is not easy to do without access to a large user base. In his talk Ben presented work on using logged online user data to evaluate personalization offline. He discussed work on user simulation in the context of evaluating system effectiveness. He also described work on using historical logged data to re-weight search outputs for evaluation, focusing on how to collect this data to arrive at unbiased conclusions.

4 Contributed Papers

In this section we provide brief overviews of the five contributed papers accepted for publication at the workshop.

The paper by Sanna Kumpulainen, Hugo Huurdeman and Heikki Keskustalo [10], focused on collaborative research tasks. They claimed that the motivating task behind information seeking of a user or a group of users involves stages, during which the task understanding and the information needs evolve. In particular, in a collaborative context the authors considered collaborative activities related to an information intensive task as task stages, and described how the increasing understanding of the task affects the task performance. Their findings are based on a small-scale study of the information interactions among participants in a history research project; the observations and analysis of the information activities of historians while retrieving, accessing, and analyzing collections of historical documents are based on Vakkari's model, which was previously only used in learning tasks by single users. The main preliminary findings show that task stages exist and that collaborative tasks evolve similarly to individual tasks; as a consequence the authors claim that personalization would benefit of extensions accounting for task stages in collaborative research work tasks.

In his paper, Jacek Gwizdka [6] explored the use of neuro-physiological (NP) signals (collected for example via fMRI and EEG) to provide metrics useful in evaluating personalization in IR (EPIR). In fact, these signals could be used to infer cognitive and affective reactions to and perception of information. He explained that while it is clear that the NP signals are inherently individual and as such could provide personalised metrics, what needs to be investigated is which specific NP measurements can be used for this purpose. To this aim two challenges are identified: 1) to identify high-level evaluation constructs for use in EPIR, and 2) how to map these constructs onto NP measures. Concerning the first challenge, He indicated the following possible constructs: relevance, learning, user experience and affect. While various works in the literature have addressed the issue of inferring relevance from NP measures, investigations related to the detection of separate manifestations of relevance via different NP measures are scarce and should be investigated. Concerning the mapping

of learning and of user experience and affect into NP measures, Jacek explained that careful experimental designs are needed.

Sanda Erdelez and Isa Jahnke [5] focused on serendipity from a socio technical perspective in their paper.In particular they claim that personalized information systems could have an impact on the loss of genuine serendipity, which over the last twenty years has been recognized as an important aspect of human information interaction. They raise issues concerning the control of personalization (what type of personalization takes place and where?) and the potential loss of serendipity with the consequence of bringing "fake serendipity". To face this issue, the authors claim that new sociotechnical heuristics will be required to explore the major challenges of 'Personalized Information Systems and Serendipity' as a sociotechnical problem.

Sherrie Hall, Rachel Champoux, Sara Garver, Caroline Harriott, and Krysta Chauncey [7] claim that using knowledge about a user can be exploited in the processes of information presentation and personalized interface design, with the aim of mitigating the users cognitive overload. More specifically, they hypothesize that personality traits modulate people's understanding of the world and are useful in adaptive interface design; in particular, they focus on the trait known as *Need for Cognition (NFC)* in their study. Their study is related to a data analysis task (users needed to identify key information among data for decision-making within a restricted time frame) and aimed to understand co-adaptive interface design (both the human and software agent adjust their behaviour and information presentation based on context, state, trait, and task). Their preliminary results show how High and Low NFC participants exhibit behavioural differences with respect to adaptation.

The paper by Gareth Jones, Gabriella Pasi, Andrea Angiolillo and Camilla Sanvitto [8] outlined a proposal for the extension of the Cranfield test collection based evaluation paradigm to incorporate details of individual searchers and their search behaviour to enable investigation and evaluation of personalisation methods within the IR process. Their paper describes a strategy for personalised data collection, and introduces a tool for the evaluation of personalised search results. The method described is being used within the PIR-CLEF track at CLEF 2018, following a pilot trial operation at CLEF 2017 [11].

5 Panel and Breakout Working Groups

One of the key motivations for the WEPIR 2018 meeting was to engage the community in a discussion of relevant topics. The workshop proved to be a great success in this regard. There was enthusiastic participation in the question and answer sessions following all presentations with a wide range of thought provoking questions and follow up discussions. The level of engagements was such that all of the 20 plus workshop attendees contributed to these discussions at some point in the presentation sessions. The presentations were following by a panel session and a set of breakout working groups on topics agreed by the workshop participants.

5.1 Panel

The discussion sessions were proceeded by a panel session with invited panelists: Michael Cole, LexisNexis, USA, Noriko Kando, National Institute of Informatics, Japan and Paul Thomas, Microsoft, Australia. Each panelist introduced issues and ideas which they believed to be relevant and important to the topic of the workshop.

5.2 Working Group Reports

After discussion among the workshop participants, three breakout working groups were formed focusing broadly on the following topics: measurement in evaluating PIR, shared understanding evaluation of PIR among researchers, and exploring search context within evaluation of PIR. The following provides brief summaries of the outcomes of the discussions from each of these groups.

Measurement The topic under discussion was to consider measurements that may be relevant within evaluation of personalisation in search. Three main areas were identified: the relationship between user search task type, personalisation, and successful completion of the task; the user's satisfaction with the level of completion of the task, and the relationship of this with personalisation of the search process; and, the relationship between the searcher's state of knowledge of the topic under investigation and its development during the search process and personalisation in the search process.

Understanding In order to develop an agreed understanding of personalisation in search and its evaluation, it is important to have a shared understanding of the topics under investigation. This group identified the following areas for consideration: consistent use of language to express concepts relevant to the evaluation of personalisation in search; an agreed set or sets of variables which have an effect of personalisation in search; an agreed set of tools for measurement of personal features as used within the search process to ensure validity across studies; and, focus on statistically valid analyses in quantitative studies in PIR. It is important to have statistically robust and meaningful methods and analysis in order be able to compare results across studies, and a shared understanding of these will help to ensure consistent and comparable experimental results. Studying and drawing on relevant literature and expertise from other fields to accelerate progress in research in PIR.

Context All search takes place in the context within which the searcher is currently operating in terms of their activities, location, temporal setting and their physiological state. Depending in their current information need some of all of these and other elements of context may impact on user search intent. Another proposal was to determine the searcher's search intent and its potential relationship with context, and explore the impact of individual context factors on search effectiveness and their importance for personalisation in the IR process. Certain contextual factors may be persistent and while others will be much more transitory, these will have differing significance in PIR with different needs for frequency of measurement. All of these issues need to be better understood with PIR and its evaluation.

6 Concluding Remarks

The organisers received a large amount of very positive feedback at the conclusion of WEPIR 2018. Participants were keen to keep in contact, and we are minded to organise another meeting to further explore the topics raised in this workshop in the future.

7 Acknowledgements

This workshop was partially supported by Science Foundation Ireland as part of the ADAPT Centre (Grant No. 13/RC/2106) (`www.adaptcentre.ie`), and by the US National Science

Foundation under grant IIS-1423239. The workshop chairs greatly appreciate the enthusiasm with which the participants approached this workshop and their contributions in particular to the reporting of the outcomes of the breakout groups.

References

[1] M. Agosti, N. Fuhr, E. Toms, and P. Vakkari. Evaluation methodologies in information retrieval. Dagstuhl Seminar 13441, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Schloss Dagstuhl, Germany, 2014.

[2] N. J. Belkin, S. Dumais, N. Kando, and M. Sanderson. Whole-session evaluation of interactive information retrieval systems. NII Shonan Meeting Report 2012-7, National Institute of Informatics, Japan, Tokyo, Japan, 2016.

[3] O. Conlan, L. Kelly, K. Koidl, S. Lawless, K. Levacher, and A. Staikopoulos, editors. *EvalUMAP2016: Towards Comparative Evaluation in the User Modelling, Adaptation and Personalization*, Halifax, Canada, 2016.

[4] S. Dumais and N. J. Belkin. The TREC interactive tracks: Putting the user into search. In E. M. Voorhees and D. K. Harman, editors, *TREC. Experiment and evaluation in information retrieval*, pages 123 – 152. MIT Press, Cambridge, MA, 2005.

[5] S. Erdelez and I. Jahnke. Personalized systems and illusion of serendipity: A sociotechnical lens. In *Proceedings of WEPIR 2018*, New Brunswick, USA, 2018.

[6] J. Gwizdka. Neuro-physiological data as a source of evaluation metrics for personalized IR. In *Proceedings of WEPIR 2018*, New Brunswick, USA, 2018.

[7] S. Hall, R. Champoux, S. Garver, C. Harriott, and K. Chauncey. Designing the user experience of a co-adaptive data analytics interface in response to user trait. In *Proceedings of WEPIR 2018*, New Brunswick, USA, 2018.

[8] G. J. F. Jones, G. Pasi, A. Angiolillo, and C. Sanvitto. A proposed method for laboratory-based evaluation of personalised information retrieval. In *Proceedings of WEPIR 2018*, New Brunswick, USA, 2018.

[9] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2011, pages 1053–1062, Beijing, China, 2011. ACM.

[10] S. Kumpulainen, H. Huurdeman, and H. Keskustalo. Personalization needs extension towards task stages in collaborative research work tasks. In *Proceedings of WEPIR 2018*, New Brunswick, USA, 2018.

[11] G. Pasi, G. J. F. Jones, S. Marrara, C. Sanvitto, D. Ganguly, and P. Sen. Overview of the CLEF 2017 personalised information retrieval pilot lab (PIR-CLEF 2017). In *Proceedings of CLEF 2017*, Dublin, Ireland, 2017. Springer.

[12] C. Sanvitto, D. Ganguly, G. J. F. Jones, and G. Pasi. A laboratory-based method for the evaluation of personalised search. In *Proceedings of The Seventh International Workshop on Evaluating Information Access (EVIA 2016)*, Tokyo, Japan, 2016.