Third Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2021)

In Memoriam Séamus Lawless

Gareth J. F. Jones
ADAPT Centre, School of Computing
Dublin City University
Dublin 9, Ireland
Gareth.Jones@dcu.ie

Nicholas J. Belkin
School of Communication & Information
Rutgers University,
New Brunswick, NJ, USA
belkin@rutgers.edu

Noriko Kando
National Institute of Informatics (NII)
Tokyo, Japan
kando@nii.ac.jp

Gabriella Pasi
Department of Informatics, Systems & Communication
University of Milano-Bicocca
Milan, Italy
pasi@disco.umib.it

ABSTRACT

The Third WEPIR 2021 workshop builds on the success of the first two WEPIR meetings held at CHIIR 2018 and CHIIR 2019. WEPIR 2021 again brings together researchers from different backgrounds interested in continuing to explore and advance the evaluation of personalisation in information retrieval. Similar to the first two workshops, WEPIR 2021 has a strong emphasis on active participation by workshop attendees. This was very successfully achieved in the first two workshops by the use of workshop breakout groups exploring topics related to personalisation and information retrieval, and the evaluation of personalisation in information retrieval in general, with subsequent report back to the workshop as a whole.

However, a key difference for WEPIR 2021 is that while the first two workshops focused on developing and articulating principles and ideas relating general topics relating to these topics, identified as interesting and important by the attendees at the workshops, WEPIR 2021 focuses breakout discussion on a number of relevant specific use cases of the evaluation of personalisation in information retrieval. A use case is assigned to each breakout group with the plan being to have more than one group work on each use case. The task for each group is to identify specific relevant factors relating to the use case in terms of user activities, data to be collected, ethical issues, and evaluation metrics. Groups will make reports of their discussions to the assembled workshop in the final session with discussion of the alternative solutions relating to the same use case, and contrasting the issues raised by the different use cases. The overall goal of the workshop is to work towards developing a general set of principles and guidelines for addressing the evaluation of specific instances of the use of personalisation in information retrieval tasks. This is consistent with the activities and goals of the first two workshops, but represents a significant progression of the activities towards concrete outcomes of benefit to those exploring personalisation in search, the issues arising in its evaluation, and how researchers might go about tackling this in specific situations.

1 INTRODUCTION

One of the key goals of information retrieval research is to advance the development of search applications which enable searchers to satisfy their information needs more effectively and efficiently. Given that the information need underlying their engagement with a search application relates to a personal need for information, it makes intuitive sense that in addition to the searcher's stated search request, a search application should make use of all available information about the searcher and the context in which the search is being carried out in order to return content most likely to be useful or relevant to this user.

Information relating to the user of a search application can be gathered from logs of their previous search activities, and also from monitoring their other online activities with various applications and their context. A user can also be requested to explicitly provide information to complete a personal profile. Given the very large amounts of information available from these different sources, an important research question is how to utilize it within the search process [3]. It might for example be used to populate personal profiles of characteristics and interests or more general user models, which can then be used within a personalised information retrieval application to return documents more likely to be relevant to the user than those returned by an equivalent, but non-personalised search application.

There are many ways in which personal information might be modelled and represented using 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, carefully planned evaluation strategies are required. These need to consider both the algorithmic aspects of the information retrieval process, and also the user-centered or interactive elements of search and the user experience of using a search application.

Evaluating the use of personalsation models in classic "single" shot information retrieval settings could amount to simply incorporating alternative personsalisation models into the process for a given query. However, much more interesting is the incorporation of personalisation in session-based settings incorporating multiple queries expressing the evolution of an information need as the searcher progresses through the session, and further the use of personalisation across multiple sessions where the user model is updated in response to developments in the searcher's interests. Evaluation in this setting is a highly complex problem.

In addition to the individual use of entirely individual information for personsalisation, we can also consider the use of group-level personsalisation, where the activities and experiences of multiple users are combined. Individuals can then be identified with a group best matching their interests and/or experience or knowledge levels, or this group information can be used to smooth the limited personal information relating to a search task for indivudual users.

The first WEPIR workshop at CHIIR 2018 brought together a group of interested reseachers to open a broad discussion of the issues relating to the evaluation of personalisated information retrieval applications within the CHIIR community, as summarized in the WEPIR 2018 report in *SIGIR Forum* [10]. The second WEPIR workshop at CHIIR 2019 extended these discussions to explore more specific issues relating to individual and group personalisation of the user search experience, and the user experience of the instantiation on personalisation in interfaces of search applications and its evaluation. Further details are provided in the WEIR 2019 report *SIGIR Forum* [11]. WEPIR 2020 further develops these activities by focusing on the exploration of the use and evaluation of personalisation in a number of specified use cases, to move towards concrete proposals for evaluation methodologies, metrics, etc. for individual search applications.

2 BACKGROUND

Prior to the establishment of the WEPIR workshop series, a number of onging and earlier initiatives and workshops have focused on topics relevant to WEPIR. While each of these has aspects relevant to WEPIR 2021, none of them directly addresses the focus of WEPIR 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 [6], 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 [12], 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 this proposed workshop. The International Workshop on Conversational Approaches to Information Retrieval held at SIGIR 2017 and SIGIR 2018 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 sought to develop a framework for the repeatable evaluation of algorithms for personalized search, and for the evaluation of user models [7][8][9]. PIR-CLEF 2017, PIR-CLEF 2018 and PIR-CLEF 2019 focused on a benchmark web search task that provided user data gathered during a single search session. These data relate to various activities undertaken during their search session by each participant, including details of relevant documents as marked by the searchers [5]. PIR-CLEF 2019 expanded the scope of this investigation by the inclusion of a task exploring the evaluation of personalisation in a medical search task.

Unlike the information retrieval 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 [4], and is currently being held on an annual basis.

The WEPIR 2018 and WEPIR 2019 workshops featured invited keynotes, a small number of short paper presentations, and extended sessions of breakout group discussions. At WEPIR 2018 these breakout groups focused on the topics of: *measurement, understanding and context* [10], and at WEPIR 2019 on the topics of *Personalisation for Individuals*, *Personalisation for Groups*, *Presentation of Personalisation to Users* and *Evaluating the Presentation of Personalisation*. Each workshop attracted more than 20 participants from diverse backgrounds who engaged very actively in lively discussions throughout the day. There was a strong consensus at the end of each meeting that it would be valuable to hold further workshops focusing on this topic.

3 PROVISIONAL WORKSHOP PLAN

Personalisation of the user search experience is an important topic in potentially enhancing the effectiveness of many search applications. However, understanding and evaluating its impact on either user acceptance of an application or its absolute contribution to the identification of relevant or useful items often raises significant challenges, and there is a lack of generally accepted methodologies for personalisation in information retrieval applications.

WEPIR 2021 builds on the activities and outcomes of the previous workshops to move beyond identification and discussion of general topics relating to personalisation in information retrieval and its evaluation, to the development of methods to tackle evaluation of personalisation for several specific use cases. Details of the proposed use cases are given below. The goal of the workshop is to seek to create plans for evaluation for these specific use cases, but also to move towards more general guidelines which can be applied to the evaluation of personalisation in other search tasks.

WEPIR 2021 will take place via Zoom and other collaboration systems, over the period 08:00 – 23:30 UTC on 19 March 2021, with the following structure:

- An initial opening session providing an introduction to the workshop, followed by a Keynote talk on methods of evaluation (2 hours);
- An initial session of short paper presentations, and explanation of the objectives of the working breakout groups to examine individual specific use cases, introduction of the use cases for personalised search and formation of breakout groups with assignment of a use case to each group. Assuming a participation level similar to the previous WEPIR workshops (20-30 participants), we plan to have at least two groups working independently on each use case. We expect this to provide interesting scope for discussion examining the alternative perspectives, ideas and methodologies introduced by the different groups of researchers considering the same problem (2 hours).
- A second opening session, followed by a second Keynote talk on methods of evaluation (2 hours);
- A second session of short paper presentations and establishment of working breakout groups (2 hours);
- Breakout sessions with groups working on developing an evaluation plan for their assigned use case. The members of the groups will schedule their (virtual) meetings according to their own constraints; these meetings will take place over the period 11:00-19:00 UTC.
- A closing session, which consists of: each breakout group reporting back in their agreed plan for the evaluation of their assigned use case; open discussion examining the findings of the breakout groups; and, wrap up and consideration of potential follow ups that emerge from the discussions.

We expect to submit a report of the workshop to SIGIR FORUM, but we will also examine with the group whether there is potential for the development of other publications arising from workshop discussions.

3.1 Use Cases

- Use Case 1: Museum Visitors: Consideration of how the experience of visitors to a museum may be personalised to their individual interests and levels of knowledge. This topic will also enable the exploration of the potential for group personalisation, and the issues of physical engagement with objects and use of mobile platforms arising from visitors moving around a physical museum. It might also explore the potential for user experience of virtual museums integrating materials from multiple physical museums.
- Use Case 2: Medical Search: Medical search is one of the most popular search topics for users of web search engines and specialist information portals. Users approach these with highly varied information needs and medical knowledge. Examination of search logs of medical portals reveals that user queries are generally topically ambiguous in terms of what the searcher's underlying information need is, and what type of documents will be suitable for an individual user.
- Use Case 3: Web Search: While the use of personalisation in web search is a long-standing topic, there is still little agreement on methodologies for enabling the large scale comparative analysis of alternative personalisation methods or how the use of their methods impacts on the user's experience of search or its overall desirability in terms of addressing the diversity of user information needs.

4 ORGANISERS

The organisers of the WEPIR 2021 workshop have a broad range of relevant expertise in the topics of the workshop including benchmark evaluation task development, interactive information retrieval, search algorithm design, and personalised information retrieval, as well as extensive experience in organising and running successful workshops.

Gareth Jones is Professor of Computing in the School of Computing, and a Principal Investigator in the ADAPT Research Centre, Dublin City University, Ireland, His research focuses on multiple topics in information retrieval including adaptive search, multimedia information retrieval (particularly for spoken content), multilingual information retrieval, interactive

and algorithm search for lifelogging, A particular focus in much of this work has been the development of evaluation frameworks including task design, test collection specification and construction, and the introduction of new task specific evaluation metrics. With Gabriella Pasi led the PIR-CLEF Lab for the comparative evaluation of personalisation algorithms in information retrieval. He is co-founder of the MediaEval multimedia benchmark evaluation initiative, and since 2002 has coordinated a variety of benchmark tasks at CLEF, FIRE, NTCIR and TRECVid. He regularly serves on the programme committees of leading conferences in information retrieval, multimedia, natural language processing and human-computer interaction. He served as Programme co-Chair for CIKM 2010 amd ECIR 2011, and as General Co-Chair for SIGIR 2013 and CLEF 2017.

Nicholas Belkin is a Distinguished Professor Emeritus in the Department of Library and Information Science at Rutgers University. He was the initiator of the Interactive Track in TREC, and variously chaired and participated in that Track for its twelve years. He was also a key participant in the TREC Session Track, devising the task classification system used in that Track. He was an organizer of the NII Shonan Seminar on Evaluation of Whole Session Information Retrieval, and of the Dagstuhl Seminar on Interactive Information Retrieval. His research group has been engaged in studies of personalization of interactive information retrieval since 2010, with funding from the US Institute for Museum and Library Services, the US National Science Foundation, and Google. His proposal that the best criterion for evaluation of interactive information retrieval is usefulness, rather than relevance, has been adopted, and validated, by several different research groups throughout the world. Professor Belkin holds the ACM SIGIR Salton Award, the ASIS&T Award of Merit, and has been the Chair of the ACM SIGIR, and the President of the ASIS&T.

Noriko Kando is Professor at the National Institute of Informatics (NII), Tokyo. Japan. Since 1999 she has been the central coordinator of the NTCIR benchmark evaluation task and conference series, which has now reached 15 editions. Additionally, she has organised a large number of conferences and workshops. Her research focuses on Evaluation of Information Access Technologies, Language Understanding and Information Access Technologies, Exploratory Search, Community Oriented Information Access System – Cultural Heritage, User Interface and Cross-lingual Information Access. She is currently Principal Investigator of a project in Japan exploring personalised search for museum archives.

Gabriella Pasi is Professor at the University of Milano-Bicocca, Italy where she leads the Information and Knowledge Representation, Retrieval and Reasoning Laboratory (IKR3 Lab) within the Department of Informatics, Systems and Communication. Her research activities concern various topics related to modelling and designing flexible and context-aware systems for the management and access to huge collections of information items (such as Information Retrieval Systems and Recommender Systems). With Gareth Jones she led the PIR-CLEF Lab for the comparative evaluation of personalisation algorithms in information retrieval. She has also contributed to the organization of several international events, in roles of both General and Program Chair (e.g. for ECIR 2018). She is Associate Editor or member of the Editorial Board of several International Journals in her domains of expertise.

REFERENCES

[1] Maristella Agosit, Norbert Fuhr, Elaine Toms and Pertti Vakkari. 2014. *Evaluation methodologies in information retrieval*. Dagstuhl Seminar 13441, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

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

[3] Paul N. Bennett, Filip Radlinski, Ryen W. White and Emine Yilmaz. 2011 Inferring and Using Location Metadata to Personalize Web Search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011).* ACM, Bejing, China, 135-144.

[4] Owen Conlan, Liadh Kelly, Kevin Koidl, Seamus Lawless, Killian Levacher, and Athanasios Staikopoulos (Eds.). 2016. *EvalUMAP2016: Towards Comparative Evaluation in the User Modelling, Adaptation and Personalization*. Halifax, Canada.

[5] C.Sanvitto, D.Ganguly, G.J. F.Jones, and G.Pasi. 2016. 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.

[6] Susan Dumais and Nicholas J. Belkin. 2005. The TREC interactive tracks: Putting the user into search. In *TREC. Experiment and evaluation in information retrieval*, Ellen M. Voorhees and Donna K. Harman (Eds.). MIT Press, Cambridge, MA, 123 – 152.

[7] G.Pasi, G.J.F.Jones, K.Curtis, S.Marrara, C.Sanvitto, D.Ganguly, and P.Sen. 2018. Overview of the CLEF 2018 Personalised Information Retrieval Lab (PIR-CLEF 2018). In *Proceedings of CLEF 2018*. Springer, Avignon, France.

[8] G.Pasi, G.J.F.Jones, L.Goeuriot, L.Kelly, S.Marrara, and C.Sanvitto. 2019. Overview of the CLEF 2019 Personalised Information Retrieval Lab (PIR-CLEF 2019). In *Proceedings of CLEF 2019*. Springer, Lugano, Switzerland.

[9] G.Pasi, G.J.F.Jones, S.Marrara, C.Sanvitto, D.Ganguly, and P.Sen. 2017. Overview of the CLEF 2017 Personalised Information Retrieval Pilot Lab (PIR-CLEF 2017). In *Proceedings of CLEF 2017*. Springer, Dublin, Ireland.

[10] Gareth J. F. Jones, Nicholas J. Belkin, Seamus Lawless, and Gabriella Pasi. 2018. Report on the CHIIR 2018 Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2018). *SIGIR Forum*, 52,1 (2018), 129-134.

[11] Gareth J. F. Jones, Nicholas J. Belkin, Seamus Lawless, and Gabriella Pasi. 2019. Report on the CHIIR 2019 Second Workshop on Evaulation of Personalisation in Information Retrieval (WEPIR2019). *SIGIR Forum*, 53,1 (2019), 29-37.

[12] Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. 2011. Evaluating Multi-query Sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*. ACM, Beijing, China, 1053–1062.