Report on the CHIIR 2019 Second Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2019)

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 Second Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2019) was held in conjunction with the ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR 2019) in Glasgow, Scotland. WEPIR 2019 followed on from the first WEPIR held at CHIIR 2018. The purpose of the workshop was again to bring together researchers from different backgrounds, interested in advancing the evaluation of personalisation in information retrieval. The workshop focused on further development of a common understanding of the challenges, requirements and practical limitations of personalisation in information retrieval and its evaluation.

1 Introduction

Information retrieval (IR) systems seek to enable users to satisfy their information needs. Given that the user's information need is personal, it is desirable for IR systems to be able to return information which is most likely to be useful to this user personally. Identifying information useful to individual searchers requires search applications to incorporate information relating to the user into the search process. There are many ways in which personal information might be used within some form of user model, and how this model might be incorporated into the IR process. In order to determine how best to implement a personalised IR application, an evaluation strategy is required. The Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR) series is intended to provide a forum for researchers interested in the topic of Personalisation in Information Retrieval (PIR) to explore relevant evaluation strategies.

A first WEPIR was held in conjunction with CHIIR 2018 [8]. This meeting provoked very lively discussion among the participants, and it was agreed that a second meeting to

explore this topic further should be organised. Since CHIIR 2019 was likely to again attract participants interested in this topic, the second meeting WEPIR 2019 was organised at CHIIR 2019 in Glasgow, U.K.

The workshop began with a keynote address by Ian Ruthven, which was followed by presentation of 5 peer reviewed contributed papers. The second half of the workshop was given over to a breakout session exploring themes relevant to the workshop identified in a breakout planning discussion, concluding with a final reporting back session by the breakout groups.

2 Background

Prior to the first WEPIR in 2018, a number of activities had focused on topics relevant to WEPIR. But while each of these had aspects relevant to WEPIR, none of them directly addressed the issues or encompassed the scope of this workshop. We provide here a brief overview of some of these related activities.

The key relevant activity from the perspective of the user is the Interactive Track at the TREC conferences, which ran for twelve years [5], and is of relevance to this workshop for several reasons. It developed methods for evaluating various aspects of system performance over entire search sessions, a crucial aspect of evaluation of personalisation. One of its main findings was the difficulty, perhaps impossibility, of applying the general TREC/Cranfield evaluation model to the dynamic situation of interactive IR, 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 IR 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 [3], 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 IR is also related to the topic of WEPIR. A number of workshops have been held in conjunction with recent international conferences in IR and artificial intelligence, and these meetings have addressed some issues relating to personalisation, but discussion of evaluation has been very limited.

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 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 currently included as a full task at CLEF [10].

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

3 Keynote

Professor Ian Ruthven, of the University of Strathclyde, opened WEPIR 2019 with an invited keynote presentation entitled *Resonance and the Experience of Relevance*. In his presentation, Ian introduced the concept of resonance, as a complement to that of relevance, and as an important factor in personalisation of IR, and its evaluation. He began by situating resonance as one pole of an "Axis of Relevance". At one pole, the general concept of relevance is more objective in nature, and more dependent on cognitive factors; at the other, relevance is more subjective in nature and more dependent on non-cognitive factors. The former is characterized by our normal concepts of topical and pragmatic relevance; the latter is characterized by the concept of resonance; along the axis, relevance is viewed as a mixture of the two.

Ian explicated the concept of resonance through examples of its use in a variety of fields and contexts, including: Physical resonance, meaning amplification or matching of forces; and Limbic resonance, meaning sharing of emotional states. In particular, within the Limbic view, Cognitive resonance is a (perceived) alignment with particularly salient understandings and beliefs of the other(s), and Emotional resonance is a felt alignment with passions, desires, aspirations. He went on to describe the ways in which this idea of resonance could influence personalization of IR, in ways not accounted for by traditional concepts of relevance, and to discuss how the effect of such personalization could be evaluated. This was an extremely stimulating presentation, which lead to a good deal of immediate discussion, especially with respect to designing to enhance resonance; whether resonance is a necessary component of personalization; and, whether resonance is an aspect of usefulness. In fact, the force of Ian's keynote resonated throughout the remainder of the Workshop; as such, it was an outstanding opening to the proceedings. The reader is referred to the presentation itself, at the WEPIR 2019 website (http://www.ir.disco.unimib.it/wepir2019/download/) for details, and the opportunity to resonate, or not, with these stimulating ideas.

4 Contributed Papers

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

In their paper titled *Interactive Search Profiles as a Means of Personalisation*, Maram Barifah and Monica Landoni [2] sought to analyze search log files to verify both how informative they are with respect to user system interactions during search, and how they can be fruitfully exploited in defining user models in digital library applications. In this context, and as an outcome of their analyses, the authors propose what they call *interactive search profiles (ISPs)* which can be used to personalise the interface of a digital library system to provide improved search experience.

The definition of novel interactive museum user guides that are able to support the user's personalized discovery and learning was addressed in the paper titled *Finding the Connection between Artifact and Personal Knowledge of Museum Visitor*, by Zehua Yang,

Yusuke Yamamoto, Takehiro Yamamoto, Noriko Kando and Hiroaki Ohshima [14]. The personalized approach proposed integrates the user's active exploratory search-and-browsing in both the virtual and the physical museum space to recommend artifacts, together with exploitation of the relationships between them. In the proposed method, the connections between the artifacts of interest and each user's knowledge are found by using Wikipedia link structures.

The paper titled *Personalization through Search Roles in Collaborative Search Scenarios*, authored by Stefanie Elbeshausen, Thomas Mandl and Christa Womser-Hacker [6] tackled the interesting issue of personalization in collaborative search. To this purpose five role-specific behavior patterns were identified and employed to personalize the search process according to the user's abilities and skills. The authors observe that this kind of personalization should help to raise the users' satisfaction by assigning sub-tasks according to preferences.

How to capture the user's goals and interests during an exploratory search task is the issue considered in the paper titled *Evaluation of Rich and Explicit Feedback for Exploratory Search* by Esben Sørig, Nicolas Collignon, Rebecca Fiebrink and Noriko Kando [13]; the authors observe that when interacting with new documents, people use annotations as an aid to understanding. These annotations, in particular the usage of highlights, can provide rich information to understand a reader's interests. Highlights constitute a form of feedback leading to improvements in exploratory search with simulated users when compared to relevance feedback. In the paper an evaluation platform is presented which had been developed by the authors to test the effect of annotation feedback on search performance, user experience and user behavior.

In her position paper titled *Evaluating Personalised Information Retrieval: a Perception of Trust*, Frances Johnson [7] presented an interesting analysis of the evaluation of personalised search from a user perspective, and made an analysis of evaluation of PIR. The focus was on the effect of personalisation on the searchers perception of a better IR interaction; to this aim the components of user evaluation were discussed and the searcher's perception of usefulness identified to specifically evaluate the effect of personalisation. In particular, it was suggested that a perception of trust and its criteria provide a framework for evaluation.

5 Breakout Working Groups

Following on from the high level of participant engagement at WEPIR 2018, one of the key objectives of WEPIR 2019 was again to engage the community in discussion of topics relevant to evaluation of PIR. We were very pleased that WEPIR 2019 again 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 oral presentations were followed by poster presentations of these papers during the lunch session which enabled interested participants to question the authors more closely. The second half of the workshop after lunch was devoted to a breakout discussion session.

After consultation with participants, four breakout working groups were formed focusing broadly on the following topics: personalisation for individuals, personalisation for groups, presentation of personalisation to users, and evaluating the presentation of personalisation. Although the groups were assigned to explore different topics, there was a surprising amount of overlap between the issues presented in the feedback reports at the end of the workshop,

and while the following summary outlines the main points raised only once, many of them appeared in more than one topic report.

5.1 Personalisation for Individuals

One of the motivating objectives of personalisation in IR is to reduce the cognitive load on the users and to increase their overall satisfaction with the search process. Adapting the response of an IR system to an individual user requires some form of user model. While this is widely accepted, there is no general consensus among researchers of what form this model should take or how it should be acquired. This group determined that model acquisition should be an incremental process, with the degree of personalisation being determined by the richness of the model. Based on this idea, they distinguished four levels of personalsiation:

- No personalisation: no adjustment of the IR system for the individual user
- Weak personlisation: limited group level adjustment, e.g. based on language preference, location or demonagrpahic group.
- Strong personalisation: targeting adjustment towards a small group of individuals, e.g. a family group.
- Ultimate personalisation: personalisation of the search at the level of the individual user.

The move between the different types of personalisation could then be a smooth transition as the user model develops with the degree of personalisation potentially under the control of the user with the option to roll it back should the user wish to.

The group was clear that evaluating how personalisation affects the user experience is crucial to building high quality IR systems. A particular concern was the need to identify situations of *over-personalisation* which can degrade the user's experience of a search system. They introduced the term *forced personalisation* for situations where the user is made to use a personalised search system where the personalisation, rather than helping the user, makes a system more difficult or perhaps impossible to use, and can lead to filter bubble effects in the content returned to the user.

With regard to the evaluation of personalisation, they suggested that a method might be developed to evaluate the quality of the user model itself, rather than search effectiveness using the model, as a direct measure of the potential of the user model within a personalisation process. Within a personalisation search process, they suggested to explore the evaluation of the degree to which the user's expectation of system personalisation had been met. Rather than looking directly at the search effectiveness of a system, they were concerned to measure or capture the response of the user to the personalisation of the search system. e.g. by changes in their actions or by their biometric response.

5.2 Personalisation for Groups

While classically, when referring to personalisation, people's first reaction is to think of adaptation of systems to the needs of an individual user, there is often value in taking account of a group to which the user may belong. As noted by the previous group, adaptation based on group membership can be considered a form of weak personalisation.

Individuals may belong to groups based on physical attributes, e.g. time, location, or personal characteristics, e.g. age, or shared professional or personal interests. The most obvious way to think about this is probably in terms of adaptation of a search system to the group, but there is also the potential for members of groups to be working together in a collaborative search endeavor.

There are issues to consider in terms of how users become members of groups and how this information is used. For example, are users placed into groups because of their attributes prior to search, or perhaps they could be grouped based on common patterns of search interests, or some combination of both of these. Once placed into groups. what information should be shared between the users as attributes of the group? This raises issues of privacy of user information, and the potential for users to grant permission for the sharing of information or at least for the information to be aggregated anonymously within a group model for system adaptation. As well as sharing personal attributes, one idea for group personalisation was the sharing or suggestion of queries and search results.

The group also gave some consideration to the topic of evaluation. They proposed simple measures such as whether users elected to continue to use a group-based search environment once they had experienced it, directly surveying to explore user satisfaction and response to this setup, and seeing to what extent they took up suggested queries or search results from other users.

5.3 Presentation of Personalisation to Users

This group examined the topic of how personalisation impacts on search results, and to what extent the user is made aware of the presence of personalisation in the results that are being presented to them. The opportunities for and suitability of the application of personalisation may vary between settings. For example museums and libraries are very different environments, similarly the device being used or physical setting may influence presentation preferences, e.g. mobile vs desktop, home vs on the move.

A wide range of questions were identified in terms of how the user is made aware of and given control of personalisation.

- Does the user know if it is on or off? Can they change this? How?
- Is it based on a profile of the user's personal information? Can they see their profile? Can they edit it?
- Is it based on the user's history? Do they have access to the history? Can they edit it?
- How does the user know which search results appear due to personalisation effects? Can they control the influence of personalisation to see more or less of specific type of personalised result?
- Are aspects of the overall presentation of the results personalised? How is this indicated to the user? Can the user control this?

With regard to the evaluation of personalisation, the group identified two main elements: assess whether the personalised or non-personalised system is better, although exact methods for achieving this were not explored; examine whether the user's behavior and interactions change over time with the inclusion of personalisation, again methods to do this were not developed.

5.4 Evaluating the Presentation of Personalisation

The final group focused on the evaluation of the presentation of personalisation in terms of measuring its effectiveness and efficiency. They considered evaluation from the perspectives of the system and the user. In terms of data collection, they considered the use of explicit methods, e.g. use of questionnaires, and implicit methods, e.g. tracking user interactions. They proposed that both of these strategies should be included to provide a more complete picture of the personalized experience than either one when it is used alone.

They highlighted the importance of identifying the aspects by which systems can be compared and evaluated. They also noted two main considerations should be to determine or measure in some sense: what users know about the system, and what the system know about the user who is using it.

There are a number of other issues which researchers should be aware of when seeking to evaluate the impact of personalisation on an IR system. These include: the user's prior knowledge of the system, which may bias their use of the system, e.g. based on extensive experience with an existing search tool. Users aware that a system is trying to personalise their experience may behave in a certain way to "teach" it what they believe to be desirable information for it to be able to improve their search experience. There are questions about whether or not users should be formally made aware of the peronsalisation, and if so how this should be done.

In terms of the system's knowledge of the user, the more the system knows about the user, the more it can personalise its presentation to them. Part of the exploration of the effectiveness of personalisation then, is to specify the knowledge that the system has about the user at the start, and as the personalisation process proceeds, based on interactions between the system and the user, and to explore its effective use for personalisation.

The group identified a number of factors which may be impacted by the inclusion of personalisation within the search process, and should thus be considered in the evaluation of personalisation in the presentation of the system to the user.

- Trustworthiness: Does the user trust the system to behave reliably? If they trust the system, they are more likely to be forgiving when it makes mistakes. Does personalisation affect the trustworthiness of the system?
- Mental Model: Users have a mental model of the system. How does personalisation impact on their mental model? Does this make their use of the system more or less effective?
- Cognitive Load: Does the inclusion of personalisation decrease or increase the mental cost of the using the system? Intuitively, if the load is higher, the user would be less likely to want to use it.
- Perceived Performance: Regardless of how well a system behaves objectively, it is important to examine the perceived performance from the perspective of the user, since they are not always correlated.
- Controlability: To what extent can the user control the system? Being able to change system behavior may help to reinforce trust, reduce cognitive load, improve alignment between the user's mental model and the actual operation of the system, and improve overall performance.
- Explainability: Explainability is currently a hot topic in research in artificial intelligence in general, and is an important topic in personalisation. It should not only be able to

answer the question "Why did the system return this result?", but also "Why did the system choose to present this result in this way?". These factors can interact in evaluation of a personalised information retrieval system, for example how is trust impacted by the explainabillty and controlability of results?

- Retrieval Performance: Consideration of the effect of personalisation on the standard retrieval effectiveness of an IR system in terms of measures relating to precision and recall.

5.5 Breakout Review

One of the perhaps surprising outcomes of the breakout discussions was that the groups spent much of their time developing a shared understanding among group members of the issues relevant to the topic under discussion in terms of personalisation. The groups all engaged energetically in their discussions and there is no doubt that all of them progressed in development of the topic under discussion. However, there was little detail reported by way of discussion on the actual topic of evaluation by each group, except for group actually tasked to look at evaluation issues. This is perhaps indicative that while many researchers are interested in personalisation in IR, there has been relatively little discussion of this topic by the broader community, and it makes sense that they should be clear on the relevant issues within personalisation as a topic, before they can think in detail about issues relating to evaluation. Vibrant and productive as they were, the main conclusion of these discussions in terms of evaluation, would appear to be that more discussion is needed.

6 Concluding Remarks

The organisers received a large amount of positive feedback at the conclusion of WEPIR 2019. There was clearly great and diverse interest in the topic of this workshop, and the much work remains to be done to develop a better understanding of the many issues raised at both WEPIR 2018 and WEPIR 2019, and there is certainly scope for further meetings examining evaluation of personalisation in information retrieval.

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] M. Barifah and M. Landoni. Interactive search profiles as a means of personalisation. In *Proceedings of WEPIR 2019*, Glasgow, Scotland, U.K., 2019.

[3] 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.

[4] 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.

[5] 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.

[6] S. Elbeshausen, T. Mandl, and C. Womser-Hacker. Personalization through search roles in collaborative search scenarios. In *Proceedings of WEPIR 2019*, Glasgow, Scotland, U.K., 2019.

[7] F. Johnson. Evaluating personalised information retrieval: a perception of trust. In *Proceedings of WEPIR 2019*, Glasgow, Scotland, U.K., 2019.

[8] G. J. F. Jones, N. J. Belkin, S. Lawless, and G. Pasi. Report on the CHIIR 2018 workshop on evaluation of personalisation in information retrieval (WEPIR 2018). *SIGIR Forum*, 55(1):129–134, 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] G. Pasi, G. J. F. Jones, K. Curtis, S. Marrara, C. Sanvitto, D. Ganguly, and P. Sen. Overview of the CLEF 2018 personalised information retrieval pilot lab (PIR-CLEF 2018). In *Proceedings of CLEF 2018*, Avignon, France, 2018. Springer.

[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.

[13] E. Sørig, N. Collignon, R. Fiebrink, and N. Kando. Evaluation of rich and explicit feedback for exploratory search. In *Proceedings of WEPIR 2019*, Glasgow, Scotland, U.K., 2019.

[14] Z. Yang, Y. Yamamoto, T. Yamamoto, N. Kando, and H. Ohshima. Finding the connection between artifact and personal knowledge of museum visitor. In *Proceedings of WEPIR 2019*, Glasgow, Scotland, U.K., 2019.