Evaluation of Personalised Information Retrieval at CLEF 2018 (PIR-CLEF)

Gabriella Pasi[1], Gareth J. F. Jones[2], Keith Curtis[2], Stefania Marrara[3(✉)], Camilla Sanvitto[1], Debasis Ganguly[4], and Procheta Sen[2]

[1] University of Milano Bicocca, Milan, Italy
pasi@disco.unimib.it
[2] Dublin City University, Dublin, Ireland
[3] Consorzio C2T, Milan, Italy
stefania.marrara@consorzioc2t.it
[4] IBM Research Labs, Dublin, Ireland

Abstract. The series of Personalised Information Retrieval (PIR-CLEF) Labs at CLEF is intended as a forum for the exploration of methodologies for the repeatable evaluation of personalised information retrieval (PIR). The PIR-CLEF 2018 Lab is the first full edition of this series after the successful pilot edition at CLEF 2017, and provides a Lab task dedicated to personalised search, while the workshop at the conference will form the basis of further discussion of strategies for the evaluation of PIR and suggestions for improving the activities of the PIR-CLEF Lab. The PIR-CLEF 2018 Task is the first PIR evaluation benchmark based on the Cranfield paradigm, with the potential benefits of producing evaluation results that are easily reproducible. The task is based on search sessions over a subset of the ClueWeb12 collection, undertaken by volunteer searchers using a methodology developed in the CLEF 2017 pilot edition of PIR-CLEF. The PIR-CLEF test collection provides a detailed set of data gathered during the activities undertaken by each subject during the search sessions, including their search queries and details of relevant documents as marked by the searchers. The PIR-CLEF 2018 workshop is intended to review the design and construction of the collection, and to consider the topic of reproducible evaluation of PIR more generally with the aim of improving future editions of the evaluation benchmark.

1 Introduction

The PIR CLEF Lab organized within CLEF 2018 has the aim of providing a framework for the evaluation of Personalised Information Retrieval (PIR). PIR systems are aimed at enhancing traditional IR systems to better satisfy the information needs of individual users by providing search results that are not only relevant to the query in general, but specifically to the user who submitted the query. In order to provide a personalised service, a PIR system leverages various kinds of information about the users and their preferences and interests,

P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 335–342, 2018.
https://doi.org/10.1007/978-3-319-98932-7_29

which are also inferred through a variety of interactions of the user with the system. The information gathered is then represented in a user model, which is typically employed to either improve the user's query or to re-rank retrieved results list obtained using the standard query, so that documents that are more relevant to the user, are presented in the top positions of the list.

In the literature, the issue of evaluating the effectiveness of personalised approaches to search has been the source of previous investigations, generally within the scope of research related to interactive information retrieval. The notion of relevance is user centered, and can vary during a search session, depending both on the task at hand and on the user's interactions with the search system. Existing work on the evaluation of PIR has investigated this issue under different perspectives. A category of approaches (the prominent ones) has relied on user-centered evaluations, mostly based on user studies; this approach involves real users undertaking search tasks in a supervised environment, and by posing the user at the centre of the evaluation activity can produce relevant and informed feedbacks. However, while this methodology has the advantage of enabling the detailed study of the activities of real users, it has the significant drawback of not being easily reproducible, thus greatly limiting the scope for algorithmic exploration. Among some previous attempts to define PIR benchmark tasks based on the Cranfield paradigm, the closest experiment to the PIR Lab is the TREC Session track[1] conducted annually between 2010 and 2014. This track focused on stand-alone search sessions, where a "session" is a continuous sequence of query reformulations on the same topic, along with any user interaction with the retrieved results in service of satisfying a specific information need; however no details of the searcher undertaking the task have been made available. Thus, the TREC Session track did not exploit any user model to personalise the search experience, nor did it allow user actions over multiple search session to be taken into consideration in the ranking of the search output.

The PIR-CLEF 2018 Lab provided search data from a single search session gathered by the activities of volunteer users within the context of a search carried out in a user selected broad search category. The data collected were the same as those for the earlier Pilot Lab in 2017 [9]. We plan in the future to gather data across multiple sessions to enable the construction and exploitation of persistent user behaviour across the multiple search sessions focusing on the same topical area, in the same manner as user searching consistently within a topical area of ongoing interest.

PIR-CLEF 2018 thus provides an evaluation framework and test collection to enable research groups working on PIR to both experiment with and provide feedback on our proposed PIR evaluation methodology.

The remainder of this paper is organised as follows: Sect. 2 outlines existing related work, Sect. 3 provides an overview of the PIR-CLEF 2018 task, Sect. 3.2 discusses the metrics available for the evaluation of the task, and Sect. 5 concludes the paper.

[1] http://trec.nist.gov/data/session.html.

2 Related Work

Recent years have seen increasing interest in the study of contextual search: in particular, several research contributions have addressed the task of personalizing search by incorporating knowledge of user preferences into the search process [2]. This user-centered approach to search has raised the related issue of how to properly evaluate the effectiveness of personalized search in a scenario where relevance is strongly dependent on the interpretation of the individual user. The essential question here is, what is the impact on search effectiveness which arises from the inclusion of personal information relating to the preferences of the individual user. To this purpose several user-based evaluation frameworks have been developed, as discussed in [3].

A first category of approaches aimed at evaluating PIR systems is focused on performing a user-centered evaluation by providing a kind of extension to the laboratory based evaluation paradigm. The TREC Interactive track [4] and the TREC HARD track [5] are examples of this kind of evaluation framework. These tracks aimed at involving users in interactive tasks to get additional information about the user and the query context. The evaluation was done by comparing a baseline run ignoring the user/topic metadata with another run considering it.

The more recent TREC Contextual Suggestion track [6] was proposed with the purpose of investigating search techniques for complex information needs that are highly dependent on both context and the user's interests. Participants in the track were given, as input, a set of geographical contexts and a set of user profiles that contain a list of attractions the user has previously rated. The task was to produce a list of ranked suggestions for each profile-context pair by exploiting the given contextual information. However, despite these extensions, the overall evaluation Was still system controlled and only a few contextual features were available in the process.

TREC also introduced a Session track [7] the focus of which was to exploit user interactions during a query session to incrementally improve the results within this session. The novelty of this task was the evaluation of system performance over entire sessions instead of a single query.

However, the above attempts had various limitations in satisfactorily injecting the user's behaviour into the evaluation; for this reason the problem of defining a standard approach to the evaluation of personalized search is a hot research topic, which needs effective solutions.

A first attempt to create a collection satisfactorily accounting for individual user behaviour in search was done in the FIRE Conference held in 2011. The Personalised and Collaborative Information Retrieval track [8] was organised with the aim of extending a standard IR ad-hoc test collection by gathering additional meta-information during the topic development process to facilitate research on personalised and collaborative IR. However, since no runs were submitted to this track, only preliminary studies have been carried out and reported using it.

As introduced above, within CLEF 2017 we organised the PIR-CLEF pilot study for the purpose of providing a forum to enable the exploration of the evaluation of PIR [9]. The Pilot Lab provided a preliminary edition of the 2018

PIR-CLEF Lab. One of the achievements of the PIR-CLEF 2017 Pilot Task was the establishment of an evaluation benchmark combining elements of a user-centered and the Cranfield evaluation paradigm, with the potential benefits of producing evaluation results that are easily reproducible. The task was based on search sessions over a subset of the ClueWeb12 collection, undertaken by 10 users by using a clearly defined and novel methodology. The collection was defined by relying on data gathered by the activities undertaken during the search sessions by each participant, including details of relevant documents as marked by the searchers. An important point is that the collection was developed but not used by any group participating at the pilot task. For this reason we were able to use this data collection as the develop dataset for the CLEF 2018 PIR-CLEF task This dataset was distributed to the 16 groups registered to the Lab. We have also prepared a second collection for PIR CLEF 2018, as well as a system able to perform a comparative evaluation of the algorithms developed by the participating groups.

3 Overview of the PIR-CLEF 2018 Task

As described in the previous sections, the goal of the PIR-CLEF 2018 Task was to investigate the potentiality of using a laboratory-based methodology to enable a comparative evaluation of PIR methodologies. The collection of data used during both PIR-CLEF 2017 and PIR-CLEF 2018 was carried out with the cooperation of volunteer users. In each case, the data collection was organized into two sequential phases:

- *Data gathering.* This phase involved the volunteer users carrying out a task-based search session during which the activities of the user were recorded (e.g., formulated queries, bookmarked documents, etc.). Each search session was composed of a phase of query development, refinement and modification, and associated search with each query on a specific topical domain selected by the user, followed by a relevance assessment phase where the user indicated the relevance of documents returned in response to each query and a short report writing activity based on the search activity undertaken. Further details of this procedure are provided in [1].
- *Data cleaning and preparation.* This phase took place once the data gathering had been completed, and did not involve any user participation. It consisted of filtering and elaborating the information collected in the previous phase in order to prepare a dataset with various kinds of information related to the specific user's preferences. In addition, a bag-of-words representation of the participant's user profile was created to allow comparative evaluation of PIR algorithms using the same simple user model.

For the PIR-CLEF 2018 Task we made available the user profile data and raw search data produced by guided search sessions undertaken by 10 volunteer users created for the IT-CLEF 2017 pilot, as detailed in Sect. 3.1. The data provided included the submitted queries, baseline ranked lists of documents retrieved

using a standard search system in response to each query, the items clicked by the user in response to this list, and document relevance information provided by the user on a 4-grade scale. Each session was performed by the users on a topic of their choosing, and search was carried out over a subset of the ClueWeb12 web collection.

The aim of the task was to use the provided information to improve the ranking of the search results list over a baseline ranking of documents judged relevant to the query by the user who entered the query.

The data was provided in csv format to the registered participants in the task. Access to the search service for the indexed subset of the ClueWeb12 collection was provided by Dublin City University via an API.

3.1 Dataset

To create datasets for distribution to the task participants, the data collected from the volunteer users was extracted and stored in csv files, and provided to the Lab participants in a zip folder.

Table 1. The PIR-CLEF dataset

cvs file	Content
cvs1	Info about the query session
cvs2	User's search log
cvs3	Relevance assessment of documents
cvs4	User's personal info
cvs5	TREC-style topic description
cvs6a	Simple user profile
cvs6b	User profile with stop words removal

As shown in Table 1, the file *user's session* (csv1) contains the information about each phase of the query sessions performed by each user. It also contains information about the user carrying the search including username, query_session ID and category, task and several timestamps of the session.

The file *user's log* (csv2) contains the search logs of each user, i.e. every search event that has been triggered by a user's action.

The file *user's assessment* (csv3) contains the relevance assessments of a pool of documents with respect to every single query developed by each user to fulfill the given task.

The file *user's info* (csv4) contains some personal information about the users such as age range, gender, occupation or native language.

The file *user's topic* (csv5) contains TREC-style final topic descriptions about the user's information needs that were developed in the final step of each search session, including also a short description provided by the searcher giving details

of the topic about which they were searching and a description of which documents are relevant to the topic and which are not.

The file *simple user profile* (csv6a) for each user contains simple profiles computed as *bag of words* (simple version - the applied indexing included tokenization, shingling, and index terms weighting).

The file *complex user profile* (csv6b) contains, for each user, the same information provided in csv6a, with the difference that the applied indexing was enriched by also including stop word removal.

The source used to extract the information employed to construct the two user profiles is the set of documents that the participant has assessed as relevant at the end of the tasks. The user's log file (cvs2) contains for each user all the queries.

Participants had the possibility to contribute to the task in two ways:

- The two user profile files (csv6a and csv6b) provide bag-of words profiles for the volunteer users, extracted by applying different indexing procedures to the considered documents. Participants could compare the results obtained by applying their personalisation algorithm on these queries with the results obtained and evaluated by the users on the same queries (and included in the user assessment file csv3). Their search had to be carried out on the ClueWeb12 collection, by using the API provided by DCU. Then, by using the 4-graded scale evaluations of the documents (relevant, somewhat relevant, non relevant, off topic) provided by the users and contained in the user assessment file csv3, it was possible to compute evaluation metrics for the created ranked lists. Note that documents that do not appear in csv3 were considered non-relevant.
- The challenge here was to use the raw data provided in the files csv1, csv2, csv3, csv4, and csv5 to create user profiles. In the approaches proposed in the literature, user profiles are formally represented as bags of words, as vectors, or as conceptual taxonomies, generally defined based on external knowledge resources (such as the WordNet and the ODP - Open Directory Project). The task here was more research oriented: to examine whether the information provided in test collection is sufficient to create a useful user profile. Also to consider whether there is information not present in the current test collection that could be included to improve the profile.

In the Lab we encouraged participants to be involved in this task by using existing or new algorithms and/or to explore new ideas. We also welcomed contributions that make an analysis of the task and/or of the dataset.

3.2 Performance Measures

At this first edition of the Lab, well known information retrieval metrics, such as Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG) were used to evaluate partcipants' results. However, a key objective of PIR-CLEF is to examine new methods of evaluating PIR, particularly within our Cranfield

based framework. In the pursuit of this we have developed a tool to enable comparative analysis of retrieval results for multiple runs across a session which is being used for explorative analysis of runs carried out using the PIR-CLEF collections. Further details on this tool are available in [10].

4 Towards More Realistic Evaluation of PIR

The PIR-CLEF 2018 Task gathered data from the volunteer searchers over only a single search session, in practice a user exploiting a certain information need is generally expected to gather information across multiple sessions. Over the course of these sessions the searcher will have multiple topics associated with their informations. Some topics will typically recur over a number of sessions, and while some search topics may be entirely semantically separate, others will overlap, and in all cases the users knowledge of the topic will progress over time and recall of earlier sessions may in some cases assist the searcher in later sessions looking at the same topic. Obviously, a personalisation model should imitate this behaviour. How to extend the data gathering methodology to this more realistic and complex situation requires further investigation.

There are multiple issues which must be considered, not least how to engage volunteer participants in these more complex tasks over the longer collections periods that will required. Given the multiple interacting factors highlighting above, work will also be required to consider how to account for these in the design of such an extended PIR test collection and the process of the information collection, to enable meaningful experiments to be conducted to investigate personalisation models and their use in search algorithms.

The design of the PIR-CLEF 2018 task makes the additional simplifying assumption of a simple relevance relationship between individual queries posed to the search engine by the retrieved documents. However, it is observed that users often approach an IR system with a more complex information seeking intention which can require multiple search interactions to satisfy. Further we can consider the relationship between the information seeking intention as it develops incrementally during the multiple search interactions and item retrieved at each stage in terms of usefulness to the searcher rather than simple relevance to the information need [11]. However, to operationalise these more complex factors in the development of a framework for evaluation of PIR is clearly challenging.

5 Conclusions and Future Work

This paper introduced the PIR-CLEF 2018 Personalised Information Retrieval (PIR) Workshop and the associated Task. The paper first introduced relevant existing work in the evaluation of PIR. The task is the first edition of a Lab dedicated to the theme of personalised search, after a successful pilot held at CLEF 2017. This is the first evaluation benchmark in this field based on the Cranfield paradigm, with the significant benefit of producing results easily reproducible.

An evaluation using this collection has been run to allow research groups working on personalised IR to both experience with and provide feedback about our proposed PIR evaluation methodology. While the Task moves beyond the state-of-the-art in evaluation of PIR, it nevertheless makes simplifying assumptions in terms of the user's interactions during a search session; we briefly considered these here, and how to incorporate these into more evaluation of PIR that is closer to real-world user experience will be the subject of further work.

References

1. Sanvitto, C., Ganguly, D., Jones, G.J.F., Pasi, G.: A laboratory-based method for the evaluation of personalised search. In: Proceedings of the Seventh International Workshop on Evaluating Information Access (EVIA 2016), a Satellite Workshop of the NTCIR-12 Conference, Tokyo Japan (2016)
2. Pasi, G.: Issues in personalising information retrieval. IEEE Intell. Inform. Bull. **11**(1), 3–7 (2010)
3. Tamine-Lechani, L., Boughanem, M., Daoud, M.: Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowl. Inf. Syst. **24**(1), 1–34 (2009)
4. Harman, D.: Overview of the fourth text retrieval conference (TREC-4). In: Proceedings of the Fourth Text REtrieval Conference (TREC-4), Gaithersburg, Maryland (1995)
5. Allan, J.: HARD track overview in TREC 2003: high accuracy retrieval from documents. In: Proceedings of The Twelfth Text REtrieval Conference (TREC 2003), Gaithersburg, Maryland, USA, pp. 24–37 (2003)
6. Dean-Hall, A., Clarke, C.L.A., Kamps, J., Thomas, P., Voorhees, E.M.: Overview of the TREC 2012 contextual suggestion track. In: Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012), Gaithersburg, Maryland (2012)
7. Carterette, B., Kanoulas, E., Hall, M.M., Clough, P.D.: Overview of the TREC 2014 session track. In: Proceedings of The Twenty-Third Text REtrieval Conference (TREC 2014), Gaithersburg, Maryland, USA (2014)
8. Ganguly, D., Leveling, J., Jones, G.J.F.: Overview of the personalized and collaborative information retrieval (PIR) track at FIRE-2011. In: Majumder, P., Mitra, M., Bhattacharyya, P., Subramaniam, L.V., Contractor, D., Rosso, P. (eds.) FIRE 2010-2011. LNCS, vol. 7536, pp. 227–240. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40087-2_22
9. Pasi, G., Jones, G.J.F., Marrara, S., Sanvitto, C., Ganguly, D., Sen, P.: Overview of the CLEF 2017 personalised information retrieval pilot lab (PIR-CLEF 2017). In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 338–345. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_29
10. Pasi, G., et al.: Overview of the CLEF 2018 personalised information retrieval pilot lab (PIR-CLEF 2018): methods for comparative evaluation of PIR. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France (2018)
11. Belkin, N.J., Hienert, D., Mayr-Schlegel, P., Shah, C.: Data requirements for evaluation of personalization of information retrieval - a position paper. In: Proceedings of Working Notes of the CLEF 2017 Labs, Dublin, Ireland (2017)