Overview of the CLEF 2019 Personalised Information Retrieval Lab (PIR-CLEF 2019)

Gabriella Pasi[1], Gareth J. F. Jones[2], Lorraine Goeuriot[3], Liadh Kelly[4], Stefania Marrara[5(✉)], and Camilla Sanvitto[1]

[1] University of Milano Bicocca, Milan, Italy
[2] Dublin City University, Dublin, Ireland
[3] Universit Grenoble Alpes, Grenoble, France
[4] Maynooth University, Maynooth, Ireland
[5] Consorzio C2T, Milan, Italy
stefania.marrara@consorzioc2t.it

Abstract. The Personalised Information Retrieval Lab (PIR-CLEF 2019) lab is an initiative aimed at both providing and critically analysing the evaluation of Personalization in Information Retrieval (PIR) applications. PIR-CLEF 2019 is the second edition of the Lab after the successful Pilot lab organised at CLEF 2017 and the first edition of the Lab at CLEF 2018. PIR-CLEF 2019 provided registered participants with two tracks: the Web Search Task and the Medical Search Task. The Web Search Task continues the activities introduced in the previous editions of the PIR-CLEF Lab, while the Medical Search Track focuses on personalisation within an ad hoc search task introduced in previous editions of the CLEF eHealth Lab.

1 Introduction

The PIR-CLEF Lab at CLEF 2019 provides a framework for the evaluation of Personalised Information Retrieval (PIR). The application of personalisation in IR seeks to enhance traditional IR systems to better satisfy the information needs of individual users. PIR systems are intended to provide search results that are not only relevant to the query in general, but more specifically to the specific interests of the user who submitted the query. In order to provide a personalised service, a PIR system can leverage various types of information about the current user and their preferences and interests. These can be stated directly by the user, or may be inferred through a variety of interactions of the user with the system. This information is then represented in some form of user model, which can be employed in the search process with the objective of improving the search results for this user. This typically operates either by seeking to improve the user's query or by re-ranking a set of retrieved results so that documents more relevant to the user are presented in the top positions of a retrieval list.

F. Crestani et al. (Eds.): CLEF 2019, LNCS 11696, pp. 417–424, 2019.
https://doi.org/10.1007/978-3-030-28577-7_31

Evaluating the effectiveness of personalised approaches to search has been investigated for many years within studies of interactive IR In this work, the notion of relevance has been user centered with potential variation during a search session, depending both on the task at hand and on the user's interactions with the search system. This work has mostly been based on user studies; this approach involves real users undertaking search tasks in a supervised environment. By placing the user at the centre of the evaluation activity these studies have produced valuable insights and feedback. 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. This greatly limits the scope for algorithmic exploration of technologies for search personalisation. 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 sessions to be taken into consideration in the ranking of the search output.

The PIR-CLEF 2019 Lab had 29 registrations, and offered two distinct tasks: the Web Search Task and the Medical Search Task to evaluate personalised search.

The Web Search Task aimed to explore personalisation in Web search sessions based on user profiles and activity with the current and previous search sessions by this user [1]. Task participants were provided with user profile data and logs of search activity with the objective of improving search effectiveness over that achieved in the logged search sessions. In particular, the participants were provided with a test dataset and with a training dataset, delivered in two stages. This task was a direct extension of the tasks offered in PIR-CLEF at CLEF 2017 [2] and CLEF 2018 [3].

The Medical Search Task addresed medical search, which is one of the most common interests of users of search engines. For this year's pilot task, the challenge offered to participants was to work on the task of generating PIR techniques for queries posed by patients on viewing their discharge summaries, where the discharge summaries are used in this personalisation process. Optionally additional resources (ontologies) could also be used in their IR techniques. Participants were invited to submit any type of run they wanted to, so long as it was personalised in some way. This pilot challenge used test collections originating from CLEF eHealth 2013 and 2014 IR challenges [4,5].

The remainder of this paper is organised as follows: Sect. 2 outlines existing related work, Sect. 3 provides an overview of the PIR-CLEF 2019 tasks, and Sect. 4 concludes the paper.

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

2 Related Work

Recent years have seen increasing attention in the study of context in search. Of particular interest here is the personalisation of search by incorporating knowledge of user preferences into the search process [6]. This focus on the incorporation of the individual user into the search process has raised the related issue of how to properly evaluate the effectiveness of personalised search in a scenario where relevance is strongly dependent on the interpretation of the individual user. To this end, several user-based evaluation frameworks have been developed, as discussed in [7].

A first category of approaches aimed at evaluating Personalised Information Retrieval Systems (PIRS) which focus on performing user-centered evaluation by providing a kind of extension to the laboratory based evaluation paradigm. The TREC Interactive track [8] and the TREC HARD track [9] are examples of this kind of evaluation framework. These aimed at involving users in interactive tasks to get additional information about them and the query context. The evaluation was done by comparing a baseline run ignoring the user/topic metadata with another run in which it is considered.

The more recent TREC Contextual Suggestion track [10] 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 containing a list of attractions that the user had 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 [11], the focus of which was to exploit user interactions during a query session to incrementally improve the results within that session. The novelty of this task was the evaluation of system performance over entire sessions instead of a single query.

The above tasks have various limitations in their injection of user behaviour into the evaluation process; for this reason the problem of defining a standard approach to the evaluation of personalised search is a hot research topic, which needs effective solutions.

A first attempt to create a collection satisfactorily accounting for the user behaviour in search was done in the FIRE Conference held in 2011. The Personalised and Collaborative Information Retrieval track [12] 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 were carried out and reported using it.

Within CLEF 2017 we launched the PIR-CLEF benchmark with a pilot study and workshop (PIR CLEF 2017), for the purpose of providing a forum for the exploration of the evaluation of PIR. The PIR-CLEF 2017 Pilot Task was a Web Search task which sought to combine user-centered methods with the

Cranfield evaluation paradigm, with the key potential benefit of producing evaluation results that are easily reproducible. The Pilot task was based on search sessions over a subset of the ClueWeb12 document collection, undertaken by 10 users by using a clearly defined and novel methodology. This collection was distributed to the participants of the PIR-CLEF 2018 [3]. A second collection was also prepared for the PIR-CLEF 2018 lab, but this was not used by any participants. PIR-CLEF at CLEF 2019 continued with the Web Search Task, but also introduced a Medical Search Task.

3 PIR-CLEF 2019 Search Tasks

In this section, we give details of the two tasks offered at PIR-CLEF 2019. We look first at the Web Search Task and then give details of the Medical Search Task.

3.1 Web Search Task

The PIR-CLEF 2019 Web Search task used both of the collections previously developed for PIR-CLEF 2017 and PIR-CLEF 2018. The 2017 data collection was released to PIR-CLEF 2019 participants first as a training dataset, with the 2018 collection released later as a test dataset. The data collection and processing are described in detail in [3]. Here we give summary details of the collection procedure.

- *Data gathering.* This phase involved a group of volunteer users carrying out a task-based search session. Each session was performed by the user on a topic of her choice selected from a provided list of broad topics, and search carried out over a subset of the ClueWeb12 web collection. During this session the activities performed by 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.
- *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.

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

The task data provided to registered participants consisted of a set of csv files. They were also provided with access to the search service used in the user search sessions, provided via an API by Dublin City University.

The data provided included the submitted queries, the baseline ranked lists of documents retrieved in response to each query by using a standard search system, the items clicked by the user in the result list, and the documents' relevance assessments provided by the user on a 4-grade scale. The data was extracted and stored in csv format in 7 files in a zip folder. Full details of the files are given in [3].

We encouraged participants to be involved in the 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.

The metrics and methodology used to evaluate and analyze PIR tasks pose significant challenges, which is one of the key motivations underlying the development of the PIR-CLEF Lab. It is not at all obvious how we might properly compare and contrast the behaviour of alternative methods of integrating personalisation into search sessions. While we can begin by using standard metrics, such as Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG) these are not sufficient to enable a detailed session based analysis of PIR methods.

As a starting point for the development of formal methodology for analysis and evaluation of our framework for laboratory-based evaluation of PIR in the Web Search task, we developed a prototype evaluation tool described in [3]. This tool has been tested with sample data collected using our session-based framework, and the intention is to analyze participant submissions to the PIR-CLEF 2019 Web Search task using this tool, both to enable detailed analysis of their submissions, but also to enable us to refine the features of the analysis tool itself.

3.2 Medical Search Task

For the new pilot Medical Search task, we challenged participants to work on the task of generating personalised retrieval techniques for the queries posed by patients on viewing their discharge summaries. The discharge summaries and optionally other external resources were to be used in this personalisation process. For this challenge a large collection of web pages was provided, along with patient queries and associated discharge summaries. This test collection was generated in the CLEF eHealth 2013 and 2014 IR challenges [4,5]. Participants were invited to submit any type of run they wanted to the challenge, so long as it was somehow personalised.

The data set for this task consists of a set of medical-related documents, provided by the Khresmoi project[2]. This collection contains documents covering a broad set of medical topics, and does not contain any patient information. The documents in the collection come from several online sources, including the

[2] http://www.khresmoi.eu.

Health On the Net organisation certified websites, as well as well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia). The topic set consists of 50 topics which were previously distributed with CLEF eHealth 2013 IR challenge and 50 topics which were distributed with CLEF eHealth 2014 IR challenge.

The 50 2013 topics were manually generated by medical professionals from highlighted disorders identified in annotated medical discharge summaries. In 2014, the 50 topics were manually generated by medical professionals from the main disorder diagnosed in the discharge summary. A mapping between queries and task 1 matching discharge summary is provided, the participants can get access to the discharge summary (from the MIMIC II database), explained below.

Topics consist of:

- *Title:* text of the query,
- *Description:* longer description of what the query means,
- *Narrative:* expected content of the relevant documents,
- *Profile:* main information on the patient (age, gender, condition),
- *Discharge_summary:* ID of the matching discharge summary

For the 2019 challenge, participants were challenged to use the discharge summaries for personalisation.

The *discharge summaries* consist of deidentified clinical free-text notes from the MIMIC II database, version 2.5[3]. Notes were authored in the ICU setting and note types include discharge summaries, ECG reports, echo reports, and radiology reports (for more information about the MIMIC II database, we refer the reader to the MIMIC User Guide).

The PIR-CLEF organisers did not provide direct access to the discharge summaries, participants were required to follow MIMIC II guidelines to access it[4].

Relevance judgements (qrel files) created by the CLEF eHealth challenge are also used in this year's challenge were provided. Details on how the qrel files were generated are available in [4,5].

Given this year's pilot Medical search task did not offer an interactive personalisation element, standard IR evaluation metrics were used for this task in 2019. Specifically, the focus was on P@5, P@10, NDCG@5, NDCG@10. Evaluation metrics are computed using the trec_eval tool[5].

4 Conclusions and Future Work

This paper described the Web Search and Medical Search Tasks offer at PIR-CLEF 2019. The Web Search task extends the work of the previous PIR-CLEF activities at CLEF 2017 and CLEF 2018, while the Medical Search Task builds

[3] http://mimic.physionet.org/.
[4] https://mimic.physionet.org/gettingstarted/access/.
[5] http://trec.nist.gov/trec_eval/.

on previous work in the CLEF eHealth Lab. The purpose of the PIR-CLEF Lab is to enable research groups to work on comparative evaluation methods for the introduction of personalisation in IR and to study its evaluation. Unfortunately, while almost 30 groups registered to participant in the PIR-CLEF 2019 Lab, none of them returned results for either of the available tasks.

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., 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
3. Pasi, G., et al.: Overview of the CLEF 2018 personalised information retrieval lab (PIR-CLEF 2018). In: Proceedings of CLEF 2018, Avignon, France (2018)
4. Goeuriot, L., et al.: ShARe/CLEF eHealth evaluation lab 2013, task 3: information retrieval to address patients' questions when reading clinical reports. In: CLEF 2013 Online Working Notes, 8138 (2013)
5. Goeuriot, L., et al.: ShARe/CLEF eHealth evaluation lab 2014, task 3: user-centred health information retrieval. In: Proceedings of CLEF 2014 (2014)
6. Pasi, G.: Issues in personalising information retrieval. IEEE Intell. Inform. Bull. **11**(1), 3–7 (2010)
7. 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)
8. Harman, D.: Overview of the fourth text retrieval conference (TREC-4). In: Harman, D.K., (eds.) TREC, volume Special Publication 500–236. National Institute of Standards and Technology (NIST) (1995)
9. 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)
10. Dean-Hall, A., Clarke, C.L.A., Kamps, J., Thomas, P., Voorhees, E.M.: Overview of the TREC 2012 contextual suggestion track. In Voorhees and Bucklan
11. 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)
12. 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
13. Villegas, M., Puigcerver, J., Toselli, A.H., Sanchez, J.A., Vidal, E.: Overview of the ImageCLEF 2016 handwritten scanned document retrieval task. In: Proceedings of CLEF 2016 (2016)
14. Robertson, S.: A new interpretation of average precision. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), pp. 689–690. ACM, New York (2008)

15. Angiolillo, A.: Comparative Evaluation of Personalised Search Systems. Università degli Studi di Milano Bicocca, Milano (2017)
16. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. (TOIS) **27**(1), 2 (2008)
17. Bai, Q., Chen, J., Hu, Q., He, L.: ECNU at CLEF PIR 2018: evaluation of personalized information retrieval. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France (2018)
18. Andreu-Marín, A., Martnez-Santiago, F., Ureña-López, L.A., Díaz-Galiano, M.C.: PIR Based in explicit and implicit feedback. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France (2018)