Evaluation of Personalised Information Retrieval at CLEF 2017 (PIR-CLEF): towards a reproducible evaluation framework for PIR

Gabriella Pasi [1], Gareth J. F. Jones[3], Stefania Marrara[1], Camilla Sanvitto[1], Debasis Ganguly,[2] Procheta Sen,[3]
[1] University of Milano Bicocca, Italy
[2] IBM Research Labs, Dublin, Ireland
[3] Dublin City University, Dublin, Ireland

Abstract. The Personalised Information Retrieval (PIR-CLEF) Lab workshop at CLEF 2017 is designed to provide a forum for the exploration of methodologies for the repeatable evaluation of personalised information retrieval (PIR). The PIR-CLEF 2017 Lab provides a preliminary pilot edition of a Lab task dedicated to personalised search, while the workshop at the conference is intended to provide a forum for the discussion of strategies for the evaluation of PIR and extension of the pilot Lab task. The PIR-CLEF 2017 Pilot Task is the first 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 10 users by using a clearly defined and novel methodology. The collection provides data gathered by the activities undertaken during the search sessions by each participant, including details of relevant documents as marked by the searchers. The PIR-CLEF 2017 workshop is intended to review the design and construction of this Pilot collection and to consider the topic of reproducible evaluation of PIR more generally with the aim of launching a more formal PIR Lab at CLEF 2018.

1 Introduction

The objective of the PIR-CLEF Lab is to develop and demonstrate the effectiveness of a methodology for the repeatable 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 but also to the specific user who submitted the query. In order to provide a personalised service, a PIR system maintains information about the user and their preferences and interests. These personal preferences and interests are typically inferred through a variety of interactions modes between the user with the system. This information is then represented in a user model, which is used to either improve the user's query or to re-rank a set of retrieved results list so that documents that are more relevant to the user are presented in the top positions of the ranked list.

Existing work on the evaluation of PIR has generally relied on a user-centered approach, mostly based on user studies; this approach involves real users undertaking search tasks in a supervised environment. 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 and does not support the extensive exploration of the design and construction of user models and their exploitation in the search process. These limitations greatly restrict the scope for algorithmic exploration in PIR. This means that it is generally not possible to make definitive statements about the effectiveness or suitability of individual PIR methods and meaningful comparison between alternative approaches.

Among existing IR benchmark tasks based on the Cranfield paradigm, the closest task to the evaluation of PIR is the TREC Session track[1] conducted annually between 2010 and 2014. However, this track focused only 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 topical information need, for which no details of the searcher undertaking the task are 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. In FIRE another attempt was made to set a framework for the evaluation of Personalized Search, under controlled experimental settings [8].

The PIR-CLEF 2017 Pilot Task provides search data from a single search session with personal details of the user undertaking the search session. This test collection was created using the methodology described in [1]. Since this was a pilot activity we encouraged participants to attempt the task using existing algorithms and to explore new ideas. We also welcomed contributions to the workshop examining the specification and contents of the task and the provided dataset.

The PIR-CLEF 2017 workshop at the CLEF 2017 Conference brought together researchers working in PIR and related topics to explore the development of new methods for evaluation in PIR.

The remainder of this paper is organised as follows: Section 2 outlines existing related work, Section 3 provides an overview of the PIR-CLEF 2017 Pilot task, and Section 5 concludes the paper.

2 Related Work

Recent years have seen increasing interest in the study of contextualisation in search: in particular, several research contributions have addressed the task of personalising search by incorporating knowledge of user preferences into the search process [2]. This user-centred approach to search has raised the related issue of how to properly evaluate search results in a scenario where relevance is

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

strongly dependent on the interpretation of the individual user. For this purpose several user-based evaluation frameworks have been developed, as discussed in [3].

A key issue when seeking to introduce personalisation into the search process is the evaluation of the effectiveness of the proposed method. A first category of approaches aimed at evaluating personalised search systems attempts to perform a user-centred evaluation provided 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, which aimed at involving users in interactive tasks to get additional information about them and the query context being formulated. 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 context and users interests. As input, participants in the track were given 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] whose focus 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.

For all these reasons, 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 in support of PIR research 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.

3 Overview of the PIR-CLEF 2017 Pilot Task

The goal of the PIR-CLEF 2017 Pilot Task was to investigate the use of a laboratory-based method to enable a comparative evaluation of PIR. The pilot collection used during PIR-CLEF 2017 was created with the cooperation of volunteer users, and was organized into two sequential phases:

- *Data gathering.* This phase involved the volunteer users carrying out a task-based search session during which a set of 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.

For the PIR-CLEF 2017 Pilot Task we made available the user profile data and raw search data produced by guided search sessions undertaken by 10 volunteer users as detailed in section 3.1.

The aim of the task was to use the provided information 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 Pilot Task data were provided in csv format to 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

For the PIR-CLEF 2017 Pilot Task we made available both user profile data and raw search data produced by guided search sessions undertaken by 10 volunteer users. 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. Each session was performed by the user on a topic of her choice selected from a provided list of broad topics, and search was carried out over a subset of the ClueWeb12 web collection.

The data has been extracted and stored in csv format as detailed in the following. In particular 7 csv files were provided in a zip folder. The file *user's session* (csv1) contains the information about each phase of the query sessions performed by each user. Each row of the csv contains:

- username: the user who performed the session
- query_session: id of the performed query session
- category: the top level search domain of the session
- task: the description of the search task fulfilled by the user
- start_time: starting time of the query session
- close_time: closing time of the search phase
- evaluated_time, closing time of the assessment phase

– end_time: closing time of the topic evaluation and the whole 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 users action. The file row contains:

– username: the user who performed the session
– query_session: id of the query session within the search was performed
– category: the top level search domain
– query_text: the submitted query
– document_id: the document on which a particular action was performed
– rank: the retrieval rank of the document on which a particular action is performed
– action_type: the type of the action executed by the user (query submission, open_document, close_document, bookmark)
– time_stamp: the timestamp of the 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:

– username: the user who performed the session
– query_session: id of the query session within the evaluation was performed
– query_text: the query on which the evaluation is based
– document_id: the document id for which the evaluation was provided
– rank: the retrieval rank of the document on which a particular action is performed
– relevance_score: the relevance of the document to the topic (1 off-topic, 2 not relevant, 3 somewhat relevant, 4 relevant).

The file *user's info* (csv4) contains some personal information about the users:

– username
– age_range
– gender
– occupation
– native_language.

The file *user's topic* (csv5) contains the TREC-style final topic descriptions about the users information needs that were developed in the final step of each search session:

– username, the user who formulated the topic
– query_session, id of the query session which the topic refers to
– title, a small phrase defining the topic provided by the user
– description, a detailed sentence describing the topic provided by the user
– narrative, a description of which documents are relevant to the topic and which are not, provided by the user

The file *simple user profile* (csv6a) for each user contains the following information (simple version - the applied indexing included tokenization, shingling, and index terms weighting):

- username: the user whose interests are represented
- category: the search domain of interest
- a list of triples constituted by:
  - a term: a word or n-grams related to the users searches
  - a normalised_score: term weight computed as the mean of the term frequencies in the users documents of interests, where term frequency is the ratio of the number of occurrences of the term in a document and the number of occurrences of the most frequent term in the same document.

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:

- username, the user whose interests are represented
- category, the search domain of interest
- a list of triples constituted by:
  - term, a word or a set of words related to the users searches
  - normalised_score,

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

- the two user profile files (csv6a and csv6b) provide the bag-of words profiles of the 10 users involved in the experiment, extracted by applying different indexing procedures to the documents. The user's log file (cvs2) contains for each user all the queries she formulated during the query session. The participant 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). The 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 Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG). Note that documents that do not appear in csv3 were considered non-relevant. As further explained in section 3.2, these metrics were computed both globally (in the literature they are just AP and NDCG) and for each user query individually, by then taking the mean.
- The challenge here was to use the raw data provided in csv1, csv2, csv3, csv4, and csv5 to create user profiles. A user profile is a formal representation of the user interests and preferences; the more accurate the representation of the user model, the higher is the probability to improve the search process. 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 request here was more research oriented: are the provided information sufficient to create a useful profile? Which information is missing? The outcome here was a report up to 6 pages discussing the theme of user information to profiling aims, by proposing possible integrations of the provided data and by suggesting a way to collect them in a controlled Cranfield style experiment.

Since this was a pilot activity 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 preliminary stage, well known information retrieval metrics, such as Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG) can be considered to benchmark the participants' systems. However, new metrics should be investigated to evaluate the task of personalised search.

4 Towards More Realistic Evaluation of PIR

The PIR-CLEF 2017 Pilot Task gathered data from the volunteer searchers over only a single search session, in practice a personalisation model is generally expected to gather and exploit information across multiple sessions for the searcher. 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. 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 2017 Pilot 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 an a more complex information seeking intention which can require multiple search iteractions to satisfy. Further we can consider the relationship between the information seeking intention as it develops

incrementally during the multiple search iteractions 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 2017 Personalised Information Retrieval (PIR) Workshop and the associated Pilot Task. The paper first introduced relevant existing work in the evaluation of PIR. The Pilot task is the preliminary edition of a Lab dedicated to the theme of personalised search that is planned to officially start at CLEF 2018. This is the first evaluation benchmark in this field based on the Cranfield paradigm, with the significant benefit of producing results easily reproducible. A pilot 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 Pilot 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. C. Sanvitto, D. Ganguly, G. J. F. Jones and G. Pasi *A Laboratory-Based Method for the Evaluation of Personalised Search.* Proceedings of the Seventh International Workshop on Evaluating Information Access (EVIA 2016), a Satellite Workshop of the NTCIR-12 Conference, June 7, 2016 Tokyo Japan.
2. G. Pasi. *Issues in personalising information retrieval.* IEEE Intelligent Informatics Bulletin, 11(1):37, 2010.
3. L. Tamine-Lechani, M. Boughanem, and M. Daoud. *Evaluation of contextual information retrieval effectiveness: overview of issues and research.* Knowledge and Information Systems, 24(1):134, 2009.
4. D. Harman. *Overview of the fourth text retrieval conference (TREC-4).* In D. K. Harman, editor, TREC, volume Special Publication 500-236. National Institute of Standards and Technology (NIST), 1995.
5. J. Allan. *HARD track overview in TREC 2003: High accuracy retrieval from documents.* In Proceedings of The Twelfth Text REtrieval Conference (TREC 2003), pages 2437, Gaithersburg, Maryland, USA, 2003.
6. Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Paul Thomas, and Ellen M. Voorhees. Overview of the TREC 2012 contextual suggestion track. In Voorhees and Bucklan.
7. B. Carterette, E. Kanoulas, M. M. Hall, and P. D. Clough. *Overview of the TREC 2014 session track.* In Proceedings of The Twenty-Third Text REtrieval Conference (TREC 2014), Gaithersburg, Maryland, USA.

8. Debasis Ganguly, Johannes Leveling, and Gareth J. F. Jones. Overview of the personalized and collaborative information retrieval (PIR) track at FIRE-2011. In Prasenjit Majumder, Mandar Mitra, Pushpak Bhat- tacharyya, L. Venkata Subramaniam, Danish Contractor, and Paolo Rosso, editors, Multilingual Information Access in South Asian Lan- guages - Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers, volume 7536 of Lecture Notes in Computer Science, pages 227240. Springer, 2011.
9. M. Villegas, J. Puigcerver, A. H. Toselli, J.A. Sanchez and E. Vidal. *Overview of the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task.* In Proceedings of CLEF 2016.
10. S. Robertson. A new interpretation of Average Precision. In Proceedings of the International ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). pp.689-690. ACM, New York, NY, USA (2008).
11. N.J.Belkin, D.Hienert, P. Mayr-Schlegel and C.Shah1, Data Requirements for Evaluation of Personalization of Information Retrieval A Position Paper, In Proceedings of Working Notes of the CLEF 2017 Labs. Dublin, Ireland.