A Proposed Method for Laboratory-Based Evaluation of Personalised Information Retrieval

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

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

Andrea Angiolillo
Department of Informatics, Systems & Communication
University of Milano-Bicocca
Milan, Italy
a.angiolillo@campus.unimib.it

Camilla Sanvitto
Department of Informatics, Systems & Communication
University of Milano-Bicocca
Milan, Italy
c.sanvitto@campus.unimib.it

ABSTRACT

Comparative evaluation of Information Retrieval Systems (IRSs) using publically available test collections has become an established practice in Information Retrieval (IR). This approach using the popular Cranfield evaluation paradigm enables researchers to compare alternative methods of addressing a specific IR task. This strategy has not to date been applied to the evaluation of Personalised Information Retrieval (PIR), for which evaluation has generally focused on user-based studies. While such studies these experiments provide many valuable insights, they do not enable the systematic comparative evaluation of the combination of IR methods and user models in alternative algorithmic approaches to PIR. In an attempt to enable such comparative invesigations, we introduce an experimental framework for the creation of test collections to facilitate repeatable laboratory-based evaluation of PIR, together with a prototype evaluation tool to analyze results collected using these collections.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

personalised information retrieval; laboratory-based evaluation; test collection development

ACM Reference Format:
Gareth J. F. Jones, Gabriella Pasi, Andrea Angiolillo, and Camilla Sanvitto. 2018. A Proposed Method for Laboratory-Based Evaluation of Personalised Information Retrieval. In *Proceedings of Workshop on Evaluating Personalised Information Retrieval (WEPIR 2018).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

WEPIR 2018, March 2018, New Brunswick, NJ, USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Recent years have seen growing interest in personalisation of search services within both the research community and private companies. The original search paradigm of "one size fits all" has been replaced by services which take account of user preferences and behaviours. A natural question which arises when seeking to incorporate personalisation into search is: what mechanisms should be used to exploit information relating to personal features and preferences and previoius search hisotry to achieve the greatest information retrieval (IR) effectiveness for the user?

The standard approach to assessing the effectiveness of traditional static information retrieval systems (IRSs) is to adopt a laboratory-based method using the Cranfield paradigm. Following this methodology, researchers assess effectiveness of an IRS via experiments on test collections in a controlled, laboratory-like setting. This setup ensures that evaluations obtained by different research teams are consistent and repeatable, and enables different IR methods to be compared directly. The test collections used in these evaluations typically consists of a document set, search queries (generally referred as as *topics//* oftne designed or chosen according to some criteria, and relevance data for each topic indicating documents deemed to fulfull the user information need of the topic. These relevance assessments are created manually by judging a selected subset of retrieved items for each topic, and for reasons of scale, often provided by third partie uninvolved in the topic creation process.

While this approach has proved invaluable in supporting research into many IR tasks, in its standard form it has shortcomings which prevent its use to support research into Personalised IR Systems (PIRSs). The most fundamental of these problems is that it ignores the individual user engaged in search process, and the concepts of their background and search preferences and experiences and the context of the search relating to a specfic topic in the evaluation process. Hence, the standard test collections currently available for conducting experiments lack suitable data to support the evaluation of personalised search. For this reason, the evaluation of PIRSs has generally relied on user-centred approaches, mostly based on user studies, i.e. experiments that involve real users in a supervised environment. While this kind of evaluation has the advantage of accounting for the subjectivity of real users, it has the significant

drawback of not being easily reproducible, meaning that it is not possible to conduct extensive comparative evaluation of alternative approaches to PIR. Thus, the development of an evaluation methodology with corresponding publically available test collection that enables repeatable evaluation of personalised search algorithms and potentially to simulate user behaviour would be highly beneficial to the IR research community.

In this paper we outline a novel method for the creation of test collections to support PIR research and introduce a prototype evaluation tool to enable comparative laboratory-based evaluation of PIR methods. The procedure described in this paper was first implemented for the pilot instance of the PIR-CLEF task at CLEF 2017 [5]. The procedure is being followed again with adjustments for the PIR-CLEF task at CLEF 2018 using the analysis tool outlined at the end of this paper.

2 DESIGN OF TEST COLLECTION FOR EVALUATION OF PIR

Our test collections for evaluation of PIR provides the traditional components needed in a laboratory-based evaluation experiment, including topic statements of user information need and corresponding manual relevance judgments for these topics. Since we wish to be able to support a wide range of users with diverse knowledge and interests, while creating a generally available test collection for repeatable experiments, for our current work we use the ClueWeb12, large crawl of over 730 million Web pages as the document collection [1].

These standard IR test collection components are accompanied by a new set of user-related information for modeling which introcorporates personal information, previous search activity and context into the evaluation.

- **user personal information**: including gender, age range, native language, occupation.
- **search logs**: log of the user's interactions with a search engine while performing search tasks.
- **documents of interest to user**: documents from the collection accessed by the user.

In order to include personal relevance data into the test collection, as part of the data collection process, participants are required to provide relevance assessments of a subset of the top ranked documents retrieved in response to their request.

In order to support investigation of direct comparison of personalisation methods incorporated into the search process. our collection also includes a bsic bag-of-words personal profile based on the participant's previous sesrch activities.

2.1 Data gathering procedure

To develop our test collection as outlined above, we adopt the following process for gathering the data.A more detailed description of this procedure can be found in [6].

The collection procedure is divided into two phases:

(1) **data gathering**: This phase involves a group of volunteer users carrying out a task-based search session during which a set of activities performed by the user are recorded (e.g, formulated queries, bookmarked documents, etc.). Each search session is composed of a phase of query development, refinement and modification; each query that the user formulates within this session is evaluated by the search engine, and the results presented to the participant in a SERP with which can interact and refine their query for a further search operation, until they decide to end their session. At the end of the session this is followed by a relevance assessment phase where the participant indicates the relevance of documents returned in response to each of their queries, and completes a short report writing activity based on the search activity undertaken.
(2) **data cleaning and preparation**: This phase takes place once the data gathering has been completed, and does not involve any user participation. It consists of filtering and elaborating the information collected in the previous phase to prepare a dataset with various kinds of information related to the specific user's preferences. In addition, the bag-of-words representation of the participant's user profile is created to allow comparative evaluation of PIR algorithms using the same simple user model.

2.2 Data gathering

The data gathering phase is performed by a groups of volunteer participants in a controlled way to ensure the quality of the entire process. Search activities are preceded by the gathering of each user's personal information, such as gender, age, native language, and job. This is following by the search phase in which the participants carry out a series of task-based sessions with all of their activities being recorded.

2.2.1 Phase 1.1: Topic development. The first phase of a task session is the development of an information need or topic to enable the participant to gain knowledge on a specific subject by performing searches. Since the main objective of the collection is to capture the personal interests of the participants, this process of information generation has been designed in such a way that it allows for personalisation.

Search category selection. The participant first selects a search category from among a predefined set, such as art, books, movies, music, sport, travel. Using high level search categories such as these allows us to categorise the collected data and capture important details about the user's topical interests within their chosen category area. After selecting the category the participant is given a search task, which is an assignment that has to be completed by finding information through an interactive search phase using a provided search engine, we refer to this as the "search session".

Since the objective of this work is to collect a rich set of information about the participants and their interests, search tasks are defined based on examples of informational and exploratory tasks [2, 4]. Users that follow this kind of task are more likely to submit a good number of queries during each search session than when they are given a task of finding information about specific facts.

Another important characteristic of the tasks for this search activity is that they cannot be either too specific nor too vague. A task which is too specific may force participants to search for topics they are not interested in; while a task which is too vague may

lead to confused and random searches. Therefore, for this work the tasks have been defined to strike a balance between being closely focused needs and those allowing freedom of interpretation by the users. A further desirable feature for the tasks is not to be linked to current events or situations because the ClueWeb 12 document collection derives from 2012.

For example, the following is one of the tasks assigned for the search category *travel*: "You are a travel lover and it is now time to plan your coming *vacation trip*. You have always wanted to visit your destination (city or country of your choosing) and now you finally the chance to do so. Find out more about attractions you'd like to visit, accommodation options and how to get there, restaurants and pubs, etc, and write a few lines about your findings."

Search session. After being assigned a task, the participant performs a search session to gather information and complete the assignment. Therefore, a search session is an iterative activity in search topic development to gain knowledge on the chosen subject. The participant creates a sequence of incrementally developing queries expressing their information needs is generated, each of which is presented to the retrieval system.

Knowledge is gained through this iterative process of query reformulation and/or development and subsequent browsing of retrieval results. During this process the participant formulates any number of text queries s/he wishes and visits any or all the retrieved documents s/he chooses.

- **Query formulation and retrieval:** The participant submits a query to the search engine, in response to which the system returns a ranked list of search results, computing using a ranking algorithm such as language modelling [3]. For each result title, URL, and a preview snippet are shown to the participant in a standard form SERP page.
- **Search result browsing:** The participant browses the search results by visiting any of the retrieved documents that they wish to and examining their content. Additionally, the participant can bookmark the documents s/he wants to refer to later.

During each search session all interactions with the system are recorded in a search log. The log contains events for:

- submission of queries,
- actions on documents and their rank:
  - opening a document
  - closing a document
  - bookmarking a document
  - unbookmarking a document
  - opening a new tab
  - scrolling.

For each event a timestamp is also logged. This enables us to compute the dwell time on a document, which can constitute important information when studying user behaviour. Moreover, bookmarks indicate documents that the user deems important with respect to their search and wants to be able to refer to later. This can suggest the user's interests or perhaps the extent of their knowledge on the topical area.

The search session ends when the user decides that s/he has gathered sufficient information to complete the task.

2.2.2 Phase 1.2: Final topic formulation. During the second phase of data gathering the participant creates a TREC-style topic description of the final topic, where a *final topic* is the user's information need behind the last query submitted during the search session. The report on the final topic includes "title", "description" and "narrative" fields, describing the information need by a phrase, a full sentence, and a description of the type of content that the user deems relevant and non-relevant to the topic respectively. Additionally, the participant is required to submit a summary of her/his findings with regards to the accomplished search task. This is done to give the participants a concrete goal and to ensure reliability of the collected material.

2.2.3 Phase 1.3: Relevance assessment. The final phase of the task session is relevance assessment, where the participant is asked to judge the relevance of a set of sampled results for each topic that s/he has developed during the search session.

During the relevance assessment phase the participant is shown in sequence each query s/he submitted and a set of search results sampled from a corresponding set of results for each one. The set of results for assessment is selected from the results produced by multiple retrieval algorithms using a stratified sampling method called *2strata strategy* [7]. This method ensures an exhaustive assessment of the small initial stratum for the ranked retrieval list for each retrieval method and a moderate assessment of the second stratum.

For each query a final set of results for assessment is made of:

- all the documents in ranks 1-10 (first stratum),
- 9 random documents in ranks 11-100 (10% sample of second stratum),
- all clicked documents for the query.

The inclusion of the visited documents in the assessment set allows us to ensure that we gather relevance judgement information for these documents.

The participant expresses the perceived usefulness of each sampled document to the information need specified in the query according to the following 4-point relevance scale:

- *off-topic*: the document subject has nothing to do with the current topic;
- *not relevant*: the document subject is related to the current topic but its content is not useful to the participant's information need;
- *somewhat relevant*: the document subject is related to the current topic and its content is slightly useful to the participant's information need;
- *relevant*: the document subject is related to the current topic and its content is useful to the participant's information need.

As in standard IR test collections, It is assumed that documents not included in the assessment set are not relevant to the topic.

Using a 4-point scale enables us to evaluate search in terms of graded or binary relevance, in the latter case by converting to a binary scale.

2.3 Simple User profile representation

A PIR test collection can be used to two purposes: i) to provide search teams working on user modelling with a rich set of information related to real user when performing a search task; ii) to allow search teams working on the definition of innovative search algorithms to have a collection of task related queries, and associated relevant material related to real users. In the latter case, if teams are not interested in defining a specific user model, they can use the simple bag of words user model that we provide with a PIR collection.

This profile is defined based the collected user-related information, without any user participation. The output is a set of basic formal representations of the user's topical interests for each completed task session.

3 COMPARATIVE EVALUATION OF PIR RUNS

Comparing effectiveness and differences in search behaviour between runs carried out using different PIR methods, whether significantly different or differing only in small variations in the settings of the personalisation method, requires careful consideration.

Since existing tools for the evaluation of IR runs do not support the detailed analysis of search behaviour across a search session, we are developing an evaluation tool we support this comparative evaluation of personalised information retrieval systems. The tool takes as input multiple standard TREC format results files from across multiple session runs for a topic set created by the user carrying out the session, with standard form qrel relevance judgments. Each session result set corresponds to the search output based on a single search methodolgy potentially incorporating a personalisation model. The tool then compares the ranked results lists calculated for the query entered at each stage of the session. Essentially the tools seeks to enable us to address the question: "which retrieval method is more effective at retrieving relevant documents at this point in the search session?".

A little consideration makes clear that this can actually be a rather complex question to address. For example, one method may retrieve a document deemed relevant for a query later in the session by the searcher providing the data, but not viewed by them earlier in the session during data collection. Would the searcher have marked it relevant had they seen it earlier, and if so would they be still find it relevant later in the session? We are seeking to develop our tool to not just complete standard metrics based on retrieved results, but also to explore alternative search session results behaviours arising from use of personalisation methods within the search process.

Out current prototype evaluation tool produces a report showing a set of standard IR evaluation measures, including precision, recall, precision@K, NDCG, etc... These metrocs are computed and compared in such a way as to highlight the different performance between the alternative run results files for the session. The results include a set of charts which graphically display the evolution of the performance of each evaluation measure through a retrieval session. This enables the display of variations between the runs in a way that is easy to see and interpret.

4 PIR-CLEF TASK

The evaluation methodology outlined in this paper is being trialled within the Personalised Information Retrieval (PIR-CLEF) task at CLEF 2018. This follows on from a pilot PIR-CLEF task carried out at CLEF 2017 [5]. The PIR-CLEF 2017 Pilot Task made available both user profile data and raw search data produced by guided search sessions undertaken by 10 volunteer users using the procedures described in this paper. The data provided included the submitted queries, the items clicked by the user in the result list, and the document relevance assessments provided by the user on a 4-grade scale. Each session was performed by the user on a topic of their choice selected from a provided list of broad topics. Search was carried out over a subset of the ClueWeb12 web collection indexed using Lucene with ranking using a language modeling retrieval model.

Further details of the PIR-CLEF 2018 task can be found at http://www.ir.disco.unimib.it/pir-clef2018/. The data from the CLEF 2017 pilot task is being used as development data for the CLEF 2018. A new test collection is currently being developed for use as the CLEF 2018 test collection.

5 CONCLUSIONS

This paper has outlined our work towards the development of a laboratory-based approach to the evaluation of Personalised Information Retrieval systems (PIRs). Our test collections developed using this method are intended to enable the adaptation of the standard laboratory-based approach used for classic IRSs to the evaluation of PIRSs. To support the analysis of the sessions based run results collected using this method, a prototype evaluation tool combining standard IR evaluation metrics and visualisation of run results has been developed.

To validate our data collection methodology, a pilot test collection was developed with the pilot PIR-CLEF task at CLEF 2017. A more extensive realisation of the PIR-CLEF task is currently being carried out within CLEF 2018.

Acknowledgements: This research work was partially supported by Science Foundation Ireland as part of ADAPT Centre (Grant No. 13/RC/2106).

REFERENCES

[1] [n. d.]. The ClueWeb12 Dataset. http://lemurproject.org/clueweb12. ([n. d.]). Last accessed: 2016-04-03.

[2] Andrei Z. Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (2002), 3–10. https://doi.org/10.1145/792550.792552

[3] Frederick Jelinek. 1980. Interpolated estimation of Markov source parameters from sparse data. *Pattern recognition in practice* (1980).

[4] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *Proceedings of the 37th International ACM SIGIR Conference on Research; Development in Information Retrieval (SIGIR 2014)*. Gold Coast, Queensland, Australia, 607–616.

[5] Gabriella Pasi, Gareth J. F. Jones, Stefania Marrara, Camilla Sanvitto, Debasis Ganguly, and Procheta Sen. 2017. Overview of the CLEF 2017 Personalised Information Retrieval Pilot Lab (PIR-CLEF 2017). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Dublin, Ireland, 338–345.

[6] Camilla Sanvitto, Debasis Ganguly, Gareth J. F.Jones, and Gabriella Pasi. 2016. A Laboratory-Based Method for the Evaluation of Personalised Search. In *In Proceedings of The Seventh International Workshop on Evaluating Information Access (EVIA 2016) - A Satellite Workshop of NTCIR-12*. Tokyo, Japan.

[7] Ellen M. Voorhees. 2014. The effect of sampling strategy on inferred measures. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR 2014)*. Gold Coast, Queensland, Australia, 1119–1122.