A Laboratory-Based Method for the Evaluation of Personalised Search

Camilla Sanvitto
University of Milano-Bicocca
Milan, Italy
c.sanvitto@campus.unimib.it

Debasis Ganguly
Dublin City University
Dublin, Ireland
dganguly@computing.dcu.ie

Gareth J. F. Jones
Dublin City University
Dublin, Ireland
gjones@computing.dcu.ie

Gabriella Pasi
University of Milano-Bicocca
Milan, Italy
pasi@disco.unimib.it

ABSTRACT

Comparative evaluation of Information Retrieval Systems (IRSs) using publically available test collections has become an established practice in Information Retrieval (IR). By means of the popular Cranfield evaluation paradigm IR test collections enable researchers to compare new methods to existing approaches. An important area of IR research where this strategy has not been applied to date is Personalised Information Retrieval (PIR), which has generally relied on user-based evaluations. This paper describes a method that enables the creation of publically available extended test collections to allow repeatable laboratory-based evaluation of personalised search.

Keywords

personalised search; laboratory-based evaluation; test collection development

1. INTRODUCTION

A fundamental and challenging activity related to Information Retrieval Systems (IRSs) is the evaluation of their effectiveness. The most common approach undertaken to assess the effectiveness of traditional IRSs is to adopt a laboratory-based method using the Cranfield paradigm. Following this methodology, researchers assess IRS effectiveness via experiments on static test collections in a controlled, laboratory-like setting. This setup ensures that evaluations obtained by different research teams are consistently repeatable and enables different IR methods to be compared directly.

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 concept of user context in the evaluation process. Moreover, the standard 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. Although 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.

Thus, the development of a publically available test collection that enables repeatable evaluation of personalised search would be beneficial to the IR research community. Our work described in this paper proposes a novel method for creating such a test collection suitable for extending the standard laboratory-based IR evaluation methods for the evaluation of personalised search.

The paper is structured as follows: Section 2 surveys current approaches to evaluating personalised search, Section 3 presents the rationale behind our proposal for the generation of a personalised test collection, Section 4 describes the design of the experimental process for gathering the necessary data, and Section 5 concludes the paper with a brief summary and outlook.

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 about a user preferences into the search process [9]. This user-centred approach to search has raised the related issue of how to properly evaluate search results in a scenario where relevance is strongly dependent on the interpretation of the individual user. To this purpose several user-based evaluation frameworks have been developed, as discussed in [10].

A first category of attempts to perform a user-centred evaluation has provided a kind of extension to the laboratory-based evaluation paradigm. The TREC Interactive track [6] and the TREC HARD track [2] 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. 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 [4] 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.

The method most widely undertaken to qualitatively assess the effectiveness of PIRSs is user studies. During these experiments participants are asked to report their subjective judgements about the system's performance by naturally interacting with it. Due to the fact that different users are involved in the studies, the experiments are not repeatable and their outputs not generally comparable.

The lack of both a suitable test collection and a standard approach to evaluation of personalised search is a limitation to researchers. Due to this fact, our proposal is the definition of test collections, in the Cranfield style, designed to evaluate approaches to personalised search; our proposal relies on a data gathering methodology for building extended test collections.

A first attempt to create a collection in support of PIR research was done in the FIRE Conference held in 2011. The Personalized and Collaborative Information Retrieval track [5] 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.

Our proposed solution is a more complete approach to enhancing the evaluation process of personalised search by making use of a laboratory-based approach and a personalised test collection.

3. PROPOSED APPROACH

Our test collection development method is designed to take into account the issues encountered to enable current test collections to be applied to PIR. The rationale behind the design of our test collection development method is to offer research teams a means to formally define a user profile for PIR by providing accurate and tested information sources about real user preferences.

Moreover, as researchers may be interested in the evaluation of innovative personalisation algorithms, we also provide a simple keyword-based representation of the information gathered about the individual users.

The test collection is designed to provide all the traditional components needed in a laboratory-based evaluation experiment such as topics and relevance judgements based on a reference document set. 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 study we have selected ClueWeb12, large crawl of over 730 million Web pages [1].

These standard elements are accompanied by a new set of user-related information for modelling and introducing personal context in the evaluation experiment. Specifically, this personal information consists of:

- **user personal information**: including gender, age range, native language, and occupation. This information can be very useful for personalising and adapting the search process to the current user.
- **search logs**: which contain the history of the user's interactions with a search engine. Search logs carry information about both the user's topical interests and their search behaviour.
- **the user's documents of interest**: provided as useful and raw sources to extract topical user preferences.

Basic user profile representations in the form of bag-of-words are also provided with the aim of offering a basic model of the user's topical interests.

All of the above contextual information is available for exploitation in creating one or more profiles for the user. Together with both the provided search topics and relevance judgements, this can be used to conduct laboratory-based evaluation experiments of PIRSs.

4. DATA GATHERING PROCEDURE

To set up the proposed test collection, we propose a design of the experimental process that aims at gathering and producing the data needed for building the collection.

The collection procedure is divided into two phases:

1. **data gathering**: this involves a group of real users, called participants, in a series of logged activities.
2. **user profile representation**: this consists of elaborating some of the information collected in the previous phase to create bag-of-words representations of the participant user's context.

4.1 Data gathering

The data gathering phase is performed by users in a controlled way to ensure the quality of the entire process. During this phase a group of participants carry out a series of task-based sessions and all their activities are recorded. A task session takes place over 3 main phases: topic development, final topic description, and relevance assessment. These are preceded by the gathering of the user's personal information, such as gender, age, native language, and job.

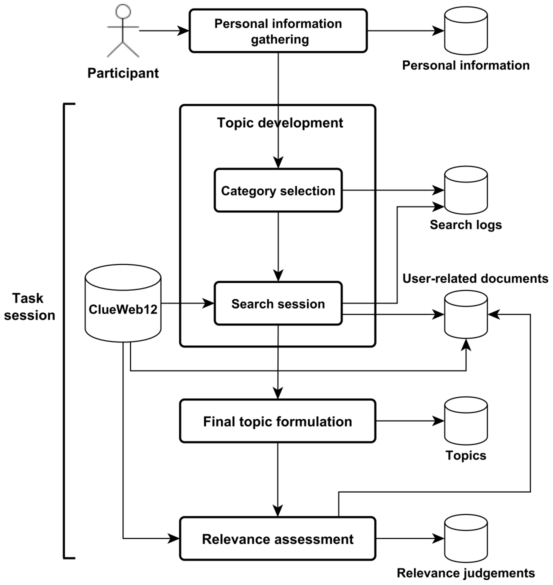

Figure 1: Data gathering process

The workflow of a single task session is represented in Figure 1 showing the items of the personalised data collection gathered during the complete process.

4.1.1 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 have been defined based on examples of informational and exploratory tasks [3, 8]. 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 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. A sequence of incrementally developing information needs is generated, each of which is presented to the retrieval system in the form of a query.

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 all the documents s/he wants.

- **Query formulation and retrieval:** the participant submits a keyword-based query to the search engine and the system returns a ranked list of search results using a ranking algorithm such as language modelling [7]. For each result title, URL, and a preview snippet are shown to the participant.
- **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.

4.1.2 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, it is required that the participant submits 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.

These topic descriptions are then used as test topics in laboratory-based evaluation experiments.

4.1.3 Phase 1.3: Relevance assessment

The final phase of the task session is the 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 each query s/he has 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* [11]. 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 gather additional important relevance judgement information of 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.

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.

4.2 User profile representation

Once the data gathering process is completed, the second phase of the experiment is performed without any user participation. The output is a set of basic formal representations of the users' topical interests for each completed task session. These representations can potentially be used as information for the creation of user profiles by researchers interested in evaluating personalised search algorithms.

Information to construct these basic representations of user's interests is the set of documents that the participant has assessed as relevant at the end of the task session. The use of these documents as information sources ensures that the model accurately represents the user's interests.

The approach chosen to model the user context is the bag-of-words representation defined as a set of words with associated weights representing the importance of the words as descriptors of the participant's interests. The computed weight for each word is the term frequency in a document defined as the ratio of the number of occurrences of the term in the document and the number of occurrences of the most frequent term in the document.

The choice of representing the information at the task session level has been made to provide flexibility in the data collection. In fact, this user representation contains all the information needed to generate the model over a different time period through aggregation or to use a topical domain filter.

5. CONCLUDING REMARKS

The idea proposed in this paper is a first step towards the adaptation of a laboratory-based approach to the evaluation of Personalised Information Retrieval Systems.

The newly defined reference data collection enables the adaptation of the standard laboratory-based approach used for classic IRSs to the evaluation of PIRSs. This approach has several advantages over previous attempts; in particular it eases the work of researchers that want to conduct evaluation tasks by providing both raw user contextual resources and predefined profile representations. Moreover, the use of a laboratory-based approach allows different research groups to reproduce and compare evaluation experiments.

Having designed the data collection strategy, the next stage of our work will be to run the data collection process. Once completed, the collected data will be processed and organised in the form of a personalised collection as outlined in this paper. To validate the utility of the collected dataset, an initial set of experiments will be carried out with a standard state-of-the-art tool for Information Retrieval.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

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

[2] J. Allan. HARD track overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of The Twelfth Text REtrieval Conference (TREC 2003)*, pages 24–37, Gaithersburg, Maryland, USA, 2003.

[3] A. Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[4] 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.

[5] D. Ganguly, J. Leveling, and G. J. F. Jones. Overview of the personalized and collaborative information retrieval (PIR) track at FIRE-2011. In *FIRE'11 Workshop*, 2011.

[6] 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.

[7] F. Jelinek. Interpolated estimation of markov source parameters from sparse data. *Pattern recognition in practice*, 1980.

[8] J. Jiang, D. He, and J. Allan. 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)*, pages 607–616, Gold Coast, Queensland, Australia, 2014.

[9] G. Pasi. Issues in personalizing information retrieval. *IEEE Intelligent Informatics Bulletin*, 11(1):3–7, 2010.

[10] 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):1–34, 2009.

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