Overview of the Personalized and Collaborative Information Retrieval (PIR) Track at FIRE-2011

Debasis Ganguly, Johannes Leveling, Keith Curtis, Wei Li, and Gareth J.F. Jones

CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
{dganguly, jleveling, kcurtis, wli, gjones}@computing.dcu.ie

Abstract. The Personalized and collaborative Information Retrieval (PIR) track at FIRE 2011 was organized with an aim to extend standard information retrieval (IR) ad-hoc test collection design to facilitate research on personalized and collaborative IR by collecting additional meta-information during the topic (query) development process. A controlled query generation process through task-based activities with activity logging was used for each topic developer to construct the final list of topics. The standard ad-hoc collection is thus accompanied by a new set of thematically related topics and the associated log information. We believe this can better simulate a real-world search scenario and encourage mining user information from the logs to improve IR effectiveness. A set of 25 TREC formatted topics and the associated metadata of activity logs were released for the participants to use. In this paper we illustrate the data construction phase in detail and also outline two simple ways of using the additional information from the logs to improve retrieval effectiveness.

1 Introduction

One major challenge in Information Retrieval (IR) is the potential to adapt retrieval results for personalized IR. Different users may enter the same query string into a search system, but their information needs can be vastly different. The notion of relevance depends upon factors such as the domain knowledge of the searcher, information gained from reading previous documents in the past, and general search behavior of a searcher, e.g. how many documents he normally reads before reformulating his search [2].

In a typical laboratory evaluation scenario of ad-hoc IR, participants are given a document collection and a set of queries (topics). The task of the participating systems is then to retrieve documents which satisfy the information need expressed in each query. Such a traditional evaluation framework does not provide enough information to facilitate personalized IR. This information includes: a) closely related topics formulated by different people with different assessments reflecting a differing notion of relevance, and b) meta-information such as the documents viewed by the users.

The process of TREC-style topic development is artificial and does not resemble iterative query reformulation in real search activities where typically a user of the search system enters an initial query, reads a few top ranked retrieved documents and reformulates the initial query until his information need is satisfied. The final query, based on

the content read thus far, retrieves one or more relevant items which satisfy his information need up to this point. Our main hypothesis is that this iterative process of topic development is more similar to the real-world search than a search based on a single topic.

To our knowledge, little or no research has addressed gathering and providing metadata for the query development process under the framework of an ad-hoc retrieval dataset.[1] The existing work on user studies for personalized IR differ in the tasks given to the users (e.g. typically search with different unrelated queries) and the document collections used for the exploration (e.g. searching the web). Our work attempts to provide a common evaluation framework to test various personalized IR systems.

The rest of the paper is organized as follows: Section 2 surveys work on user modelling and personalized IR, Section 3 presents our approach to generating user logs in a controlled environment, Section 6 outlines the planned task to be undertaken within the FIRE 2011, and Section 7 concludes the paper with a brief summary and outlook.

2 Related Work

The research question we want to explore is whether IR systems can present more relevant documents to individual searchers (hence addressing personalization) by exploiting his browsing information and that of users with similar search interests.

Recent works on the study of user search patterns include that of Kellar et. al. [6]. They report that users spend most of their time, view most pages, and extensively use the browser functions for information gathering tasks, thus establishing the need for extensive user studies of information gathering tasks. Kelly and Belkin [7] report that there is no direct relationship between the display time of a document and its usefulness, and that display times differ significantly according to a specific task and according to individual users. White and Kelly [11] show that tailoring document display time thresholds for implicit relevance feedback based on task grouping is beneficial for personalization. Liu and Belkin [8] design a method for decomposing tasks into sets of (in)dependant subtasks and show that the task stage for an independent subtask is helpful in predicting document usefulness. This is attributed to the fact that users gain knowledge across stages regarding the usefulness of documents.

The related work on user studies motivated us to generate a log of the entire topic creation process to make information about the search process available to the retrieval systems, which can help tuning IR systems to user-specific needs. The lack of availability of user logs greatly limits the research that can be undertaken in this area outside of industrial research laboratories. Our proposed methodology is designed to make a set of query logs freely available and distributable to promote personalized IR research.

The LogCLEF[2] log analysis initiative provides log data from different providers [3], but these datasets lack relevance assessments.

TREC 2010 introduced a new track called the Sessions Track[3], where the motivation is to form and evaluate a session of related queries [5]. This track involves modifying an

[1] Metadata includes all information from the search history for all query formulations.

[2] http://www.promise-noe.eu/mining-user-preference/logclef-2011/

[3] http://ir.cis.udel.edu/sessions/

initial query into a more general query, a more specific query, or one addressing another facet of the information need. Our proposed track is different because firstly, we do not manually form query variants, but expect the participants to contribute in generating search data and provide them with search logs from other participating and volunteering topic developers. Secondly, our track is not primarily concerned with query sessions, but with categorizing users based on their interests and with exploring whether individual searchers can profit from information about similar searches or users.

NTCIR-9[4] organized the Intent task, where topics are formed automatically by random sampling from Chinese web search query logs. A difference between our proposed task and the NTCIR Intent task is that the latter deals with web search and uses a bottom-up approach (starting from existing query logs), whereas we try to address elements of personalization with a top-down approach, aiming to create interaction logs.

In summary, there are two important differences compared to previous research:

1. The topic development and relevance assessment will be performed by the same person.
2. The same (static) corpus is utilized for search and logging the topic development process, because experiments which are based on web search logs are typically not reproducible due to the dynamic nature of web documents.

3 Data Construction Methodology

To promote our approach to automatic "closed-box" personalized and collaborative IR experiments and encourage researchers to use and contribute to this method, we organized a pilot track named Personalized IR (PIR)[5] in the Forum of Information Retrieval and Evaluation[6] (FIRE) 2011.

The closed set of documents, on which browsing activities were logged, is the FIRE 2011 English ad-hoc document collection comprising of news from the Indian newspaper *The Telegraph* from 2001 to 2010 and news from Bangladesh, comprising of almost 400K documents in total.

A web service[7] was developed and hosted, which was used during the topic development phase to browse through the collection and construct topics. The sequential steps towards creation of a topic are as follows. A topic developer logs into the system with a registered user ID and is henceforth refered to as a *user* of the web interface. The user then goes through a search phase (selecting the search category, submitting queries and viewing result documents) and a topic formulation and evaluation phase (summarizing the found information, formulating the final topic, and assessing relevance for documents). The system logs all these user actions. The topic development procedure is illustrated in Fig. 1. We explain each step as follows.

[4] http://www.thuir.org/intent/ntcir9/
[5] http://www.cngl.ie/Fire-PIR/
[6] http://www.isical.ac.in/~clia/
[7] http://www.cngl.ie/Fire-logs/

Category selection. The system presents a list of broad search *categories* from which the topic developer has to select one. The categories are listed below.

1. Social impact on land acquisition
2. Honour killing incidents
3. Indian cricketing events
4. Indian tourism
5. Relation of India with its neighboring countries
6. Indian political scams
7. Healthcare facilities
8. Indian paintings and painters
9. Indian traditions and customs
10. Indian armed forces
11. Indian education policy
12. Bollywood movies
13. Adventure sports
14. History of Indian vernaculars
15. Terrorist attacks

The categories were intended to represent broad search domains of news articles which a user can freely browse, gain knowledge during the search phase and finally enter his own specific query. Also deriving the topics from a pre-defined list of categories is intended to ensure development of related topics with overlapping information needs for different users.

The search categories have been selected in accordance to the TREC guidelines of topic development, which involves performing trial retrievals against the document set and choosing topics for which the result set is not too small or too large [4]. Also to ensure that we have roughly a uniform distribution of queries across these broad categories, our system dynamically adapts the list for different users e.g. if enough queries have been formed from category 1 and none from category 2 then the system removes category 1 from the list presented to a new user.

Query formulation and retrieval. After selecting a category, the user iterates through query formulations, retrieving different documents at each iteration. The retrieval engine which the system uses at the back end is Lucene.

View/browse result documents. The user can read documents retrieved in the previous step by clicking on the result URLs and can also bookmark a document to refer to it later. The user is expected to go through a series of query reformulations before he feels that he has gained sufficient knowledge to enter a topic to the system.

Topic summarization. The system presents a form where the topic developer has to enter a report/summary on the subject matter of the chosen category. The content of the summary acts as a means to ensure that the topic developer has indeed gained knowledge about the search category and that the final topic indeed is based on information

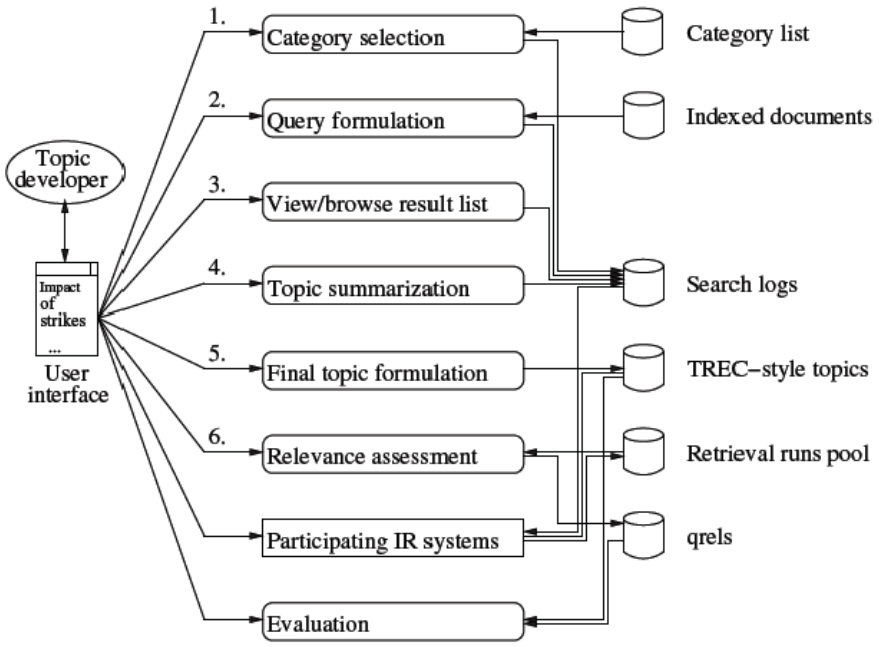

Fig. 1: Data flow diagram of the topic development phase.

from documents viewed by him. A randomly entered summary on the other hand can indicate an illformed test topic unsuitable for the evaluation experimentation.

Final topic formulation. As a next step the system asks them to form a TREC-formatted topic based on the knowledge gained thus far. This query aims at one user-specific, "personalizable" aspect of the initial search category, i.e. one particular aspect of the category that the topic developer is especially interested in. We refer to this topic as the *final topic* (because it is the final topic entered by a user for a particular search category) or a *test topic* (because this topic is released in the test topic file to be used for the final evaluation). The topic developers have to fill in the *title*, *description* and *narrative* fields for the query, describing the information need by a phrase, a full sentence, and a description of which documents are relevant and which are not. These TREC-style topics serve as input for the IR runs.

Relevance assessment. Relevance assessments are based on the pool of submissions. The developer of a topic was assigned the responsibility to mark the relevant documents according to the relevance criteria expressed in the narrative field of the topic provided by him. We aimed to investigate if there exists a *personal* notion of relevance, i.e. how often a document is relevant for two different topics belonging to the same category. Another aspect of research was to see how many of the documents bookmarked or

viewed for a long time by topic developers for a category are actually relevant for the test topic entered in that category.

4 An Example Scenario

Let us assume a topic developer selects the example topic "Terrorist attacks" from the pre-defined list of search categories. He then enters a series of queries in the system (e.g. "Terrorist attacks Kashmir", "Terrorist organizations", "Terrorism in Mumbai"), views the documents, bookmarks some of them and starts gaining knowledge about the category given to him. Figure 2 illustrates this topic development process. For the chosen initial search category, the user issues a query Q_1 and gets a ranked list of returned documents $\{D_1^1, \ldots D_m^1\}$. A subset of relevant documents (viewing or bookmarking a document might be an implicit indicator of relevance) is used to reformulate Q_1 into Q_2. The released meta-information corpus would thus contain each intermediate query Q_i, the set of top documents returned for Q_i namely $\{D_1^i, \ldots D_m^i\}$ and the actions of the user.

After going through a few iterations, the user then fills up the topic summary and submits his topic titled "26/11 Mumbai attack" with an appropriate description and narrative. Later on he also has to assess documents for relevance for this topic.

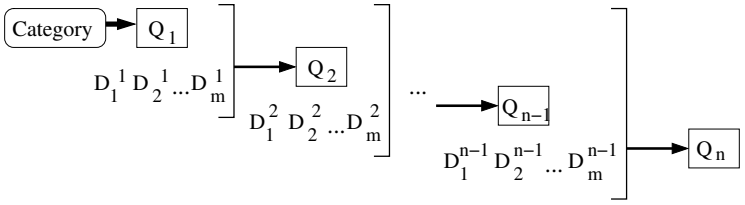

Fig. 2: Structure of the meta-information.

Consider another topic developer who selects the same topic. He also browses documents through the system and eventually ends up with a topic titled "Ajmal Kasav trial". Now we see that the topic "Ajmal Kasav trial" is related to the query "26/11 Mumbai attack" since Ajmal Kasav was the person convicted of the murders on the 26th Nov night in Mumbai.

The assumption is that the set of documents that the first user views or bookmarks can also be the potential relevant candidates for the second query. Also the principle of recommender systems can be applied here where we predict the relevance of a document by its popularity (of viewing or bookmarking) among different users executing similar queries. Building up on the hypothesis that information obtained from one user might benefit satisfying the information need(s) of the other, the intended challenge for the participants was to develop ideas of how to increase retrieval effectiveness for all searchers with similar search intentions by exploiting the browsing history of all such searchers.

<top>
<num>1</num>
<username>*******</username>
<categoryname>Adventure sports</categoryname>
<title>rock climbing india</title>
<desc>The intent is to find general information about the rock climbing sport in India.</desc>
<narr> Relevant documents are documents that give information or news about the rock climbing sport in India. Relevant documents will also be ones that discuss achievements of rock climbers who climbed mountains in India.
</narr>
<top>

5 Data Details

5.1 Test Topics

Twenty-five TREC formatted test topics having two additional tags of *username* and *categoryname* were released. For each topic, the string enclosed within the *username* tag denotes the registered user name of the developer for this topic and the *categoryname* tag contains the name of the category selected while developing it. An example topic is shown in below.

5.2 Search Logs

A single line in the log file represents a search event by a user, where a search event can be either a click on the URL or on the bookmark button corresponding to a document, returned as part of a ranked list in response to a query execution. The logs are formatted as comma separated values and have the following structure: *user name, category, query name, document name, rank of this document, action performed on this document*, and *time stamp*. The first field i.e. the name of the user serves as an identifier to trace the originator of the event. The second field i.e. the name of the category is used to identify the top level search category from which the event was generated. This is particularly useful in restricting investigation of browsing history within a single search session for a single user or for a group of users who chose the same search category. The next field i.e. the *query name* is the seach string for which this particular event (click or bookmark) occurred. The fourth and fifth fields are the name and the retrieval rank of the document which was clicked or bookmarked. The sixth field distinguishes between the two types of action possible which is one of *resultclick*, denoting that the user clicked on the URL of this document, and *bookmark*, which indicates that the user bookmarked this document for refering to it later. The last field records the time stamp of the event.

6 Retrieval Methodology

Twenty-six participants registered for the PIR track but unfortunately not a single one of them submitted any run for the track. We therefore decided to generate three baseline runs demonstrating the potential usefulness of the collected log data.

The retrieval model used for all the baseline runs is Language Modeling (LM) implemented within SMART[8]. We used the standard SMART stopword list and the default stemmer of SMART, which is a variant of the Lovin's stemmer. The LM implementation of SMART employs a Jelinek Mercer smoothing [10]. The smoothing parameter λ was set to 0.3, which is the optimal value for the ad-hoc task on FIRE 2010 English collection.

The first baseline named BL_1 is a simple ad-hoc IR run using the titles of queries. The run involves a pseudo-relevance feedback (PRF) step of query expansion using $R = 10$ i.e. assuming that top 10 documents are pseudo-relevant, and using $T = 10$ i.e. adding 10 selected terms from these documents to the query. The term selection is based on LM term scores as defined by Ponte in [9]. The second baseline BL_2 was generated using the additional intermediate queries that a user entered in the search system before formulating the final test query, whereas the third baseline BL_3 uses the clicked documents by a user.

In order to extract out the information about the intermediate query titles and the viewed documents efficiently at the test query execution phase, we preprocessed and organized the log data into a two level hash indexed data structure. This is because given a test query string and the associated identity of the user and category of this query, respectively refered to as *current user* and *current category*, we would quickly want to narrow down on the subset of logging events which is useful for this test query. The top level of the log data structure is thus indexed by the category name, which quickly narrows down to the search to the current category. The next level of indexing is on the user name which extracts out logs only for the current user. The log records at the leaf level of this bi-level index is organized as a list, where each list element stores an intermediate query name entered into the search system and a pointer to the list of retrieved document names in response to that query. Figure 3 shows a schematic organization of the data structure. We build up this in-memory data structure only once in the preprocessing stage. We iterate through each log record from the CSV and insert it in its proper place in the bi-level hash table. Refering to the Figure 3, if we want to get log information for a test query whose category name is $\mathtt{C}_5$ and user name is $\mathtt{U}_3$ we first query the left-most hash table with $\mathtt{C}_5$, follow the pointer and reach another hash table where we query with $\mathtt{U}_3$. Following the pointer we reach the activity records for that user who in this particular example has entered only one query $\mathtt{Q}_3$ and clicked on documents $\mathtt{D}_7$ and $\mathtt{D}_8$.

With this description of the organization of the log data in memory we are now ready to outline the methodology for generating the baseline runs BL_2 and BL_3. BL_2 only uses the intermediate query strings i.e. the query strings marked as $\mathtt{Q_i}$s in Figure 3. For a test query we retrieve the intermediate query strings for the current user and the current category. This precisely constitutes the list of queries entered into the search system by the current user during the search session which led to the development of the final test query. We add these query strings into the test query title and report the retrieval run as BL_2. The rationale behind this approach is that the intermediate query strings serve to act as the actual intent of the user during the search session.

[8] `ftp://ftp.cs.cornell.edu/pub/smart/`

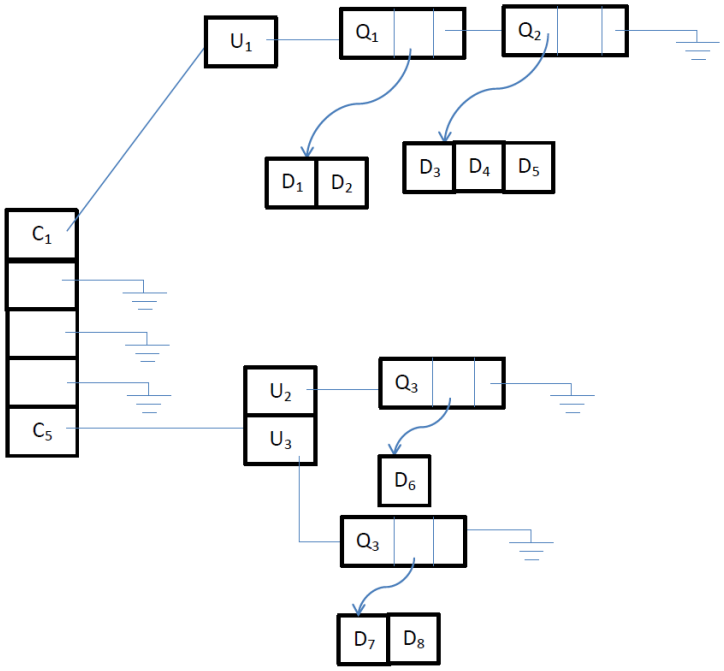

Fig. 3: Data structure for efficient log processing.

For BL_3 we use the clicked documents in addition to the top ranked documents of an initial retrieval run as the potential set of pseudo relevant documents. The rationale is that viewed documents are likely to be relevant for the intermediate query and in turn for the final test query assuming that the final test query reformulation has been affected by the contents of the viewed documents.

6.1 Results

In the absence of any submitted runs, at the time of writing this paper we have been able to complete only a rapid evaluation by restricting manual assessments to the pool of top 30 documents obtained from the three retrieval runs BL_1, BL_2 and BL_3. We thus report average precisions at fixed cut off values of 5 and 10. Table 1 shows the fixed point cut off precision values averaged over 25 test topics. We see that using the log information can increase the precision within top 10 documents. Expanding the current query by words from previous queries has shown a positive effect and using clicked documents for PRF has been able to demonstrate an increase in P@5.

Due to the lack of participation in the PIR task, we do not have sufficient log data to obtain conclusive results. Our preliminary evaluation effort nonetheless is indicative of

Table 1: Retrieval Results.

Run Name	P@5	P@10
BL_1	0.60	0.59
BL_2	**0.68**	0.60
BL_3	0.64	0.52

a possible trend in retrieval effectiveness. We need more time to generate the complete set of manual relevance assessments and compute the standard evaluation metric MAP.

We will investigate about the reasons for the lack of participation from the participants. One plausible reason can be that the domain of news articles is not very suitable for navigational searches. A collection of informative articles such as the Wikipedia might fit into the task based exploratory search. As a future work we intend to use the INEX ad-hoc collection, which comprises of the full Wikipedia collection [1] for building up log data to make the search task more interesting for the topic developers.

7 Conclusions and Outlook

The proposed methodology can act as the first stepping stone towards evaluation of different retrieval systems under the same test bed of user generated logs. The log generation process has been designed to address aspects of personalization by capturing individual information needs for a broad search category. The history of the documents viewed prior to developing the final topic makes the topic development process transparent to a retrieval system.

We have shown that navigational search sessions can be generated through task based browsing activities in a constrained domain or category and have also demonstrated that such activity logs can potentially be leveraged upon to improve retrieval precision at top ranks.

In the absence of any submitted retrieval run, we generated two retrieval runs one involving intermediate queries for expansion of the final test query, and the other using viewed documents for PRF. The results should motivate participants to submit their own runs at next year and thus contribute towards the development of the track. Participants from related tracks such as the Sessions Track at TREC, LogCLEF at CLEF, and the intent finding track at NTCIR may also be interested to participate in this track. The task this year focused on English, but as the methodology is language-independent, it can be applied for Indian languages also used at FIRE (e.g. Bengali or Hindi) if enough interest can be raised from FIRE participants.

Acknowledgments

This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie`). We would like to express sincere gratitude to Rami Ghorab, Trinity College Dublin, for

making the IR framework available to us. The authors are also grateful to Maria Eskevich, Mohd. Javed, Yalemisew Abgaz and Sudip Naskar from Dublin City University and to Gaurav Arora from DAIIT, India for contributing to test topic development.

References

1. P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the inex 2010 ad hoc track. In *INEX*, pages 1–32, 2010.
2. M. J. Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989.
3. G. M. Di Nunzio, J. Leveling, and T. Mandl. Multilingual log analysis: LogCLEF. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Murdoch, editors, *ECIR 2011*, volume 6611 of *LNCS*, pages 675–678. Springer, 2011.
4. D. Harman. Overview of the third text retrieval conference (TREC-3). In *TREC*, 1994.
5. E. Kanoulas, P. Clough, B. Carterette, and M. Sanderson. Session track at TREC 2010. In *SIMINT workshop SIGIR '10*. ACM, 2010.
6. M. Kellar, C. R. Watters, and M. A. Shepherd. A field study characterizing web-based information-seeking tasks. *JASIST*, 58(7):999–1018, 2007.
7. D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR '04*, pages 377–384. ACM, 2004.
8. J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks: the roles of task stage and task type. In *SIGIR'10*, pages 26–33. ACM, 2010.
9. J. M. Ponte. *A language modeling approach to information retrieval*. PhD thesis, University of Massachusetts, 1998.
10. J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR-98*, pages 275–281. ACM, 1998.
11. R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *CIKM 2006*, pages 297–306. ACM, 2006.