View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DCU Online Research Access Service

Overview of the CLEF 2018 Personalised Information Retrieval Lab (PIR-CLEF 2018)

Gabriella Pasi[1], Gareth J. F. Jones[2], Keith Curtis[2], Stefania Marrara[3],
Camilla Sanvitto[1], Debasis Ganguly[4], Procheta Sen,[3]
[1] University of Milano Bicocca, Italy
[2] Dublin City University, Dublin, Ireland
[3] Consorzio C2T, Milan, Italy
[4] IBM Research Labs, Dublin, Ireland

Abstract. At CLEF 2018, the Personalised Information Retrieval Lab (PIR-CLEF 2018) has been conceived to provide an initiative aimed at both providing and critically analysing a new approach to the evaluation of personalization in Information Retrieval (PIR). PIR-CLEF 2018 is the first edition of this Lab after the successful Pilot lab organised at CLEF 2017. PIR CLEF 2018 has provided registered participants with the data sets originally developed for the PIR-CLEF 2017 Pilot task; the data collected are related to real search sessions over a subset of the ClueWeb12 collection, undertaken by 10 users by using a novel methodology. The data were gathered during the search sessions undertaken by 10 volunteer searchers. Activities during these search sessions included relevance assessment of a retrieved documents by the searchers. 16 groups registered to participate at PIR-CLEF 2018 and were provided with the data set to allow them to work on PIR related tasks and to provide feedback about our proposed PIR evaluation methodology with the aim to create an effective evaluation task.

CEUR-WS.org/Vol-2125/invited_paper_12.pdf

1 Introduction

The PIR CLEF Lab organized within CLEF 2018 aims to provide a framework for the evaluation of Personalised Information Retrieval (PIR). PIR systems seek to enhance traditional IR systems to better satisfy the users information needs by providing search results that are not only relevant to the query but more specifically to the interests of the user who submitted the query. In order to provide a personalised service, a PIR system can leverage various kinds of information about the current user and their preferences and interests. These can be stated directly or be inferred through a variety of interactions of the user with the system. This information is then represented in a user model, which can be employed to either improve the user's query or to re-rank a set of retrieved results list so that documents more relevant to the user are presented in the top positions of the list.

Evaluating the effectiveness of personalised approaches to search has been investigated for many years within studies of interactive information retrieval.

In this work, the notion of relevance has been user centered with potential variation during a search session, depending both on the task at hand and on the user's interactions with the search system. This work has mostly based on user studies; this approach involves real users undertaking search tasks in a supervised environment. By placing the user at the centre of the evaluation activity these studies have produced valuable insights and feedback. However, while this methodology has the advantage of enabling the detailed study of the activities of real users, it has the significant drawback of not being easily reproducible, thus greatly limiting the scope for algorithmic exploration of technologies for search personalisation. Among some previous attempts to define PIR benchmark tasks based on the Cranfield paradigm, the closest experiment to the PIR Lab is the TREC Session track[1] conducted annually between 2010 and 2014. This track focused on stand-alone search sessions, where a "session" is a continuous sequence of query reformulations on the same topic, along with any user interaction with the retrieved results in service of satisfying a specific information need; however no details of the searcher undertaking the task have been made available. Thus, the TREC Session track did not exploit any user model to personalise the search experience, nor did it allow user actions over multiple search session to be taken into consideration in the ranking of the search output.

The PIR-CLEF 2018 Lab has provided search data gathered in search sessions carried out ten volunteer users: the provided data were originally collected for the Pilot Lab run in 2017. We plan in the future to gather data across multiple sessions to enable the construction and exploitation of persistent user behaviour data collected from the user across the multiple search sessions. This year the data were provided to the 16 groups registered to task. with the objective of allowing them to attempt the proposed tasks. An evaluation using this collection was run to allow research groups working on PIR to both experience with and provide feedback about our proposed PIR evaluation methodology. Two papers were submitted and accepted for presentation at the workshop related to the Lab; both papers report on the usage of the collected data to perform different tasks; the work reported in these papers is summarized later in this overview. The papers give some useful suggestions to improve the data gathering process, which will give rise to interesting discussions during the Lab.

The remainder of this paper is organised as follows: Section 2 outlines existing related work, Section 3 provides an overview of the PIR-CLEF 2018 task, Section 4 discusses the metrics available for the evaluation of the task, Section 5 overviews papers submitted by task participants, and Section 6 concludes the paper.

2 Related Work

Recent years have seen increasing interest in the study of contextual search: in particular, several research contributions have addressed the task of personalizing search by incorporating knowledge of user preferences into the search process

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

[2]. This user-centered approach to search has raised the related issue of how to properly evaluate the effectiveness of personalized search 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 [3].

A first category of approaches aimed at evaluating personalized search systems (PIRS, Personalized Information Retrieval Systems) are focused on performing a user-centered evaluation by providing 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. 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 both context and users interests. Participants in the track were given, as input, 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] the focus of which was to exploit user interactions during a query session to incrementally improve the results within that session. The novelty of this task was the evaluation of system performance over entire sessions instead of a single query.

However, the above tasks have various limitations to the satisfactory injection of user behaviour into the evaluation proces; for this reason the problem of defining a standard approach to the evaluation of personalized search is a hot research topic, which needs effective solutions.

A first attempt to create a collection satisfactorily accounting for the user behaviour in search was done in the FIRE Conference held in 2011. The Personalised and Collaborative Information Retrieval track [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.

Within CLEF 2017, we launched the PIR-CLEF benchmark with a pilot study and workshop (PIR CLEF 2017). for the purpose of providing a forum for the exploration of the evaluation of PIR. The Pilot Lab provided a preliminary edition of the 2018 PIR-CLEF Lab. One of the achievements of the PIR-CLEF 2017 Pilot Task was the setting up of an evaluation benchmark which seeks to combine user-centered methods with the Cranfield evaluation paradigm, with the key potential benefit of producing evaluation results that are easily reproducible.

The Pilot task was based on search sessions over a subset of the ClueWeb12 document collection, undertaken by 10 users by using a clearly defined and novel methodology. The collection was defined by relying on data gathered from activities undertaken during the search sessions by each participant, including details of relevant documents as marked by the searchers. An important point to be outlined is that the collection was prepared but not used by any group participating at the pilot task. For this reason at PIR-CLEF 2018 we relied on this data collection. We distributed it to the 16 groups registered to the Lab. We have also prepared a second collection, as well as a prototype system for the comparative evaluation of systems developed by participating groups. The data collection is described in more detail in Section 3.1.

3 Overview of the task

The goal of the PIR-CLEF 2018 Task is to investigate the use of a laboratory-based method to enable comparative evaluation of PIR methods. The collection defined within the PIR-CLEF 2017 Pilot Study and the PIR-CLEF 2018 Lab 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 Task we made available the user profile data and raw search data produced by guided search sessions undertaken by the 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 Task data was 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 2018 Task we made available both user profile data and raw search data produced by guided search sessions undertaken by the 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 carried out over a subset of the ClueWeb12 web collection.

The data was extracted and stored in csv format in 7 csv files in a zip folder which was provided to participants. The details of the contents of the csv files are as follows:

csv 1: The file *user's session* 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.

csv 2: The file *user's log* contains the search logs of each user, i.e. every search event that was 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.

csv3: The file *user's assessment* 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).

csv4: The file *user's info* 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

csv6: The file *simple user profile* 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.

csv6b: The file *complex user profile* 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,

Task participants had the possibility of contribute in two different ways:

- the two user profile files (csv6a and csv6b) provide bag-of words profiles for each of the 10 volunteer searchers in the data collection. The profiles were created by applying different indexing procedures to the documents that the searcher assessed as relevant during the search session. The searcher's log file (cvs2) contains all the queries she formulated during the query session. Task participants could compare the results obtained by applying their personalisation algorithms on these queries with the results obtained and evaluated by the searchers 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) using the standard NIST trec_eval tool. Note that documents that do not appear in csv3 were considered non-relevant.
- 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 by the participant 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.

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.

4 Evaluation and Analysis

The metrics and methodology used to evaluate and analyze the PIR-CLEF task pose significant challenges. It is not at all obvious how we might properly compare and contrast the behaviour of alternative methods of integrating personalization into search sessions. While we can start off using stand metrics, such as Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG) for individual these will not be sufficient to enable a detailed session based analysis.

As a starting for point for the development of formal methodology for analysis and evaluation of our framework for laboratory-based evaluation of PIR, we have developed a prototype evaluation tool which we describe in the remainder of this section.

4.1 PIR Evaluation Tool

Our proposed evaluation tool is designed to provide a repeatable approach for the evaluation of PIR. A few description of the analysis, design and implementation of the current version of this tool is provided in [11].

This evaluation tool consists of three sequential phases:

1. **File Extraction**: During this phase the data is extracted from standard TREC format results files created for each retrieval query operation. The extracted data is stored in efficient data structures to accelerate the next steps.
2. **Metric Calculation**: During this phase a set of standard IR evaluation measures are calculated, including Precision, Recall, Precision@K, NDCG, etcṄovel approaches for defining a retrieval session are also defined.
3. **Output Generation**: During this phase the evaluation tool produces a report which shows a set of standard IR evaluation measures. These evaluation measures are computed and compared in such a way as to highlight different performance measures between alternative results files. Additionally, a set of charts are included in results which graphically display the evolution of the performance of each evaluation measure through a retrieval session.

We now describe each phase in more detail.

File Extraction In this phase all information contained in the results files is extracted and efficient data structures are created to accelerate the evaluation process. These files contain the required information to estimate the performance of a system, therefore the files need to be created in a specified format:

- TREC format results file: containing the runs performed by a PIR system using the personalised data collection. This tool supports concurrent execution of multiple results files.
- Relevance judgments: containing the relevant documents mark by the volunteer searchers.
- Search logs: containing all searcher activities recorded during the search sessions.
- Commands file: allowing for the results to be tailored.

TREC Format Results File: The evaluation tool requires TREC format results files containing the ranked lists computed by the PIR system. The TREC format has been chosen because it a well known in the IR community. Results files must have the following fields: user id, topic id, query id, document id, rank, name.

Relevance Judgments File: The relevance judgments file contains the relevant documents marked by participants during the third phase of the PIR-CLEF experiment. This file has the following fields: user id, topic id, query id, document id, relevance score.

Search Logs: The search logs contain the user information gathered during the search sessions. This information is used by the PIR system to create the user representation that will be used in the retrieval process and by the evaluation tool to simulate the user behaviour.

Commands File: This allows for the tailoring of the evaluation process by specifying which measures and charts are to be computed by the tool.

Metric Calculation During this phase metrics are computed to enable the performance of the PIR systems to be evaluated and compared. The evaluation tool first computes per-query measures which allow the evaluation of the effectiveness considering a single ranked list. It then computes novel approaches for evaluation of retrieval sessions considering multiple ranked lists.

We next describe the process undertaken by the evaluation tool computing per query measures. We then detail the approaches defined for the evaluation of retrieval sessions.

Per-Query Measures The evaluation tool computes the following measures for each query in the personalised data collection: Precision, Recall, Precision@K, Recall@K, F-measure, R-precision, Average Precision, and NDCG.

The computation process of each metric follows these steps:

1. A ranked list set containing the ranked lists related to the queries are extracted from the result file provided by the PIR system, so as to have a ranked list for each query.
2. The relevance judgments set containing the relevance judgment documents for each query is extracted from the relevance judgments file.
3. The evaluation measure is calculated by comparing the retrieved documents list and the relevance judgments set for each query.

For each triple measure-user-topic, the evaluation tool generates both a line chart and a bar chart to show the evolution of the measure through the retrieval session as well as to compare the performance of the different input algorithms.

The created charts have a y-axis, which represents the measure value, and an x-axis, which represents the queries belonging to the same topic sorted by time using the timestamp contained in the search logs.

Session Measures Real users often begin an interaction with a search engine with a query which they need to reformulate one or more times. The ability of the PIR system to improve results after query reformulation cannot be easily assessed by the measures normally used for measuring system effectiveness, but requires new approaches taking into consideration the user behaviour triggered by the system. Three alternative approaches are proposed to simulate user behaviour through a retrieval session.

1. **Using logs file**: User behaviour is simulated using information contained in the search logs.

2. **Looking for non-relevant document**: Ranked lists extracted from the result file are considered to simulate the user path through the session.
3. **Using the probabilistic distribution**: User behaviour is defined using the probability distribution defined in [12].

Using only the logs file: If you have a set which contains the queries performed by a user for a topic, and a ranked list related to the queries extracted form the result file, this is a decision point provided by the search logs. The information contained in the ranked list represents the behaviour of the participant who performed the search session in the experiment. The evaluation of a session is carried out by applying the following steps:

1. Each ranked list is cut at position j, where j indicates the position of the last document opened by the user.
2. The new ranked lists are merged to build the session ranked list.
3. Precision and Recall are computed using the session ranked list derived in the previous steps.
   This approach has the drawback that is assumes the user behaviour triggered by the PIR system which retrieved the ranked list, is the same as the one that caused by the IR system used to gather the data in the PIR-CLEF experiment. But user behaviour is strongly related to the retrieval system due to the position of the relevant document in the ranked list.

Looking for non relevant document: This approach cuts the ranked lists at the position of the first non relevant document found after the decision point. This way the documents in the ranked list are thought to simulate user behaviour, making it dependent on the PIR system which retrieved the ranked lists. This is based on the following assumptions:

1. The decision point suggests the number of documents that the user considers in the ranked list.
2. The user continues to look at the ranked list in the case of the last document opened being relevant.

This approach considers the documents in the ranked list. The procedure ensures that different ranked lists lead different user behaviour to make the evaluation process as realistic as possible. Despite the higher personalization of the user behaviour, this approach is still based on the decision points in the search logs. Consequently, an approach that does not use the information form the search logs has been designed to provide a simulation of the user behaviour as independently as possible from the retrieval system used.

Using a probabilistic distribution: The probabilistic distribution defined by [12] was used to simulate user behaviour to be independent of the search logs. In this work they suggested that a user progress from one document in the ranked list to the next with probability p, and end their examination of the ranking at that point with probability 1 - p, making the following suggestions:

1. Each decision point is made independently of the current depth reached in the ranking, independently of previous decisions, and independently of whether or not the document examined is relevant or not.
2. The user always looks at the first document, looks at the second with probability p, and looks at the third with probability p^2, and at the ith with probability $p^{i\text{-}1}$.

The user model proposed is a reasonable approximation of how people consult ranked lists and similar behaviour has been observed in user experiments.

This evaluation tool produces a report showing a set of standard IR evaluation measures, computed and compared to highlight the different measures between the input algorithms. A set of charts is also generated to display the evolution of evaluation measures through a session.

The evaluation tool generates a report with the following fields:

1. An initial row providing all the information about the query:
   (a) User ID
   (b) Topic ID
   (c) Query ID
2. Measure name: The name of the calculated measure.
3. Algorithm name: The retrieval algorithm used for the run.
4. Measure value: Represents the performance of the algorithm which retrieved the ranked list used to compute the measure.
5. S_1 - S_i: It represents the difference between the algorithm with the best performance for the measure considered and the algorithm considered.

5 Participant Submissions

Two participant papers were accepted for presentation at the PIR-CLEF session at the CLEF 2018 conference. The papers have made use of the provided data collection to examine two different tasks, and both present interesting and useful findings and suggestions on how to improve the PIR-CLEF dataset.

The paper titled *ECNU at CLEF PIR 2018: Evaluation of personalised information retrieval* [13] presents a study exploring the potential of the dataset provided by PIR CLEF. The authors report in their paper two different experiments based on two distinct baselines, i.e. “query level baseline” and “session level baseline”; in the first baseline each single query in a session is evaluated independently, while in the second one all queries in a session are summed up to define a single query. The authors then report on the experiments they made, in which they applied two methods for query expansion and an approach to define a “topic sensitive user model” based on search sessions. In the reported discussion, the following suggestions to improve the dataset are made: i) to provide more numerous relevance labels, ii) to increase the number of provided user-related information (i.e. to have information related to more users profiles), and iii) to define richer query logs.

The paper titled *PIR based on explicit and implicit feedback* [14] addresses Task 2 of the PIR-CLEF lab, i.e. user profiling. More specifically, the authors explored the use of explicit and implicit feedback to define user profiles. Concerning explicit feedback, the subjective relevance judgments provided by the searchers for a given set of documents is employed to train a text classifier, thus exploiting the PIR task as a text classification task. Concerning implicit feedback, the correlation between information that is inferred form the data and relevance judgments provided by the users has been analyzed. Several analyses and useful remarks are reported in the paper.

6 Conclusions and Future Work

In this work the PIR-CLEF 2018 Personalised Information Retrieval task was presented. This task is the first edition of a lab dedicated to the theme of personalised search after the successful pilot held at CLEF 2017. This is the first evaluation benchmark in this field based on the Cranfield paradigm, with the significant benefit of producing results easily reproducible. The PIR-CLEF 2018 workshop has provided a Lab task based on a test collection that has been generated by using a well defined methodology. An evaluation using this collection has been run to allow research groups working on PIR to both experience with and provide feedback about our proposed PIR evaluation methodology. We also introduced our current work on a prototype system for the comparative analysis of PIR systems across search sessions.

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. A.Angiolillo, Comparative Evaluation of Personalised Search Systems, Universit degli Studi di Milano Bicocca, Milano, Italy, 2017.
12. A. Moffat and J. Zobel, Justin, Rank-biased precision for measurement of retrieval effectiveness, ACM Transactions on Information Systems (TOIS), 27(1), ACM 2008.
13. Q. Bai, J. Chen, Q. Hu and L. He. ECNU at CLEF PIR 2018 : Evaluation of Personalized Information Retrieval, In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, 2018.
14. A. Andreu-Marn, F. Martnez-Santiago, L. A. Urea-Lpez and M. C. Daz-Galiano. PIR Based in Explicit and Implicit Feedback, In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, 2018.