PERSON: Personalized Information Retrieval Evaluation Based on Citation Networks

Shayan A. Tabrizi[a], Azadeh Shakery[a,b,*], Hamed Zamani[c], Mohammad Ali Tavallaei[d]

[a] *School of ECE, College of Engineering, University of Tehran, Tehran, Iran*
[b] *School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran*
[c] *Center for Intelligent Information Retrieval, College of Information and Computer Sciences, University of Massachusetts Amherst, MA 01003*
[d] *Institute of Applied Intelligent Systems, University of Tehran, Tehran, Iran*

Abstract

Despite the importance of personalization in information retrieval, there is a big lack of standard datasets and methodologies for evaluating personalized information retrieval (PIR) systems, due to the costly process of producing such datasets. Subsequently, a group of evaluation frameworks (EFs) have been proposed that use surrogates of the PIR evaluation problem, instead of addressing it directly, to make PIR evaluation more feasible. We call this group of EFs, *indirect evaluation frameworks*. Indirect frameworks are designed to be more flexible than the classic (direct) ones and much cheaper to be employed. However, since there are many different settings and methods for PIR, e.g., social-network-based vs. profile-based PIR, and each needs some special kind of data to do the personalization based on, not all the evaluation frameworks are applicable to all the PIR methods. In this paper, we first review and categorize the frameworks that have already been introduced for evaluating PIR. We further propose a novel indirect EF based on citation networks (called *PERSON*), which allows repeatable, large-scale, and low-cost PIR experiments. It is also more information-rich compared to the existing EFs and can be employed in many different scenarios. The fundamental idea behind PERSON is that in each document (paper) d, the cited documents are generally related to d from the perspective of d's author(s). To investigate the effectiveness of the proposed EF, we use a large collection of scientific papers. We conduct several sets of experiments and demonstrate that PERSON is a reliable and valid EF. In the experiments, we show that PERSON is consistent with the traditional Cranfield-based evaluation in comparing non-personalized IR methods. In addition, we show that PERSON can correctly capture the improvements made by personalization. We also demonstrate that its results are highly correlated with those of another salient EF. Our experiments on some issues about the validity of PERSON also show its validity. It is also shown that PERSON is robust w.r.t. its parameter settings.

Keywords: Personalized search, Evaluation, Citation networks, Personalization

*Corresponding author
Email addresses: s.tabrizi@ut.ac.ir (Shayan A. Tabrizi), shakery@ut.ac.ir (Azadeh Shakery), zamani@cs.umass.edu (Hamed Zamani), tavallaie@alumni.ut.ac.ir (Mohammad Ali Tavallaei)

1. Introduction

The diversity of users and their information needs makes personalized information retrieval (PIR) a necessity in Web-based information retrieval (IR) systems. However, since evaluating the performance of PIR systems depends on the users' opinions and interests, the Cranfield paradigm-based evaluation [1] is not sufficient anymore for this task. On the other hand, evaluating PIR methods by real users in real scenarios is very costly and is not scalable. Therefore, evaluating such systems is a challenging task.

The difficulty of evaluating PIR methods can be discussed from three perspectives: *i)* From the judgments perspective, there is no globally correct judgment. Judgments differ for each user, and thus we have to deal with a two dimensional space (users and documents) instead of a one dimensional space (documents), in which the size of the users dimension is as large as the number of all human beings. This makes data gathering too expensive and seriously challenges the generalizability of the results, according to the severe sparseness of the space; *ii)* From the users perspective, the user whose judgments we have must be known. We need some sort of information about the users to be able to provide the information to the PIR methods being compared and observe how well each of them can use the information to personalize the results. This information is hard to gather. Even if it is available, publishing it publicly is not possible, in many circumstances, because of the privacy concerns. On the other hand, even supposing we could obtain the information for a number of people, we could not do so for all the people. This brings up the issue of how well our sample represents people of different kinds; *iii)* From the PIR methods perspective, different methods demand different resources to perform the personalization. For example, a social network (SN) of users is required by SN-based PIR methods, while a textual profile of users is required by a profile-based PIR method. The demand of resources causes many of the proposed EFs to be inapplicable to many PIR methods since they cannot provide the required information. The inapplicability makes evaluating a large number of PIR methods on a common dataset very hard, and this by itself makes comparing different kinds of PIR methods extremely complicated. This is one important reason why we do not observe many research studies focused on thoroughly studying and comparing the performances of different PIR approaches, unlike in many other research fields.

The above difficulties in PIR evaluation has given rise to several frameworks for personalized retrieval evaluation, each one having its own pros and cons. We have divided them into two categories: direct and indirect evaluations. In direct evaluations, users themselves participate in evaluating personalized search systems. For instance, they may be asked to fill questionnaires or participate in interviews to express their opinions about retrieval systems (e.g., [2]). In indirect evaluations, on the other hand, PIR systems are evaluated using a surrogate problem similar to PIR. For example, the data of users' taggings in a folksonomy [3] website are used to simulate search operations (e.g., [4]). We discuss indirect evaluation thoroughly in Section 2.2. Although direct EFs can be more accurate compared to indirect ones, they are highly expensive in terms of time, cost, and human resource. Therefore, using large-scale and flexible indirect EFs is inevitable in many circumstances. To the best of our knowledge, this paper is the first paper on PIR evaluation that seriously considers different indirect EFs and surveys them. For another survey on PIR evaluation, see [5].

There are several approaches to indirect evaluation of PIR methods. We have categorized indirect EFs into five categories: category-based evaluation (e.g., [6–9]), interaction simulation (e.g., [10, 11]), play-count-based evaluation (e.g., [12]), folksonomy-based evaluation (e.g., [4, 13–17]), and desktop search evaluation (e.g., [18–22]). These frameworks have been previously used as surrogates to evaluate PIR systems (except the last one, desktop search evaluation, as will be discussed in Section 2.2.5). All of these frameworks have several simplifying assumptions to make the evaluation possible. However, as described above, due to the fact that personalization is commonly performed based on some information of the users and different PIR methods demand different information items, not all of these evaluations are applicable to all PIR methods (See [23, 24]). For instance, some PIR methods do the personalization based on a SN of users (e.g., [8]), while some of these EFs (e.g., interaction simulation) do not necessarily have the SN of users.

Our Work. In this paper, we propose an **information-rich evaluation framework** that is suitable for evaluating PIR methods with different information needs. By "information-rich" we mean having many information items (such as SN, user profiles, keywords, document categories, time) that can potentially be provided to different PIR methods to be used for personalization.

The proposed framework is based on citation networks. The main idea behind the proposed EF is that the documents (papers) cited in a document d are potentially related to d from the perspective of d's authors. In other words, the documents cited in d could be considered as relevant documents to a query generated from document d for the authors of d. According to this idea, we generate a number of queries from scientific publications and use them to evaluate PIR methods. This evaluation framework, which is called *PERSON*[1], allows repeatable, large-scale, and low-cost PIR experiments. This framework is also rich in information items. For example, the co-authorship network can be considered as a SN of users, or the documents' keywords can be obtained from the dataset. This information richness is discussed more in Section 3.2.

It is important to bear in mind that PERSON does not intend to completely replace direct evaluation (direct user feedback), rather it is a low-cost and flexible alternative to it. Although PERSON can give us much information about the performance of PIR systems, it is still highly important to gather real users' feedback. However, when a user study is not possible due to the lack of time or resources, or when the PIR methods change frequently (e.g., in the research and development phase), PERSON would be an excellent choice. Furthermore, even when directly studying users is possible, PERSON can be used to limit the number of PIR methods that users should evaluate, e.g., through parameter tuning. This can make user studies easier and more worthwhile.

To examine our proposed EF, we use a cleaned version of AMiner's citation network V2 dataset[2] [25] containing approximately 600,000 scientific publications. We conduct quite a few experiments to study the reliability and validity of PERSON.

Research Questions. In order to validate PERSON, we address the following research questions throughout this paper:

1. Each PIR method is, in the first place, an IR method. Can PERSON correctly

[1] PErsonalized Retrieval evaluation baSed On citation Networks

[2] `https://aminer.org/citation`

rank non-personalized IR methods according to their retrieval performances? Is PERSON consistent with basic IR heuristics [26]?
2. Can PERSON be used to evaluate personalized IR methods? Are its results consistent with those of human judgments?
3. Can PERSON be used to evaluate SN-based PIR methods? Is co-authorship network a proper source of information for personalization?
4. There are several issues that may challenge the validity of PERSON (e.g., Does not the noisy nature [w.r.t. the judgments] of our defined query [title of the searcher's paper] make it uninformative and useless in the search? See Section 4.5 for the list of issues discussed). Do these issues question the validity of our framework?
5. Some of the documents PERSON considers relevant may be indeed irrelevant and vice versa. Do these misjudgments make PERSON's evaluations unacceptable?
6. Is PERSON robust w.r.t. its parameter settings?

In summary, our extensive experiments indicate that PERSON is a reliable and valid way of evaluating PIR methods. Table 11 illustrates the key findings of our experiments.

Contributions. The contributions of this paper can be summarized as follows:

1. We provide a survey of the previous personalized search evaluation frameworks with a novel categorization of them. To the best of our knowledge, this paper is the first work that seriously considers different indirect EFs and surveys them;
2. We propose a novel EF based on datasets of scientific publications that makes evaluating personalized search methods possible without any user involvement. The EF allows repeatable, large-scale, and low-cost PIR experiments. An important characteristic of the proposed EF is that it is more information-rich compared to the existing EFs and can be employed in many different scenarios;
3. We conduct many experiments to study the reliability and validity of the proposed framework from different perspectives.

Outline. The remainder of this paper is organized as follows: Section 2 reviews the existing evaluation frameworks for PIR systems; PERSON is further introduced and discussed in Section 3; PERSON is then evaluated in Section 4; We finally conclude our paper and discuss possible future directions in Section 5; To make the paper flow smoother, we explain several reproducibility details in a separate appendix (Appendix A).

2. Personalized Search Evaluation Frameworks

Evaluating PIR systems is a challenging task because of to the reasons explained in the previous section. Therefore, various frameworks have been so far proposed for evaluating PIR methods. In this section, we provide a new classification of the existing PIR evaluation frameworks. We divide the frameworks into two main categories–direct evaluation and indirect evaluation–which are in turn divided into several categories. In the following, we discuss these categories and highlight their strengths and weaknesses. Note that the objective here is not to mention every single paper that has employed some EF, but to discuss different categories of EFs, while giving some illustrative examples.

2.1. Direct Evaluation

In the first category of EFs, humans are involved in the PIR evaluation process. In fact, they either implicitly or explicitly determine which documents are relevant and

which ones are not. Direct frameworks are supposed to be the most accurate EFs since they directly evaluate PIR performance by humans (i.e., real users of PIR). However, these frameworks are either highly expensive in terms of time, cost, and human resource or not easily applicable for research purposes [27].

Direct EFs are categorized into two different types [27]: offline and online. In offline EFs, experts or actual users are asked to explicitly evaluate the results of retrieval systems. Conversely, in online EFs, the interactions of users with retrieval systems are used to estimate the performance of the retrieval systems. In theory, offline EFs may be more accurate than the online ones since in offline EFs judgments are explicitly determined by users, while in online EFs judgments must be estimated from the interactions of users, such as users' click logs (although in practice this is questionable since laboratory behaviour of users may not be consistent with their real behaviours [27]). On the other hand, online frameworks are often easier to use, but they normally are not publicly accessible [27] (mostly the users' profiles are not published in click log datasets because of privacy concerns).

In the following, we review the existing offline and online direct frameworks for evaluating PIR systems.

2.1.1. Relevance Judgment

Relevance judgment is an offline EF in which users explicitly judge the documents for each query (e.g., [27–33]). Using these judgments, some metrics like MAP, ERR [34], and NDCG [35] are calculated and these metrics are used to compare the performances of different PIR methods. Usually, some methods such as pooling [36, 37], intelligent topic selection [38], or filtering based on citations [39, 40] are used to limit the number of documents to be judged in order to reduce the amount of human work needed. This kind of evaluation is best suited for creating standard evaluation testbeds, but is very costly to be performed. Because of the high cost, this EF is generally information-poor, e.g., the corresponding datasets are small or do not have users' SN. It is noteworthy that in some works (e.g., [6, 41]), a number of evaluators are employed and are asked to assume themselves as users with particular profiles and judge the documents from their perspectives.

2.1.2. Side-by-side Evaluation

Side-by-side evaluation is an offline EF which is used for comparing the results of two or more retrieval systems. The strength of this framework is that users directly decide which retrieval system performs better . Hence, there is no need to consider certain evaluation metrics (e.g., MAP or NDCG), which per se impose certain biases to the evaluation results. A weakness of this framework is that users might consider only a few of the top-retrieved results, and this may bias the evaluation towards the high-ranked documents. This bias may not be acceptable in all scenarios, especially when recall is more important. In addition, although this EF may require less human work compared to the relevance judgment, it is still costly and time-consuming. Moreover, new judgments must be made for each new PIR method being compared, which makes this EF absolutely unscalable, specially in tuning the PIR methods' parameters (in which many configurations must be compared). This framework is used for example in [42].

2.1.3. Click-based Evaluation

Evaluating PIR systems based on clicks is one of the online EFs which was previously used, e.g., in [43–47]. This framework considers a click as an indicator of relevancy, although there are different ways to interpret it as a quantitative relevancy score. Sometimes, other information about the interactions of the users such as mouse movements and dwell-time is also used. For example, [47] considers a document as relevant iff it is clicked and the click either is followed by no further clicks for 30 seconds or is the last click in the session. Although the information about the users' clicks and their profiles can be easily accessed by a search engine, it generally is not publicly accessible. This framework also has two main drawbacks: (*i*) users' behaviours depend on the ranked list generated by the search engine, and (*ii*) users often click on a few documents, and thus information about the relevancy of other documents (especially those that are not in the top-retrieved ones) is not available.

2.1.4. Interleaved Evaluation

This online EF [48] (used, e.g., in [27]) combines the ranked lists generated by two (or more) retrieval systems and anonymously shows them to the user. It then evaluates them with considering the users' clicks on the results of each of them. Several variants of interleaved evaluation have been proposed, e.g., [49–53]. Radlinski et al. [53] showed that interleaved evaluation is more sensitive to changes in ranking quality than metric-based evaluation (scoring each IR method individually with some absolute metrics). Using this framework is again costly for research purposes since generally a researcher cannot change the results of an industrial search engine and get the users' feedback. In practice, a researcher probably needs to build a browser plugin to intervene in users' searches and collect the data. Obviously, finding enough users willing to install and use the plugin can be absolutely difficult and even impossible for a large number of users. Interleaved evaluation also needs new judgments for each new PIR method being compared, which makes it hard to be used for parameter tuning, although some extensions of it are proposed to address this issue (e.g., [50, 52]).

2.1.5. User Study

In this EF (e.g., [2]), real users use a search system and after that, they fill a questionnaire or participate in an interview. The results of these questionnaires and interviews are further used for evaluating retrieval systems. Table 1 illustrates some sample questions from [2]. This EF is highly expensive in terms of time, cost, and human resources. It also needs new studies for each new PIR system being compared.

2.2. Indirect Evaluation

In the second category of EFs, PIR systems are evaluated using a problem similar to the personalized search. Although, these problems differ from the personalized search, they can be used as surrogates for it to make its evaluation more feasible. In the following, we review these EFs.

The first four frameworks have been previously used as surrogates to evaluate PIR systems, while, to the best of our knowledge, desktop search evaluation has not been used for PIR evaluation. However, since desktop search is highly related to personalization [20], we believe it can be considered as a surrogate for PIR.

Table 1: Sample questions for a user study (from [2]).

What is your overall experience with systems using ranked outputs and full-text databases, such as Google? 1-7, 1 is very experienced, 7 is no experience
When faced with a search problem do you tend to: (a) Look at big picture first, (b) Look for details first, (c) Both
How satisfied are you with the overall results for this task using OmniSeer? 1-7, 1 most satisfied, 7 least satisfied

2.2.1. Category-Based Evaluation

Some EFs consider the underlying categories of a documents collection for relevance judgment. For example, [6] proposes ASPIRE. ASPIRE uses a collection whose documents are classified into several areas of interest or categories (e.g., sports, technology, politics, etc.). Each simulated user is associated with one or more of these categories and the documents in the categories are used to extract the user's profile. The paper states that any query can be used but recommends to use queries formulated by real users. Based on these settings, the paper suggests to consider a document as relevant iff it belongs to the user's categories and has been retrieved by a baseline IR method among the first *topkRel* results. It is noteworthy that the collection's documents may have manually assigned categories or can be categorized by a clustering process. Therefore, almost any collection can theoretically be used in this framework, although the paper only experiments the evaluation performance on a manually categorized dataset.

A drawback of this framework is the use of the baseline IR method, which can bias the results towards the PIR methods with characteristics similar to the baseline method. Another attribute of this framework is that it assumes that the documents not belonging to the user's categories are irrelevant. Although this assumption may be reasonable in many of searches, it is not true in all searches. In fact, one important point that differentiates IR from recommender systems is that in IR users may search for information out of their expertise/interests (or in general, characteristics), while in recommender systems the relevant recommended items are normally related to the user's expertise/interests (or characteristics). This framework is only appropriate for searches that are related to the expertise/interests (or characteristics) of users, while our framework can be used to evaluate occasional searches by considering authors' papers that are out of their main fields of research as query papers.

This paper is especially remarkable since it thoroughly studies and validates the reliability of the EF, while many other EFs neglect that important part and just propose an EF and use it in some application. Similarly, in our paper, we try to conduct several experiments to prove the validity of PERSON, although our experiments are different from theirs since they carried out a user study and we validate PERSON by other studies. Another paper in this category is [7], which considers ODP[3] categories for relevance judgment. [8, 9] also take a similar approach. They exploit the YouTube video categories as the evaluation categories.

[3] http://www.dmoz.org

2.2.2. Interaction Simulation

In interaction simulation (e.g., [10, 11]), a user and his interactions with the system are simulated according to a well-defined retrieval scenario and then used to evaluate PIR methods. This type of evaluation, unlike most of the others, considers some series of interactions for evaluation instead of a set of independent searches. Thus, this framework can be used to evaluate the ability of personalization methods to comply with the users' short-time needs. For example, [11] uses a dataset with known relevance judgments and based on that, simulates different styles of interaction. For instance, one style is to only traverse relevant information and another one is to traverse a combination of relevant and irrelevant information combined in some randomized manner. These simulations are then used to evaluate implicit feedback models. The drawback is that using this framework requires designing and implementing the simulations and making sure they are good representatives of the users' behaviours.

2.2.3. Play Count for Evaluation

In some music (or video) websites, like last.fm[4], each user can assign a tag to each music (video) item. Khodaei and Shahabi [12] proposed an evaluation framework based on the last.fm data, which can also be used for similar websites. The main idea behind their framework is to consider tags as the queries and the number of times each music is played by a user as the relevance score of that music for the user. More precisely, they consider the set of tags assigned to a music by users as a document. The friendship network is also considered as the SN used for personalization. In addition, they randomly choose one to three tags from the list of all tags, as the query, and a random user from the list of all users with a minimum of four friends as the searcher. As the judgments, they select music containing one or more query terms and order them based on the number of times the searcher has played each of them without skipping to the next music (*playcount*). The top k results are considered as relevant documents and the *playcount*s are used as relevance scores. They also filter out queries for which no results are generated.

Although in this framework relevance scores for relevant documents are personalized, these scores are independent of the queries. In other words, relevance scores of the relevant documents are solely determined based on the user. This can be considered as a major weakness of this EF.

2.2.4. Folksonomy-based Evaluation

Recently, folksonomy-based EF has attracted much attention because of its ease of access and also accompanying a SN of users (e.g., [4, 13–17, 41, 54]). This framework uses folksonomy websites (e.g., Delicious[5]) to create a PIR evaluation collection. In folksonomies, each user can assign one or more tags to each item (e.g., webpage). The main idea behind the folksonomy-based EF is to use each tag as a query and consider the items tagged by that tag as the relevant documents from the viewpoint of the tagger user. The cost of creating such personalized search collections is very low, because of the easy access to the folksonomies' data. To the best of our knowledge, no considerable

[4] `http://last.fm`
[5] `http://delicious.com`

study on the reliability of this framework is conducted. This EF is the most similar existing EF to ours. See Section 3.1.3 for a comparison.

2.2.5. Desktop Search Evaluation

Desktop search, that is searching for files (or items in general) in one's personal computer, is one of the real problems that attracts much attention due to the increasing amount of data in personal computers [18]. Several papers (e.g., [18–22]) consider the problem of evaluation in desktop search.

To the best of our knowledge, no dedicated work is focused on using desktop search evaluation for evaluating PIR; however, it can potentially be considered as a surrogate problem of PIR evaluation since desktop search is highly related to personalization [20] and it is essentially personalized. More precisely, different PIR methods can be used for the task of retrieving personal items and their evaluation results on the task are considered as their evaluation on PIR. The point is that for each personal computer the user is known and his contents and search history can be used for personalization.

However, despite the fact that desktop search has several similarities with personalized search, there are four main differences between desktop search and personalized web search: (*i*) Unlike web search, in desktop search document collections are not shared among users and each user has his own collection; (*ii*) The goal of desktop search is to find relevant information among different types of items, such as documents, emails, and presentations. Meta-data for each of these items are usually available; (*iii*) In desktop search, users often try to find known items; however, there are tremendous numbers of documents on the Web that users are not aware of and want to discover some relevant ones; (*iv*) There are lots of invaluable features in web search, such as links and anchor texts, which are missing in desktop search.

In conclusion, we think that using desktop search evaluation as a surrogate of PIR evaluation is theoretically possible and may be beneficial in some circumstances. But, comprehensive studies on its effectiveness and on the impact of the above differences on its evaluation results must be conducted.

3. PERSON: Personalized Retrieval Evaluation Based on Citation Networks

As pointed out in Section 2, direct EFs mainly are expensive and not scalable or are not easily accessible for research purposes. This makes indirect EFs a necessity in PIR evaluation. On the other hand, not all the existing indirect EFs are applicable to all PIR methods. For instance, several of the aforementioned indirect EFs are not usable for evaluating SN-based PIR methods since they are not accompanied by a SN of users. Another example is evaluating PIR methods that take the temporal dimensions of users (e.g., drift of users' preferences over time) into account. Not all the aforesaid indirect EFs have the profiles of the users over time.

Regarding all the above factors, we propose a novel information-rich indirect EF which makes evaluating PIR systems that need various information items possible. To this aim, we employ citation networks of academic publications for personalized search evaluation. The basis of our framework is that when a user writes a paper, the references are related to the paper from the author's point-of-view. Therefore, assuming user (author) u wrote a document (paper) d that references a set of documents R and assuming that q is a proper query representation of d, documents $d' \in R$ can be

considered as relevant documents to q from the u's perspective. We call d a *query paper* hereafter. Also, in the rest of this section, by "relevant document" we mean a document that is considered relevant in PERSON (as opposed to a truly relevant document), unless otherwise stated.

Although the general idea of PERSON is intuitive, there are several questions regarding its implementation and also its validity. For example, how to extract proper queries from the papers? Or a number of cited papers might be indeed irrelevant to the query; do not these papers make the evaluation process flawed? In the rest of this section, we discuss a number of these questions and then experimentally answer the rest of them in the next section.

3.1. PERSON's Components

PIR evaluation needs at least four different kinds of information: document collection, queries, relevance assessments, and some information about the users. In the following, we describe PERSON's components providing this information.

3.1.1. Document Collection

In PERSON, we use the papers of a scientific publications dataset as the document collection. However, unlike documents used in a typical text retrieval problem, the papers are structured and have different parts with specific meanings (title, abstract, authors, keywords, etc.). Therefore, we need to extract textual representations of the papers. Different textual representations of the papers are possible.

Two basic ways of extracting textual representation of a paper are *abstract-based representation* and *content-based representation*. The former only considers the abstract of a paper as its textual representation and the latter uses all of the main contents of a paper (excluding authors, keywords, etc.) as the representation. Content-based representation is not feasible in many cases since the full contents of a large number of papers are barely accessible, while their abstracts are usually much more easily accessible. Choosing which representation to use also depends on whether we need to evaluate PIR methods on short documents (use abstract-based representation) or long ones (use content-based representation). However, many other representations are possible. For example, one might consider using the main contents of a paper except the related works section and argue that related works may be very diverse and is not necessarily directly related to the gist of the paper. In this paper, since we do not have access to the papers' full contents, we use a modified version of abstract-based representation. Since title is an absolutely important piece of information about a paper, we concatenate it with the abstract and use the result as the textual representation. We call this form of representing a paper *modified abstract-based representation*.

3.1.2. Query Extraction

As pointed out above, PERSON requires extracting a query from each query paper. Since publications datasets contain several information items, i.e. are information-rich, different query extraction schemes are possible. Here, we discuss several possible choices for the query extraction:

- *Title-based scheme:* The title of each paper typically contains the main and the most important message of the paper, and thus can be considered as a proper query representation of the paper. Statistics of our dataset show that the papers' titles

contain 7.1 ± 2.4 terms after removing stop words. Considering queries with five or more terms as verbose [55], the statistics show that the title-based scheme is an appropriate way of obtaining verbose queries. Moreover, using this scheme, short queries with three or four terms are frequent, but proper queries with only one or two terms are rare. As a conclusion, the basic version of the title-based scheme is able to extract queries with more than two terms, although modifying it to produce very short queries is possible (e.g., through selecting more important terms), which we leave for future work;

- *Keyword-based scheme:* Keywords of a paper can be considered as a representation of it, and thus can be used as its representative query. This type of query can be obtained in two different ways: by automatically extracting keywords (e.g., by [56], which selects keywords based on a graph-based ranking model for text) or by using authors' defined keywords. As a comparison the former approach is usable even when the dataset does not contain the keywords data while the latter is expected to yield a more accurate representation of a paper. As a complementary note, keywords can be used to evaluate PIR methods in an analogous manner to the folksonomy-based EF. In other words, a keyword an author a has assigned to a paper p can be considered as a query and paper p is considered as the corresponding relevant document. However, this is out of the scope of this paper.
- *Abstract-based scheme:* If in any application we have to deal with very long queries, we can use the abstracts as the queries, because an abstract is supposed to be a good summary of a paper. Statistics of our dataset show that papers' abstracts contain 79.7 ± 44.1 terms after removing stop words.
- *Anchor-Text-based scheme:* Another approach to generate a query, is to use the texts around a citation in a paper as the query corresponding to that citation. The reason why we used the word generate instead of extract is to emphasize that this scheme is fundamentally different from the previous ones, in the sense that it results in a different query for each reference. In fact, using this scheme, multiple queries are generated from each query paper (some queries might be the same since they may have similar anchor texts) and for each query only one document is relevant (the cited paper corresponding to that query). Therefore, this approach requires different experimental settings and evaluation metrics. Anchor-text-based scheme is superior to the previous schemes in the sense that queries generated based on it are expected to be more relevant to their corresponding citations compared to a single query extracted based on the whole paper. This approach has a similar rationale to the use of anchor texts in web retrieval (e.g., [57]). Even, Brin and Page [58] point out that the anchor texts often provide a more accurate description of a page than the page itself. So, using anchor texts to generate queries is quite rational in PERSON (considering that the information is provided by the authors and are thus personalized). However, the problem is that the full contents of papers are often unavailable, as in our employed dataset, and even having the full texts of papers, automatically recognizing citations and their corresponding references is a complicated task. Therefore, even though this scheme can yield more precise queries, it is impractical with most of the datasets.
- *Manual scheme:* The previous schemes allow PERSON to be used fully unsupervised. However, since the most time-consuming part of a traditional PIR evaluation is obtaining the relevance judgments, one may use PERSON only to obtain the relevance

judgments and formulate the queries manually. In this scheme, the evaluator manually formulates the query corresponding to each paper in order to obtain more precise queries that are more similar to the real-world ones. This can be seen as a trade-off between the required effort and the accuracy of the queries. A similar approach is also used in [6], in which manually formulated queries are used in an indirect EF. Manual scheme may also be used after initial experimentations (e.g., to select the competitive methods and tune the parameters) are done in a fully-unsupervised manner, to select the best performing method.

- *Hybrid schemes:* Query extraction schemes are not limited to the aforementioned ones. Many hybrid schemes can be proposed based on the information available in a paper. For example, one may use a combination of title and keywords; another may consider a base query (e.g., title) and expand it with anchor text to simulate interactions of a user expanding his initial query (We call this "user interaction simulation"). Manual supervision may also be used to filter out improper queries. These and other hybrid schemes may be preferable in some applications.

3.1.3. Relevance Assessments

The general idea of PERSON's relevance judgments is discussed at the beginning of Section 3. In the following, we first explain it in more detail. Then, we discuss the validity of PERSON and its fundamental assumption.

The general idea of PERSON's relevance judgments is to consider references as personalized relevant documents. More precisely, if an author u wrote a paper and q is a proper query representation of the paper, references of the paper are considered as the relevant documents for query q from the perspective of u. Furthermore, some heuristics can be employed to improve the results:

- *Inappropriate relevants:* If a document that is considered relevant by PERSON is not retrieved by any of the PIR methods being compared, we mark it as "inappropriate" and consider it irrelevant when calculating the performance metrics. The intuition behind this heuristic is that not all the references of a paper are directly relevant (from an IR perspective) to it [39, 40] (e.g., some papers in the related works section). Therefore, we use the fact that none of the PIR methods could retrieve a reference as an indicator that the reference may not have any direct similarity (social and textual) to the query paper, and thus is probably irrelevant indeed. We discuss this in more detail below, when we discuss the main assumption of PERSON. We call the references retrieved by at least one of the PIR methods being compared *appropriate relevants*.
- *Inappropriate searches:* If none of the PIR methods being compared retrieve a relevant document in a search, we mark the search as *inappropriate* and ignore it. Although not ignoring it also does not change the comparison results (since all PIR methods will score zero in the evaluation for the search), ignoring inappropriate results helps highlighting the differences (e.g., if the query generated from a paper is not a good indicator of its references, and thus none of the PIR methods could retrieve a relevant result, its corresponding search is ignored in the significance tests).
- *Inappropriate queries:* We use a heuristic to filter out improper queries. We ignore the searches for which the query paper itself is not retrieved by any of the PIR methods being compared. We call the queries of these searches *inappropriate queries* since they probably are not appropriate summaries of the query papers. Using this

heuristic did not seem to have a substantial impact on our results (In an experiment we conducted, only one of the 2,000 queries was filtered out by this heuristic).

- *Ignoring self-citations:* We have the option to, or not to, ignore self-cited references to reduce the bias incurred by the authors' tendency to cite their own works. It is important to note that if ignoring self-citations is applied, all the papers of the authors of the query paper must be filtered out from the collection (e.g., through filters[6] in Lucene) in order not to penalize the methods ranking those papers higher. Whether to use this option or not depends on the nature of the task in hand and also the existence of methods that can exploit self-citations to unfairly achieve better results among the methods being compared. In our experiments, we did not ignore these references to provide more possibility of improvement for personalized methods.
- *Publication-date-based filtering:* Since in PERSON the references of a query paper are considered relevant to its corresponding query, almost all of the relevant documents in PERSON are published before that paper. Therefore, we almost have no judgments for the papers published after the query paper. This knowledge can be used to reduce the noise of our EF: We only consider the papers published not after the query paper in the retrieval, and thus mitigate the error incurred by the truly relevant documents that are considered irrelevant in PERSON. In a publications dataset, a large portion of the documents relevant to a query paper are published after it on average. Thus, applying this heuristic can dramatically reduce the noise of the relevance judgments in PERSON.

However, there is a main concern about the validity/fairness of PERSON and its fundamental assumption: Is it a valid/fair assumption to consider all and only the references of a paper as the relevant documents to the query extracted from the paper? We address this question in the following:

- Indeed, all references of a paper are **not** necessarily relevant (from an IR perspective) to that paper and vice versa [39, 40]. However, the important point is that although the aforesaid assumption is not true in general, the presence of a positive correlation between citing a paper and its relevancy cannot be ignored. In addition, it is not required that every single cited paper, which is considered as relevant in PERSON, be truly relevant. That is because we just need to compare the results of different PIR methods and mistakenly considering some irrelevant document as relevant is the same for all the PIR methods; therefore it does not violate the fairness of our comparisons, when our evaluation measure is averaged over many queries. We will experimentally show this in Experiment VII of Section 4. Moreover, according to the inappropriate searches heuristic, if none of the PIR methods being compared retrieve a relevant document, we ignore the search. Also, in Experiment VI of Section 4, we will show that retrieving a document marked by pure chance as relevant (without considering any textual nor social similarity) is very unlikely. So, if a search is marked as appropriate, it is very probable that for each appropriate relevant at least one of the PIR methods could capture a meaningful similarity (textual or social) between the pair <*query*, *searcher*> and it. Thus, it is reasonable to consider the appropriate relevants as real relevants.
- Lee and Croft [59] employ a similar idea to build a (non-personalized) web test

[6] `https://lucene.apache.org/core/6_6_0/core/org/apache/lucene/search/BooleanClause.Occur.html`

collection using the data of community question answering (CQA) platforms. They mention that answerers in CQA sometimes include URLs in their answers to provide more information. They propose to consider CQA questions as queries and the associated linked webpages as relevant documents. Obviously, this method, similar to PERSON, can have considerable noise in its judgments. However, in their experiments, they evaluate four different IR methods and show that "the relative effectiveness between different retrieval models is consistent with previous findings using the TREC queries". They state the large number of queries generated, which is also the case in PERSON, as the reason for why the noise in the judgments does not invalidate the results.

- Carterette et al. [60] demonstrate that evaluation over more queries with, up to a point, incomplete judgments is as reliable as evaluation over fewer queries with complete judgments. In PERSON, evaluations are conducted over thousands of queries which can considerably neutralize the effect of noise in the judgments. This is also validated in the experiments of Section 4. Moreover, in [60], about 1,800 queries is considered as a large number of queries. However, PERSON can be used to evaluate the systems over several hundred thousands or even millions of queries, although in practice our experiments show that its results are almost stable after evaluating about 1000 queries.
- Studying the relationship between each paper and its references in other researches also shows the validity of this basic assumption. Deng et al. [61] exploit heterogeneous bibliographic networks for expertise ranking. They hypothesize *document consistency hypothesis*, in which they say: it is reasonable to assume the neighbors of a document are similar to it. This hypothesis is similar to our assumption that the content of a document is probably similar to that of its neighbors in a citation network. The paper then validates document consistency hypothesis, which can be interpreted as a validation of our assumption.
- The rationale of the folksonomy-based EF is very similar to ours. They both consider the *extension* (paper or tag) assigned to a document by a user as his query and the document itself as the relevant result. Obviously, despite the simplification assumptions made in both, they are still sufficiently invaluable due to the lack of enough data for PIR evaluation. However, folksonomy-based framework is better for evaluating shorter queries, while PERSON is more suitable for longer ones.
- It is worth noting that all indirect EFs (See Section 2.2) have similar simplifying assumptions. For example, the play-count-based framework of [12] calculates the relevance score of each relevant music for a query by counting the number of times the music is listened to by the user, which is obviously not true since it does not take the query into account. As another example, folksonomy-based EF considers the pages a user has tagged as relevant to those tags and the other pages as irrelevant. This assumption is also obviously not true since there may be many other relevant pages that the user has not tagged. However, according to the lack of standard and one-size-fits-all evaluation resources for personalized search, indirect EFs, including our proposed one, are much beneficial despite their simplification assumptions.

3.1.4. User Profiles

All PIR methods require some information about users to personalize the results based on it. One major drawback of many EFs (direct and indirect) is that they

can provide little information to the PIR methods. This makes them impractical for evaluating many PIR methods. PERSON is outstanding in the sense that it can provide different kinds of information about users, and thus can be used to evaluate a wide range of PIR methods. Here, we discuss some information that PERSON can provide for personalization:

- *Social Network:* Several PIR methods need a SN of users to perform personalization based on it (e.g., [12]). In PERSON, the co-authorship network of authors can trivially be used as a SN of users. Even, if some PIR method needs more SNs of users (e.g., need a multilayer network [62]), networks based on authors' affiliations or their geographical locations or their publication venues can be used. Moreover, if a PIR method requires a SN with textual edges [63], the co-authorship network with the co-authored papers as edge labels can be used.
- *User Profile:* Many PIR methods need users' profiles (e.g., [29]). PERSON can provide the profiles based on the users' papers. This can be done in different granularities; the profiles can include abstracts, full contents, or keywords of the users' papers. Therefore, the profiles with different granularities can be used to evaluate PIR methods in different applications.
- *Temporal Dimension of Users:* The publication dates of papers can be used to evaluate PIR methods' abilities to deal with temporal dimension of users, e.g., the changes of the users' preferences over time (the drift of preferences) [64].
- *Search History:* Some PIR methods perform personalization based on the user's long-term search history (e.g., [65]). Since each paper of an author can be used to simulate a search operation in PERSON (with considering the appropriate relevants as the positive user feedback to the corresponding query), papers of the author can be used to simulate his search history. The simulated history may then be used for personalization. We call this approach "search history simulation" hereafter. However, to make sure that a simulated search history has the required characteristics for personalization, it needs to be thoroughly examined. We leave this examination for future work.

3.2. Comparison with Other EFs

As explained in Section 2, there are many approaches to evaluate PIR methods. However, what makes PERSON stand out from the others is that there are various kinds of information in a scientific publications dataset which makes PERSON highly flexible. Here, we summarize the information items in publications datasets that can potentially be used in the evaluation process. Afterwards, we compare PERSON with other EFs.

- *Documents with different granularities:* As described in Section 3.1.1, documents with different granularities can be used in PERSON. This makes PERSON usable in many different applications.
- *Queries with different granularities:* Several schemes for extracting queries from a query paper are possible in PERSON, as described in Section 3.1.2. This allows PERSON to use queries with different granularities in the evaluation process and makes it usable in many different applications. However, since we only focus on the title-based query extraction scheme in our experiments, the quality of evaluation

based on other schemes must be assessed more thoroughly in future work. This includes assessing the evaluation results for very short queries (one or two terms).

- *Social networks of users:* Several SNs may be extracted from the data of a publications dataset; however, our focus in this paper is on the co-authorship network. See Section 3.1.4 for more details.
- *User profiles:* As described in Section 3.1.4, profiles with different granularities can be used in PERSON. This makes it applicable to many different applications. It is worth noting that in some EFs (e.g., the folksonomy-based framework), the

Table 2: Comparing different EFs for personalized search. The features we think are more important are emboldened.

Feature	Offline Direct	Online Direct	Category Based	Interaction Simulation	Desktop Search	Play Count	Folk. based	PERSON
Number of Queries	small	**large**	**large**	$\triangle$	small	**large**	**large**	**large**
Building Cost	high	**low**	**low**	medium	medium	**low**	**low**	**low**
Public Accessibility	**yes**	no	**yes**	$\triangle$	no	**yes**	**yes**	**yes**
Direct User Evaluation	**yes**	**yes**	no	no	no	no	no	no
Joint User-Query Relevancy	**yes**	**yes**	no	$\triangle$	**yes**	no	**yes**	**yes**
Multiple Interactions	no	**yes**	no	**yes**	*	no	no	$\oplus$
Short Query	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**$^{\circ}$
Verbose Query	**yes**	**yes**	**yes**	$\triangle$	no	no	no	**yes**
Social Network	*	*	no	-	no	**yes**	**yes**	**yes**
Multiple Social Networks	no	no	no	-	no	no	no	**yes**
Document Category	no	no	**yes**	-	no	*	no	**yes**
Query Category	**yes**	no	?	-	no	no	no	**yes**
Keywords	no	no	-	-	no	no	no	**yes**
Document Time	-	**yes**	-	-	**yes**	no	**yes**	**yes**
Query Time	no	**yes**	-	$\triangle$	**yes**	no	**yes**	**yes**
Profile Documents with Multiple Granularities	no	no	-	-	no	no	**yes**	**yes**
User-based Profiles	*	no	no	-	no	**yes**	no$^{\diamond}$	**yes**
Search History	no	**yes**	$\star$	$\triangle$	**yes**	no	$\star$	$\star$
Dataset Type	**general**	**general**	**general**	**general**	specific	specific	specific	specific$^{\bullet}$
Small Profiles	**yes**	**yes**	no	**yes**	*	**yes**	**yes**	**yes**
Unbiased EF	**yes**	**yes**	no	$\triangle$	**yes**	no	**yes**	**yes**
Non-Personalized IR Evaluation	**yes**	**yes**	no	$\triangle$	**yes**	no	**yes**	**yes**
Is Validated?	**yes**	**yes**	**yes**	no	**yes**	no	no	**yes**

$^{\triangle}$ Depends on the approach used for the simulation.
* The framework does not necessarily have that property and providing it in the framework is not trivial but may be possible.
$^{\oplus}$ May be possible through user interaction simulation (See Section 3.1.2).
$^{\circ}$ Although, NO for very short queries when using title-based scheme, as used in our experiments.
$^{-}$ Depends on the dataset used. For some datasets (e.g., a publications dataset), it may be possible.
$^{?}$ Depends on the approach used for obtaining queries.
$^{\diamond}$ Although tags are chosen by the users, the pages they tag are not written by themselves.
* May be possible through search history simulation (See Section 3.1.4).
$^{\bullet}$ However, PERSON may be used in datasets having links among their documents and with known documents' authors and is not limited to scientific papers datasets. See Section 5 for more details.

documents used for constructing the profile of a user are not necessarily written by the user himself (e.g., the user has only tagged the documents in [17]). This indicates an over-reliance on the documents for profile construction since the actual information we have about the user is that he thinks the document is related to the tag and we cannot certainly consider all of the document's contents as the user's preferences (For example, the user might have tagged a page with "scam"). However, in PERSON, the profiles are constructed based on the papers the user has authored (or co-authored). Therefore, we expect that the profiles constructed based on the users' papers are more indicative of the users' preferences.

- *Temporal dimension of users:* The publication dates of papers can be used to capture the temporal dimension of users. See Section 3.1.4 for details.
- *Venues:* The venues that papers are published in, such as conferences, workshops, and journals, can be considered for both results analysis and evaluating domain-specific retrieval.
- *Keywords:* In addition to using keywords for query extraction, they can be used as documents' keywords if a PIR method needs the keywords of documents.
- *Classification of papers:* Most of the well-written papers are classified into one or several categories (e.g., categories of the ACM CCS classification tree[7]) by their authors. These categories can be used for several tasks, like evaluating domain-specific retrieval, document clustering/classification, or query classification. The categories may also be used, in a similar manner to the category-based evaluation (See Section 2.2.1), to make PERSON's judgments more accurate. For example, one can consider only the references that are in the same top-level category of the query paper as relevant, or give them higher relevance scores. This can be viewed as a combination of the category-based framework and PERSON.

The above information items make PERSON an information-rich, flexible, and general framework for evaluating PIR methods. In Table 2, we summarize the items and compare them to those of the existing EFs. The values of this table are based on a basic form of each EF and a normal setting of web retrieval evaluation. Also, since different versions of the category-based EF may be possible, we consider ASPIRE [6] as the representative method for the purpose of filling this table. In the following, we provide a brief explanation of the features in the table:

1. *Number of Queries:* The number of queries the EF can provide for evaluation.
2. *Building Cost:* The cost of gathering the data, implementing the required models and preparing the testbed.
3. *Public Accessibility:* Can the necessary data for performing the evaluation be publicly published? The main concern here is the users' privacy.
4. *Direct User Evaluation:* Do real users evaluate the PIR problem (not a surrogate problem)?
5. *Joint Query-User Relevancy:* Are the relevant documents determined by a combination of query and searcher or are they determined solely based on the searcher (and possibly filtered after retrieval)? For example, in a play-count-based EF, the relevancy of a track is determined by the number of times the user (searcher) has played it, independently from the query (No query-dependent relevancy). However,

[7] http://www.acm.org/about/class/class/2012

after retrieval, only the tracks played by the user which are retrieved are considered as relevant. On the other hand, e.g., in folksonomy-based EF, the relevancy of a document is determined based on the tagger (searcher) and the tag (query) he assigned to the document. Thus, folksonomy-based EF has query-dependent relevancy.

6. *Multiple Interactions:* Can the EF evaluate the behavior of a PIR method in search sessions with multiple interactions of the user?
7. *Short Query:* Can the EF evaluate the performance of a PIR method for short queries.
8. *Verbose Query:* Can the EF evaluate the performance of a PIR method for verbose queries.
9. *Social Network:* Can the EF provide a SN of users for PIR methods?
10. *Multiple Social Networks:* Can the EF provide more than one SN of users for PIR methods?
11. *Document Category:* Are documents categorized into pre-defined categories?
12. *Query Category:* Are queries categorized (or can be easily categorized) into pre-defined categories?
13. *Keywords:* Are keywords of each document available?
14. *Document Time:* Is the publication time of a document known?
15. *Query Time:* Is the issue time of a query known?
16. *Profile Documents with Multiple Granularities:* Does the EF have the profile documents in multiple granularities, and thus can it be used to evaluate PIR methods in different settings (requiring different profile document lengths)?
17. *User-based Profiles:* Are the profiles of the users constructed based on the texts they authored (or co-authored) or are they constructed based on others' texts that the searcher has shown some interest in (e.g., tagged it with some tag).
18. *Search History:* Can the EF provide the search history of users to the PIR methods so they can personalize the results based on the history?
19. *Dataset Type:* Is the EF only applicable to a specific type of datasets (e.g., scientific papers datasets) or it can be applied to a wide range of datasets.
20. *Small Profiles:* Can the EF be used to evaluate cases when the user profile is small or can it just be used to evaluate PIR for users with very large profiles? This is important especially for evaluating PIR methods under cold-start conditions. A *yes* in the table indicates that evaluation on both small and large profiles is possible, while a *no* indicates that the EF can only be used on very large profiles.
21. *Unbiased EF:* Are the relevant documents in the judgments determined by considering the ones a baseline (probably non-personalized) retrieval method retrieves? For example, in the category-based evaluation, a document is considered relevant iff it belongs to the user's categories and has been retrieved by a baseline method among the first *topkRel* results. This can bias the judgments towards that particular baseline method which can make a problem if the methods being compared are of different natures. For example, one reranks the results of the baseline method and the other is not dependent on the baseline and uses some other retrieval basis.
22. *Non-Personalized IR Evaluation:* Can the EF be used to evaluate non-personalized IR methods?
23. *Is Validated:* Is there, to the best of our knowledge, any considerable examination on the validity of the EF?

From Table 2, it can be seen that PERSON is superior to many other EFs in different aspects of the comparison. However, one obvious deficiency of PERSON is that when using the title-based query extraction scheme, as in our experiments, proper queries with only one or two terms are rare, and thus they cannot be reliably evaluated. Selecting important terms of title or using other query extraction schemes may be used to evaluate very short queries. However, this first requires examining the reliability and validity of these solutions, which we defer for future work. Another deficiency of PERSON is that, at least at first glance, it is applicable only to the scientific papers datasets. However, in Section 5, as a direction for future work, we describe how it may be also applicable to some other types of datasets.

4. Experiments

In this section, we describe our experiments on the proposed EF and analyze the results. To make sure the reader can capture the essence of the results and is not confused with irrelevant reproducibility details, we describe the details in a separate section (Appendix A). We number the experiments to make them more easily referable in the appendix. In the following subsections, we first introduce the dataset we use in our experiments. We then describe the experimental setup. Afterwards, in each subsection, we analyze PERSON from a particular perspective. We address the research questions mentioned in Section 1 in this section. We finally summarize the results in Section 4.7.

4.1. Dataset

To examine the proposed EF, we use the AMiner's citation network V2 dataset [25] containing approximately 1.4 million scientific publications. We cleaned the dataset and wrangled it into another dataset containing over 600k papers. Detailed pre-processing and data cleaning steps as well as the dataset statistics are reported in Appendix A.1. The dataset[8] and the related codes[9] are freely available for research purposes.

4.2. Experimental Setup

In the experiments, we use a basic setting of PERSON. We use the title-based scheme (See Section 3.1.2) to extract queries from papers. We also use modified abstract-based representations of papers as documents. In addition, we exploit the inappropriate relevants, inappropriate searches, inappropriate queries, and publication-date-based filtering heuristics (See Section 3.1.3).

All of the results reported are based on a total of 2,000 queries, unless otherwise stated. The numbers of appropriate searches are mentioned in Appendix A. We consider the first author of each document as the person who issued its corresponding query since he is expected to have the most contribution to the paper.

In the experiments, we use three widely used evaluation measures in the field of information retrieval: mean average precision (MAP), precision of the top k retrieved documents (P@k), and normalized discounted cumulative gain (NDCG) [35]. MAP and NDCG give us estimations of the overall retrieval performance and P@k gives us

[8] Available at `https://figshare.com/articles/PERSON_Dataset/3858291`.

[9] An updated version of the codes as well as a tutorial on how to practically use PERSON are available at `https://github.com/shayantabrizi/PERSON`.

some sense of how a user perceives the performance of retrieval. We use NDCG@k also for the parameter robustness analysis, where we need to measure the performance of PIR methods at different values of k. We consider $k = 100$ for NDCG@k, unless otherwise stated. We also consider the robustness index (RI) [66] in Experiment IV (Section 4.4) for compatibility of our results with those of the paper that we compare the results with.

For each query, 100 documents are retrieved to compute the measures, unless otherwise stated. The effect of varying this value is discussed in Section 4.6. For statistical testing, we use one-tailed paired Student's t-test with 99% confidence for comparing measures and one-tailed Kendall's rank correlation coefficient [67] (τ_B) with 99% confidence for comparing rankings. The experiments are done using Apache Lucene 6.6.0[10], unless otherwise stated.

4.3. Evaluating Non-Personalized IR Methods

Experiment I. In this subsection, we answer the first research question raised above: Can PERSON correctly rank non-personalized IR methods according to their retrieval performances? Is PERSON consistent with basic IR heuristics [26]?

To this end, we first study the effect of popular IR heuristics in the proposed EF to see whether it is consistent with the basic IR axioms or not. We employ the vector space model (VSM), as implemented in Apache Lucene 6.6.0.

We evaluate the performance of several VSM-based IR methods using PERSON. We consider three dimensions for IR methods in this experiment: *i)* term frequency (TF) weighting formulation used; *ii)* using inverse document frequency (IDF) weighting or not; *iii)* using document length normalization or not. We experiment with three well-known TF weighting formulations: *i)* BinaryTF (term occurrence); *ii)* RawTF (term count); and *iii)* LogTF (logarithm of term count). We show a retrieval method by $\mathcal{M}_{norm,idf,tf}$ in which $norm \in \{\text{yes}, \text{no}\}$ indicates whether document length normalization is used or not, $idf \in \{\text{yes}, \text{no}\}$ indicates whether IDF weighting is employed or not, and $tf \in \{\text{binary}, \text{raw}, \text{log}\}$ indicates the TF weighting formulation used.

Denoting the performance of a method, $\mathcal{M}$, by $f(\mathcal{M})$, according to the idea behind TFC1 (Term Frequency Constraint 1) of [26] (that "the first partial derivative of the formula w.r.t. the TF variable should be positive"), it is expected that $f(\mathcal{M}_{x,y,\text{binary}}) < f(\mathcal{M}_{x,y,\text{raw}})$ since RawTF takes term counts into account. In this inequality and the following ones, we use $x \in \{\text{yes}, \text{no}\}$, $y \in \{\text{yes}, \text{no}\}$, and $z \in \{\text{binary}, \text{raw}, \text{log}\}$ to show all inequalities obtained by substituting these variables with all their possible values. In addition, according to the idea of TFC2 it is expected that $f(\mathcal{M}_{x,y,\text{raw}}) < f(\mathcal{M}_{x,y,\text{log}})$ since the axiomatic analysis of information retrieval heuristics [26] shows that a good TF weighting formula should satisfy the concavity condition. Also, according to the idea of TDC (Term Discrimination Constraint), it is expected that $f(\mathcal{M}_{x,\text{no},z}) < f(\mathcal{M}_{x,\text{yes},z})$ since the occurrences of more discriminative terms can be a better signal of the relevancy of the document and according to the idea of LNC1 (Length Normalization Constraint 1), it is expected that $f(\mathcal{M}_{\text{no},y,z}) < f(\mathcal{M}_{\text{yes},y,z})$ since longer documents should not receive unfairly high scores compared to the shorter ones. Thus, a total of 24 inequalities are expected to hold.

[10] `https://lucene.apache.org`

Table 3: Evaluation results for different VSM-based retrieval methods (Experiment I). All the improvements (made by using a better TF function, adding IDF, or adding document length normalization) are significant at $p < 0.01$. The results corresponding to the inequalities that do not hold are marked.

Length Normalization	IDF	TF Weighting	#	NDCG	MAP	P@10
No	No	Binary	1	0.181	0.088	0.047
		Raw	2	0.157[1]	0.062[1]	0.036[1]
		Log	3	0.227	0.113	0.060
	Yes	Binary	4	0.243	0.123	0.064
		Raw	5	0.234[4]	0.103[4]	0.059[4]
		Log	6	0.295	0.151	0.077
Yes	No	Binary	7	0.203	0.104	0.054
		Raw	8	0.249	0.126	0.066
		Log	9	0.281	0.152	0.074
	Yes	Binary	10	0.258	0.134	0.068
		Raw	11	0.304	0.158	0.080
		Log	12	**0.333**	**0.180**	**0.087**

The results of this experiment are reported in Table 3. It can be seen that for all the three evaluation measures the mentioned inequalities hold except for $f(\mathcal{M}_{\text{no},y,\text{binary}}) < f(\mathcal{M}_{\text{no},y,\text{raw}})$. Also, $\mathcal{M}_{\text{yes,yes,log}}$ achieves the best results according to all the measures, which is expected. Note also that all the improvements (made by using a better TF function or using IDF or using document length normalization) are significant at $p < 0.01$.

Experiment II. Although the above results look impressing, we conducted another experiment that resulted in interesting insights into the results. We tested if $f(\mathcal{M}_{\text{no},y,\text{binary}}) < f(\mathcal{M}_{\text{no},y,\text{raw}})$ holds on other datasets and according to the traditional Cranfield-based evaluation framework. We made use of two standard evaluation datasets: AP (Associated Press 88-89, topics 51-200) and CLEF356 (CLEF 2003,5,6 ad-hoc track collection, topics 141-200 & 251-350)[11] (See Table A.14 for the collections statistics). The results of this experiment are reported in Table 4[12]. Interestingly, it can be seen that for all the measures $f(\mathcal{M}_{\text{no,no,binary}}) < f(\mathcal{M}_{\text{no,no,raw}})$ does not hold in both the datasets and $f(\mathcal{M}_{\text{no,yes,binary}}) < f(\mathcal{M}_{\text{no,yes,raw}})$ does not hold in CLEF356. As a conclusion, it seems that the idea of considering term frequencies is valid only when length normalization is employed. This may be explained by the fact that comparing the frequency of terms in documents with different lengths may be unfair. Therefore, after ignoring the corresponding inequalities, the results of PERSON are fully consistent with the expected results according to all the measures.

Experiment III. To be more confident about the results of PERSON in non-personalized settings, we also evaluate more sophisticated non-personalized retrieval methods with PERSON. We use the language modeling framework [68] with the Dirichlet prior smoothing [69] ($\mu = 400$) as the baseline (LM). We use two pseudo-relevance-

[11] There are several larger and newer collections for evaluating ad-hoc retrieval systems, such as ClueWeb and GOV2. Note that our purpose here is to just study these equations, and thus the employed collections are sufficient for this purpose.

[12] This experiment is done using the Lemur Toolkit v4.12.

Table 4: Evaluation results for different VSM-based retrieval methods without length normalization on AP and CLEF356 datasets (Experiment II).

Dataset	IDF	TF Weighting	NDCG	MAP	P@10
AP	No	Binary	0.187	0.073	0.201
		Raw	0.167	0.065	0.195
	Yes	Binary	0.223	0.099	0.236
		Raw	0.258	0.120	0.289
CLEF356	No	Binary	0.293	0.162	0.167
		Raw	0.185	0.070	0.128
	Yes	Binary	0.328	0.180	0.174
		Raw	0.273	0.119	0.173

Table 5: Evaluation of more sophisticated non-personalized retrieval methods using PERSON (Experiment III). All the improvements (made by using a better pseudo-relevance feedback model or considering topics) are significant at $p < 0.01$.

#	Method	NDCG	MAP	P@10
1	LM	.348	.189	.092
2	Log-Logistic	.359	.195	.095
3	LL+Rel	.366	.202	.099
4	LBDM	**.378**	**.204**	**.100**

feedback-based methods for comparison. The first method is Log-Logistic [70] and the second method is LL+Rel [71]. The second method integrates the retrieval scores in the Log-Logistic pseudo-relevance feedback model, and is expected to obtain better results than Log-Logistic. In addition, we use LDA-Based Document Model (LBDM) [72] as a representative method integrating topic-model-based scores in the retrieval formula, which is expected to obtain better results compared to the baseline. We use LDA [73] (Number of topics = 100) to extract the topics. The results of the experiment are reported in Table 5. The results show that PERSON can reveal the effect of pseudo-relevance feedback in retrieval. Moreover, it can be seen that PERSON correctly gives higher scores to the method using a better PRF model (LL+Rel). Similarly, LBDM's results show that PERSON can reveal the improvement of considering topics in retrieval, as previously shown in [72]. Also, all the improvements (made by using a better pseudo-relevance feedback model or considering topics) are significant at $p < 0.01$.

To summarize, the results show that PERSON's evaluations are consistent with the basic IR axioms. Moreover, our experiments indicate that the better a retrieval method is, the higher score it obtains in PERSON, and thus PERSON is successful at comparing different retrieval methods, at least for non-personalized ones.

4.4. Personalization Effect

In this subsection, we address the second and the third research questions raised above: *i)* Can PERSON be used to evaluate personalized IR methods? Are its results consistent with human judgments? *ii)* Can PERSON be used to evaluate SN-based PIR methods? Is co-authorship network a proper source of information for personalization?

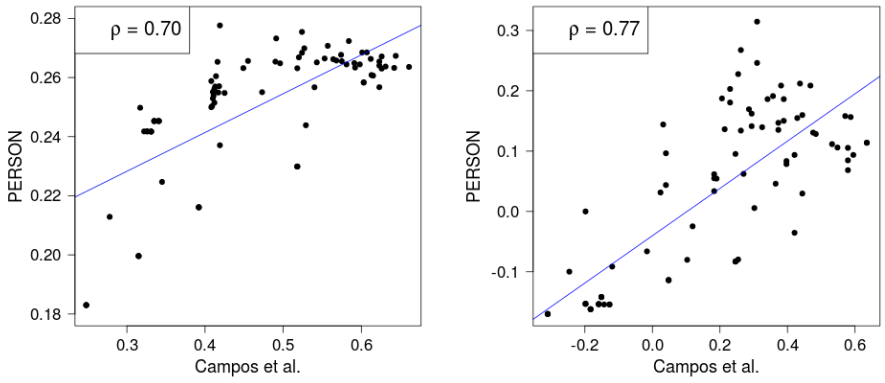
In the following experiments, we study the performance of a number of PIR methods using PERSON. We consider the Lucene's implementation of the language modeling

Table 6: Results for evaluating different Campos methods (Experiment IV). The differences between the NDCG values and the baseline NDCG that are significant are denoted by using "*". μ and σ denote average and standard deviation values, respectively.

k	p_0	QE	NQE	HRR	SRR	IRR	I-HRR	p-HRR
NDCG@50								
5	0.33	**0.230***	0.267*	0.264*	0.258*	0.258*	0.250*	0.242*
5	0.66	**0.230***	0.266*	0.263*	0.263*	0.264*	0.253*	0.242*
5	0.99	**0.230***	0.244	0.257*	0.264*	0.265*	0.255*	0.242*
10	0.33	0.216*	0.272*	0.267*	0.261*	0.261*	0.250*	0.242*
10	0.66	0.216*	0.265*	0.263*	0.266*	0.268*	0.256*	0.242*
10	0.99	0.216*	0.237*	0.257*	0.264*	0.266*	0.257*	0.242*
20	0.33	0.200*	0.275*	0.269*	0.264*	0.265*	0.251*	**0.245**
20	0.66	0.200*	0.259*	0.263*	**0.268***	0.271*	0.260*	**0.245**
20	0.99	0.200*	0.225*	0.255	0.265*	0.267*	0.257*	**0.245**
40	0.33	0.183*	**0.278***	**0.270***	0.266*	0.266*	0.254*	**0.245**
40	0.66	0.183*	0.250	0.263*	**0.268***	**0.273***	**0.265***	**0.245**
40	0.99	0.183*	0.213*	0.255	0.265*	0.266*	0.255	**0.245**
μ		0.209	0.258	0.263	0.264	0.266	0.255	0.243
σ		0.018	0.021	0.005	0.003	0.004	0.004	0.002
Baseline		0.249						
RI								
5	0.33	**-0.083**	0.112	0.085	0.114	0.114	0.034	-0.154
5	0.66	**-0.083**	0.046	0.030	0.094	0.105	0.136	-0.154
5	0.99	**-0.083**	-0.080	-0.035	0.068	0.106	0.187	-0.154
10	0.33	-0.114	0.160	0.128	0.156	0.158	0.055	-0.154
10	0.66	-0.114	0.062	0.079	0.131	0.155	0.203	-0.154
10	0.99	-0.114	-0.080	0.006	0.094	0.135	0.228	-0.154
20	0.33	-0.142	**0.162**	**0.147**	**0.209**	**0.212**	0.095	**-0.153**
20	0.66	-0.142	0.044	0.054	0.150	0.186	0.268	**-0.153**
20	0.99	-0.142	-0.091	-0.025	0.084	0.140	0.246	**-0.153**
40	0.33	-0.170	0.144	0.134	0.186	0.191	0.169	-0.162
40	0.66	-0.170	0.000	0.031	0.141	0.181	**0.315**	-0.162
40	0.99	-0.170	-0.100	-0.066	0.062	0.096	0.209	-0.162
μ		-0.127	0.032	0.047	0.124	0.148	0.179	-0.156
σ		0.034	0.101	0.070	0.046	0.038	0.085	0.004

framework with the Dirichlet prior smoothing ($\mu = 100$) as the baseline. We call this method LM hereafter.

Experiment IV. In this experiment, we consider several profile-based PIR methods– QE, NQE, HRR, SRR, IRR, I-HRR, p-HRR–proposed in [29]. We call these methods *Campos methods*. These methods use user profile terms to perform the personalization. However, they use four different approaches to personalization: QE and NQE use query expansion; HRR, SRR, and IRR rerank the original search results according to the results of NQE; I-HRR reranks the results of NQE according to the original query results; and p-HRR (which represents a bad performing method) reranks the original query results according to the results of a query composed of only the profile terms. The important point is that these methods cover a variety of personalization approaches. For details of each of these methods, see [29].



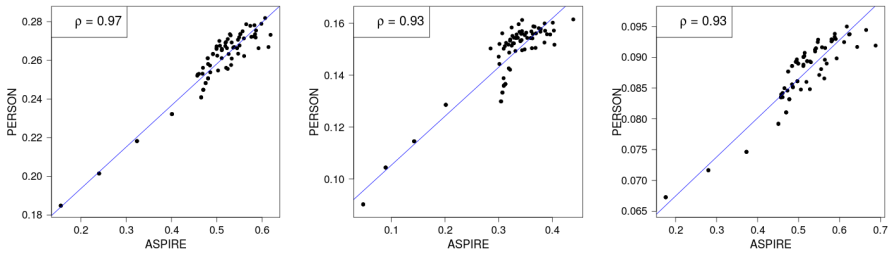
(a) NDCG **(b)** RI

Figure 1: Scatter plots of our results in Experiment IV (Table 6) against those of Campos et al. [29] and the corresponding least squares lines.

All of these personalized methods need a basis retrieval method. We use the aforementioned baseline (LM) as the basis method. Also, these methods have two main parameters: the number of expanded terms, k; and p_0 which controls the importance of the profile terms with respect to the original query terms. We test the methods with different values of these parameters. Table 6 depicts PERSON's results. This table is analogous to Table 3 of [29]. In Table 6, we denote the differences between the NDCG values and the baseline's NDCG that are significant by using "*". It is important to note that the results in Table 3 of [29] are based on human judgments.

Our results are not fully commensurable with those of [29] since they represent the performances of the PIR methods with two different implementations and on two different datasets. But, if both of the tables are really representing the performance of different PIR methods, we expect to observe to some extent similarities between them. Subjectively comparing the results, there are so many similarities. For example: QE and p-HRR mostly perform worse than the baseline, while the others improve it; QE is highly sensitive to increase in k; NQE is very sensitive to increase in p_0; p-HRR is very insensitive to the parameters. However, the best performing method in [29] is HRR, while the best one is NQE in PERSON. Interestingly, this can reasonably be explained by the way relevance judgments are collected in [29]. For each query, they consider the first 50 results retrieved by both the non-personalized baseline and HRR (but not NQE) in the judgment process. Considering the fact that unjudged documents are deemed irrelevant, the judgments are obviously biased towards HRR and thus it is probable that the reason why HRR obtained higher scores in [29] is the unfair bias.

To compare the results objectively, we calculated the Pearson correlation coefficient between each entry of our table (12 parameter configurations for each of the 7 Campos methods) to the corresponding one in the Campos's and obtained $r = 0.70$ for NDCGs and $r = 0.77$ for RIs. Figure 1 illustrates the scatter plots of the data and the corresponding least squares lines. It is obvious that a considerable amount of positive



(a) NDCG **(b)** MAP **(c)** P@10

Figure 2: Scatter plots of the PERSON's results against those of ASPIRE in Experiment V according to different evaluation measures.

correlation is observed. Considering the fact that the implementations and the datasets are different, these high correlations show that the performances of the PIR methods almost follow the same dynamics in both of the evaluation approaches. This supports the claim that PERSON correctly evaluates the performance of PIR methods.

Experiment V. To be more confident about our conclusion, we implemented ASPIRE [6] (See Section 2.2.1) to compare PERSON's results with those of ASPIRE in our dataset. The choice of ASPIRE among different EFs is important in two aspects: *i)* The paper is especially remarkable since it thoroughly studies and validates the reliability of ASPIRE, as opposed to some other EFs; *ii)* The paper evaluates ASPIRE according to human judgments, and thus the consistency of our results with those of ASPIRE supports the consistency of our results with human judgments.

In this experiment, we use our dataset as the collection for ASPIRE in order to compare the EFs on the same dataset. Although Vicente-López et al. [6] use manually formulated queries and manually assigned categories in their experiments, they state that other types of queries or other types of category assignment (e.g., by clustering) can be used. We use the title-based scheme of query extraction to obtain the queries and use text clustering for category assignment. To cluster the documents, we employ topic modeling [74]. Loosely speaking, we use LDA [73] (Number of topics = 100) to extract the topic distributions of documents and consider each topic as a cluster. Each document is then assigned to the topic which has the highest probability in the document. As a result, the papers of the dataset are partitioned into the unsupervisedly-extracted categories. These categories are used in ASPIRE to both model user profiles and generate relevance judgments. For more details on the implementation, see Appendix A.7.

Figure 2 illustrates the scatter plots of PERSON's results versus ASPIRE's results w.r.t. different measures. The figures are based on 12 parameter configurations (the same configurations as Experiment IV) for each of the 7 Campos methods. The results show that the measures are highly correlated between PERSON and ASPIRE. Also, it can be seen that the Pearson correlation coefficient for NDCG is increased from 0.70 in Figure 1 to 0.97 in Figure 2, in which evaluations are performed on the same dataset. The high correlation between PERSON's results and those of ASPIRE (which is shown to be consistent with human judgments) supports the validity of PERSON in evaluating PIR.

Experiment VI. We perform another experiment to see whether PERSON is successful at ranking different personalization methods for a fixed profile parameter configuration. In the Campos methods, there are two parameters k and p_0. We use 12 different combinations of parameter values ($k \in \{5, 10, 20, 40\}$ and $p_0 \in \{.33, .66, .99\}$), similar to the original paper [29]. For each of the configurations, we evaluate the baseline (LM) and the Campos methods (QE, NQE, HRR, SRR, IRR, I-HRR, and P-HRR) with both PERSON and ASPIRE. Assuming that ASPIRE's results are consistent with human judgments, we consider them as ground truth and examine how much the final rankings of the methods in PERSON are similar to those of ASPIRE. For comparison, we use Kendall's rank correlation coefficient, τ. Kendall's τ ranges between -1 and $+1$. A value of $\tau = -1$ indicates a total disagreement; $\tau = +1$ indicates a total agreement; and $\tau = 0$ indicates that the rankings are uncorrelated. To obtain an expected upper bound on the τ value for each profile configuration, we also use ASPIRE on two different query sets and calculate the rank correlation coefficient between the two evaluation results. We refer to the rank correlation coefficients between the results of ASPIRE and those of PERSON as τ_{PERSON} and the rank correlation coefficients between the results of the two ASPIRE runs as τ_{ASPIRE}.

Table 7 shows the rank correlation coefficients for different parameter configurations and according to the different measures. In addition, τ values of experiments in which the null hypothesis (that the rankings are uncorrelated) is rejected are emboldened. It can be seen that not only the null hypothesis is rejected in most cases, but indeed high values of τ are usually obtained. For example, average τ for the NDCG measure over all the configurations is 0.775. To have some sense of this value, it can be compared to the average (over different parameter configurations) τ values of 0.797 (for BM25 model) and 0.754 (for VSM model) reported in [6] for the rank correlations between the ASPIRE evaluations and the user study. [6] does not report τ values for MAP or P@10, so we do not have a reference point to compare with; But, except for a few exceptions, τ values for these measures are also high. Moreover, about the exceptions, a possible explanation is that in these cases the differences among the PIR methods are rather small which make discriminating among them more difficult (but also less important). The reason is that if we take a deeper look at the results, we notice that the only two τ_{PERSON} values that are less than 0.5 correspond to the values of τ_{ASPIRE} that are considerably smaller than the most occurring value of 1.0. Therefore, in these cases the differences may be subtle, such that minor perturbations in the evaluation measures (e.g., caused by using two different sets of queries in ASPIRE) change the ranks of the methods. However, even without considering this possible explanation, overall the τ values are generally high, which supports the validity of PERSON as an evaluation method for comparing the performances of different PIR methods.

In some situations, only finding the best performing method is important and finding the total ranking of the methods is not required. To assess the performance of PERSON in such a task, we assume, for each configuration and each measure, the top ranked method in ASPIRE is in fact the best method. We average the ranks of the supposedly best methods in PERSON over different configurations to have an estimation of how PERSON ranks the best methods. We show this average with $\bar{\mathcal{R}}_{top}$. In this experiment, $\bar{\mathcal{R}}_{top}$ can take values of 1–8. From Table 7, it can be seen that on average the best methods are ranked close to the top in PERSON according to all the measures. This

Table 7: Results for Experiment VI. τ_{ASPIRE} denotes the rank correlation coefficient between the evaluation results of ASPIRE on two different query sets. τ_{PERSON} denotes the rank correlation coefficient between the results of PERSON and those of ASPIRE. $\bar{\mathcal{R}}_{top}$ denotes the average, over different parameter configurations, rank of the top ASPIRE result in the PERSON's ranking. $\bar{\mathcal{R}}_{top}$ can take values of 1–8.

Measure	k	5			10			20			40		
	p_0	**.33**	**.66**	**.99**	**.33**	**.66**	**.99**	**.33**	**.66**	**.99**	**.33**	**.66**	**.99**
NDCG	τ_{ASPIRE}	**.93**	**.93**	**1.0**	**1.0**	**.93**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	τ_{PERSON}	**.79**	**.86**	**.64***	**.86**	**.86**	**.57***	**.86**	**.79**	**.64***	**.93**	**.79**	**.71**
	$\bar{\mathcal{R}}_{top}$ **[1-8]**	2.08											
MAP	τ_{ASPIRE}	**1.0**	**1.0**	**.86**	**1.0**	**1.0**	**.93**	**1.0**	**1.0**	**.93**	**1.0**	**1.0**	**1.0**
	τ_{PERSON}	**.71**	**.79**	.29	**.79**	**.79**	.43	**.79**	**.79**	.50	**.79**	**.79**	.50
	$\bar{\mathcal{R}}_{top}$ **[1-8]**	2.17											
P@10	τ_{ASPIRE}	**.98**	**1.0**	**1.0**	**.98**	**.98**	**.93**	**1.0**	**1.0**	**.93**	**.93**	**1.0**	**1.0**
	τ_{PERSON}	**.83**	**.63***	.50	**.83**	**.78**	**.57***	**.98**	**.91**	**.79**	**.98**	**.84**	**.86**
	$\bar{\mathcal{R}}_{top}$ **[1-8]**	2.17											

supports the validity of PERSON in the task of finding the best performing PIR method.

Experiment VII. In another experiment, we study whether PERSON can be used to rank different profile configurations for a fixed PIR method. Being able to correctly rank different configurations for a method is especially important since it allows the EF to be used for tuning the parameters of a PIR method; Parameter tuning may be one of the best use cases of an indirect EF, for which collecting user-based judgments without having access to a search engine data (e.g., in some academic researches) is very difficult, if not impossible. For each PIR method (QE, NQE, HRR, SRR, IRR, I-HRR, and P-HRR), we evaluated the method with different combinations of parameter values ($k \in \{5, 10, 20, 40\}$ and $p_0 \in \{.33, .66, .99\}$) as well as the baseline (LM) with both PERSON and ASPIRE. We examine how much the final rankings of PERSON are similar to those of ASPIRE (τ_{PERSON}). Similar to the previous experiment, to obtain an expected upper bound on the τ value for each PIR method, we also used ASPIRE on two different query sets and calculated the rank correlation coefficient between the two evaluation results (τ_{ASPIRE}). Table 8 displays the results. The τ values of experiments in which the null hypothesis (that the rankings are uncorrelated) is rejected are emboldened. Again, it can be seen that not only the null hypothesis is rejected in most cases, but indeed high values of τ are usually obtained. For example, average τ for the NDCG measure over all the methods is 0.82. To have some sense of this value, it can be compared to the average (over different methods) τ values of 0.73 (for BM25 model) and 0.64 (for VSM model) reported in [6] for the rank correlations between the ASPIRE evaluations and the user study. Table 8 also shows $\bar{\mathcal{R}}_{top}$ for different measures. In this experiment $\bar{\mathcal{R}}_{top}$ can take values of 1–13. It can be seen that on average the top results are ranked rather close to the top in PERSON according to all the measures. This supports the validity of PERSON in the task of finding the best parameter configuration for a PIR method.

Experiment VIII part I. We also consider a simple SN-based personalization method. The method [12] (with $\alpha = .85$) is based on a convex combination of *social* (personalization aspect, based on the SN) and *textual* (query similarity aspect) scores.

Table 8: Result for Experiment VII. $\bar{\mathcal{R}}_{top}$ denotes the average, over different PIR methods, rank of the top ASPIRE result in the PERSON's ranking. In this experiment, $\bar{\mathcal{R}}_{top}$ can take values of 1–13. For the definitions of τ_{ASPIRE} and τ_{PERSON}, see Table 7.

Measure	τ	QE	NQE	HRR	SRR	IRR	I-HRR	P-HRR
NDCG	τ_{ASPIRE}	**1.0**	**.95**	**.97**	**1.0**	**1.0**	**.97**	**1.0**
	τ_{PERSON}	**1.0**	**.85**	**.59**	**.87**	**.79**	**.82**	**.82**
	$\bar{\mathcal{R}}_{top}$ **[1-13]**	1.57						
MAP	τ_{ASPIRE}	**1.0**	**.97**	**1.0**	**.97**	**.95**	**.92**	**1.0**
	τ_{PERSON}	**1.0**	**.72**	**.51**	**.82**	.26	.23	**.64**
	$\bar{\mathcal{R}}_{top}$ **[1-13]**	3.86						
P@10	τ_{ASPIRE}	**1.0**	**.97**	**1.0**	**.95**	**.97**	**.99**	**1.0**
	τ_{PERSON}	**1.0**	**.49***	**.51**	**.87**	**.87**	**.83**	**.82**
	$\bar{\mathcal{R}}_{top}$ **[1-13]**	1.57						

Table 9: Evaluation of different retrieval methods using PERSON (Experiment VIII). The results are segmented into different parts for referability. All the differences are significant at $p < .01$.

Part	#	PIR Method	NDCG	MAP	P@10
I	1	LM	0.274	0.144	0.083
	2	Social-Textual	**0.316**	**0.170**	**0.097**
	3	IRR	0.306	0.157	0.091
II	4	Random	0.001	0.000	0.000
III	5	Social	0.097	0.036	0.023
	6	Profile	0.131	0.055	0.039

The social score for a document d, a query q, and a searcher u is calculated according to Eq. (1), in which *urf* represents the relatedness of a user to the searcher, *uaf* represents the importance/relevance of a user to a document, *uwf* represents the overall importance of a user, and U_d is the set of users with some actions (authoring a paper, in our context) on d. For more details on the social score, see Appendix A.8. We call this method Social-Textual.

$$social(d, q; u) = \sum_{v_i \in U_d} urf(u, v_i) \times uaf(v_i, d) \times uwf(v_i) \tag{1}$$

Table 9 depicts the results of the baseline algorithm (LM), Social-Textual, and the best (according to MAP) Campos method (IRR, $k = 20, p_0 = 0.66$). It can be seen that the SN-based method outperforms the baseline. Note also that all the differences in this table are significant at $p < .01$. This shows that not only PERSON can correctly capture the improvement of a SN-based personalization method, but also using the co-authorship network as the required SN is in fact rewarding and a reasonable choice.

To summarize the findings of this subsection, the results show that PERSON can be used for comparing the performances of different PIR methods and its results are consistent with those of the traditional Cranfield-based EF.

4.5. Validating the Results

In the next set of experiments, we focus on validating the proposed EF. To do so, we investigate three hypotheses that might challenge our proposed EF and check whether

they are rejected or not. We expect them to be rejected.

1. The defined query (title of the searcher's paper) is too noisy (w.r.t. the judgments) to provide any information about the relevancy of the retrieved papers;
2. The personalized methods perform better simply because the information they use for personalization (e.g. profile of the user or his position in SN) is a good indicator of the defined relevancy (being cited) and the defined query (title) is not indeed that important;
3. Since citation is used as an indicator of relevance, there are definitely many irrelevant documents (from the viewpoint of a human judger) that are considered relevant. Also, there are some relevant documents that are considered irrelevant. These make the judgements too noisy, and thus the comparisons based on them are uninformative and may be unfair.

To study these hypotheses, we conduct several experiments. Since, each of these experiments may reject several of the hypotheses, we explain them one by one and at the end of each one make the hypotheses it rejects explicit:

Experiment VIII part II. In the first experiment, we implement a retrieval method (Random) retrieving completely random documents. As can be seen in Table 9, the method results in almost zero in all the measures. This shows that the previous retrieval methods (Part I) perform by far better than Random. Also, as mentioned above, the differences are significant. This indicates that the title of a paper is in fact a strong source of information about its citations, i.e. the relevant documents, and thus the title-based scheme is indeed an appropriate query extraction scheme. As an interesting conclusion of this experiment, we can state that the retrieved documents considered relevant by PERSON, are expected to be truly relevant. This is because the probability of retrieving a random document is too low, and when a document, that is deemed citable by the author, is retrieved by an IR method, it is expected that the IR method could actually capture a meaningful relevancy between it and the query paper. This experiment rejects hypothesis 1.

Experiment VIII part III. In this experiment, we retrieve the results only based on the social score (Social in Table 9) of [12] (i.e., ignoring the query similarity aspect). We explained the social score in Section 4.4. From Table 9, it can be seen that the social score cannot be a good indicator of relevancy by itself, and indeed we require a real PIR method, combining the query similarity and the searcher's preferences, to obtain outstanding results. This is an important characteristic of a valid PIR evaluation framework, which discriminates it from a recommender system evaluation framework. This rejects hypotheses 1 and 2.

In the next experiment, we retrieve the results solely based on the profile of the user (constructed as in the Campos methods with $k = 20$), ignoring the query (Profile in Table 9). Again, Table 9 shows that this method also fails to achieve acceptable results, and thus the query text is indeed pertinent to the relevancy of results. Therefore, in PERSON, none of the mentioned information about users (SN position and user profile) are sufficient for determining the relevant results, and we need the query to capture the user's information need, as in a typical PIR evaluation framework. Thus, hypotheses 1 and 2 are rejected.

Experiment IX. To study hypothesis 3, we conduct an experiment to verify that the comparisons based on PERSON are indeed informative and fair. We expect that the

Table 10: Analysis of the effect of noise in relevance judgements (Experiment IX). #AR indicates the number of unjudged or irrelevant documents considered as relevant. %RR indicates the percentage of relevant documents considered as irrelevant. %WJ indicates the percentage of wrong performance judgments.

Dataset	**Metric**	**#AR**	0	20000	50000	100000	0	0	0	5000	20000	50000
		%RR	0	0	0	0	0.25	0.5	0.75	0.5	0.5	0.75
AP	P@10	DIR	0.44	0.50	0.60	0.78	0.32	0.22	0.11	0.23	0.29	0.49
		JM	0.41	0.47	0.58	0.76	0.30	0.20	0.10	0.22	0.28	0.48
		%WJ	-	0	0	10	0	5	11	1	12	13
	MAP	DIR	0.39	0.40	0.48	0.62	0.31	0.23	0.15	0.22	0.25	0.40
		JM	0.35	0.37	0.45	0.61	0.28	0.21	0.13	0.21	0.24	0.39
		%WJ	-	0	0	0	0	0	3	0	0	0
CLEF356	P@10	DIR	0.37	0.45	0.56	0.74	0.28	0.19	0.10	0.21	0.27	0.47
		JM	0.34	0.42	0.53	0.73	0.26	0.18	0.09	0.20	0.26	0.45
		%WJ	-	0	0	2	0	0	8	1	10	5
	MAP	DIR	0.44	0.35	0.42	0.56	0.37	0.30	0.18	0.23	0.23	0.36
		JM	0.41	0.33	0.40	0.55	0.35	0.28	0.17	0.21	0.22	0.35
		%WJ	-	0	0	2	0	3	23	0	0	0

misjudged documents do not change the relative performance (averaged over queries) of the IR methods being compared since relevant documents are selected solely based on the underlying problem and independent of any special IR method, and the chance of performance increase or decrease by the misjudgments is similar for them. To validate this, we consider some standard evaluation datasets containing human relevance judgments and inject some misjudgments in their judgments and investigate whether the misjudgments change the relative performance of the IR methods or not. More precisely, we randomly select some unjudged or irrelevant documents and consider them as relevant. In addition, we randomly remove some documents from the set of relevant documents. We make use of the same datasets used in Experiment II (See Table A.14). Language modeling framework with KL-divergence retrieval model is employed as the basis for our comparisons and two IR methods DIR (The LM basis using Dirichlet smoothing [69] with $\mu = 1000$) and JM (The LM basis using Jelinek-Mercer smoothing [69] with $\lambda = .5$) are compared[13]. These methods are chosen such that they have comparable performances but one (DIR) performs significantly better on the datasets.

Table 10 demonstrates the results of the experiment. #AR (Added Relevants) shows the number of unjudged or irrelevant documents considered as relevant. %RR (Removed Relevants) shows the percentage of relevant documents considered as irrelevant. We also define a wrong performance judgment as obtaining a performance comparison inconsistent with the non-noisy setting. In other words, obtaining a result showing that JM performs better than or equal to DIR is a wrong performance judgment (since in the non-noisy setting DIR outperforms JM). We denote the percentage of wrong performance judgments in 100 runs of each noisy setting by %WJ. In this table, we use modified MAP (in which relevant documents not retrieved by any of the IR methods being compared are considered irrelevant. See Appendix A.4 for more details) and P@10 (averaged over all queries) as the performance measures. We have also reported

[13]This experiment is done using the Lemur Toolkit v4.12.

the performance measures averaged over 100 runs.

From Table 10, it can be seen that not only the averages of the performance measures are consistent with the initial measures after adding noise, but also wrong judgments are very rare. We obtained similar results for other methods and also on the CLEF 2001-2 Ad-hoc Track collection (topics 41-140), which we do not mention here for brevity. This shows that when the number of queries is high relative to the amount of noise in the judgments, the effect of noise almost disappears when the performance measure is averaged over all queries. This is inline with the finidings of Carterette et al. [60] that evaluation over more queries with, up to a point, incomplete judgments is as reliable as evaluation over fewer queries with complete judgments.

One might argue that the reason why the noises do not considerably impact the evaluation results in this experiment is that the misjudged documents are selected completely at random. However, although this argument is valid and noisy judgments that are highly biased towards one of the IR methods being compared can change the comparison results, it is important to note that in PERSON the relevant documents are not selected according to any special IR method and are selected by humans. Thus, we expect that the results are not remarkably biased towards any of the IR methods being compared, in many cases. Moreover, our previous experiments showed a significant amount of consistency with the traditional Cranfield-based EF, and therefore we believe that assuming the judgments are not remarkably biased towards any of the IR methods being compared is a reasonable assumption in many circumstances. In addition, in this experiment, AP and CLEF356 datasets have 150 and 160 queries, respectively, which are much lower than the number of queries we use in our framework (which can be much more than 1000). Thus, we expect that the noise effect is even more decreased in PERSON by the large number of queries. This supports the fairness and informativeness of PERSON and rejects hypothesis 3.

4.6. Parameter Robustness Analysis

In this subsection, we study the impact of the parameters of PERSON on the results. There are two main parameters to be investigated: the number of retrieved results considered in computing the metrics, k; and the number of queries, $\#Q$ (More precisely, $\#Q$ denotes the number of appropriate searches in the following). For brevity, we focus on the NDCG measure. It is noteworthy that in the following experiments, to study the effect of k in the evaluations, we study NDCG@k with different values of k for simplicity. Although this is not exactly similar to changing k in PERSON (because of the inappropriate relevants, inappropriate searches, and inappropriate queries heuristics), we do not expect to observe systematically different results compared to when different values of k are used in PERSON.

Experiment X. Part I. We first study the sensitivity of the final ranking of PIR methods to the parameters in PERSON. As a sample, we first use the results of Experiment VII to analyze the sensitivity. As a reminder, in the experiment, for each PIR method (QE, NQE, HRR, SRR, IRR, I-HRR, and P-HRR), we evaluated the method with different combinations of parameter values ($k \in \{5, 10, 20, 40\}$ and $p_0 \in \{.33, .66, .99\}$) as well as the baseline (LM) with both PERSON and ASPIRE. To analyze the robustness w.r.t. $\#Q$, we compare the ranking of PIR methods obtained in PERSON after each $\#Q$ with the final ranking after evaluating all the queries by Kendall's rank correlation

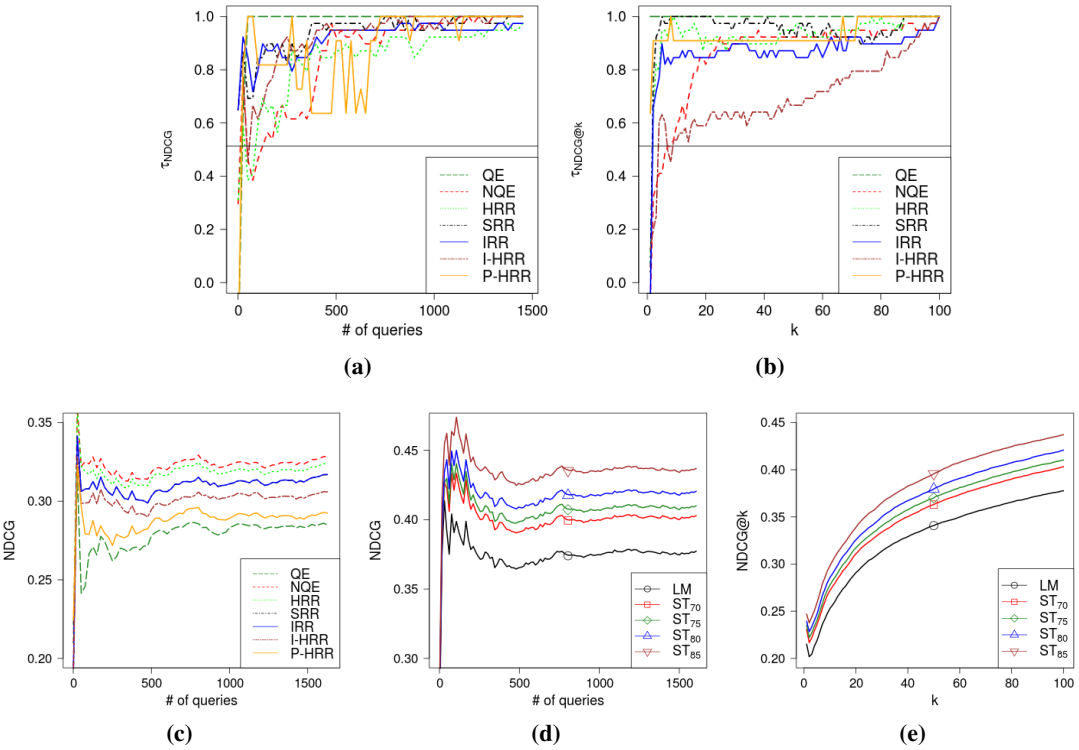


Figure 3: (a) Kendall's τ between the ranking obtained in PERSON after each $\#Q$ (the number of appropriate searches) with the final ranking after evaluating all the queries (based on Experiment VII). (b) Kendall's τ between the rankings of PIR methods for different values of k with the final ranking for $k = 100$ (based on Experiment VII). (c) NDCG values w.r.t. $\#Q$ in Experiment VI. (d) NDCG values w.r.t. $\#Q$ in Experiment X. Part III. (e) NDCG@k for different values of k in Experiment X. Part III.

coefficient. Figure 3a displays Kendall's τ w.r.t. $\#Q$ for NDCG. The 99% confidence thresholds (according to Kendall's τ_A) of rejecting the null hypothesis that the rankings are uncorrelated are also specified. It can be seen that after evaluating a few hundred queries the rankings of almost all the methods pass the null hypothesis rejection threshold and after about 1000 queries the rankings are almost identical to the final rankings. This shows that for large enough values of $\#Q$ PERSON can be deemed robust w.r.t. $\#Q$. It is noteworthy that in this experiment the methods to be ranked differ only in the parameter settings and in many cases the differences are subtle (e.g., see Table 6). Thus, correctly ranking them despite the subtle differences is indeed a hard task. In the following, we observe that in other easier experiments the convergence to the final ranking is obtained with much less queries.

In addition, Figure 3b displays Kendall's τ between the rankings of PIR methods obtained in PERSON according to NDCG@k for different values of k with the final ranking for $k = 100$. It can be seen that for values of k as small as 20 the null hypothesis rejection threshold is passed and indeed the rankings are almost identical

to the final rankings except for I-HRR which takes somehow longer to reach its final ranking. Again, we mention that this task is indeed a hard task and we will show that in easier tasks the convergence to the final rankings is obtained even at lower values of k. Therefore, the choice of k is also not critical in the performance of PERSON and for large enough values of k (e.g., $k > 40$) PERSON seems to be robust.

Experiment X. Part II. To observe whether only the rankings of the PIR methods remain almost unchanged by increasing the number of queries or indeed the absolute values of evaluation measure are stabilized, we display the values of NDCG w.r.t. the number of queries for different methods in Figure 3c. These figures are based on the results of Table 7 for $k = 5$ and $p_0 = .33$. We report the values for only one parameter setting for brevity. Figure 3c demonstrates that not only the relative ranks of the methods are stable, but also their NDCG values are stabilized after evaluating some hundred queries. In fact, NDCG values remain almost unchanged after about 250 iterations. Therefore, for large enough values of $\#Q$, even the absolute values of NDCG can be deemed meaningful in PERSON.

Experiment X. Part III. To check whether PERSON is also robust for the SN-based PIR method used in our experiments (Social-Textual), we perform another experiment. We evaluate the LM baseline as well as Social-Textual with different parameter values and compare their results. We call the methods: LM, ST_{55}, ST_{65}, ST_{75}, and ST_{85} (in which ST_x is Social-Textual with parameter $\alpha = x$). Figure 3d illustrate the performances of the methods w.r.t. $\#Q$. It can be seen that in this easier task, the NDCG values are relatively stable after a low number of queries (possibly less than 100). Also, Figure 3e illustrates NDCG@k for different values of k. It can be seen that the ranking of the methods are almost unchanged for different values of k, and thus for large enough values of k (e.g., $k > 40$) the choice of k does not seem to be critical in the final ranking of the methods in PERSON.

4.7. Summary

In this subsection, we summarize the key findings of the experiments explained in this section in Table 11.

In conclusion, PERSON can be deemed as a fair and informative way of comparing PIR methods. Moreover, considering the lack of enough resources for PIR evaluation and taking the accessibility of publications datasets and their rich information items into account, the proposed framework is extremely rewarding.

5. Conclusions and Future Work

In this paper, we studied the problem of evaluating personalized retrieval systems. To this aim, we first categorized and reviewed the frameworks previously used in the literature for evaluating PIR. We further proposed an indirect framework for PIR evaluation based on citation networks, which is an information-rich EF and is thus superior to many other EFs in its applicability to different scenarios.

To evaluate the proposed EF, we constructed a dataset by performing data cleaning on AMiner's citation network V2 dataset. The constructed dataset is freely available for research purposes. The experiments showed that the results obtained by the proposed evaluation framework match those obtained by the traditional Cranfield-based

Table 11: A summary of the key findings of the experiments.

Exp. I&II	• PERSON is consistent with the traditional Cranfield-based evaluation in comparing basic VSM-based IR methods.
Exp. III	• PERSON can correctly expose the positive effect of pseudo-relevance feedback in two feedback methods. The better method also obtained higher scores; • PERSON can correctly expose the positive effect of considering topics in retrieval in the method of [72].
Exp. IV	• PERSON can expose the positive effect of personalization in the Campos PIR methods; • PERSON's results were also positively correlated with those of the original paper of Campos et al. [29].
Exp. V	• PERSON's results for different Campos methods were highly correlated with those of ASPIRE in our dataset.
Exp. VI	• PERSON's results for ranking PIR methods for a fixed profile parameter configuration were positively correlated with those of ASPIRE; • In the task of finding the best PIR method for a fixed profile parameter configuration, the top ASPIRE result was ranked close to the top in PERSON, on average.
Exp. VII	• PERSON's results for ranking different profile configurations for a fixed PIR method were positively correlated with those of ASPIRE; • In the task of selecting the best profile configuration for a PIR method, the top ASPIRE result was ranked rather close to the top in PERSON, on average.
Exp. VIII Part I	• PERSON can capture the positive impact of personalization in a SN-based PIR method (Social-Textual); • Using the co-authorship network as a source of information for personalization is rewarding and a reasonable choice.
Exp. VIII Part II	• PERSON's results are far from random; • The documents considered as relevant by PERSON which are retrieved by one of the IR methods are expected to be truly relevant.
Exp. VIII Part III	• The generated query is a vital source of information for retrieving references of the query paper (i.e., relevant documents in PERSON); • Neither profile similarity nor social similarity are sufficient for a good retrieval performance, when measured by PERSON. Indeed, we need a real PIR method, combining query similarity and the searcher's preferences, to obtain remarkable results. This is an important characteristic of a valid PIR EF, which discriminates it from a recommender system evaluation framework.
Exp. IX	• The presence of noise in PERSON's relevance judgments is not a major problem since the results are averaged over many queries.
Exp. X	• PERSON is a reliable and robust EF. For large enough values of parameters, changing the parameters does not impact the evaluation rankings substantially; • After a sufficient number of queries, the absolute values of the evaluation measures are almost unaffected by increasing the number of queries.

evaluation. Our experiments also indicated that PERSON can recognize the superiority of personalized retrieval methods over simple retrieval systems. In addition, we conducted several experiments on the performance of PERSON in comparing different PIR methods and demonstrated its validity. We also studied the robustness of PERSON w.r.t

its parameters and showed that it is highly robust. Overall, our extensive experiments demonstrated that PERSON is a fair and valid way of evaluating PIR methods.

It is important to note that PERSON is a complement to, and not a replacement for, direct evaluation. Although PERSON is greatly informative about the performance of PIR systems, it is certainly important to study real users. However, when a user study is not possible due to the lack of time or resources, or when the PIR methods change frequently (e.g., in the research and development phase), PERSON would be an affordable and reliable alternative. Moreover, even when studying real users is feasible, PERSON can be used to lessen the number of PIR methods users should evaluate, e.g., through tuning the parameters by PERSON. This can make user studies less costly and more worthwhile.

Because of the novelty of the idea of PERSON, it has great potential for extension. Some possible directions for future work are:

1. Our framework and its idea may also be used in other datasets having links among their documents and with known documents' authors and is not limited to scientific papers datasets. For example, if a dataset of webpages contains the authors of the pages (e.g., a blogs dataset), it can be used in PERSON. This can be done by considering the authors as the searchers and selecting an appropriate query extraction scheme such as anchor-text-based scheme. The hyperlinks can then be considered as personalized relevance judgments. Community question answering websites are also another candidate for being used in PERSON (See [59]). However, the validity of the results in datasets other than scientific papers datasets needs to be thoroughly analyzed, and is an interesting direction for future work.
2. As described in Section 3.1.2, the title-based scheme for query extraction cannot be used for evaluating very short queries (one or two terms). Modifying this scheme to support very short queries (e.g., by selecting only the most important terms) is valuable future work.
3. In this paper, we only studied the title-based scheme for query extraction in our experiments. Studying the other suggested schemes or proposing new ones is an interesting research direction. Especially, the anchor-text-based scheme is of particular interest since it is expected that it extracts queries that are more relevant to the references.
4. In our experiments, we used the modified abstract-based representation for representing papers. Studying the content-based representation with datasets in which full texts of papers are available (e.g., PMC Open Access Subset[14]) can be invaluable future work, which per se can help in evaluating PIR methods on long texts.
5. In our experiments, we compared PERSON's results for several PIR methods with those of [29], that are based on human judgments. We also compared PERSON's results with those of ASPIRE which is shown to be consistent with human judgments. However, conducting a user study and getting users' feedback can shed more light on the characteristics of PERSON's evaluations.
6. Author names appear in different forms in papers, e.g., "John Smith" and "J. Smith". In our experiments, we did not disambiguate author names. Using datasets with disambiguated names or employing author name disambiguation methods (See [75,

[14] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Table A.12: Dataset statistics before pre-processing.

Total number of papers (documents)	1,397,240
Total number of authors	1,073,322
Total number of co-authors	2,627,939
Total number of citations	3,021,489

76] for more information) can result in more authentic user profiles, and thus in more accurate results.

7. Currently PERSON cannot be used to evaluate PIR methods that personalize the results based on the search history of the users. One attractive direction for future work is to use search history simulation (See Section 3.1.4) to make the evaluation of such methods possible.

A. Reproducibility Details

In this appendix, we explain the reproducibility details of our experiments. We first describe some details of compiling the dataset and the set of data cleaning steps performed. Then, we describe some general considerations on implementing PERSON. Finally, we explain the experiment-specific details.

A.1. Dataset

AMiner's citation network V2 dataset was not directly usable in our task due to the large number of missing values. For example, the abstract or the references information of many papers were not available. Thus, we performed several data cleaning steps before conducting our experiments. The statistics of the dataset before pre-processing are depicted in Table A.12 and the statistics of the cleaned dataset are reported in Table A.13. More details on the cleaned dataset are illustrated in Figure A.4. The major pre-processing steps taken are:

1. stripping diacritics from titles and abstracts. E.g., 'à' was replaced by 'a';
2. removing some typical noises from the beginning of abstracts (e.g., "abstract:" and "abstract—");
3. removing papers whose titles start by "LIST OF," "INCLUDING ANY IMPLIED," or "Proceedings of the .* Winter Simulation Conference" regular expressions;
4. removing the papers whose titles contained less than or equal to two terms, their abstracts contained less than or equal to 25 terms (excluding stop words), or had no authors. Papers for which sums of the lengths of their titles and abstracts were less than or equal to 200 characters were also removed;
5. merging duplicate papers: papers whose titles, after removing spaces, dashes, dots, apostrophes, and commas and being converted to lower-case, were equal were considered duplicates. In merging, we considered the titles, abstracts, authors, and references separately. The longer version of these contents (or the one having more entries in the authors and the references) was selected for the merged paper. If the contents of a merged paper were not solely from one initial paper, it was marked as *merged*. Merged papers were not used as query papers;
6. constructing the co-authorship graph and removing the papers and authors not in the giant component of the graph.

Table A.13: Cleaned dataset statistics.

Total number of papers (documents)	616,889
Average title length	9.6 ± 3.4
Average abstract length	143.8 ± 81.7
Total number of citations	1,426,867
Total number of authors	558,898
Average number of authors	2.9 ± 1.6
Total number of co-authorships	2,438,267
Average number of references*	5.5 ± 5.3

* Only references to the papers within the dataset; averaged over papers with at least one such reference.

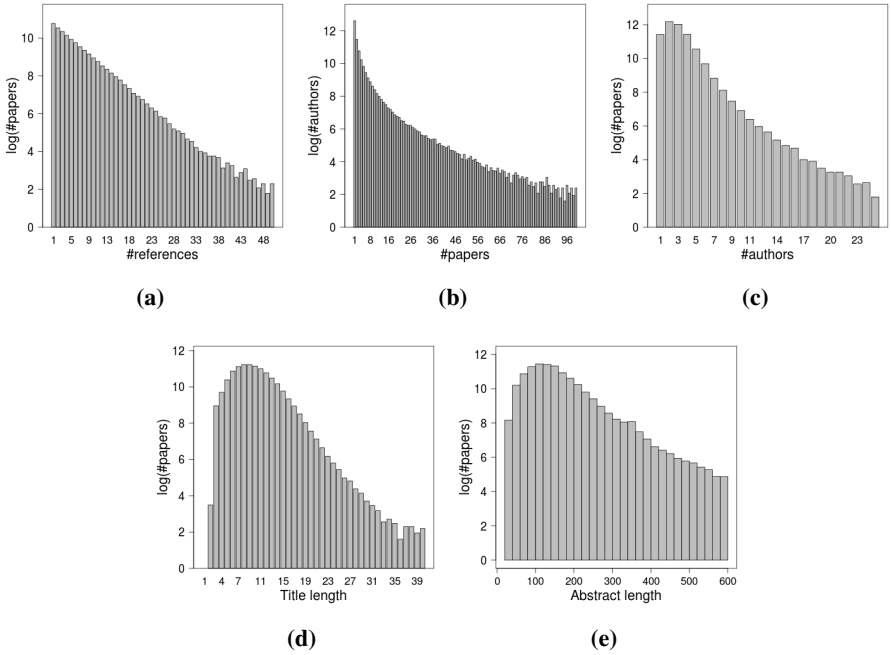


Figure A.4: Dataset statistics after removing stop words (not including outliers). (a) Logarithm of the number of papers with x references (Only references to the papers within the dataset). (b) Logarithm of the number of authors having x papers. (c) Logarithm of the number of papers with x authors. (d) Logarithm of the number of papers with titles of length x. (e) Logarithm of the number of papers with abstracts of length x (binned).

A.2. PERSON Implementation

In our implementation of PERSON, we used the KStem stemmer [77], which is known ([78]) to be less aggressive than the Porter stemmer [79]. We also used the list of English stop words available in MALLET [80] v2.0.7.

In query extraction, we did not consider papers having less than or equal to five

Table A.14: The collections statistics for Experiments II and IX.

ID	Collection	Queries (title only)	Avg. Doc. Length	#qrels
AP	Associated Press 88-89	TREC 1-3 Ad-Hoc Track, topics 51-200	287	15838
CLEF356	Los Angles Times 94 & Glasgow Herald 95	CLEF Ad-Hoc Track 2003,5,6, topics 141-200 & 251-350	313	4327

references[15]. Furthermore, we removed the special terms and symbols that have some meaning in Lucene's query parser (e.g., '*&&*' and '*AND*') from the extracted queries. As mentioned above, we also did not consider the papers marked as merged in the query extraction (See Appendix A.1). In addition, we removed the query paper from the retrieval results of PIR methods, if any of them retrieved it, before calculating the performance metrics since a paper cannot cite itself. Moreover, we used only the years of the publications for publication-date-based filtering.

A.3. Experiment I

In Experiment I, we had several options for implementing LogTF and LogIDF weighting functions. We used the following ones in it:

$$LogTF = \log(RawTF + 1), \tag{A.1}$$

$$LogIDF = \log(\frac{\#\ of\ docs + 1}{RawDF + 1}) + 1 \tag{A.2}$$

Also, the default document length normalization formula of Lucene is used. This experiment resulted in 1666 appropriate searches.

A.4. Experiments II and IX

In these experiment, we used Lemur Toolkit 4.12 for retrieval and trec_eval 9.0 for evaluation. In addition, "-M 100" option is used in trec_eval to obtain more comparable results. The statistics of the datasets used in these experiments are illustrated in Table A.14.

It's important to mention that in Experiment IX, a modified version of MAP is used. To explain this, let us take a look at the MAP formula, Eq. (A.3), in which N is the number of queries, Q_j is the number of relevant documents for query j, and $P(doc_i)$ is precision at i-th relevant document. The modified MAP differs from MAP in that for each query it considers the number of relevants that are actually retrieved by at least one of the IR methods as Q_j instead of the total number of relevants. The point is that when the judgments are noisy (i.e., some relevant-considered documents may be irrelevant and vice versa), the values of Q_js may not be accurate. Since Q_js indeed act as weights of $P(doc_i)$, changes in them can change the relative importance of finding a relevant document for a query in comparison to other queries' relevants. Now, consider a situation in which we are comparing two PIR methods, A and B, for which $\text{MAP}_A > \text{MAP}_B$ according to true judgments. In a noisy setting, judgments differ from the true judgments and thus the relative importance of finding each relevant document is

[15]Usually queries with less than or equal to five relevant documents are not considered in computing the MAP value [81].

changed. The point is that even if there is no difference between the performances of the two methods in the two settings, i.e. all of the retrieved relevants are marked as relevant in both judgments and none of the documents incorrectly marked as relevant in the noisy judgments are retrieved, we may obtain MAP values such that $\text{MAP}_A < \text{MAP}_B$. For example, suppose a query having only one real relevant document. If A retrieves it at the top of the results list while B does not retrieve it at all, this increases MAP of A considerably but decreases that of B. Now consider a noisy setting in which still the only retrieved relevant is the one previously retrieved by A and none of the false relatives are retrieved, but Q_j for that query is now ten instead of one. This makes the MAP measure, substantially underestimate the relative superiority of A in that query. Therefore, MAP_A may be lower that MAP_B when averaged over all queries, although both methods retrieved exactly the same relevant documents in both settings. In fact, our results showed that the changes in the MAP values where mostly determined by the changes in the number of relevant documents, instead of exactly which documents were misjudged. In other words, repeating the same experiment with the same number of misjudged documents but with different misjudged document resulted in rather the same outcomes. This first caused us to report a modified version of MAP in Experiment VII. In the modified MAP, relevant documents not retrieved by any of the IR methods being compared are considered irrelevant. Second, it validates our inappropriate relevants heuristic (Section 3.1.3) for considering the documents not retrieved by any of the PIR methods as irrelevant in calculating the measures in PERSON. Using the modified MAP can also make the way metrics are calculated in this experiment more similar to that of PERSON, and thus make our conclusions more generalizable to PERSON.

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(doc_i) \tag{A.3}$$

A.5. Experiment III

In this experiment, in both the pseudo-relevance feedback methods [70, 71], we considered $\beta = .15$, *the number of added terms* = 25, $c = 4.0$, and *the number of feedback documents* = 10.

To extract the topics for LBDM, we used MALLET [80] v2.0.7. We imported the data using "–keep-sequence –remove-stopwords" options and "–num-topics 100 –optimize-interval 10 –optimize-burn-in 20 –use-symmetric-alpha false" were used for training the topics. We used the learned topics in Equation 8 of [72] to implement LBDM.

This experiment resulted in 1527 appropriate searches.

A.6. Experiment IV

There were several details in implementing the Campos methods and using them:

1. In generating the profiles, only terms with a DF greater than 100 were considered;
2. Campos et al. did not make the IDF and TF weighting functions used explicit. We used LogIDF (as defined in Eq. (A.4)) and RawTF;

$$LogIDF = \log(\frac{\#\ of\ docs}{RawDF + 1}) + 1 \tag{A.4}$$

3. We did not use the query paper in the generation of the profile of the user. Otherwise, we would have unfairly provided extra information about the query to the Campos methods. Consequently, in the experiments including at least one of the Campos methods (Experiments IV-VIII), we considered a paper as a query paper only if its first author had at least one paper other than it in PERSON;
4. We did not use the papers written by the searcher that were relevant to the search being performed (i.e. are cited by the query paper) in the profile generation. Thus, in Experiments IV-VIII, if the author did not have any other papers than these papers and the query paper, we marked the query as inappropriate in PERSON.
   This experiment resulted in 1783 appropriate searches.

A.7. Experiments V-VII

For PERSON, the details of the Campos methods in these experiments is similar to those of Experiment IV, while for ASPIRE the points numbered with 3 and 4 are not considered. Also, publication-date-based filtering is not employed in ASPIRE because it does not require that. Moreover, for both PERSON and ASPIRE, we used TF-IDF weighting in both ordering the profile terms and weighting them. Also, similar to [29], we used $k = 50$ in NDCG@k in these experiments. For ASPIRE, similar to the original paper [6], we used $topkRel = 2 \times topkEval$, in which *topkEval* is the number of retrieved documents considered in the evaluations.

In addition, to have more meaningful categories, we filtered out some general topics. More precisely, since we used an asymmetric Dirichlet prior over document-topic distributions to obtain better topics [82], some of the learned topics mainly represent general words and not coherent topics. To find these topics, we sorted the learned topics according to their α values and identified that the top 9 topics with the highest α values are general topics (General topics tend to have high α values since they appear in lots of documents). For example, the top words in one of the general topics were "method base propose result paper analysis improve show approach experiment". We ignored the general topics in these experiments (i.e., each document was assigned to its non-general topic with the highest probability and also in the query extraction we did not consider a document whose most probable topic was a general topic as a query).

A.8. Experiment VIII

We used the implementation of Social-Textual that the authors kindly provided to us. The implementation only considers users who have a limited distance from the searching user. We used two as the limit (i.e. only the searcher's friends and friends of friends were considered).

In addition, we used a user action function (*uaf*) that is inversely related to the number of authors of a paper:

$$uaf(u_i, o_k) = \begin{cases} \left(\frac{1}{\#\text{ of authors of } o_k}\right)^{½}, & \text{if } u_i \text{ is an author of } o_k \\ 0, & \text{otherwise.} \end{cases} \tag{A.5}$$

Also, user weight function (*uwf*) was defined as Eq. (A.6), similar to the original implementation.

$$uwf(u_i) = \log\left(1 + \min\left(\frac{\#\text{ of co-authors of } u_i}{100}, 1\right)\right). \tag{A.6}$$

Also, note that Social-Textual requires a basis retrieval method. We used the baseline method (LM) as the basis method.

This experiment resulted in 1783 appropriate searches.

A.9. Experiment X

In this experiment, to make the results look better, we did not draw the points for all x values on the x-axis in the plots. We drew every n-th point, in which n may be different for each figure. Also, in Experiment X Part I, we compared the ranking of methods obtained in PERSON after each $\#Q$ with the ranking at $\#Q = 1470$. This number is the minimum number of appropriate searches across different PIR methods.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive comments. We would also like to thank Ali Montazeralghaem for his helps in implementing some of the methods used in our experiments. This research was in part supported by a grant from the School of Computer Science, Institute for Research in Fundamental Sciences (IPM), and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] E. M. Voorhees, The philosophy of information retrieval evaluation, in: Proceedings of the 2002 Cross Language Evaluation Forum, 2002, pp. 355–370.

[2] E. Santos, J. Q. Zhao, H. Nguyen, H. Wang, Impacts of user modeling on personalization of information retrieval: An evaluation with human intelligence analysis, in: Proceedings of the Fourth Workshop on Evaluation of Adaptive Systems, 2005, pp. 27–36.

[3] H. Xie, X. Li, T. Wang, R. Y. Lau, T.-L. Wong, L. Chen, F. L. Wang, Q. Li, Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy, Information Processing & Management 52 (1) (2016) 61–72.

[4] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, SoPRa: A new social personalized ranking function for improving web search, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013, pp. 861–864.

[5] L. Tamine-Lechani, M. Boughanem, M. Daoud, Evaluation of contextual information retrieval effectiveness: Overview of issues and research, Knowledge and Information Systems 24 (1) (2010) 1–34.

[6] E. Vicente-López, L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, A. Tagua-Jiménez, C. Tur-Vigil, An automatic methodology to evaluate personalized information retrieval systems, User Modeling and User-Adapted Interaction 25 (1) (2015) 1–37.

[7] A. Sieg, B. Mobasher, R. Burke, Web search personalization with ontological user profiles, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, 2007, pp. 525–534.

[8] L. Gou, H.-H. Chen, J.-H. Kim, X. L. Zhang, C. L. Giles, SNDocRank: A social network-based video search ranking framework, in: Proceedings of the International Conference on Multimedia Information Retrieval, 2010, pp. 367–376.

[9] L. Gou, H.-H. Chen, J.-H. Kim, X. L. Zhang, C. L. Giles, SNDocRank: Document ranking based on social networks, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 1103–1104.

[10] D. Petrelli, On the role of user-centred evaluation in the advancement of interactive information retrieval, Information Processing & Management 44 (1) (2008) 22–38.

[11] R. W. White, I. Ruthven, J. M. Jose, C. J. V. Rijsbergen, Evaluating implicit feedback models using searcher simulations, ACM Transactions on Information Systems 23 (3) (2005) 325–361.

[12] A. Khodaei, C. Shahabi, Social-textual search and ranking, in: Proceedings of the First International Workshop on Crowdsourcing Web search, 2012, pp. 3–8.

[13] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, A. Vakali, Using social annotations to enhance document representation for personalized search, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013, pp. 1049–1052.

[14] D. Vallet, I. Cantador, J. M. Jose, Personalizing web search with folksonomy-based user and document profiles, in: Proceedings of the 32nd European Conference on Information Retrieval, 2010, pp. 420–431.

[15] Q. Wang, H. Jin, Exploring online social activities for adaptive search personalization, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 999–1008.

[16] S. Xu, S. Bao, B. Fei, Z. Su, Y. Yu, Exploring folksonomy for personalized search, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 155–162.

[17] D. Zhou, S. Lawless, V. Wade, Improving search via personalized query expansion using social media, Information Retrieval 15 (3) (2012) 218–242.

[18] S. Chernov, P. Serdyukov, P.-A. Chirita, G. Demartini, W. Nejdl, Building a desktop search test-bed, in: Proceedings of the 29th European Conference on Information Retrieval, 2007, pp. 686–690.

[19] S. Chernov, E. Minack, P. Serdyukov, Converting desktop into a personal activity dataset, in: Proceedings of 9th Russian National Research Conference on Digital Libraries, 2007, pp. 280–283.

[20] Sergey, G. Demartini, E. Herder, M. Kopycki, W. Nejdl, Evaluating personal information management using an activity logs enriched desktop dataset, in: In Proceedings of the 3rd Personal Information Management Workshop, 2008.

[21] D. Elsweiler, D. E. Losada, J. C. Toucedo, R. T. Fernandez, Seeding simulated queries with user-study data for personal search evaluation, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 25–34.

[22] J. Kim, W. B. Croft, Retrieval experiments using pseudo-desktop collections, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 1297–1306.

[23] M. R. Ghorab, D. Zhou, A. O'connor, V. Wade, Personalised information retrieval: Survey and classification, User Modeling and User-Adapted Interaction 23 (4) (2013) 381–443.

[24] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, Social networks and information retrieval, how are they converging? a survey, a taxonomy and an analysis of social information retrieval approaches and platforms, Information Systems 56 (2016) 1–18.

[25] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 990–998.

[26] H. Fang, T. Tao, C. Zhai, A formal study of information retrieval heuristics, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 49–56.

[27] N. Matthijs, F. Radlinski, Personalizing web search using long term browsing history, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011, pp. 25–34.

[28] J. Teevan, S. T. Dumais, E. Horvitz, Personalizing search via automated analysis of interests and activities, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 449–456.

[29] L. M. de Campos, J. M. Fernandez-Luna, J. F. Huete, E. Vicente-Lopez, Using personalization to improve XML retrieval, IEEE Transactions on Knowledge and Data Engineering 26 (5) (2014) 1280–1292.

[30] A. Younus, C. O'Riordan, G. Pasi, A language modeling approach to personalized search based on users' microblog behavior, in: Proceedings of the 36th European Conference on Information Retrieval, 2014, pp. 727–732.

[31] X. Shen, B. Tan, C. Zhai, Implicit user modeling for personalized search, in: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005, pp. 824–831.

[32] D. Vallet, P. Castells, M. Fernandez, P. Mylonas, Y. Avrithis, Personalized content retrieval in context using ontological knowledge, IEEE Transactions on Circuits and Systems for Video Technology 17 (3) (2007) 336–346.

[33] D. Ganguly, J. Leveling, G. J. F. Jones, Towards evaluation of personalized and collaborative information retrieval, in: The First Workshop on Personalised Multilingual Hypertext Retrieval (PMHR 2011), 2011, pp. 42–49.

[34] O. Chapelle, D. Metlzer, Y. Zhang, P. Grinspan, Expected reciprocal rank for graded relevance, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 621–630.

[35] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20 (4) (2002) 422–446.

[36] J. Zobel, How reliable are the results of large-scale information retrieval experiments?, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 307–314.

[37] D. E. Losada, J. Parapar, A. Barreiro, Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems, Information Processing & Management 53 (5) (2017) 1005–1025.

[38] M. Kutlu, T. Elsayed, M. Lease, Intelligent topic selection for low-cost information retrieval evaluation: A new perspective on deep vs. shallow judging, Information Processing & Management 54 (1) (2018) 37–59.

[39] A. Ritchie, S. Teufel, S. Robertson, Creating a test collection for citation-based IR experiments, in: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp. 391–398.

[40] A. Ritchie, S. Robertson, S. Teufel, Creating a test collection: Relevance judgements of cited & non-cited papers, in: Large Scale Semantic Access to Content (Text, Image, Video, and Sound), 2007, pp. 251–261.

[41] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, A. Vakali, Persador: Personalized social document representation for improving web search, Information Sciences 369 (2016) 614–633.

[42] P. Thomas, D. Hawking, Evaluation by comparing result sets in context, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, 2006, pp. 94–101.

[43] Z. Dou, R. Song, J.-R. Wen, A large-scale evaluation and analysis of personalized search strategies, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 581–590.

[44] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, H. Wang, Enhancing personalized search by mining and modeling task behavior, in: Proceedings of the 22Nd International Conference on World Wide Web, 2013, pp. 1411–1420.

[45] P. N. Bennett, F. Radlinski, R. W. White, E. Yilmaz, Inferring and using address metadata to personalize web search, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 135–144.

[46] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, X. Cui, Modeling the impact of short- and long-term behavior on search personalization, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 185–194.

[47] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, B. Billerbeck, Probabilistic models for personalizing web search, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, 2012, pp. 433–442.

[48] T. Joachims, Evaluating retrieval performance using clickthrough data, in: Text Mining, Physica/Springer Verlag, 2003, pp. 79–96.

[49] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, M. de Rijke, Multileaved comparisons for fast online evaluation, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014, pp. 71–80.

[50] A. Schuth, R.-J. Bruintjes, F. Buüttner, J. van Doorn, C. Groenland, H. Oosterhuis, C.-N. Tran, B. Veeling, J. van der Velde, R. Wechsler, D. Woudenberg, M. de Rijke, Probabilistic multileave for online retrieval evaluation, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, pp. 955–958.

[51] K. Hofmann, S. Whiteson, M. de Rijke, A probabilistic method for inferring preferences from clicks, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 249–258.

[52] K. Hofmann, S. Whiteson, M. de Rijke, Estimating interleaved comparison outcomes from historical click data, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 1779–1783.

[53] F. Radlinski, M. Kurup, T. Joachims, How does clickthrough data reflect retrieval quality?, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008, pp. 43–52.

[54] M. R. Bouadjenek, A. Bennamane, H. Hacid, M. Bouzeghoub, Evaluation of personalized social ranking functions of information retrieval, in: Proceedings of the 13th International Conference on Web Engineering, 2013, pp. 283–290.

[55] M. Gupta, M. Bendersky, Information retrieval with verbose queries, Foundations and Trends in Information Retrieval 9 (3-4) (2015) 209–354.

[56] R. Mihalcea, P. Tarau, TextRank: Bringing order into texts, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404–411.

[57] V. Dang, B. W. Croft, Query reformulation using anchor text, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, 2010, pp. 41–50.

[58] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1-7) (1998) 107–117.

[59] C.-J. Lee, W. B. Croft, Building a web test collection using social media, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013, pp. 757–760.

[60] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, J. Allan, Evaluation over thousands of queries, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 651–658.

[61] H. Deng, J. Han, M. R. Lyu, I. King, Modeling and exploiting heterogeneous bibliographic networks for expertise ranking, in: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, 2012, pp. 71–80.

[62] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter, Multilayer networks, Journal of Complex Networks 2 (3) (2014) 203–271.

[63] C. Bouveyron, P. Latouche, Z. Rawya, The stochastic topic block model for the clustering of vertices in networks with textual edges, Statistics and Computing (2016) 1–21.

[64] J. Tan, Discusses of user interest model in personalized search, International Journal of Advancements in Computing Technology 5 (1) (2013) 619–626.

[65] M. Speretta, S. Gauch, Personalized search based on user search histories, in: The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005, pp. 622–628.

[66] C. Carpineto, G. Romano, A survey of automatic query expansion in information retrieval, ACM Computing Surveys 44 (1) (2012) 1–50.

[67] M. G. Kendall, Rank Correlation Methods, 4th Edition, Griffin, 1970.

[68] J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 275–281.

[69] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to information retrieval, ACM Transactions on Information Systems 22 (2) (2004) 179–214.

[70] S. Clinchant, E. Gaussier, Information-based models for ad hoc IR, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 234–241.

[71] A. Montazeralghaem, H. Zamani, A. Shakery, Axiomatic analysis for improving the log-logistic feedback model, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016, pp. 765–768.

[72] X. Wei, W. B. Croft, LDA-based document models for ad-hoc retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp. 178–185.

[73] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[74] C. C. Aggarwal, C. Zhai, A Survey of Text Clustering Algorithms, Springer, 2012, pp. 77–128.

[75] A. A. Ferreira, M. A. Gonçalves, A. H. Laender, A brief survey of automatic methods for author name disambiguation, SIGMOD Record 41 (2) (2012) 15–26.

[76] S. Khan, X. Liu, K. A. Shakil, M. Alam, A survey on scholarly data: From big data perspective, Information Processing & Management 53 (4) (2017) 923–944.

[77] R. Krovetz, Viewing morphology as an inference process, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993, pp. 191–202.

[78] J. Xu, W. B. Croft, Corpus-based stemming using cooccurrence of word variants, ACM Transactions on Information Systems 16 (1) (1998) 61–81.

[79] M. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[80] A. K. McCallum, Mallet: A machine learning for language toolkit, http://mallet.cs.umass.edu (2002).

[81] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[82] H. M. Wallach, D. Mimno, A. McCallum, Rethinking LDA: Why priors matter, in: Proceedings of the 22Nd International Conference on Neural Information Processing Systems, 2009, pp. 1973–1981.