**User Modeling and User-Adapted Interaction manuscript No.**
(will be inserted by the editor)

# An Automatic Methodology to Evaluate Personalized Information Retrieval Systems

**E. Vicente-López · L.M. de Campos · J.M. Fernández-Luna · J.F. Huete · A. Tagua-Jiménez · C. Tur-Vigil**

Received: Jun 26, 2013 / Accepted: date

**Abstract** Due to the information overload we are faced with nowadays, personalization services are becoming almost essential, in order to find relevant information tailored to each individual or group of people with common interests. Therefore, it is very important to be able to build efficient and robust personalization techniques to be part of these services. The evaluation step is a crucial stage in their development and improvement, so much more research is needed to overcome this issue. We have proposed an automatic evaluation methodology for personalized information retrieval systems (ASPIRE), which joins the advantages of both system-centred (repeatable, comparable and generalizable results) and user-centred (considers the user) evaluation approaches, and makes the evaluation process easy and fast. Its reliability and robustness have been assessed by means of a user-oriented evaluation. ASPIRE may be considered as an interesting alternative to the costly and difficult user studies, able to discriminate between either different personalization techniques or different parameter configurations of a given personalization method.

**Keywords** Information Retrieval · Personalization · Evaluation Framework.

"This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is free of plagiarism or self-plagiarism as defined in Springer's Policy on Publishing Integrity"

E. Vicente-López · L.M. de Campos · J.M. Fernández-Luna · J.F. Huete
Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S.I. Informática y de Telecomunicación, CITIC-UGR, Universidad de Granada, 18071-Granada, Spain
E-mail: {evicente,lci,jmfluna,jhg}@decsai.ugr.es

A. Tagua-Jiménez · C. Tur-Vigil
Servicio de Publicaciones Oficiales, Parlamento de Andalucía, 41009-Sevilla, Spain.

## 1 Introduction

Nowadays we are immersed in a world where the possession and the accurate management of information is crucial. With different implications, this affirmation is true for all levels of our society ranging from governments, companies or organizations to every single citizen. We currently live in the so-called 'knowledge society'.

Digital information allows easy and efficient storage, access or modification processes over it. Thus, almost every existent non-digital information source, as might be music, video, books, document collections, etc., are being progressively digitized, and of course, all new created information is digital. This fact is leading to an exponential increase of digital information in recent years, especially owing to Internet.

In order to be able to find relevant results within this huge amount of information, the use of Information Retrieval Systems (IRS) has become a must. This kind of systems retrieves results based only on the information contained in the user issued query, which almost always is composed by a small set of keywords in natural language. This strategy has achieved good retrieval results in the last years. However, the IRS returns the same results for the same query, independently of the user. But considering that user queries are usually short and ambiguous [44], together with the previously cited exponential increase of information, a new approach is required. An approach in which the user itself, and not only the issued query, is considered as an important part within the retrieval process. Personalization [24,26,45] is this new approach, and hot arising research area, whose main objective is to retrieve results closer to the user in order to better cover his/her specific information needs. The introduction of personalization into the retrieval process allows some potential advantages, such as the disambiguation of short queries, the increase of top results precision, the improvement of domain specific retrieval tasks, and even, the inclusion of social behaviours, as it happens with recommender systems.

Any personalized system faces up to three main different stages: 1) starting on how to acquire and represent the information about the user context, which is usually stored as the user profile, 2) continuing on how to exploit this user profile information in order to retrieve the most relevant results, which satisfy the user information needs, and finally 3) how to best present the previous retrieved results to minimize the user effort and maximize the user satisfaction with the system.

Concretely, this article focuses on the evaluation of the second personalization stage. Every system needs to be evaluated in order to measure how good the given system does the task it was designed for. The evaluation of personalized IRS is very difficult due to their complexity, since usually there is an underlying implemented personalization method, which tends to have many configuration parameters to be adjusted, and subjective components as relevance assessments come into play, between other issues. For these reasons, and the involved potential costs of evaluation, most personalized IRS do not con-

duct real world experiments for their validation [53]. However, the evaluation step is crucial in the development of any personalized IRS.

To evaluate traditional system-centred IRS, where the user is not an integral part of the retrieval process, an evaluation framework based on the Cranfield paradigm [19] is normally used. This evaluation methodology aims to ensure repeatable and controlled experiments between different IRS, extracting comparable measures and generalizable conclusions about them. System-centred evaluation methodologies have contributed to have a very good performance of general IRS nowadays. But looking to this approach from a more practical point of view, it is actually very limited, because it does not consider anything about the IRS final real users. The IRS under this approach are not able to adapt their results to their users, whose activities are complex and subjective by nature [6]. Pursuing the inclusion of the user context into the evaluation process, several user-centred evaluation frameworks have been developed. They can be classified into three main categories [48]: extensions to the laboratory-based Cranfield paradigm, contextual simulations and the obvious user studies. The word context is defined in [1] as: "Any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or physical or computational object". Under our methodology we always refer to context as the user interests and preferences. In other parts of the article, such as in Section 2, context may refers to user interests and preferences but also to time, location, used device, the user background knowledge, etc.

Actually, as it is stated in [22], it is not only about developing user-centred IRS and the corresponding evaluation frameworks, forgetting the system-centred side, but to combine both evaluation perspectives: get the best of the system-centred approximation, adapting it in order to also get the best final user satisfaction in their daily experience with the system.

With the previous intention in mind, we have developed an automatic evaluation methodology to evaluate the retrieval effectiveness of any personalized IRS. This methodology will be denoted as ASPIRE, acronym of "Automatic Strategy for Personalized Information Retrieval systems' Evaluation". ASPIRE combines the repeatable, comparable and generalizable main advantages of system-centred approaches, together with the inclusion of the user context into the evaluation of the retrieval process, which is the main benefit of user-centred approaches. At the same time, ASPIRE avoids the interaction with real users, which is the cause of the user studies evaluation lack of control, not repeatable, not comparable and not generalizable results. ASPIRE also allows to avoid the difficulty and big associated costs of congregating several users, in some cases experts, for a long time to perform the evaluation process, including questionnaires, interviews, etc. The combination of the system-centred and user-centred evaluation frameworks advantages allows a fast and easy testing of a personalized IRS.

Thus, given two or more personalization approaches, one of the main goals of our methodology is to reach a ranking of them that is close to the one that would have been obtained by the same methods using real user interactions.

In a similar way, another ASPIRE goal is to be used to set up the best configuration parameters for a given personalization technique. To accomplish both goals, the use of traditional evaluation frameworks is very difficult, or most of the times impossible, mainly due to the required user effort, in terms of time spent making personalized judgements over a set of topics (the assessments vary with the users). However, these judgements represent a key component to obtain the desired ranking. The use of our proposed methodology turns this process into an easy, low effort and low-cost task.

As any other system or methodology, ASPIRE will be also evaluated in order to check its validity for the evaluation of personalized IRS. ASPIRE mainly tries to be an alternative to user studies which, if performed in a controlled way, are the best real user-centred evaluation methodology. Therefore, we compare its results against the results obtained from a real user study.

It should be clear that ASPIRE does not pretend to completely replace user studies, rather it should be considered as an alternative to them. Although ASPIRE joins the advantages from both system-centred and user-centred evaluation approaches, it is of paramount importance to collect qualitative information about the IRS from real users [52]. ASPIRE pretends to be an alternative to user studies for the evaluation of personalized IRS, specially in their first development stages or when the user study is not feasible due to any circumstances, such as the lack of resources or time. ASPIRE also helps to make final user studies experimentation more worthwhile, by limiting the number of personalized IRS configurations that users must evaluate.

The remainder of the article is structured as follows. Section 2 gives an overview of the different IR evaluation strategies existing in the literature. Section 3 describes our proposed automatic evaluation methodology, ASPIRE. Section 4 shows its use and validation through the definition of some metrics and Section 5 shows the comparison of ASPIRE results with the ones obtained from a user study and with a state-of-the-art approach, using several retrieval models. Finally, Section 6 finishes with some general conclusions and proposals for future work.

## 2 Information Retrieval Systems Evaluation

Evaluation is the process that measures how good any system does the task it was designed for. Traditional IRS have an underlying retrieval model which tries to return a ranked set of relevant results based on a given query. To do that task, the retrieval model needs a representation of the documents and the query itself. The retrieval model tries to match both representations to select and rank those documents which better satisfy the user information needs, represented by the issued query. These systems are considered as system-centred approaches. However, personalized IRS are those in which, besides to the issued query, additional information about the user context is considered in the retrieval process. These IRS are considered as user-centred approaches.

We next make a summary of the evaluation characteristics of system-centred and user-centred approaches, and a brief literature review, focusing on the latter ones, which is the approach this article is focused on.

### 2.1 Evaluation approaches for system-centred IRS

Historically, to evaluate system-centred IRS, an evaluation methodology based on the Cranfield paradigm [19] is used. The evaluation framework consists of a test collection composed by a document collection, a set of well-defined queries and a set of manually assigned relevance assessments (the relevant documents associated to each query). The main evaluation metrics of this paradigm are the well-known precision and recall, and some others generally based on them. The manually assigned relevance assessments are used by those metrics, in order to get useful evaluation values to compare the retrieval effectiveness of the different models or systems.

The Cranfield laboratory-based evaluation framework has been used for many years in the IR area. It has a number of advantages that have allowed the IR continuous development, obtaining better retrieval results each time. Some of these advantages stand out: 1) it allows *repeatable* and *comparable* evaluation experiments. The ability to repeat an experiment is considered as a key characteristic of any empirical study [36]. Systems which are not able to be evaluated under this kind of evaluation frameworks have big problems in their development and improvement. For example, the evaluation results obtained from the inclusion of a possible system improvement will not be comparable with previous results, since both of them will have been obtained under different circumstances, making impossible to discern whether the improvement is real or not. And 2), under the same experimental conditions the findings are *generalizable*. If relevance assessments are large enough, test collections are reusable [54]. This characteristic allows new systems and models to be evaluated with the same documents and queries, applying the same relevance assessments.

Traditional evaluation frameworks designed to evaluate system-centred IRS are not suitable to evaluate the new cognitive side derived from the introduction of the user in the retrieval process. Both, the retrieval and the evaluation, ignore the influence of the previous user's cognitive side in the whole retrieval process [28]. Some examples of the previous affirmation are the following: system-centred IRS assume that users' requests are a good representation of the user, which leads to serious problems on the reliability of real life relevance assessments [50]. Similarly, system-centred IRS assume that user's requests are well defined, while they actually are almost always ambiguous. Therefore, the extracted conclusions are not always truly generalizable [6].

### 2.2 Evaluation approaches for user-centred IRS

As it may be derived from above, an alternative to the Cranfield evaluation laboratory-based model is needed for evaluating user-centred IRS. There are several evaluation frameworks, but not any standard, for user-centred systems. For this reason, we next give an overview of the most important evaluation strategies, being mainly guided by the broad survey done in [48].

Considering the new characteristics within a user-centred evaluation approach, the evaluation process could be separated in two main different steps: the user profile evaluation step and the retrieval process evaluation step.

In the user profile evaluation step, the main objective is to measure the user profile accuracy. User profiles are viewed as the representation of user models. There are two main user profile representations in order to store the information about the user: a set of weighted keywords [46] and rich semantic based structures, sometimes enhanced with the use of ontologies [49]. The question is: in what degree the user profile representation, which models the user, is a reliable portrait of this user? As usual, there are no standard evaluation metrics to answer to the previous question. Some approaches are given in [23,31], where it is also shown how these metrics are weakly dependent on the selected user profile representation.

Although a good representation of the user profile, in terms of having the biggest accuracy representation with respect to the real user tastes and preferences is very important, this is beyond the scope of this article, which is focused on the evaluation of the personalized IRS retrieval effectiveness. Nevertheless, once chosen, we can evaluate how to better use it in order to increase the personalized IRS retrieval effectiveness, therefore selecting the best user profile configuration for this purpose.

There are three main evaluation frameworks for user-centred IRS:

***Extensions to the laboratory-based evaluation***: these extensions were the first attempt to perform a more user-centred evaluation framework. They model a small interaction between the system and the user, including some metadata about the user (e.g. genre or location) and the query (e.g. purpose). With the inclusion of real users, they try to make the evaluation process more realistic and relatively controlled. TREC interactive track [27] and HARD track [2] are examples of this kind of evaluation frameworks, which mainly compare a baseline run ignoring the user/query metadata with another run considering it.

However, these extensions to the laboratory-based evaluation framework are still system controlled, their evaluation metrics are still based on precision and recall, and their capability of capturing the contextual aspects is very limited in all the process, allowing a restricted personalized evaluation.

***Contextual simulations***: they simulate users and user-system interactions through a well defined retrieval scenario. They are also called hypothesis-based evaluation studies in [35], since they call hypothesis to the well defined

retrieval scenarios. A scenario represents a possible user-system interaction, which the retrieval model should consider to provide better results to the given simulated user.

Contextual simulations have been proposed to attenuate the limitations of laboratory-based evaluation extensions. Instead of modeling a minimal interaction, contextual simulations are able to model different user interactions, used retrieval strategies, and external factors, which could influence in the user interaction decisions.

Thanks to the ability of simulating user-system scenarios, contextual simulations are used for two main purposes:

- *The development of IRS measuring the contextual retrieval effectiveness*: in [42], the authors present a personalized search system which builds models of user context as ontological profiles. These user profiles are built assigning user interest scores, derived in an implicit way, to concepts from the ODP ontology. Since the user behaviour changes dynamically, an spreading activation algorithm is used to maintain these interests updated. They show that reranking the search results, based on the ontological based user profile, helps to present the most relevant results to the user.
- *The development of the associated interfaces design*: in [51], the authors develop search interfaces and search scenarios, which interact with different retrieved information such as the title, the summary or a sentence in context, with the objective to test several implicit feedback models.

The following are the most common contextual simulation characteristics: the document collection is usually provided by a controlled IR framework such as TREC, or it is based on a set of online open source web pages. The topics depend on the document collection, and therefore, they may be provided by the controlled framework, or they may be even automatically generated without any user involvement as in [42]. The relevance assessments are given by the controlled framework, or depend on whether the document is classified or belongs to the concept or user interest being simulated. Users are simulated by hypothetical context situations. Usual evaluation metrics are precision and recall at cut-off n, and the standard mean average precision (MAP). Finally, the evaluation is performed without the interaction of real users, and performing a comparison between a run only involving the query and a final run with the query personalized using the context.

Contextual simulations are worthwhile since they are less resources consuming than experiments with real users, and they provide comparable evaluation results within the same retrieval scenarios. However, they also have some disadvantages, such as the topics and collection might not be interesting to many searchers, the effort to define the retrieval scenarios or hypotheses is usually hard, and the relevance assessments are fixed by assessors, which probably have a very different background knowledge with respect to the real users.

***User studies***: they are the best evaluation method from the qualitative point of view, since the IRS effectiveness is directly evaluated by real users in

a real retrieval environment. User studies include all user interactions with the IRS and several ways of feedback, such as questionnaires, interviews, as well as a constant monitoring of the user behaviour with respect to the preassigned search task within the IRS.

User studies are commonly performed by simulating work situations [8, 40], which pursue to involve individual users into a similar environment and information need search tasks, with respect to their daily job. For this reason, the search task must be appropriate for the user related to his/her experience with the given task. Users range from expert users (project team members, technical employees,...) to common users (normal citizens, children,...).

This evaluation approach also has some disadvantages, such as its enormous time and resources requirements, which limit its realization most of the times. In order to perform a user study, real users must be involved. We must think that all the work a user must do in a user study is very hard and time consuming. They normally have to spend several hours performing searches, reading a lot of information, checking the relevant documents found, etc., and when they finish, they usually have to fill one or more questionnaires. There is another inconvenient if the user study involves expert users. They usually are busy people, which not only limit their availability, but also keep them out of their jobs meaning a lose of money for them and/or their companies or institutions. Additionally, sometimes a physical place and computational resources are also required to accommodate the users during the user study.

Performing a user study is very difficult due to all of the previous characteristics, but there is another issue to take into account, the balance between control and realism. The experiments are not repeatable if different users are involved in different user studies, or even if not always the same users are involved in the same tasks, because of the individual differences between them, such as their background knowledge about the task, their intelligence or their familiarity with the search interface, for example. The previous fact makes very difficult to discern the influence of the system evaluated variables over the overall retrieval effectiveness, making the conclusions not generalizable. To try to diminish the previous problems, several recommendations are given in order to ensure a minimal reliability of user study experimental findings in [6, 18]: the more users the better, minimal interactions between users, users must ignore which system is being evaluated, permute order of search tasks between users, run a pilot study before the main study, etc.

The following are the most common user study characteristics: the web conforms the document collection, being consulted through a public search engine. Topics are created, selected from a predefined set, or both, by the users. Relevance assessments are based on the user click-through data or explicitly made by them. Almost anybody could be a user but, most of the times, they have some knowledge about the IRS or the collection. Some evaluation metrics used in user studies are MAP, precision at cut-off n or NDCG. In the user study evaluation protocol, the user interacts with the system performing search tasks in different domains, performs the topics as stated above, and judges relevant documents among the first results list. All this information is

stored in a log file for performing the evaluation. This log file usually consists of the user's queries, retrieved results, clicked results or relevance assessments, and any other important contextual information.

Both system-centred and user-centred IRS evaluation frameworks (especially the user studies in the latter case), are the two poles of the evaluation range. Both of them have merits that could be exploited at different stages of the IRS design [35]. In order to guarantee the main IRS technical objectives, system-centred evaluation is more suitable in their first design stages, allowing to have controlled and repeatable experiments. However, user-centred approaches are more suitable for further stages, since they introduce the search context into the IRS design, allowing the evaluation of the system improvement from the user dynamic perspective. Another claim about the advantages of using both approaches along the different stages of any IRS is done in [22], where the authors affirm that the use of both approaches offers more information about the system real performance than any of them separately.

## 3 ASPIRE: the proposed Automatic Strategy for Personalized Information Retrieval systems Evaluation

As seen in Section 2, there are several evaluation methodologies for personalized IRS, but there is still no agreement about the definition of a standard evaluation framework and the evaluation metrics to be used, since all the previous methodologies have different disadvantages. As the issue of evaluating personalization is significant and the evaluation of personalized systems is a crucial stage in their development, much more research is necessary to overcome this issue.

The proposed automatic methodology for the evaluation of personalized IRS aims to join the advantages of system-centred and user-centred evaluation approaches. In particular, ASPIRE produces repeatable, comparable and generalizable results (main characteristics of system-centred approaches), which allows an iterative evaluation process, which in turn, lets a fast and easy IRS development. At the same time, ASPIRE is designed to evaluate different personalized techniques, which allows the integration of the user context within the retrieval process (main characteristic of user-centred approaches). ASPIRE evaluation results offer a compromise among the quantitative and controlled results of system-centred approaches and the qualitative results of user studies.

Since ASPIRE is an automatic evaluation methodology, where no interaction with real users is required, it is framed within the personalized evaluation category of contextual simulations. ASPIRE shares a lot of features with contextual simulations. But, at the same time, it has some features which facilitate its execution among other contextual simulation approaches, and allows its evaluation results to meet the compromise between the advantages of system-centred and user-centred approaches.

ASPIRE pretends to be a configurable tool which allows to select the best personalized approach between any two or more number of them, according to the selected evaluation metric, or even to select, within the same personalization approach, its best configuration parameters set up. Additionally, it transforms all this process into an automatic process, selecting the best personalized approach with low effort and cost. This last feature represents an important advantage against other personalized evaluation frameworks, especially those where real users are involved, but also with respect to the own contextual simulations it belongs to, where an exhaustive well defined retrieval scenario must be specified in order to simulate users and user-system interactions.

We next give the characteristics of ASPIRE:

1. ***Document collection***: we can use any document collection, with the only requisite that the documents in this collection, or at least a subset of them, must be able to be classified into different areas of interest or categories. This classification could be explicitly performed by another system component, or the own document collection may be already implicitly classified. An explicit example would be that the documents had some associated tags (e.g. from a controlled vocabulary) either manually or automatically assigned. In the case of this last approach, a first step based on a clustering process could be used to find clusters of similar documents according to their contents. Later a classification process could assign new documents to the corresponding clusters. An implicit example would be that the documents were already classified by its own nature, such as a newspaper, where its news are classified into sections, which represent its different areas of interest or categories (e.g. sports, international,...), or the case of the collection of documents that we have used in our experiments, where each one belongs to a specific parliamentary committee (economy, agriculture, employment,...).
2. ***User Profiles***: no real user interaction with the IRS is required, users are simulated. Each of these simulated users is associated with one or more areas of interests of the document collection. Consequently, we assume that each simulated user will be interested in the topics of the documents which compose the selected area(s) of interest. There are several ways of representing the interests of a (simulated) user by means of a user profile. The most common is to use a set of weighted terms. We use this strategy extracting the set of profile terms from the content of the documents corresponding to the area(s) of interest associated with the simulated user. This can be done by means of an automatic learning process of the most representative terms of these documents (based, for example, on term frequency (tf) and/or inverted document frequency (idf)).
3. ***Queries***: any query can be used, although we advise to use queries formulated by real users of the document collection (obtained from a log file). An heterogeneous set of queries should be used for better retrieval evaluation, representing a trustworthy sample of real user information needs.

4. ***Relevance assessments***: one of the main drawbacks of almost all evaluation systems is the need to have previously assigned relevance assessments for each query or to have real users judging the relevance degree of documents for a given query. ASPIRE avoids this problem by using a procedure that simulates the documents which are relevant for a given query and a given user profile along the way. We do it in the following way: we run the given query against the non personalized IRS and we obtain a ranked list of results. A document will be considered relevant for a given user profile (and the query) if it belongs to the area(s) of interest this user profile represents and it has been retrieved by the IRS among the first *topkRel* results. The intuition behind this procedure is that if a document is among the first ones retrieved by the system for the query and it also belongs to the area(s) of interest associated to the simulated user, then probably this document simultaneously is about the topics of the query and those of interest for the user. Hence, it is relevant for the query from the specific point of view of the simulated user. *topkRel* should be a relatively low threshold, because it would be quite uncommon that a truly relevant document appears in a very lower position in the ranking, let us say, in the position 1000 (if we use a high value of *topkRel* then probably we would introduce many false positive relevant documents for the given query). We will see in Section 5.2 why *topkRel* is important for the relevance assessments criteria.

Any personalization technique can be evaluated, provided that is compatible with the user profile representation being considered. In this sense, any evaluation metric can be used, although we consider particularly valuable the use of rank-based evaluation metrics, as for example, NDCG (Normalized Discounted Cumulative Gain) [30], which was designed to measure a ranking quality by computing the normalization of the weighted sum of the degree of relevance of ranked documents. It assumes that highly relevant documents are more useful if they are on the top of a ranking, and, at the same time, more useful that marginally relevant documents.

An important parameter for the computation of NDCG is the number of results being considered for the evaluation of the performance of the system, named *topkEval*. In some way *topkEval* represents the number of documents a user could evaluate during his/her interaction with the system. In this sense, we suggest that *topkRel* should be greater than the threshold *topkEval*, i.e. $topkRel > topkEval$. By means of this fact, we give the opportunity to the personalization techniques to push up some potentially relevant results into the *topkEval* range.

The proposed evaluation framework, although being a contextual simulation, has some advantages over them. ASPIRE is mainly developed to test personalized IRS retrieval effectiveness. Contextual simulations are more devoted to evaluate the interactions between the user and the IRS. This feature allows the evaluation of the IRS interfaces effectiveness, but this is not the aim of ASPIRE. This extra capability of contextual simulations comes at an associated cost. They need a deep definition of the retrieval scenario, defined

a priori by a sequence of user interactions with the IRS, which are not always easy to define. In contrast, ASPIRE only needs a classifiable document collection (as mentioned earlier, if the collection is not preclassified, a clustering process could be used to define the categories) and a set of common queries for this collection. That is to say, ASPIRE allows the evaluation and improvement of personalized IRS with a very low effort under a completely automatic evaluation process.

### 3.1 Related work

There are not so many approaches similar to ASPIRE but some in the same line. We outline some of them in increasing order of similarity with our approach.

One of the first approaches was [32], whose main purpose is to take into account the dynamic interests of users within the user modeling. The authors have developed a simulation based information filtering system to overcome difficulties on studies where user factors, such as the environment conditions or their current mood, can impact interests. This system uses an approach known as reinforcement learning for user modeling. Some different scenarios are performed with this system to examine model accuracy and filtering effectiveness. Another approach is [51], which evaluates relevance feedback algorithms using searcher simulations through different interfaces, with the intention of determining which of these models to use in the final version of the interface. The search interfaces provide interactions as a source of evidence for the models, using viewed results as the indicator of relevance. The searcher simulations allow them to have more controlled experimental conditions and to model complex interactions without the need of real users. Authors specify that the conclusions derived from their work are still provisional, since they use the evaluation methodology to evaluate implicit feedback models, but the methodology itself is not validated. This is exactly what we are doing in this article, validating our proposed methodology with a real user study. The first reference is focused on user modeling improvement, and the second one on evaluating how search interfaces provide more relevant information to relevance feedback algorithms, both focusing on different aspects with respect to us. In addition, they must define the user-system interactions which will compose the retrieval scenario.

In [20] an evaluation protocol for session-based personalization (searching a sequence of related queries, i.e, short-term personalization) is proposed. The profiles are based on the topics provided by the TREC HARD collection. The user profile is simulated for each topic using a set of documents returned by the system, which have previously been judged as relevant by TREC assessors. The queries in a session are built by selecting the top terms associated to subtopics (subsets of relevant documents for the topic). The emphasis of the evaluation is put in the delimitation of the session boundaries. The main differences with our proposal, in addition to the focus on short-term personalization, are that

they use simulated queries instead of real ones (as also done in [42]) and that relevance judgements are not simulated but real.

The closer approach to ASPIRE is [42]. The authors build user context models as ontological profiles, assigning implicitly learned interest scores to existing concepts of a domain ontology (ODP). As these interests are dynamic, a spreading algorithm is used to maintain them updated along time. Their aim is to demonstrate that re-ranking improves the disambiguation of user query intent. They use a document collection of 10,226 documents indexed under various ODP concepts leading to a training, profile and test sets of documents. Depending on the query, each document in the collection is considered relevant if it is classified under the concept being simulated, and not relevant otherwise. They automatically build four variations of keyword queries using the top terms associated to the concept/user interest being simulated. The search results were retrieved using a cosine similarity measure for matching, using top-n recall and precision metrics.

Although Sieg et al. approach [42] and our proposed methodology seems very similar, they actually have several differences, such as: they focus more in the user profiles than in the retrieval effectiveness of different personalized techniques; they disambiguate ambiguous queries more than personalize them; the way they build the queries based on an unique concept of the ODP ontology, which may not represent real user information needs; in contrast, we use real queries suitable to be evaluated under more than one document collection area of interest. And maybe the two most important differences: we verify our evaluation methodology validity with a real user study, and we design a generic and automatic evaluation methodology, with the intention to be easily used under very different evaluation situations for any personalized IRS approach.

There are also some interesting works, not focused on personalized but on general retrieval systems evaluation, which also propose methods without using real relevance judgements. In [33], a pool of documents is generated from the top $b$ documents returned by each of the systems being evaluated for a given query. These documents are ranked according to their similarity with the query using a vector space model, and the top $s$ documents of this ranking are assumed to be the (pseudo) relevant documents for the query. In [43] the (pseudo) relevant documents are extracted from the pool randomly, using a simple model for how relevant documents occur in the pool. In both cases the correlation between the rankings of retrieval systems using these simulated judgements and using human judgements was not very high[1] in experiments with TREC collections.

In general, the stability of system rankings is measured by using Kendall's rank correlation, $\tau$. We have to be cautious when the rankings are obtained over narrow score ranges [39], since low $\tau$ values might be expected. As [34] concludes, the evaluation strongly depends on a relative small set of top-ranked results. So, whenever our automatic methodology is able to find such kind

[1] A Kendall $\tau$ correlation always below 0.5.

of documents, we might expect to obtain rankings similar to those obtained within the user study.

Finally, although not directly related to our work, there is a set of studies which consider ranking evaluation with low cost [4,16,34,7]. Some clear differences might be stated. The first one is that these papers are related to the reliability and robustness of relevance judgements to evaluate information retrieval systems. They focus on the number of queries, pool depth, etc. In these cases, all the relevance judgements are assumed to be true. Additionally, some works can be found which introduce error in the relevance assessments, studying how the system rankings were affected [43,17]. But in all the cases, they do not tackle with the problem of personalization, in such a way that what is relevant for an user might not be relevant for another one.

Another approach for low cost evaluation is the use of simulated queries using generation models that simulate a candidate query for a given set of documents, which are assumed to be relevant for that query [3,25]. Although these papers reveal interesting trends, further studies in this direction are necessary in order to provide comparable results to manually assessed judgements.

## 4 ASPIRE Use and Validation

Our main objective is to validate the reliability of the proposed methodology for the evaluation of personalized IRS. Moreover, we are going to show how ASPIRE allows to test and select the best personalization techniques from a set of different personalization techniques, also considering for each of them, the possible configuration parameters of the user profiles. This test and selection process is usually very difficult, or most of the times impossible, with traditional evaluation frameworks. However, the use of our proposed automatic evaluation methodology turns this action into a faster decision process.

In a previous paper [15], we evaluated a wide set of 13 different and heterogeneous personalization approaches using the relevance assessments from a user study. We are going to compare the results of this study with the results obtained by applying ASPIRE under the same circumstances considering different retrieval approaches, in order to provide evidences about the reliability of our automatic evaluation methodology.

We followed the advices given in [6,18] for ensuring the reliability of user study findings, so that the user study results should be considered as the real results. Therefore, the results of the personalization techniques evaluated with the ASPIRE evaluation framework should be close to them or, at least, to follow the same dynamics, in order to validate ASPIRE.

### 4.1 Experimental framework

Let us explain first some details about the experimental framework considered.

***Search engines***: We will explore our approach using three retrieval models. On the one hand, our methodology will be tested with a search engine

specifically designed for dealing with structured documents and, on the other hand, we will consider retrieval models designed for working with flat documents. The selected retrieval models are:

- Garnata: The first one is Garnata [11], which is based on probabilistic graphical models, namely Bayesian networks and influence diagrams. The theoretical basis of Garnata is explained in detail in [9,10]. Garnata is designed to work with structured information, concretely with XML documents[2]. This structured IRS has been improved and tested at three editions of the INEX Workshop ([12] describes our last participation). It has also been applied to build a real IRS for parliamentary documents [13]. After submitting a query, the system ranks a set of non-overlapping elements according to their relevance to the topic.
- BM25: The second retrieval model is based on a probabilistic retrieval approach, particularly we consider the BM25 term weighting formulas which have been used quite widely and quite successfully across a range of collections and search tasks, representing a state-of-the-art tf-idf-like retrieval function.
- VSM: The third one is a vector space retrieval model, VSM, using the implementation in the Lucene search engine[3]. The similarity function is derived from the classical cosine measure, which can also boost term weighting based on user specified requirements, e.g., the importance of the fields. Note that since the data used in the evaluation only contain one field, we do not consider the boosting factors.

***Document collection***: Since the used retrieval models work with structured and flat documents, we will consider two different versions of the Records of Parliamentary Proceedings of the regional Parliament of Andalusia (Spain). Each document contains full transcriptions of all the members of the parliament's speeches in each parliamentary session. Each record corresponds to different Committee Sessions, where each of these committees is dedicated to a specific area of interest, e.g. economy, health, education, employment, etc. We have selected a subset of 658 committee sessions from the sixth and seventh terms of office (eight years, from March 2000 to March 2008). We are going to present in detail each collection.

- Structured-XML collection: in this case, a document is marked up in XML, which follows the following structure, representing the different semantic levels of the sessions:

*session/initiative/intervention/speech/paragraph*

[2] The fact of focusing on XML information retrieval requires the adaptation of some search engine components. For example the retrievable elements are not only complete documents but document components (called structural units), which may overlap. However this does not represent any problem when using ASPIRE.

[3] Lucene is a popular open-source search software. It provides indexing and search technologies, which is frequently used by several applications all over the world, ranging from mobile devices to sites like Twitter, Apple and Wikipedia. This search engine is designed to work with plain (non-structured) documents. http://lucene.apache.org/

Note that, for example, one session could have several initiatives which at the same time could contain several interventions and so on. We have a total of 432,575 different retrievable elements (e.g. an intervention of a member of the Parliament or a paragraph within this intervention), and a size of 122MB.
- The Flat version of this collection is obtained after considering one different document for each initiative. Each document includes all the information relevant for the initiative, without any structure. In this case, we have a total of 3732 documents with a size similar to the structured collection.

Notice that for both collections each retrievable element only belongs to one committee, and therefore the collection itself is already implicitly classified, as required by ASPIRE. At this point we have to mention that there are committees that are supposed to be related only to a specific subject, as for example, "Economics", but there are others which contain documents related to several subjects, as is the case of "Culture, Sports and Tourism", which are gathered under a unique committee.

### 4.2 User study

Within the user study we have considered real users which have been instructed to assume his/her chosen profile. This profile corresponds to a person interested in documents related to the topics discussed in a specific committee. In this sense, for the same query two different users might provide different judgements. These judgements were made after a brief training phase to familiarize the users with the retrieval system interface. Note that the goal of the user study is to obtain the relevance assessments for each combination of query, committee-based profile and user. We denote each triplet of user, query and selected profile as an *evaluation triplet*.

***Queries***: we used an heterogeneous set of 23 queries formulated by real users of the document collection. In order to reduce the burden of the users in the relevance assessment step, these queries were selected taking into account that there are not a very large number of relevant documents for each query. Nevertheless, they represent a small but trustworthy sample of real user information needs. This set of queries has an average length of 2.61 terms per query, which is in the range of the average search query length studies [47].

***Users and user profiles***: the user study involved 31 users. Each user submitted one or several of the previous 23 queries to the IRS assuming the corresponding profile. Particularly, eight profiles were selected, corresponding to eight different Committees (in broad sense these committees are related to agriculture, culture, economy, education, employment, environment, health and justice). For simplicity, we talk about "pure" user profiles (representing one document collection area of interest), but we have to note that, in our experimentation, "hybrid" profiles have also been considered because committees with a mixed content were assigned to the users.

Several users were able to submit and evaluate their queries by assuming (in turn) different profiles, giving a total of 126 different evaluation triplets.

With respect to the profile's representation, we have used simple sets of weighted terms. These terms have been learned based on the content of the documents from the corresponding area of interest, following a tf*idf approach. For more details about the construction of the profiles, see [14].

***Relevance assessments***: Given an *evaluation triplet*, the user will use his/her profile to determine what constitutes a relevant (and non-relevant) document for each one of the 126 queries. We have used Garnata as the search engine, using the structured collection in XML format. The results of the interaction of the user with the system is a set of relevance assessments for each retrieved element[4].

Following the guidelines of [34], and considering the number of triplets, we ask the users to judge deep pools (under the corresponding profile) for the selected topics, ensuring that if the user start judging a query, he/she complete the assessment for it.

Particularly, a pool of up to 100 elements has been considered, pool size that has been proved to give reliable results [54]. The pool is composed by the 50 first elements retrieved by the non personalized IRS in response to the query, plus the 50 first results returned by the IRS using one personalization strategy (Hard reranking, see [14]) for the same query. We did it in this way to avoid that many possible relevant results, not appearing among the first 50 results obtained by the non personalized query, were considered as irrelevant[5]. The personalization strategy used to perform this additional evaluation was selected carefully to avoid the bias that the relevance judgements obtained with this strategy could induce on the evaluation of the other strategies[6].

The average number of relevant elements in the pool per query is 18.7, with a standard deviation of 14.2. Note that this high value for the standard deviation mainly came from the fact that we are considering different profiles for judging the same query. For instance, the query "price level rise" has an average number of relevant documents of 30, 2, 45 and 2 under the profiles of agriculture, culture, economy and education, respectively.

By means of this process, we obtain the set of relevance assessments for the Structured Collection. These assessments have been extrapolated in order to obtain the users' assessment for the Flat Collection. Particularly, we have considered that an initiative (a document in the flat collection) will be relevant if itself or any of its descendants (interventions, speeches or paragraphs) have been judged relevant by the user.

[4] It should be noticed that the user did not judge if a given retrieved result was the best possible one, but only whether or not its content was relevant to the given query and profile (binary assessments).

[5] The source of the problem is the limitation of judging only the first 50 results retrieved by the IRS, but it was necessary since the evaluation of a great number of results would require too much time and effort from the users.

[6] As Hard reranking only considers the list of results of the original non personalized query, it does not introduce any relevance assessments not present in the original results list.

### 4.3 ASPIRE relevance assessments

For our proposed methodology we used the same document collection and evaluation triplets, i.e. the 126 triplets of queries with their associated users and profiles, as in the user study, under the Garnata, BM25 and VSM retrieval models. The only difference is that the relevance assessments of each of the 126 triplets were not obtained from real users but simulated, as prescribed by ASPIRE. Briefly, we submit each original query to each retrieval system, focusing on the highest ranking results (the *topkRel* highest results). Then, each one of these results is considered relevant to this topic if and only if it belongs to the same area(s) of interest than the one(s) represented by the user profile (see Section 3). Since in the user study the users evaluated up to 100 (50+50) results, we have initially fixed *topkRel* to 100 in our experiments.

### 4.4 Validation methodology

Notice that by using ASPIRE assessments we are considering as relevant documents some good candidates (they are highly ranked and belong to the same topic of interest) but, on the other hand, some errors are included in the assessments: some relevant documents may be missed and some non relevant documents might be considered as relevant. Now, the problem is to measure the impact of these mistakes on the evaluation.

To tackle this problem we are going to consider two different criteria: on the one hand, we conduct a comparison between both (real and automatic) relevance assessments. By means of this comparison we can measure the amount of error. On the other hand, we want to evaluate whether ASPIRE is (or not) a reliable methodology to evaluate personalized IRS. In other words, if the errors made in the assessments could cause too much damage in the aggregate, which invalidates the conclusions obtained by our approach. In this sense, we will compare the retrieval performance (using the NDCG metric) obtained by the user study and by ASPIRE, under different personalization techniques and user profile configurations.

In this experimentation we have fixed the number of results being considered for the evaluation of the system performance, i.e. *topkEval*, to the first 50 results. Note that this meets our requirements, since $topkEval = 50 < 100 = topkRel$. The next section presents the metrics employed for these purposes and then, Section 4.4.2 presents the used personalization techniques and profile configurations.

#### *4.4.1 Metrics*

Firstly, we will focus on the comparison between the relevance assessments from the user study and from ASPIRE. To accomplish this goal we have considered two evaluation metrics.

– The first metric measures which percentage of the real relevance assessments associated to each evaluation triplet (*query, profile, user*) truly belongs to the documents within the corresponding user *profile*. Remember that the main assumption of ASPIRE is that the simulated relevance assessments are always extracted from the document collection area(s) of interest associated to the simulated user profile. Then, if an element in the real relevance assessments does not belong to the area(s) of interest the user profile represents, it will never be considered as a relevant result by ASPIRE, and always considered as relevant otherwise. In some sense this metric measures the confidence we can expect from the ASPIRE relevance assessment criteria.
We define this evaluation metric by the following formula:

$$\text{Conf}(q,p,u) = \text{size}(et_{q,p,u} \cap id_p)\frac{100}{size(et_{q,p,u})} \tag{1}$$

where $et_{q,p,u}$ is the set of relevance assessments of the evaluation triplet for the given query $q$, profile $p$, and user $u$; $id_p$ is the set of all documents in the collection belonging to the profile $p$.

– The second evaluation metric aims to measure the overlap degree between the relevance assessments provided by the user study and by ASPIRE. To do that, we propose to use the counterparts of the classical measures of precision (pre), recall (rec) and F from the classification field. Let $et_{q,p,u}$ be as in eq.(1) and $et_{q,p}$ be the set of simulated relevance assessments for the query $q$ and the profile $p$.

$$pre = \frac{tp}{tp+fp}, \quad rec = \frac{tp}{tp+fn}, \quad F = \frac{2 * pre * rec}{pre + rec} \tag{2}$$

where $tp$ denotes true positives (number of relevance assessments in $et_{q,p}$ which are also in $et_{q,p,u}$), $fp$ denotes false positives (number of relevance assessments in $et_{q,p}$ which are not in $et_{q,p,u}$), and $fn$ stands for false negatives (number of relevance assessments in $et_{q,p,u}$ which are not in $et_{q,p}$).
For XML documents things are a bit difficult because the elements in the sets $et_{q,p,u}$ and $et_{q,p}$ can match partially (the two elements are not identical but one contains the other). For that reason we need to use a function, $sim(A_j, U_i)$, measuring the degree of similarity between an element $A_j \in et_{q,p}$ and other element $U_i \in et_{q,p,u}$. This function is based on the distance between these elements in the XML hierarchy, obviously if there is no overlap between $A_j$ and $U_i$ the similarity is zero (see [15] for details).
Taking this similarity measure into account, the definition of $tp$, $fp$ and $fn$ in eq.(2) is:

$$tp = \sum_{sim(A_j,U_i)\neq 0} sim(A_j, U_i) \tag{3}$$

$$fp = size(\{A_j \in et_{q,p} \mid sim(A_j, U_i) = 0 \; \forall U_i \in et_{q,p,u}\}) \tag{4}$$

$$fn = size(\{U_i \in et_{q,p,u} \mid sim(A_j, U_i) = 0 \; \forall A_j \in et_{q,p}\}) \quad (5)$$

Secondly, and in order to measure the retrieval performance of a retrieval run (e.g. any personalization technique) we will use the NDCG metric [30]. This evaluation metric estimates the cumulative relevance gain observed by a user for the first documents in a retrieved list of results. Since users are prone to check only the first results, the metric includes a discounting factor to reduce the document effect as its position is down within the ranking. The metric value is normalized by the ideal ranking, that is, all relevant results are consecutive from the beginning of the list. This metric has been proved valuable for the comparison of retrieval performance between systems [38].

– The NDCG metric is calculated as follows:

$$NDCG@x = \frac{1}{N} \sum_{i=1}^{x} \frac{2^{rel(d_i)} - 1}{\log(i+1)} \quad (6)$$

where $x$ = topkEval, $N$ is the ideal DCG value for the relevant results, $i$ is the ranking position of the document being evaluated, $d_i$ is the document at position $i$, and $rel(d_i)$ is the relevance value of $d_i$. When we work with the collection exploiting the XML organization, some NDCG adaptations are required to deal with partial matchings between retrieved elements and relevance judgements. See [15] for details. For the flat document collection there is no need to make any modification in the NDCG formula.

#### *4.4.2 Personalization techniques and profile configurations*

One of the motivations of our methodology is to be able to obtain a ranking of personalized systems and/or configuration parameters for a given personalization technique, maybe with the goal of selecting the best ones. In this sense, we ask the following question: how much would the rankings change if relevance assessments were chosen using ASPIRE instead of the real user study ones? In this section we will highlight the different alternatives for personalization analysed in our experimentation.

With respect to the ranking of personalized search, we have tested the Structured-based and Flat-based retrieval models, considering their behaviour under the original query (non-personalized), denoted as (*Orig*), and also the behaviour under a set of 13 and 7 different personalization approaches, respectively. In the last case, since BM25 and VSM are flat-based, we have not considered those XML-oriented alternatives, neither those that require a modification of the search engine, as we shall explain later.

Within this wide set of techniques (explained in detail in [15]) there are approaches from the three possible retrieval stages where personalization can be applied: before the search (e.g. query expansion - *QE and NQE*), within the search (not very used yet, e.g. retrieval model modification - *NQE+m, HRR+m, SRR+m and IRR+m*), and after the search is performed (e.g. re-ranking - *HRR, SRR and IRR*). We have even included two bad performance

personalization techniques (*I-HRR and p-HRR*), which are used in [15] to demonstrate some of the other personalization techniques design decisions, and we have still considering them here to support and strength the ASPIRE reliability. Finally, we have also included two additional Content-and-Structure personalization techniques (*CAS and CAS-or*). As the reader may realize, several of the previous personalization techniques are hybridizations between some of the explained three basic approaches where personalization may be applied. We must recall that both the Content-and-Structure and within-the-search personalization techniques can not be used by BM25 and VSM. We also have to note that more important than the specific characteristics of each of the used personalization techniques, is the fact that they cover a great variety of personalization approaches.

All of the 13 previous personalization methods have an underlying common feature: in one way or another, all of them make use of an expanded query, which uses the appropriate user profile weighted terms in the expansion step. Based on this characteristic, we have tested all the personalization techniques under 12 different combinations of the two main user profile configuration parameters: the number $n$ of expanded terms used (5, 10, 20 and 40) and a normalization factor $p$ applied over their associated weights (0.33, 0.66 and 0.99), which controls the importance of the expanded terms with respect to the original query terms.

Therefore, we have a set of 157 different IRS configurations to be tested $(12 \times 13 + 1)$ under Garnata and 85 different IRS configurations to be tested $(12 \times 7 + 1)$ under each of the BM25 and VSM models. Each of these IRS configurations involves the use of the 126 evaluation triplets, which represents a total number of 19,782 and 10,710 different ranked lists of retrieved results for each evaluation approach (user and ASPIRE) ready to be evaluated using NDCG. We will only display results based on the averages of NDCG values across the 126 evaluation triplets for the 157 and 85 different IRS configurations, respectively.

## 5 Results

In this section we will first show the comparison between the user study and ASPIRE results with our proposed evaluation framework using Garnata, BM25 and VSM. Then we will show the same results comparison between the user study and a state-of-the-art evaluation approach described in [42]. All these comparisons will show how reliable and robust is ASPIRE under different scenarios.

### 5.1 User study-ASPIRE results comparison under our proposed evaluation framework

In this first section we will show both the relevance assessments and the retrieval performance evaluation comparisons, under our proposed evaluation

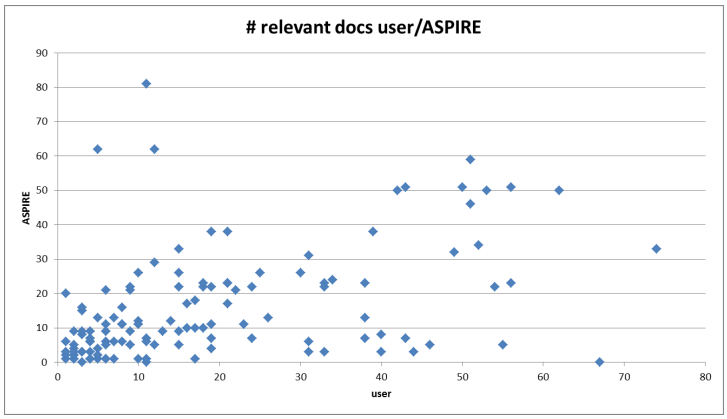


**Fig. 1** Number of user study evaluation triplets relevance assessments ($x$ axis) against ASPIRE evaluation triplets relevance assessments ($y$ axis). Each point in the graph represents an evaluation triplet.

framework. The first part will try to validate the ASPIRE relevance criteria, while the second part will show how well ASPIRE behaves with respect to the real user study evaluation results.

### *5.1.1 Relevance assessments comparison (validating ASPIRE relevance criteria)*

In this case, we will focus on the truly obtained assessments using the XML-based retrieval model, where a deeper analysis can be done. Nevertheless, similar results would be obtained when considering the unstructured retrieval systems.

We will begin this point comparing the raw number of relevant results per triplet in the real user study ($x$ axis) and the ones obtained using ASPIRE ($y$ axis), see Figure 1. From this graph, we can see that in general the automatic assessments tend to select a lower number of relevant elements (1965) than the user study (2362), although they are positively correlated with a Pearson correlation value of 0.452.

The application of the *Conf* metric (eq. 1) over the 126 evaluation triplets gives an average value of 75.93% with a standard deviation of 28.97%. This means that approximately 3 out of 4 results judged as relevant by real users are also considered as potentially relevant by ASPIRE, conforming its main assumption about relevance assessments. We believe that this is a quite good ratio. Anyway, we will see later whether the remaining 24.07% of miss-assessed relevant results will cause a proportional difference when analysing the user study-ASPIRE comparison of evaluation results or not.

The average results, across all the evaluation triplets, of precision, recall and F measure, together with the standard deviations, are displayed in Table 1. We can observe large deviations, so that the behaviour is quite different

| | pre | rec | F |
|---|---|---|---|
| $\mu$ | 0.548 | 0.513 | 0.450 |
| $\sigma$ | 0.307 | 0.334 | 0.260 |

**Table 1** Averages and standard deviations of precision, recall and F across the 126 evaluation triplets.

depending on the query and the profile being evaluated. The average values indicate that the overlap degree between real and simulated relevance assessments is around 50%. As with the previous metric, the interesting question now is whether this overlap degree is good enough to produce sufficiently close performance evaluation results for both the user study and ASPIRE approaches.

*5.1.2 Retrieval performance evaluation results comparison*

Although the previous metrics are interesting, giving a first insight of ASPIRE's quality for the generation of relevance assessments, the actually important values are those based on the comparison between the evaluation of the retrieved results using the user study and ASPIRE. These results are the ones the final users will use to judge the behaviour of any personalized IRS, and if ASPIRE is able to evaluate them pretty similar to the way the users would do it, independently of the underlying relevance assessments, ASPIRE will be a good method to simulate those users.

It is important to remark that we have evaluated a very heterogeneous set of personalization techniques, ranging from some very good to some very bad performance methods with different retrieval models. These features make the derived conclusions more robust and valuable.

This section pursues two main objectives: 1) to test whether ASPIRE should be considered as a reliable approach in the evaluation of personalized IRS, comparing its evaluation results with those obtained from the carried out user study. And, 2) to show whether ASPIRE is able to rank properly the personalization approaches, and even to rank the different profile representations for the given personalization methods, in accordance with the user study results.

**Is ASPIRE a reliable evaluation approach?**

To answer this question we will compare the performance for the different queries under a given configuration. In order to tackle this problem we run, for each configuration of personalization techniques and profile parameters, the 126 evaluation triplets, computing the average NDCG values for each run. As measure of the quality of our approach we consider whether the NDCG values obtained from the automatic evaluation might correlate (or not) with their respective values in the user study.

We shall note that for the XML-based approach we have studied 157 different runs, one for each combination, whereas for flat-based ones the number of runs is 85.

| Pearson Correlations | | | |
|---|---|---|---|
| Model | Range | Avg | SDV |
| XML | [0.450, 0.716] | 0.604 | 0.057 |
| BM25 | [0.177, 0.794] | 0.631 | 0.144 |
| VSM | [0.313, 0.797] | 0.591 | 0.137 |

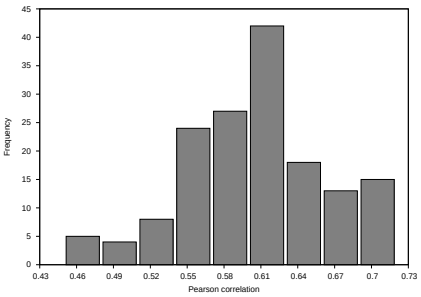


**Fig. 2** NDCG user study-ASPIRE Pearson correlations for the different runs. The table presents the correlation ranges, average values and standard deviations over the different runs. The right-hand side presents an histogram of the correlation values for the XML-based approach.

We will examine these values at two different levels of granularity. Firstly, we consider each profile combination as an independent run and compute the quality (using NDCG) of each query under this combination for both, the user-study and ASPIRE based approaches. Then, we use Pearson correlations between the NDCG values of each query to determine the resemblance between the different results. Now we are going to discuss the obtained results, presented in Figures 2 and 3. Thus, Left-hand side of Figure 2 shows the Pearson correlations between the NDCG obtained with the user study vs the ones obtained using ASPIRE. Also, for illustrative purposes, the right-hand side shows a histogram that resumes these data for the XML-based retrieval model. In all the cases, we can observe a positive correlation with an average value around 0.6 with low standard deviation. Thus, ASPIRE can be considered as a moderately good predictor of the relative performance on individual queries for the considered models.

The same correlation coefficients are plotted in Figure 3 (in the $y$ axis) against the averaged NDCG values from the user study (in the $x$ axis). The main conclusion drawn observing this figure is that the correlation values do not depend on the real (good or bad) performance of the given personalization techniques-user profile configuration parameters, except for a small number of outliers (for low Pearson correlation values).

The previous results represent an intra-configuration comparison. Now we will focus on how these configurations relate to each other. In this case, we consider an evaluation matrix where the columns represent the personalization strategies and the rows represent the profile's representation parameters. In this matrix, each cell represents the average NDCG obtained after running the 126 evaluation triplets under a given configuration, being a global measure about its quality.

Figure 4 plots the user study averaged NDCG values ($x$ axis) against the corresponding values obtained from ASPIRE ($y$ axis), for the 157 evaluation

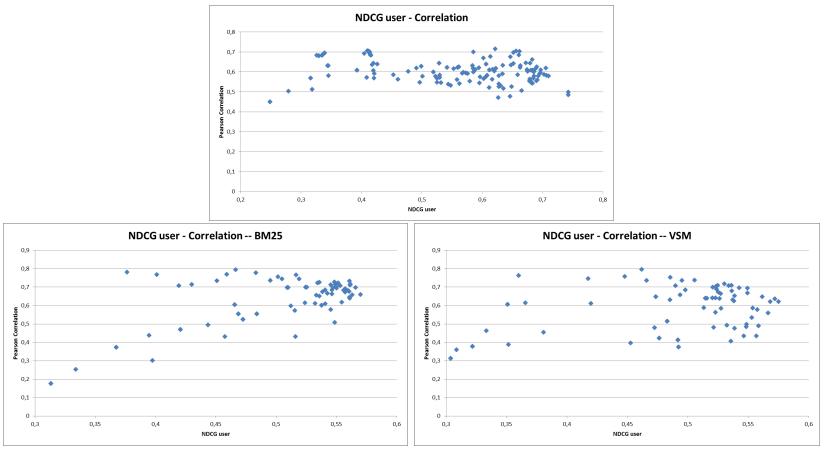


**Fig. 3** NDCG user study-ASPIRE correlations (y axis) against the averaged NDCG values from the real study (x axis). Top Figure represents the results obtained with the XML-based approach, whereas the bottom figures represents the results obtained with BM25 and VSM.

matrix cells for the XML-based approach and the 85 combinations of the BM25 and VSM retrieval models. In the figure we also show the lineal regression line and the ideal fit line, labelled as *LR (y = x)*. Some conclusions may be drawn from this figure:

1. They show how the ASPIRE results are almost always very close to those obtained from the user study. A linear regression over this data shows a R-squared values of 0.8357, 0.945 and and 0.940 for XML, BM25 and VSM, respectively. So, we have obtained very high Pearson correlation coefficients equal to 0.914, 0.972 and to 0.970 for each respective model[7].
2. Taking into account the ideal fit line, labelled as *LR (y = x)*, we can show that the ASPIRE results are very close to the real NDCG values (obtained from the user study). This is important because ASPIRE usually does not overestimate neither underestimate significantly the performance of the given personalization technique, independently of the retrieval model considered.

Taking into account all the results, we could conclude that ASPIRE is able to robustly evaluate any given personalization technique, independently of the used retrieval model.

**Stability of rankings of ASPIRE-user study personalization techniques and user profile configurations**

We are going to test whether we may trust on the systems ranking provided by ASPIRE, comparing it with that provided by the user study. Our objective is to look at the stability of rankings, rather than the absolute values

[7] The correlation values in this case are greater than those in Figure 2, because here we are correlating the averaged NDCG values for ASPIRE and the user study, not the underlying and more diverse 126 evaluation triplets of each of these combinations.

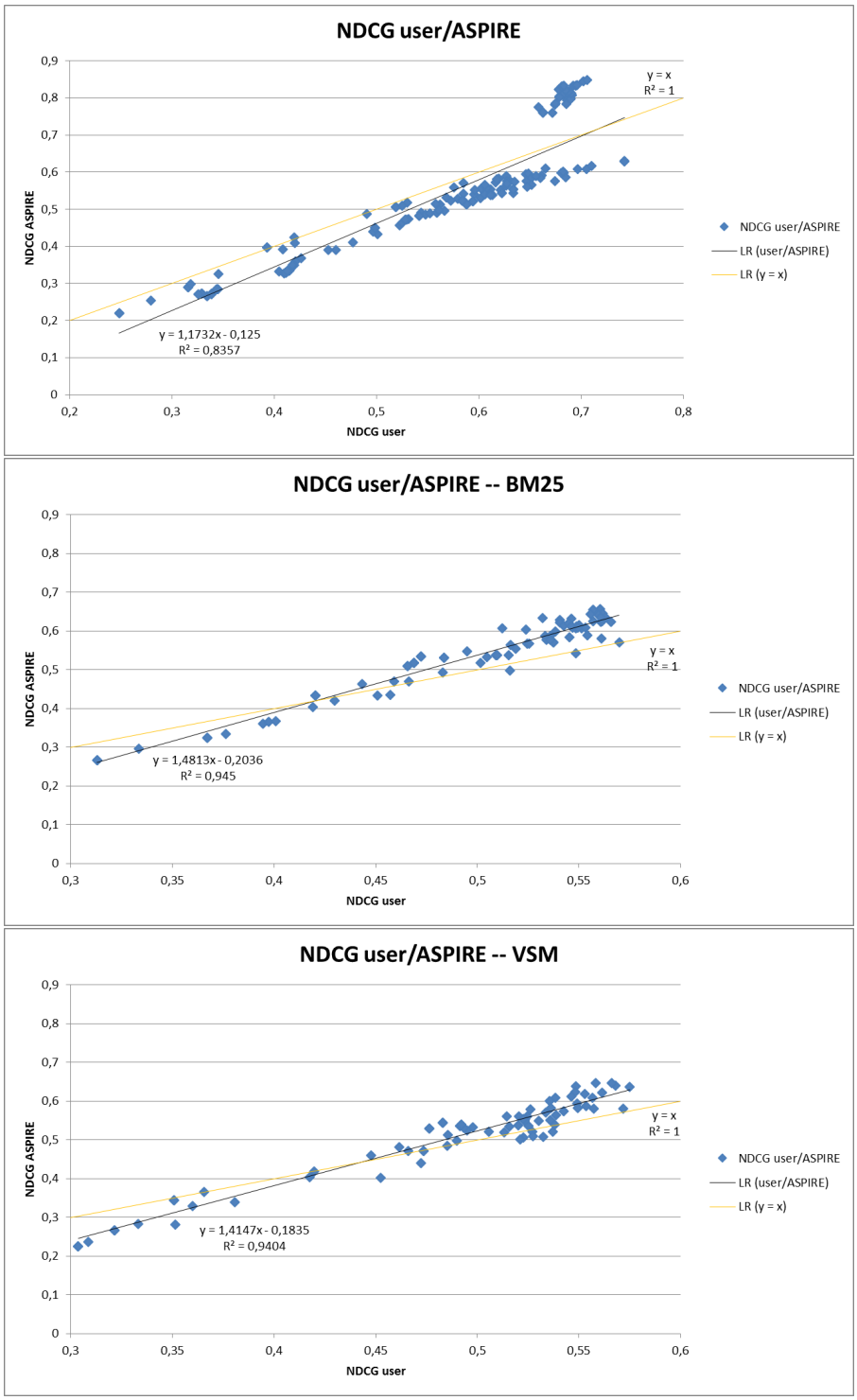


**Fig. 4** Averaged NDCG values from ASPIRE (y axis) against the averaged NDCG values from the real study (x axis), for each one of the personalization techniques-user profile configuration parameters combinations. From top to bottom we show the results obtained using XML, BM25 and VSM retrieval models.

of the NDCG measure. We use the Kendall's $\tau$ rank correlation, also known as Kendall's coefficient, to compare both rankings. Kendall correlation is a function of the minimum number of pair-wise swaps required to turn one ranking into another. If the agreement between the two rankings is perfect, then $\tau = 1$; if the disagreement is perfect, then $\tau = -1$, and if both rankings are independent then $\tau \simeq 0$.

Taking into account that people usually disagree about relevance [41], problem aggravated in the case of personalized information retrieval, and that error in the assessments could affect the system ranking [4,17], it becomes necessary to contextualize the correlation value between the automatic and human-based evaluation. With this objective in mind, we also present the comparison results of the rankings obtained by considering two different user-based sets of assessments. These sets have been obtained after randomly split our original set into two independent groups, A and B. Then, by comparing the system rankings for these groups, we might identify in which way the human differences in judgements may impact the relative system performance. This value, denoted as $\tau_{A/B}$, shall be used as reference for our comparison.

*Selecting the best personalization techniques for a fixed profile:*

As we want to select the best personalization techniques, we are going to compute the rank correlations (Kendall's $\tau$) between the real and simulated results of the personalization techniques, for each of the 12 user profile configurations, displayed in Table 2. In this table, we also present the standard deviation among the NDCG values obtained for each personalization approach for both, the user study $\sigma_{us}$ and the results obtained using ASPIRE, $\sigma_{ASP}$. These deviations are displayed to illustrate the differences in performance between the different personalization techniques: the greater standard deviations the greater the difference between the performance of the different approaches is.

In average, a Kendall $\tau$ value of 0.896 with a low standard deviation, $\sigma$, of 0.063 is obtained from the 12 user profile configurations for the XML-based model. Note that these values are quite similar to the ones obtained using the two different users' judgements with an average $\tau_{A/B}$ value of 0.924 with standard deviation of 0.039. (average values $\tau = 0.797$, $\sigma = 0.090$ and $\tau = 0.754$, $\sigma = 0.125$ are obtained for BM25 and SVM models, respectively).

These rather high values show that the ASPIRE and the user study rankings are very similar, independently of the user profile configuration being considered. This is particularly true when there exists significant differences among the different values used to obtain the rankings ($\tau$ increases with greater values of $\sigma_{us}$). This has some sense, since in the case in which we obtain lower $\sigma_{us}$ it might difficult to be sure about the obtained ranking: the differences in performance among the different methods are rather small, thus making it more difficult (but also less important) to discriminate between them. Therefore, we can be pretty sure that ASPIRE is a reliable method to discriminate

| $n$ | 5 | | | 10 | | | 20 | | | 40 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0.33 | 0.66 | 0.99 | 0.33 | 0.66 | 0.99 | 0.33 | 0.66 | 0.99 | 0.33 | 0.66 | 0.99 |
| $\tau_{A/B}$ | 0.947 | 0.896 | 0.896 | 0.974 | 0.844 | 0.974 | 0.922 | 0.922 | 0.974 | 0.896 | 0.922 | 0.922 |
| $\tau_{XML}$ | 0.947 | 0.844 | 0.922 | 0.922 | 0.974 | 0.896 | 0.740 | 0.844 | 0.922 | 0.922 | 0.870 | 0.948 |
| $\sigma_{us}$ | 0.104 | 0.105 | 0.109 | 0.112 | 0.121 | 0.128 | 0.119 | 0.133 | 0.143 | 0.135 | 0.159 | 0.168 |
| $\sigma_{ASP}$ | 0.145 | 0.150 | 0.152 | 0.155 | 0.158 | 0.163 | 0.160 | 0.170 | 0.178 | 0.171 | 0.190 | 0.199 |
| $\tau_{BM25}$ | 0.878 | 0.619 | 0.810 | 0.878 | 0.810 | 0.714 | 0.714 | 0.714 | 0.905 | 0.810 | 0.905 | 0.810 |
| $\sigma_{us}$ | 0.053 | 0.038 | 0.032 | 0.063 | 0.054 | 0.058 | 0.072 | 0.067 | 0.072 | 0.086 | 0.093 | 0.102 |
| $\sigma_{ASP}$ | 0.077 | 0.064 | 0.055 | 0.080 | 0.078 | 0.082 | 0.105 | 0.110 | 0.122 | 0.130 | 0.147 | 0.159 |
| $\tau_{VSM}$ | 0.619 | 0.524 | 0.714 | 0.810 | 1.000 | 0.810 | 0.714 | 0.714 | 0.714 | 0.905 | 0.714 | 0.810 |
| $\sigma_{us}$ | 0.053 | 0.041 | 0.036 | 0.070 | 0.072 | 0.083 | 0.076 | 0.087 | 0.098 | 0.086 | 0.102 | 0.109 |
| $\sigma_{ASP}$ | 0.071 | 0.055 | 0.057 | 0.093 | 0.101 | 0.112 | 0.113 | 0.130 | 0.142 | 0.129 | 0.153 | 0.165 |

**Table 2** Kendall $\tau$ correlations of the personalization techniques for each one of the 12 combinations of the user profiles configuration parameters: number of expanded terms, $n$, and normalization factor, $p$.

| method | QE | NQE | HRR | SRR | IRR | I-HRR | p-HRR | CAS | CAS-or | NQE+m | HRR+m | SRR+m | IRR+m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_{A/B}$ | 1.000 | 0.939 | 0.879 | 0.788 | 0.818 | 0.848 | 0.931 | 0.727 | 0.424 | 0.667 | 0.879 | 0.939 | 0.939 |
| $\tau_{XML}$ | 1.000 | 0.970 | 1.000 | 0.909 | 0.879 | 0.818 | 0.473 | 0.576 | 0.545 | 0.939 | 0.939 | 0.909 | 0.909 |
| $\sigma_{us}$ | 0.105 | 0.112 | 0.082 | 0.042 | 0.054 | 0.005 | 0.006 | 0.011 | 0.009 | 0.039 | 0.053 | 0.041 | 0.041 |
| $\sigma_{ASP}$ | 0.113 | 0.112 | 0.086 | 0.048 | 0.061 | 0.012 | 0.007 | 0.018 | 0.013 | 0.034 | 0.042 | 0.032 | 0.032 |
| $\tau_{BM25}$ | 1.000 | 0.758 | 0.424 | 0.667 | 0.718 | 0.909 | 0.667 | - | - | - | - | - | - |
| $\sigma_{us}$ | 0.087 | 0.079 | 0.016 | 0.015 | 0.016 | 0.040 | 0.007 | - | - | - | - | - | - |
| $\sigma_{ASP}$ | 0.121 | 0.101 | 0.019 | 0.031 | 0.029 | 0.065 | 0.011 | - | - | - | - | - | - |
| $\tau_{VSM}$ | 1.000 | 0.909 | 0.545 | 0.697 | 0.121 | 0.848 | 0.333 | - | - | - | - | - | - |
| $\sigma_{us}$ | 0.069 | 0.086 | 0.012 | 0.017 | 0.009 | 0.033 | 0.007 | - | - | - | - | - | - |
| $\sigma_{ASP}$ | 0.098 | 0.110 | 0.016 | 0.019 | 0.017 | 0.051 | 0.006 | - | - | - | - | - | - |

**Table 3** Kendall tau correlations of the user profiles for each one of the personalization techniques.

between (better and worse) personalization methods.

*Selecting the best user profile configurations for a fixed personalization method:*

As we want to select the best user profile configurations, now we will also calculate the rank correlations between the real and simulated results obtained by the different user profile configurations, for each of the 13 personalization techniques in the case of XML-based retrieval and the 7 personalization techniques that could be applied in the case of flat retrieval. These correlations are shown in Table 3, where we also show the standard deviation among the different values used to compute the rankings.

Focusing on XML-based model, the averaged Kendall $\tau$ value is also high, 0.836, with a standard deviation equal to 0.181, higher than in the previous case. This is due to the existence of three personalization methods where the correlations are not so high (p-HRR, CAS and CAS-or, see Table 3). As before, the performance of the different profile configuration under these methods is quite stable (the exhibit a very low value of standard deviation). In the same way, similar values has been obtained using different human judgements, with a $\tau_{A/B}$ value of 0.829 with standard deviation of 0.154.

With respect to BM25 and VSM models, we obtain averaged values of $\tau_{BM25} = 0.735$, $\sigma_{BM25} = 0.186$, and $\tau_{VSM} = 0.636$, $\sigma_{VSM} = 0.322$. The explanation is the same, ASPIRE is able to rank with guarantee the different profile's configurations only in those situations where the impact is relevant.

Therefore, it seems that when the differences in performance between the different user profile configurations of a given personalization method are important (greater standard deviation), ASPIRE is able to discriminate among them. When these differences are small (low standard deviation), then it is not so critical to accurately distinguish which are the better user profile configurations.

An alternative to tackle these situations where is difficult to rank among the different approaches could be to obtain more data by increasing the number and types of queries. Although this approach would be expensive when using real users, it is not when using ASPIRE.

## 5.2 How much topkRel matters? Following Sieg et al. guidelines.

To conclude the experimentation we would like to show the performance of ASPIRE when using the guidelines proposed by Sieg et al. [42] which is, as pointed out in Section 3.1, the closest approach to our proposal. In their work, the relevance criteria is to consider a document as relevant if it is classified under the ODP ontology concept being simulated, and not relevant otherwise. Our relevance criteria is similar, considering a document as relevant if it belongs to the area(s) of interest the given user profile represents, but not only that, but also if it has been retrieved by the IRS among the first *topkRel* results.

We have relaxed this last relevance criteria restriction (the main difference with Sieg et al. criteria) to see if it is really important or not. In order to perform this comparison we will use as ground truth the results obtained with the user study and Garnata. To simulate the lack of this restriction we have established $topkRel = 1500$, which is the Garnata maximum number of retrieved results for a given query. This approach will be denoted as ASPIRE_S, to emphasize that we are following the Sieg et al. guidelines.

In this experiment we will follow the same steps than in Section 5.1, i.e. we first perform a comparison between relevance assessments and then we will evaluate the retrieval performance. As we will see, ASPIRE_S is including a great amount of noise into the relevance assessments, being therefore hard to obtain accurate results.

### *5.2.1 Relevance assessments comparison*

As might be expected, the number of ASPIRE_S raw relevance assessments has grown a lot with the change from $topkRel = 100$ to $topkRel = 1500$, since *topkRel* limits the number of potentially relevant results. The user study number of relevance assessments is still 2,362, while the number of current ASPIRE_S relevance assessments is 17,373 (before it was 1,965). This introduces a large number of false positives in the ASPIRE_S assessments (around 85% in average) and also, the number of relevance assessments are softly correlated exhibiting a Pearson correlation value of 0.285.

| | pre | rec | F |
|---|---|---|---|
| $\mu$ | 0.190 | 0.753 | 0.221 |
| $\sigma$ | 0.276 | 0.295 | 0.245 |

**Table 4** Averages and standard deviations of precision, recall and F across the 126 evaluation triplets with $topkRel = 1500$.

In order to measure the new automatic assessments we will consider the average results across all the evaluation triplets, of precision, recall and F measure, together with the standard deviations, displayed in Table 4. We can still observe large deviations (quite different behaviour depending on the query and the profile being evaluated) together with a notable deterioration of precision and F measure. Thus, only around 20% of the simulated relevance assessments are correct. However, we can observe a much higher recall, which is reasonable since we let much more relevance assessments come into play. We will focus on the performance's comparison between the results using the user study and Sieg-based relevance criteria, which in this case seems more difficult considering the previous overlap degree and number of false positives.

### *5.2.2 Retrieval performance evaluation results comparison*

This section also pursues the same two objectives of Section 5.1.2, i.e. 1) to test whether ASPIRE_S should be considered as a reliable approach in the evaluation of personalized IRS, and 2) to show whether ASPIRE_S is able to rank properly the personalization approaches and the different profile representations.

**Is ASPIRE_S a reliable evaluation approach?**

Firstly, we consider independently each one of the 157 configurations of personalization techniques and profile parameters. For each configuration we focus on the quality of the rankings obtained for each evaluation triplet, trying to measure how the 126 ASPIRE_S-based NDCGs correlates with the NDCGs obtained using the real user assessments.

As a resume, the Pearson correlation values vary in the range $[-0.051, 0.655]$, with an average correlation equal to 0.325 and a relatively large standard deviation of 0.185. These correlation coefficients are plotted in Figure 5 (in the $y$ axis) against the averaged NDCG values from the user study (in the $x$ axis). In this figure we can see how the results of the performance comparison are almost randomly distributed in the range $[-0.051, 0.655]$ (compare this figure with Figure 3).

Once we have evaluated the intra-configuration comparison, now we will focus on how these configurations relate to each other. In this case, we consider an evaluation matrix with 157 cells, but now each of these cells represents the average NDCG obtained after running the 126 evaluation triplets under a given configuration.

Figure 6 plots the user study averaged NDCG values ($x$ axis) against the corresponding values obtained from ASPIRE_S ($y$ axis), for the 157 evaluation

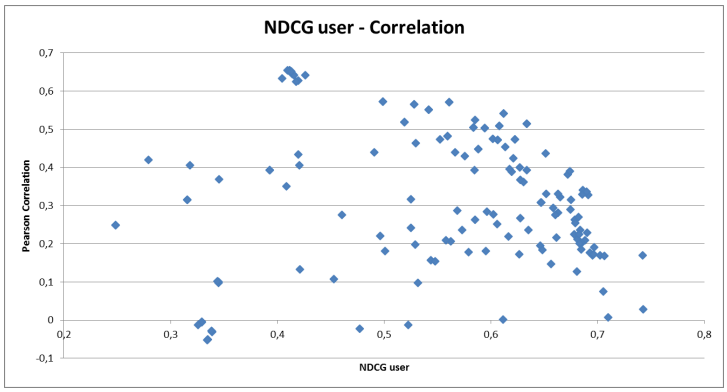


**Fig. 5** NDCG user study-ASPIRE_S correlations (y axis) against the averaged NDCG values from the real study (x axis), with $topkRel = 1500$.

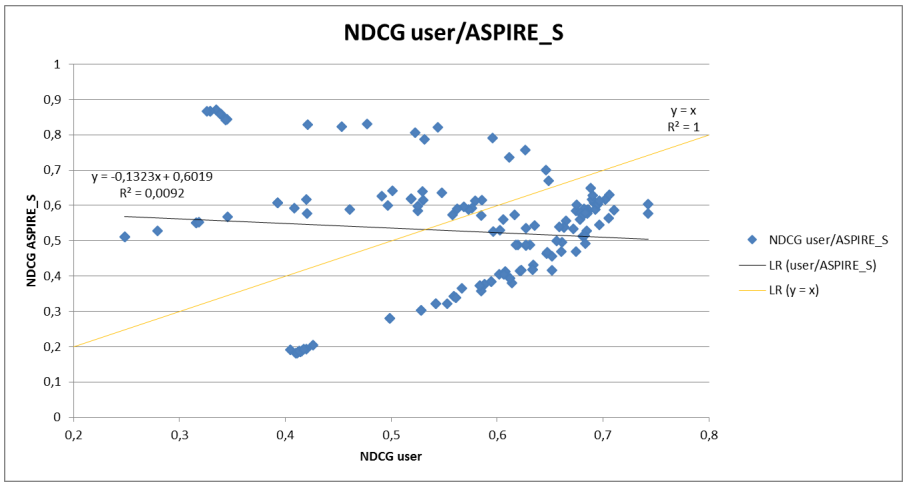


**Fig. 6** Averaged NDCG values from ASPIRE_S (y axis) against the averaged NDCG values from the real study (x axis), for each one of the 157 personalization techniques-user profile configuration parameters combinations with $topkRel = 1500$.

matrix cells (compare it with Figure 4. In this case, none of the conclusions drawn from our proposed ASPIRE framework are fulfilled since the current ASPIRE_S results are not close to the user study ones. The linear regression labelled as *LR (user/ASPIRE_S)* shows an R-squared value of 0.0092 (0.096 Pearson correlation coefficient), i.e., both variables are almost independent. We also can see that there is a big number of results on both sides of the ideal fit line, labelled as *LR (y = x)*, sometimes overestimating and sometimes underestimating the real different personalization techniques performance.

| $n$ | 5 | 5 | 5 | 10 | 10 | 10 | 20 | 20 | 20 | 40 | 40 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0.33 | 0.66 | 0.99 | 0.33 | 0.66 | 0.99 | 0.33 | 0.66 | 0.99 | 0.33 | 0.66 | 0.99 |
| $\tau$ | 0.447 | 0.221 | 0.013 | 0.247 | 0.013 | -0.091 | 0.091 | -0.143 | -0.013 | -0.091 | -0.065 | 0.039 |
| $\sigma_S$ | 0.188 | 0.192 | 0.194 | 0.192 | 0.196 | 0.197 | 0.191 | 0.193 | 0.193 | 0.192 | 0.191 | 0.188 |

**Table 5** ASPIRE_S: Kendall tau correlations of the personalization techniques for each one of the 12 combinations of the user profiles configuration parameters: number of expanded terms, $n$, and normalization factor, $p$.

| method | QE | NQE | HRR | SRR | IRR | I-HRR | p-HRR | CAS | CAS-or | NQE+m | HRR+m | SRR+m | IRR+m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 1.000 | 0.121 | -0.818 | -0.636 | -0.576 | 0.818 | -0.424 | 0.545 | 0.606 | 0.788 | 0.970 | 0.939 | 0.939 |
| $\sigma_S$ | 0.046 | 0.040 | 0.082 | 0.073 | 0.059 | 0.006 | 0.011 | 0.030 | 0.028 | 0.062 | 0.077 | 0.045 | 0.045 |

**Table 6** ASPIRE_S: Kendall tau correlations of the user profiles for each one of the 13 personalization techniques.

**Stability of rankings of ASPIRE_S-user study personalization techniques and user profile configurations using the new relevance criteria**

The next step is to test whether we may trust on the systems ranking provided by ASPIRE_S. As in Section 5.1.2 we use Kendall's $\tau$ rank correlation to look at the stability of rankings.

*Selecting the best personalization techniques for a fixed profile:*

We are going to compute the rank correlations between the real and simulated personalization techniques results, for each of the 12 user profile configurations. These correlations are displayed in Table 5. We also include the standard deviation of the NDCG over the different personalization techniques using ASPIRE_S. An averaged Kendall $\tau$ value of 0.056 with a large standard deviation of 0.172 is obtained from the 12 user profile configurations.

These low values and large deviation show that the ASPIRE_S and the user study rankings are very different, independently of the user profile configuration being considered. Therefore, we can not ensure that ASPIRE_S is a reliable method to discriminate between (better and worse) personalization methods.

*Selecting the best user profile configurations for a fixed personalization method:*

We are going to compute the rank correlations between the real and simulated different user profile configurations results, for each of the 13 personalization techniques. These correlations are shown in Table 6. In this case the averaged Kendall $\tau$ value is also low, 0.329, and the standard deviation very high and equal to 0.699. This low average value and large deviation show that there are big differences between the different personalization techniques. Although the $\sigma_S$ values are low, we can not ensure that ASPIRE_S is a reliable method to discriminate between (better and worse) user profile configurations.

To conclude this section, and considering all the results obtained during the comparison of ASPIRE_S with the results obtained using an user study (which considers the real interaction of the users with the system), we can

state that the guidelines proposed by Sieg et al. seems to be not useful for an automatic evaluation of a personalized retrieval system.

## 6 Concluding Remarks

In this paper we have faced the difficult problem of personalized IRS evaluation. Without any doubt, the inclusion of personalization is every day more and more frequent in a high variety of services. This tendency shows the importance of being able to build efficient and robust personalization techniques to be part of these services. The evaluation step of any personalized system is a crucial stage in their development and improvement. Indeed, high efforts are made to evaluate personalized systems. We have reviewed several methodologies for the evaluation of personalized IRS in the literature, but all of them have some disadvantages in one way or another.

Considering the previous facts, we have proposed an automatic evaluation methodology for personalized IRS. This methodology joins the advantages of the system-centred and user-centred evaluation approaches producing repeatable, comparable and generalizable results together with the inclusion of the user context within the evaluation process. We must specify that the proposed evaluation approach is focused in maximizing the retrieval effectiveness, leaving aside the evaluation of the user-IRS interaction. The only requirements to use ASPIRE is to have a document collection where its documents (at least part of them) are able to be classified into different categories and a suitable set of queries for this collection.

Moreover, we have validated ASPIRE by comparing its results with those obtained from a carried out user study. Not so much evaluation approaches present this validation process which, in our opinion, is a key factor to trust on the proposed evaluation methodology.

Some ASPIRE reliability metrics have been proposed regarding both, the generated relevance assessments and the evaluation of the retrieved results. Although the simulated relevance assessments are not completely similar to the ones obtained from the user study (around 75% of the real relevance assessments are compatible with the basic assumption used to generate the simulated relevance assessments, and there is an overlap degree of around 50% between the real and simulated relevance assessments), they are good enough to get very similar evaluation results. Figure 4 is very clarifying, showing Pearson correlation values greater than 0.914 between both sets of results.

We have also shown how ASPIRE may be used to select the best personalization techniques from a set of them, or the best user profile configurations for a given technique. The high correlation values between the rankings obtained, for different personalization methods and different profile configurations, by ASPIRE and the user study, give an idea of the expected reliability of these selections.

We should also mention that in our evaluation tests we have used a very heterogeneous set of personalization approaches, ranging from very good to

very bad ones, obtaining good results in all the cases. In addition, ASPIRE has been tested with Garnata, BM25 and Vector Space models showing similar results and reinforcing the fact that ASPIRE is robust and, at the same time, is independent on the type of collection used (XML or flat), being applicable in any of these circumstances. This demonstrates that ASPIRE is not only a reliable evaluation approach but also robust. In this line, we have compared also with a state-of-the-art approach, very similar to our proposal. In this case we have shown that we have to be very cautious with the results derived in the case we follow the guidelines proposed by Sieg et al.[42].

On the other hand, it should be clear that ASPIRE does not pretend to completely replace user studies, since it is very important to collect qualitative information about the IRS from real users. It rather should be considered as an easy, fast and reliable alternative to them. To have a reliable and robust evaluation methodology is a very good resource, specially indicated for the first stages in the development of personalization techniques, or when a user study is not possible due to any circumstances, such as the lack of resources or time, or for example to pre-analyze the expected performance for those queries that should be used in a real user study. ASPIRE also helps to make final user studies experimentation more worthwhile, by limiting the number of personalized IRS configurations that users should evaluate.

As future work, we would like to extend ASPIRE to also incorporate the user-IRS interactions into the automatic evaluation process.

**Acknowledgements** This paper has been supported by the Spanish "Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía" and the "Ministerio de Ciencia e Innovación" under the projects P09-TIC-4526 and TIN2011-28538-C02-02, respectively.

## References

1. G.D. Abowd, A.K. Dey, P.J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. Handheld and Ubiquitous Computing, LNCS 1707, pp. 304-307, 1999.
2. J. Allan. Hard track overview in TREC 2003: High accuracy retrieval from documents. Proceedings of the 14th Text Retrieval Conference, Gaithersburg-Maryland, USA, 2005.
3. L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. Proceedings of the 30th ACM SIGIR conference on research and development in information retrieval, Amsterdam, The Netherlands, pp. 455-462, 2007.
4. P. Baiely, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. Relevance assessment: Are judges exchangeable and does it matter?. Proceedings of the 31st ACM SIGIR conference on research and development in information retrieval, Singapore, pp. 667-674, 2008.
5. P. Borlund, and P. Ingwersen. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. Proceedings of the 21st ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia, pp. 324-331, 1998.
6. P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Journal of Information Research, 8(3):152, 2003.
7. C. Buckley, and E.M. Voorhees. Retrieval evaluation with incomplete information. Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, pp. 25-32, 2004.

8. K. Bystrom, and K. Jarvelin. Task complexity affects information seeking and use. Information Processing and Management, 31(2): 191-213, 1995.
9. L.M. de Campos, J.M. Fernández-Luna, and J.F. Huete. Using context information in structured document retrieval: An approach using influence diagrams. Information Processing and Management, 40(5): 829-847, 2004.
10. L.M. de Campos, J.M. Fernández-Luna, and J.F. Huete. Improving the context-based influence diagram model for structured document retrieval: removing topological restrictions and adding new evaluation methods. Lecture Notes in Computer Science: Advances in Information Retrieval, 3408, pp. 215-229, 2005.
11. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, and A.E. Romero. Garnata: An information retrieval system for structured documents based on probabilistic graphical models. Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Paris, France, pp. 1024-1031, 2006.
12. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C. Martín-Dancausa, and A.E. Romero. New utility models for the Garnata information retrieval system at INEX'08. Lecture Notes in Computer Science: Advances in Focused Retrieval, 5631, pp. 39-45, 2009.
13. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C. Martín-Dancausa, A. Tagua-Jiménez, and C. Tur-Vigil. An integrated system for managing the andalusian parliament's digital library. Program-Electronic Library and Information Systems, 43(2), 156-174, 2009.
14. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, and E. Vicente-López. XML search personalization strategies using query expansion, reranking and a search engine modification. Proceedings of the 28th ACM Symposium on Applied Computing, Coimbra, Portugal, pp. 872-877, 2013.
15. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, and E. Vicente-López. Using personalization to improve XML retrieval. IEEE Transactions on Knowledge and Data Engineering, to appear. DOI: 10.1109/TKDE.2013.75.
16. B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, and J. Allan. Evaluation over thousands of queries. Proceedings of the 31st ACM International Conference on Reseach and Developments in Information Retrieval, Singapore, pp. 651-658, 2008.
17. B. Carterette and I. Soboroff. The effect of Assessor Error on IR System Evaluation. Proceedings of the 33rd ACM International Conference on Reseach and Developments in Information Retrieval, Geneva, Switzerland, pp. 539-546, 2010.
18. D.N. Chin. Empirical evaluation of user models and user-adapted systems. User Modeling and User-Adapted Interaction, 11(1-2): 181-194, 2001.
19. C.W. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. Vol. 1 - Design. ASLIB Cranfield Project. Technical Report, 1966.
20. M. Daoud, L. Tamine-Lechani, and M. Boughanem. A contextual evaluation protocol for a session-based personalized search. Proceedings of the 2nd Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE) in conjunction with the 32nd European Conference on Information Retrieval, Toulouse, France, 2009.
21. G. Demartini, and S. Mizzaro. A Classification of IR Effectiveness Metrics. Lecture Notes in Computer Science: Advances in Information Retrieval, 3936, pp. 488-491, 2006.
22. A. Díaz, A. García, and P. Gervás. User-centred versus system-centred evaluation of a personalization system. Information Processing and Management, 44(3): 1293-1307, 2008.
23. C. Ding, and J.C. Patra. User modeling for personalized Web search with self-organizing map. Journal of the American Society for Information Science and Technology, 58(4): 494-507, 2007.
24. Z. Dou, R. Song, and J.R. Wen. A large-scale evaluation and analysis of personalized search strategies. Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, pp. 581-590, 2007.
25. D. Elsweiler, D.E. Losada, J.C. Toucedo, and 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, pp. 25-34, 2011.
26. M.R. Ghorab, D. Zhou, A. O'Connor and V. Wade. Personalised Information Retrieval: survey and classification. User Modeling and User-Adapted Interaction, 23: 381-443, 2013.
27. D. Harman. Overview of the fourth text retrieval conference (trec-4). Proceedings of the 4th Text Retrieval Conference, Gaithersburg-Maryland, USA, 1995.

28. P. Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. Journal of Documentation, 52(1): 3-50, 1996.
29. P. Jaccard. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société vaudoise des Sciences Naturelles, 37: 547-579, 1901.
30. K. Jarvelin, and J. Kekalainen. Cumulative gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4): 422-446, 2002.
31. F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering, 16(1): 28-40, 2004.
32. J. Mostafa, S. Mukhopadhyay, and M. Palakal. Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering. Information Retrieval, 6(2): 199-223, 2003.
33. R. Nuray, and F. Can. Automatic ranking of retrieval systems in imperfect environments. Proceedings of the 26th ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, pp. 379-380, 2003.
34. S. Pal, M. Mitra and J. Kamps. Evaluation Effort, Reliability and Reusability in XML Retrieval. Journal of the American Society for Information Science and Technology, 62(2): 375-394, 2011.
35. D. Petrelli. On the role of user-centred evaluation in the advancement of interactive information retrieval. Information Processing and Management, 44(1): 22-38, 2008.
36. V. Ramesh, Robert L. Glass, and Iris Vessey. Research in computer science: an empirical study. Journal of Systems and Software, 70(12): 165-176, 2004.
37. K. van Rijsbergen. Information Retrieval, London, England, Butterworths & Co. Ltd., 1979.
38. T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval 11(5), pp. 447-470.
39. M. Sanderson and I. Soboroff. Problems with Kendall's tau. In Proceedings of the 30th ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, pp. 839-840, 2007.
40. E. Santos Jr, Q. Zhao, H. Nguyen, and H. Wang. Impacts of user modeling on personalization of information retrieval: an evaluation with human intelligence analysts. Proceedings of the 4th workshop on the evaluation of adaptive systems (held in conjunction with the 10th International Conference on User Modeling), Edinburgh, UK, pp. 27-36, 2005.
41. T. Saracevic. Relevance: A review of and a framework for thinking on the notion in Information Science. Journal of the American Society for Information Science, 26(6): 321-343, 1975.
42. A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. Proceedings of the 16th ACM International Conference on Information and Knowledge Management, Lisbon, Portugal, pp. 525-534, 2007.
43. I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. Proceedings of the 24th ACM SIGIR conference on Research and development in information retrieval, New Orleans, USA, pp. 66-73, 2001.
44. A. Spink, S. Ozmutlu, H.C. Ozmutlu, and B.J. Jansen. US versus european web searching trends. Proceedings of the 25th ACM SIGIR conference on research and development in information retrieval, Tampere, Finland, pp. 32-38, 2002.
45. B. Steichen, H. Ashman, and V. Wade. A comparative survey of personalised information retrieval and adaptive hypermedia techniques. Information Processing and Management, 48(4): 698-724, 2012.
46. K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. Proceedings of the 13th International Conference on World Wide Web, Manhattan, NY, USA pp. 675-684, 2004.
47. M. Taghavi, A. Patel, N. Schmidt, C. Wills, and Y. Tew. An analysis of web proxy logs with query distribution pattern approach for search engines. Computer Standards & Interfaces, 34(1): 162-170, 2012.
48. L. Tamine-Lechani, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowledge and Information Systems, 24(1): 1-34, 2009.
49. X. Tao, Y. Li, and N. Zhong. A personalized ontology model for web information gathering. IEEE Transactions on Knowledge and Data Engineering, 23(4): 496-511, 2011.

50. A.H. Turpin, and H. William. Why batch and user evaluations do not give the same results. Proceedings of the 24th ACM SIGIR conference on research and development in information retrieval, New Orleans, USA, pp. 225-231, 2001.
51. R.W. White, I. Ruthven, J.M. Jose, and C.J. Van Rijsbergen. Evaluating implicit feedback models using searcher simulations. ACM Transactions on Information Systems, 23(3): 325-361, 2005.
52. R.W. White. Contextual simulations for information retrieval evaluation. Proceedings of the 2nd ACM SIGIR Workshop on Information Retrieval in Context, Sheffield, UK, pp. 27-28, 2005.
53. Y. Yang, and B. Padmanabhan. Evaluation of online personalization systems: a survey of evaluation schemes and a knowledge-based approach. Journal of Electronic Commerce Research, 6(2): 112-122, 2005.
54. J. Zobel. How reliable are the results of large-scale information retrieval experiments?. Proceedings of the 21st ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia, pp. 307-314, 1998.