Personalized Diversification of Search Results

David Vallet and Pablo Castells
Universidad Autónoma de Madrid
Escuela Politécnica Superior, Departamento de Ingeniería Informática
{david.vallet, pablo.castells}@uam.es

ABSTRACT

Search personalization and diversification are often seen as opposing alternatives to cope with query uncertainty, where, given an ambiguous query, it is either preferable to adapt the search result to a specific aspect that may interest of the user (personalization) or to regard multiple aspects in order to maximize the probability that some query aspect is relevant to the user (diversification). In this work, we refute this antagonistic view, and hypothesize that these two directions may in fact be effectively combined and enhance each other. We research the introduction of the user as an explicit random variable in state of the art diversification methods, thus developing a generalized framework for personalized diversification. In order to evaluate our hypothesis, we conduct an evaluation with real users using crowdsourcing services. The obtained results suggest that the combination of personalization and diversification achieves competitive performance, improving the baseline, plain personalization, and plain diversification approaches in terms of both diversity and accuracy measures.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval – retrieval models, information filtering.

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Diversity, personalization, search.

1. INTRODUCTION

Maximizing the total returned relevance in response to an information need has been traditionally the main fundamental principle underlying research and development in the Information Retrieval (IR) field. More recently, strands of research in the community considered other angles of the practical effectiveness of the output of retrieval systems, and the role that diversity, in particular, plays along with relevance in delivering effective value to the users of such systems [5,6,22]. The value of IR diversity is motivated by the uncertainty involved in a single query as the only evidence of the user information need [1,7,15,16]. The assumption is that different users may mean slightly (or even quite) different things by the same query expression. In the absence of further knowledge or observations about the actual user need beyond the explicit query, the goal of the diversity approach is to satisfy as many users as possible with a single result set. More specifically, diversification aims to minimize the number of totally unsatisfied users, trading degrees of satisfaction in exchange for increasing the size of the –at least minimally– satisfied population. The implicit assumption is that the loss involved in zero vs. small utility is far more significant than the loss involved between degrees of satisfaction. In other words, relevance follows a diminishing marginal returns pattern, where the gain obtained from increased relevance decreases fast with the number of relevant documents [14,21].

This view contrasts with the principles behind personalization. IR personalization takes the same starting point and addresses the same problem as diversity, where a sole query is viewed as an insufficient expression of the full and precise user need. But in contrast with diversification, the personalization approach strives to get further knowledge beyond the query, so as to explicitly tackle and overcome –to the degree this is possible– the initial uncertainty [9,11,12,18]. The uncertainty is reduced by extracting and exploiting further information from a wide variety of observations and resources, such as session feedback from the user, long-term records of user behavior, user stereotypes, social links, and other sources for user context and interest modeling. While in the diversity approach the system accepts and adapts its behavior to a situation of uncertainty, personalization tries to change this situation by enhancing the system knowledge about the user need. Rather than aiming to satisfy as many users as possible, personalization aims to build, in a way, a sense of who the user is, and maximize the satisfaction for the specific sensed user.

From these considerations, it might seem natural to see diversity and personalization as contradictory approaches. One might understand that by opting for diversity, the adaptation of results to specific users has been implicitly given up on. Conversely, if the query uncertainty is reduced by grasping the user context, the reason for diversification may have been done away with to some extent. In this paper we contend, however, that diversity and personalization are not necessarily incompatible or mutually exclusive goals, but on the contrary they may complement and in fact enhance each other.

Our research is motivated by the understanding that personalization is a process that involves itself a great deal of uncertainty, considerably higher, in fact, than ad-hoc search. The observation of implicit evidence of interest from users involves indeed much higher degrees of ambiguity and incompleteness than regular queries. The interpretation of the input for personalization is an explicit and inherent –and considerably difficult– part of the personalization task. The same observed contextual cues may fit several plausible interpretations whereby, by the same rationale as for search diversification, diversifying the system's assumptions may help minimize the ratio of severe failures in guessing the implicit user interests. Moreover, the observations of user behav-

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

ior and preferences available to the system usually cover a small fraction of the full range of user interests, whereby context representations are far more incomplete than queries as expressions of user needs. Furthermore, the same as search diversity considers notions of aspects and facets in user queries and needs, user interests are typically heterogeneous, whereby the consideration of user profile aspects for diversification can be as natural as is the introduction of query aspects.

In this paper we develop a generalization of existing diversification approaches, by adding a personalization component into them. Specifically, we introduce a user random variable in probabilistic diversification models previously developed in the literature on IR diversity [1,16]. We describe in detail a full approach to develop the resulting framework into computable components. We report experiments with real users where our framework is compared to partial approaches that apply either personalization or diversification alone. We also compare our approach to a previous approximation to diversity and personalization combined. Our approach compares favorably to all such alternatives, thus providing an empiric validation of our hypothesis.

The rest of the paper is organized as follows. In section 2, we further discuss the possible complementarity between diversity and personalization. In section 3 we propose the new personalized diversity models. In section 4 we present a possible framework instantiation of the model. In section 5, we describe the evaluation framework. Experimental results are presented in section 6, followed by a brief discussion of related work in section 7. Finally, we conclude in section 8 and indicate possible future directions of the presented work.

2. DIVERSITY AND PERSONALIZATION, AN INTEGRATED VIEW

The vision of combining diversity and personalization opens a rich area for research, barely explored to date. Diversity and personalization are complementary and can play together in different ways. The likes and preferences of people typically comprise quite diverse areas of interest, therefore not all of one's preferences should come into play in a specific retrieval task. The activation of the right preference area is critical for the performance of a personalization technique. Areas of user interest are often related to different contexts (work, leisure, time, tasks, situations, etc.). Choosing the wrong area is the typical cause of personalization intrusiveness – the Achilles' heel of personalization: the inferred preferences may be right, but they are applied at the wrong time.

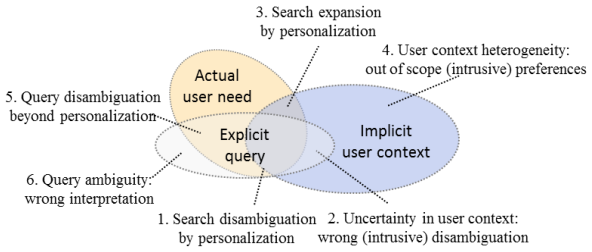

Figure 1. Diversity and personalization as complementary dimensions.

Diversity and personalization may interact and be combined in different ways to enhance each other. For instance, the personalization system may use a diversified selection of user interests, to cope with the uncertainty about which are most pertinent to the current context. Likewise, personalization can improve the effectiveness of aspect weighting in diversification, by favoring query interpretations which are predicted to be more related to each specific user. Besides interacting with each other, personalization and diversification can of course be directly applied to the search system. Which combination of the three components (diversity, personalization, and search system) is more appropriate may depend on a number of factors, including the characteristics of the query, the system knowledge about the user context, and relations between both. For instance, to the extent that user preferences are unrelated to the query, it may be advisable to apply them to the diversification component, and to little or no extent to the accuracy-seeking component. And so forth.

As a noteworthy example of the combination of diversity and personalization, diversity has been an actively researched topic in the Recommender Systems field in the last decade (see e.g. [19,20,23,24]). Recommendation is a genuine personalized retrieval task, but interestingly, one in which the explicitly query is absent, and only implicit evidence of user interests is available to the retrieval system. To this respect, our research vision is new in relation to prior work in that we deal with the problem of handling user preferences, the diversity dimension, and a query, altogether. Also interestingly, recommendation diversity has only been researched by similarity-based methods, devoid of an explicit representation of user intents (with the exception of [23]). In contrast, we shall consider an explicitly intent-oriented approach.

Figure 1 illustrates the complementarity of diversity and personalization in query-based search. Personalization is useful to focus the scope of diversification within a closer scope to general user interests (area 1 in the figure). Still, a fraction of user-neutral diversity beyond this may be appropriate (area 5). This fraction can be made variable and dependent on the degree of uncertainty in the system's knowledge about user preferences. Similarly, depending on this uncertainty, the diversification of user preferences should be more appropriate or less –e.g. diversity of personalization is less appropriate if the area 1 in the figure is much larger than the area labeled 2, and it is more beneficial in the opposite situation.

3. PERSONALIZED DIVERSITY

The approach we investigate here consists of the introduction of the user as an explicit random variable in the diversification models. We take as starting point the original formulations of two representative intent-oriented diversification algorithms, namely IA-Select [1] and xQuAD [15,16], characterized by using an explicit representation of query intents for diversification. Other diversification schemes not explicitly using aquery aspect space [5,6,21,22] might be generalized as well towards personalized versions, which we envision as future work.

As a starting point and general principle in our approach, a user random variable is introduced in every model component as a complement of the query and aspect variables in the representation of user needs. In practical terms, with respect to the original non-personalized methods, the resulting expressions give rise to new and/or additional model terms requiring handling and estimation, as we shall see. We shall deal with and provide approaches for them. On the other hand, the user-based formulation provides for considering independence assumptions at different points, which result in different variants of the framework, standing for different angles on personalized diversification, as we shall see.

3.1 Personalized IA-Select

The IA-Select and xQuAD diversification algorithms greedily rerank search results by maximizing an objective function. Based on the description in [1], the objective function of IA-Select can

be expressed as:

$$f_S(d) = \sum_c p(c|q)p(q|d)p(c|d) \prod_{d' \in S} \left(1 - p(q|d')p(c|d')\right)$$

where S is the incremental subset of diversified documents from the original result set. We generalize this formulation to a personalized version by introducing the user random variable:

$$f_S(d,u) =$$
$$= \sum_c p(c|q,u)p(q|d,u)p(c|d,u) \prod_{d' \in S} \left(1 - p(q|d',u)p(c|d',u)\right)$$

In this expression we have three main components involved:

- A personalized search system: $p(q|d,u)$.
- The personalized query aspect distribution: $p(c|q,u)$.
- The personalized aspect distribution over documents: $p(c|d,u)$.

The first component represents a baseline personalized search system. If we only have a non-personalized system, we may develop this term using Bayes' rule as:

$$p(d|q,u) = \frac{p(q|d,u)p(d|u)}{p(q|u)} \sim \frac{p(q|d)p(d|u)}{\sum_{d'} p(q|d')p(d'|u)} =$$
$$= \frac{p(d|q)p(d|u)}{\sum_{d'} p(d'|q)p(d'|u)} = C_1(q,u)\, p(d|q)p(d|u) \quad (1)$$

where we have assumed q and u are conditionally independent given a document, and we assume a uniform document prior. The expression is thus reduced to the combination of a non-personalized search system, represented by $p(d|q)$ –the same as in the original IA-Select–, and a pure personalization factor represented by $p(d|u)$. The latter factor captures how much user u is predicted to like document d (regardless of a specific search task), expressed as how likely it is to observe the document given the user. This is what a personalization system typically provides for, based on some mechanism for extracting or learning user preferences. We shall provide further details for a particular implementation of the baseline search and personalization components in our experiments in section 4.

Now the aspect-document distribution component of personalized IA-Select can be developed as:

$$p(c|d,u) = \frac{p(d|c,u)p(c|u)}{p(d|u)} \sim \frac{p(d|c)p(c|u)}{\sum_{c'} p(d|c')p(c'|u)} =$$
$$= \frac{p(c|d)p(c|u)}{\sum_{c'} p(c'|d)p(c'|u)} = C_2(d,u)\, p(c|d)p(c|u) \quad (2)$$

where we assume a uniform aspect prior, and conditional independence between documents and users given a query aspect.

A natural way to estimate the personalized aspect distribution $p(c|u)$ is by marginalization over the set of documents:

$$p(c|u) = \sum_d p(c|d,u)p(d|u) \sim \sum_d p(c|d)p(d|u)$$

where we assume conditional independence between aspects and users given a document.

Finally, for the personalized query aspect distribution $p(c|q,u)$ several derivations are possible. A convenient one is to develop $p(c|q,u)$ by marginalizing over the set of documents, because it allows taking advantage of the computation of the two previous top-level components in equations 1 and 2:

$$p(c|q,u) = \sum_d p(c|d,q,u)p(d|q,u) \sim \sum_d p(c|d,u)p(d|q,u) \quad (3)$$

where we assume the conditional independence of query aspects and queries given a user and a document. We may also consider $p(c|d,u) \sim p(c|d)$ as a simplified option (assuming conditional independence of aspects and users given a document), using the approximations to $p(c|d)$ discussed in the next section.

3.2 Personalized xQuAD

The objective function of the xQuAD algorithm is formulated in [16] as:

$$f_S(d) = (1-\lambda)p(d|q) + \lambda\, p(d,\bar{S}|q) =$$
$$= (1-\lambda)\, p(d|q) + \lambda \sum_c p(c|q)p(d|c) \prod_{d' \in S} \left(1 - p(d'|c)\right)$$

In the personalized version of this algorithm the objective function is thus generalized to:

$$f_S(d,u) = (1-\lambda)\, p(d|q,u) + \lambda \sum_c p(c|q,u)p(d|c,u) \prod_{d' \in S} \left(1 - p(d'|c,u)\right) \quad (4)$$

This expression involves three main components:

- The personalized search system: $p(q|d,u)$.
- The personalized query aspect distribution: $p(c|q,u)$.
- The personalized, aspect-dependent document distribution: $p(d|c,u)$.

The first two terms were already handled in the previous section (equations 1 and 3). Using Bayes' theorem and equation 2, it can be seen that the third term can be approximated by:

$$p(d|c,u) \sim \frac{p(c|d)p(d|u)}{\sum_{c'} p(c'|d)p(c'|u)} = C_2(c,u)\, p(c|d)p(d|u)$$

where –as in equation 2– we are assuming documents and users are conditionally independent given a query aspect. Again, the estimation of $p(d|u)$ admits different approaches depending on the available data, an example of which we develop in our experiments in section 4. In appendix A, we offer alternative developments of $p(c|q,u)$, which might be more suitable to specific frameworks.

Finally, the term $p(c|d)$ can be estimated in different ways depending on the nature of the query aspect space and available observations. For instance, if c is an Open Directory Project (ODP) category as in [1,14], we may estimate $p(c|d)$ by a text classification method as the probability that document d belongs to class c. Other estimation approaches can be applied for query aspects of a different nature. For instance, $p(c|d)$ can be computed by calling the baseline search system on c and d, as far as c belongs to the space of queries –this is the case in [16], where c are related and suggested queries from a search engine.

3.3 Discussion

The personalized versions of IA-Select and xQuAD provide a framework that supports different views on the combination between diversity and personalization. For instance, the summation $\sum_c p(c|q,u)p(d|c,u)$ in xQuAD can be seen as a marginalization over query aspects for a personalized search model $p(d|q,u)$ (assuming conditional independence of documents and queries given a user and an aspect), that is, a means to diversify user preferences over a query aspect space. This can be seen as a means to tackle the uncertainty about implicit user interests by a diversification strategy, as motivated earlier in section 2.

Conversely, $p(c|q,u)$ can be seen as an enhancement in the estimation of the aspect distribution $p(c|q)$ for diversification, by exploiting available information about user preferences for as-

pects, in order to improve the effect of diversity by making it more user-specific. Furthermore, $p(d'|c,u)$, in the novelty component, enhances the redundancy model by personalizing the utility estimation of seen documents. With slight variations in the formulation, analogous effects can be identified in IA-Select.

On the other hand, the framework flexibly supports different configurations by selectively removing the user variable (i.e. adding user-independence assumptions) from some of its components. This selective removal results into different combinations and views on personalization and diversity. For instance, if we remove personalization from the external baseline system $(1-\lambda)\,p(d|q,u)$ in xQuAD, the result is a personalized diversification of a non-personalized search system. We may do the opposite in order to have a non-personalized diversification of a personalized retrieval system. At a finer granularity, we may combine a personalized aspect distribution $p(c|q,u)$ with a non-personalized redundancy component $p(d'|c)$. And so forth. The variations could even be dynamic and selective, depending e.g. on the reliability of the system knowledge about implicit user interests, the degree of relatedness between recorded preferences and the query at hand, the vagueness of the query, and so forth.

In our experiments, we compare five framework configurations, as we shall see: a) fully personalized diversification, b) non-personalized diversification, c) non-diversified personalization, d) a plain (non-personalized, non-diversified) search system baseline, and e) personalization *after* diversification, the latter in order to compare our framework to a personalized diversification approach reported in prior work [13].

4. FRAMEWORK INSTANTIATION

In this section we introduce a particular instantiation of the proposed personalization and diversification framework. Instantiating our model involves providing estimates for: 1) the personalization component $p(d|u)$; 2) the diversity component $p(c|d)$; and the document relevance component $p(d|q)$. In the following section we offer insight on how to estimate the query relevance component, based on an external search engine, and the personalization component, from a weighted term-based representation of documents and user profiles. In section 4.2, we present estimations more specific to our framework: a social-based representation of documents and users, and the calculation of the diversification component. Finally, section 4.3 addresses a common problem that can arise from the use of external sources to estimate some components, and its implication on the xQuAD approach.

4.1 Model estimations

Similar to e.g. [4], the document generation model can be approximated from a baseline retrieval function by normalizing the score by the sum of all scores of the top *k* documents being diversified:

$$p(d|q) \sim \frac{\bar{s}(d,q)}{\sum_{d' \in R_q} \bar{s}(d',q)} \tag{5}$$

where R_q is the set of documents to be diversified for query q. In our implementation we make use of an external Web search system, estimating $\bar{s}(d,q)$ as a rank-sim normalization [10] of the baseline retrieval function, that is $\bar{s}(d,q) = 1 - \tau(d,q)/|R_q|$, where $\tau(d,q)$ is the position of document d in the order induced by the retrieval system scores $s(d,q)$ for $d \in R_q$.

The user preference model described by $p(d|u)$ can be obtained in many ways, depending on the available document features and user behavior observations. For instance, for text retrieval, we may marginalize $p(d|u)$ over words:

$$p(d|u) = \sum_w p(d|u,w)p(w|u) \sim p(d)\sum_w p(w|d)p(w|u)/p(w) \tag{6}$$

where we assume conditional independence of words and users given a document. The following estimates of the resulting distributions were tested in our experiments:

$$p(w|d) \sim \frac{tf(w,d)}{\sum_{w'} tf(w',d)} \qquad p(w|u) \sim \frac{tf(w,u)}{\sum_{w'} tf(w',u)}$$

$$p(w) \sim \frac{\sum_d tf(w,d)}{\sum_d \sum_{w'} tf(w',d)} \qquad p(d) \sim \frac{1}{|\Delta|}$$

where $tf(w,d)$ is the term count of term w in d, $tf(w,u)$ is the term count in the user profile representation, and Δ is the document collection.

Alternatively, we can estimate the user preference model by an adaptation of the BM25 probabilistic model, similar to the approach presented in [18]:

$$p(d|u) = \sum_{w:tf(w,d)>0} iuf(w)\frac{tf(w,u)\,(k_1+1)}{tf(w,u)+k_1\left(1-b+\frac{b|u|}{\mathrm{avg_u}|u|}\right)} \tag{7}$$

where $iuf(w)$ is the inverse user frequency of term w in the set of users, and $|u|$ is the size of the user profile calculated as $\sum_w tf(w,u)$. We set b and k_1 to the standard values of 0.75 and 2, respectively. In the following section we show how to implement this estimate with user and document profiles based on social annotations.

4.2 Framework-specific estimations

We propose now an instantiation of the personalized and diversification components by exploiting two sources of information. As a source for document and user representations, we use the social bookmarking site Delicious[1], which allows the collective bookmarking of Web pages by user generated tags. This collection of tag-user and tag-document relations, commonly named *folksonomy*, is used in our model to compute the values for $tf(w,u)$ and $tf(w,d)$, where in this case w represents a tag in the folksonomy and the frequency values are, respectively, the number of times a user used the tag in their profile bookmark annotations, and the number of times a tag was used (by any user) to annotate a document. The inverse frequency of a tag, $iuf(w)$ can be computed by counting how many users have a specific tag in their profile. For the latter calculation, we used a small training collection of 3,000 users, randomly sampled from Delicious. The advantages of using a source of this nature for document and user profile representation is threefold: first, it is a simple solution to having user and document profiles represented in the same term space model; second, folksonomies offer a more concise way of representing content, facilitating the computation of our model; and third, it facilitates the evaluation of our approaches, by providing us with a publicly-available user profiling resource, sparing the need for sophisticated profile learning techniques.

Another key element for our model is the estimation of the diversity component $p(c|d)$, which should capture how related a category is to a document. In our experiments we shall follow a content-based approach, by classifying Web pages into the ODP taxonomy. We first considered a Rocchio-based classification approach, as used by Agrawal et al [1], but after close

[1] Delicious, Discover yourself: http://www.delicious.com/

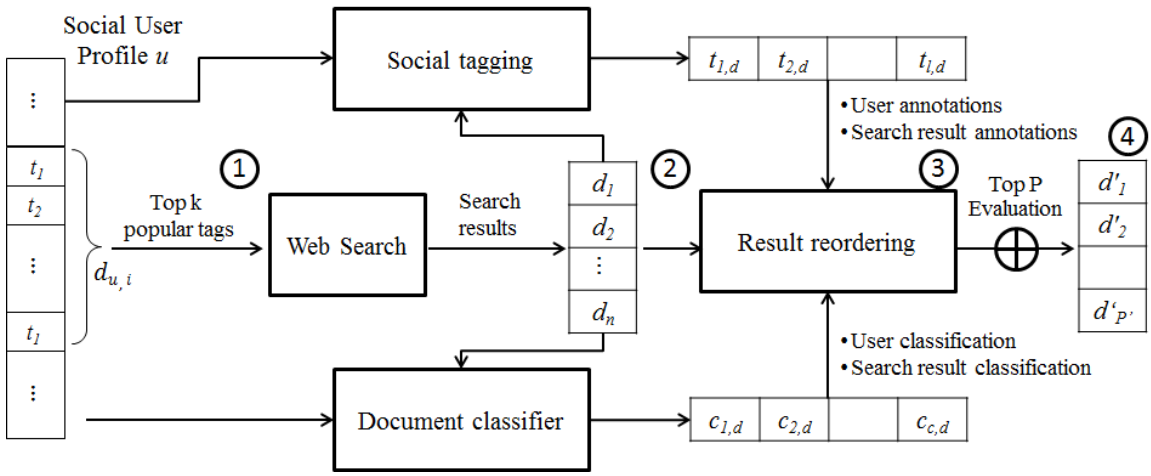

Figure 2. Process of user topic creation, annotation, and evaluation

examination and testing on a development collection, we found it too imprecise to be used as an estimator[2]. As an alternative, we made use of the Textwise[3] ODP classification service which yielded acceptable results on our development collection. The Textwise API returns up to three possible ODP classifications for a document, ranked by a score in [0,1] that reflects the degree of confidence on the classification. We used directly this value as our estimate of $p(c|d)$, after normalization

4.3 Component Normalization

Both the original and the personalized version of the xQuAD scheme use a linear combination of two components, one seeking to procure relevance, and the other seeking diversity (see equation 4). The influence of these two factors can be adjusted by varying the λ parameter. In order for this linear combination to make proper sense, the two components should be comparable –a case of the classic rank fusion problem, which requires a normalization step [10]. Our framework instantiation estimates involve heterogeneous sources and systems, such as an external baseline search system, social annotations, and an external classification service. The use of such third-party tools and resources makes normalization all the more necessary. For this purpose we use a distribution-based normalization approach described by Fernandez et al [8]. Normalization is applied on the two components (relevance and diversity) that are linearly combined, both in the original and the personalized version of xQuAD. The objective function is thus modified to:

$$f_S(d,u) = (1-\lambda)\,\Psi_1\big(p(d|q,u)\big) + \lambda\,\Psi_2\left(\sum_c p(c|q,u)p(d|c,u)\prod_{d'\in S}\big(1-p(d'|c,u)\big)\right)$$

where Ψ_1 and Ψ_2 are the score normalization functions. The normalizers use a histogram of output values from the functions to be normalized, sampling from very simple training data (e.g. no relevance judgments required). The reader can refer to [8] for more details on the normalization method.

5. EVALUATION FRAMEWORK

A proper evaluation of personalization techniques is difficult to achieve with offline data only, and in our case indeed requires relevance judgments from real users and potentially real Web search topics, as well as the availability of some form of user profile representation. We used a crowdsourcing service for this purpose, requiring users of this service –known as *workers*– to have a social profile in Delicious in order to evaluate a number of search topics. We provide further details on this in section 5.1.

We first investigated creating a set of fixed ambiguous topics using Wikipedia's disambiguation pages, which indicate concepts or topics that may have multiple meanings. This methodology was used in previous evaluation schemes in the literature [14,15]. However, after an initial analysis of the worker's feedback, we found that workers were sometimes unfamiliar with the selected ambiguous topics, and were as a consequence unable to provide consistent relevance judgments. As a solution to this problem, we investigated an automatic topic creation methodology, which uses the social profile of the worker in order to find test topics known to the worker. In order to extract topics from the Delicious profile, we adapted to our online evaluation setup an offline evaluation technique initially suggested by Vallet et al. [18].

This topic generation technique creates a test topic from a candidate Web page bookmark, randomly chosen from the worker's Delicious profile. Given the candidate bookmark, a query topic is created, using as keywords the top *K* most popular annotations (tags) assigned by all users to the same bookmark, thus obtaining a keyword based topic representation related to the bookmark. We ensured that the topics were generated by somewhat popular bookmarks by only generating topics from documents that had annotations from at least five different users.

By following the above process there is a much higher chance that the evaluation topic is known by the user (as it is related to the document saved in his profile). When analyzing the workers' feedback, we found a much more positive response to this type of profile generated topics, in terms of both knowledge of the topic, and user engagement in the evaluation process.

The complete topic and result generation process is depicted in Figure 2. The first step **(1)** is to generate a topic suitable for the user to evaluate, following the process described above. By setting the number *K* of popular tags used in the topic generation process, we were able to generate somewhat ambiguous topics, since a larger number of keywords would, in general, generate more specific topics. Given our focus on diversification and

[2] Other alternatives such a Naïve Bayes classifier or a probabilistic estimator were considered, obtaining similar results.

[3] Textwise LLC: http://textwise.com

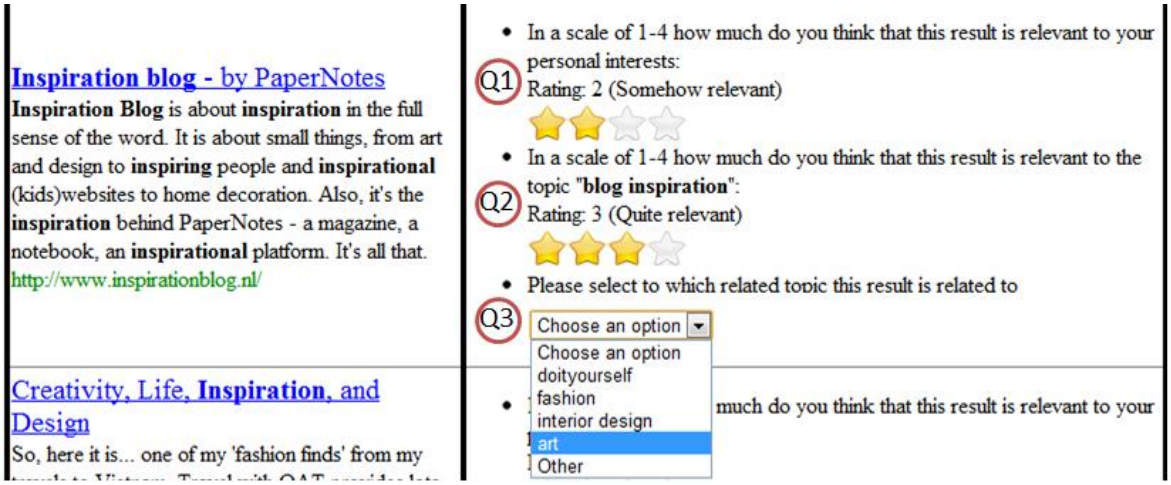

Figure 3. Snapshot of the interactive evaluation interface. Evaluation guidelines have been removed due to space restrictions

personalization on ambiguous topics, we randomly generated an equal amount of topics of size $K = 1$ and $K = 2$ since a large number of keywords would, in general, generate more specific topics. Section 6 further investigates the effect of distinct topic query sizes. The generated query topic was then sent to the Yahoo! Search API[4], from which the top N results were obtained. In our experiments, N was set to 300. We set the location parameter of the search API to the United States, as workers were selected from this region.

In step **(2)**, we extract both social and classification information from each of the documents in the result set. The social annotations were extracted directly from Delicious, by collecting the top 100 most recent annotations given by users. The document topic classification is obtained using the Textwise URL classification service, which returns up to the top 3 ODP most likely categories for each document. The ODP classes were cut down to level 3 in the ODP taxonomy (Top / Level 1 / Level 2 / Level 3). This annotation extraction and document classification process provides the information needed to compute the personalization and diversification estimates of our studied approaches, respectively. It is worth noting that these two services did not provide a complete coverage over the search results. In particular, the Delicious coverage was 57.5% over the top 300 results and went up to a 69.5% coverage over the top 20 results. Textwise had a coverage of 94.7% and returned an average of ~2 categories per document. In order to cope with the incomplete coverage of Delicious, we added a smoothing factor to the computation of $\hat{p}(d|q,u) = \lambda\, p(d|q,u) + (1-\lambda)p(d)$, with $\lambda = 0.5$, so that documents out of coverage would not get a strict zero score.

In step **(3)**, the output of each evaluated approach is used to rerank the Web search results according to personalization and/or diversification features extracted in step 2. The top P documents of each reranking technique are aggregated into a single evaluation list, which is then randomly shuffled and passed onto the evaluation framework. In our experiments, we evaluated the top $P = 5$ results of each algorithm.

5.1 Crowdsourced Evaluation

Crowdsourcing services (e.g. Amazon mechanical turk[5], Crowdflower[6]) are recently proving to be a valuable tool to optimize the necessary resources to evaluate a wide range of information retrieval tasks [2,3]. These evaluations are based on relevance judgments provided by users of the crowdsourcing service. Alonso and Baeza proposed a methodology to obtain relevance assessments for TREC related topics [2] using such crowdsourcing services. In their study, it was concluded that workers can provide relevance judgements with comparable quality to those provided by expert assessors.

Our evaluation framework, however, has to additionally consider the personalization and diversification factors involved in our research. Regarding diversification, Agrawal et al. [1] required workers to manually classify results over a predefined set of ODP categories, representing the subtopics that could possibly be related to a search query. The goal was to obtain a result/topic relevance assessment that could be used to evaluate a diversification approach in terms of both accuracy and diversification metrics. In our setup, in addition to this type of assessment, we also needed to obtain a personal relevance assessment, i.e., to which degree a search result is preferred subjectively by the assessor. This value allows us to measure the accuracy of the personalization techniques with respect to the real interest of the user. Our evaluation framework is based on prior early work [17], which focused on the acquisition of personal assessments from workers, and has been adapted to include diversification factors. The main difference between our evaluation methodology and the one presented by Alonso and Baeza [2,3] is that our assessments are subjective to the worker, thus relevance assessments cannot be compared across workers, as they evaluate specific topics related to their profile, according to their personal interests and their own notion of relevance to the topic.

In prevention for malicious workers, we implemented a simple technique based on the injection of irrelevant results in some of the evaluated topics, which were then used as quality assessment topics to detect workers that were answering questions randomly. This would not detect workers that mark all results as irrelevant, but a close inspection of the gathered results did not show this behaviour.

We also had a number of prerequisites to access the evaluation interface: 1) as mentioned earlier, workers had to have a valid Delicious account;

[4] Yahoo! BOSS API: http://developer.yahoo.com/search/boss/

[5] Amazon mechanical turk: http://www.mturk.com/

[6] Crowdflower: http://crowdflower.com/

Table 1: Diversity metric values for the evaluated approaches. Values in bold indicate the best for each metric. Underlined values indicate a statistically significant difference with respect to the baseline, double underlined values indicate, in addition, a statistical significance with respect xQuAD (Wilcoxon, p < 0.05).

	Topic relevance									**User relevance**								
	Topic			Topic (*K*=1)			Topic (*K*=2)			User			User (*K*=1)			User (*K*=2)		
	ERR-IA@5	α-nDCG@5	S-recall@5	ERR-IA@5	α-nDCG@5	S-recall@5	ERR-IA@5	α-nDCG@5	S-recall@5	ERR-IA@5	α-nDCG@5	S-recall@5	ERR-IA@5	α-nDCG@5	S-recall@5	ERR-IA@5	α-nDCG@5	S-recall@5
Baseline	0.274	0.787	0.500	0.268	0.796	0.493	0.281	0.779	0.513	0.254	0.626	0.475	0.240	0.592	0.460	0.273	0.670	0.498
IA-Select	0.262	0.758	0.463	0.251	0.748	0.451	0.274	0.769	0.475	0.237	0.582	0.426	0.211	0.532	0.398	0.268	0.642	0.458
xQuAD	0.275	0.793	0.510	0.272	0.806	0.509	0.279	0.778	0.511	0.257	0.624	0.478	0.244	0.584	0.459	0.273	0.673	0.502
$Pers_{BM25}$	0.267	0.778	0.486	0.266	0.794	0.490	0.267	0.758	0.481	0.262	0.646	0.483	**0.261**	**0.631**	0.481	0.262	0.664	0.486
$xQuAD_{BM25}$	0.272	0.794	0.516	0.266	0.798	0.509	0.279	0.788	0.526	0.258	0.631	0.492	0.244	0.597	0.470	0.273	0.671	0.520
PIA-Select	0.251	0.742	0.489	0.231	0.723	0.478	0.276	0.767	0.503	0.239	0.593	0.481	0.219	0.546	0.457	0.263	0.650	0.512
$PIA\text{-}Select_{BM25}$	0.279	0.803	**0.543**	0.269	0.793	0.515	**0.291**	**0.818**	**0.579**	**0.273**	0.656	**0.521**	0.260	0.613	**0.489**	**0.290**	0.710	**0.562**
PxQuAD	0.275	0.796	0.503	0.268	0.805	0.501	0.285	0.788	0.510	0.268	0.649	0.489	0.255	0.620	0.478	0.286	0.687	0.506
$PxQuAD_{BM25}$	**0.281**	**0.810**	0.519	**0.275**	**0.818**	**0.519**	0.288	0.802	0.520	0.267	**0.658**	0.496	0.249	0.612	0.460	0.290	**0.716**	0.542

2) the Delicious account was validated by asking the worker to add a random bookmark to their public profile; 3) the Delicious profile had to be at least 3 months old, contain at least 30 bookmarks, and the 30th most recent bookmark had to be at least 15 days old. In general, the use of the quality assessment topics and our prerequisites for accessing the evaluation interface were enough to avoid misbehaving workers.

Figure 3 shows a snapshot of the evaluation interface. The aggregated output of the evaluated approaches is presented as a search result list to the worker, who has to provide individual assessments by answering three questions:

- Q1 (user): a 4-grade scale assessment on how relevant is the result to the user's interests.
- Q2 (topic): a 4-grade scale assessment on how relevant is the result to the evaluated topic.
- Q3 (subtopic): workers assign each result to a specific subtopic related to the evaluated topic.

Q1 provides for measuring the accuracy of the evaluated approaches with respect to the user interest. Q2 is used to evaluate how relevant is the result to the overall search topic –a successful reordering technique will place results high that are assessed as both relevant to the topic and to the user's interests. Finally, Q3 provides a subjective classification of the search result to the possible subtopics related to the search topic. The interface is designed to encourage the reuse of introduced subtopics, so as to facilitate workers to provide a more concise categorization of search results. A tutorial was included in the topic description, which was helpful for workers to understand the concepts of subtopics and how this assignment is performed. We decided to take a different approach from Agrawal et al. [1], who used predefined ODP subtopics for result categorization, for the following reasons: 1) we felt that some ODP categories are difficult to understand, often not self-explanatory, and thus can be confusing for workers for annotation; 2) candidate ODP categories need to be automatically selected by the ODP classifier, which could cause difficulties for workers in case of classification errors. In Agrawal et al.'s study this problem was alleviated by having more than one worker evaluating a single topic, but in our personalized setup topics can only be evaluated by a single worker. After analyzing the workers' assessments and open comments feedback, they seemed to be comfortable with providing their own classification scheme. It is also worth noting that our query generation process produced topics that were known to the worker.

The assessment collection process spanned over a period of four weeks. During this period, we were able to collect information from 35 users that satisfied our prerequisites, collecting assessments for a total amount of 180 topics (median of 4 topics per user) and 3,800 individual results. The user profile and relevance assessment information has been anonymized and made publicly available at *http://ir.eps.uam.es/~david/persdivers/*.

6. EXPERIMENT RESULTS

Nine different approaches were evaluated with the approach described in the previous section: the baseline system, namely the original ranking returned by the Web retrieval engine; two state of the art search diversification approaches, IA-Select [1] and xQuAD [16]; a plain (in the sense of "not diversified") personalized search approach based on social tagging profiles and BM25, presented in [18] ($Pers_{BM25}$); a two stage diversification and personalization approach, suggested by Radlinski and Dumais in [13] which, in our implementation, first applies the xQuAD algorithm and then the $Pers_{BM25}$ technique ($xQuAD_{BM25}$); two different personalized versions of IA-Select, one with a probabilistic calculation of $p(d|u)$ –equation 6– (PIA-Select), and the other using BM25 (equation 7) as an alternative calculation ($PIA\text{-}Select_{BM25}$), as described in section 4.1; two different personalized versions of xQuAD, differing on the same alternative as in PIA-Select for the calculation of $p(d|u)$.

In order to evaluate for diversity, we use three well known metrics: the intent aware version of expected reciprocal rank (ERR-IA), α-nDCG [7], and subtopic recall (S-recall) [22]. For accuracy, we use nDCG and precision. User assessments were graded from one to four, where a value greater than one was assumed as relevant. The nDCG metric made use of the graded values for each of these approaches, all metrics take this cut off point.

Table 2: Accuracy metrics for evaluated approaches. Values in bold indicate the best performing values. Underlined values indicate a statistically significant difference with respect to the baseline, double underlined values indicate, in addition, a statistical significance with respect to xQuAD (by Wilcoxon, $p < 0.05$ in both cases).

	Topic relevance						**User relevance**							
	Topic		Topic (K=1)		Topic (K=2)		User		User (K=1)		User (K=2)		F(Topic,User)	
	nDCG@5	P@5	nDCG@5	P@5	nDCG@5	P@5	nDCG@5	P@5	nDCG@5	P@5	nDCG@5	P@5	nDCG@5	P@5
Baseline	0.393	0.918	0.401	0.914	0.385	0.923	0.330	0.702	0.318	0.657	0.344	0.755	0.359	0.796
IA-Select	0.363	0.911	0.367	0.905	0.359	0.917	0.294	0.664	0.277	0.617	0.314	0.720	0.325	0.768
xQuAD	0.381	0.911	0.390	0.907	0.370	0.915	0.320	0.670	0.310	0.611	0.332	0.741	0.348	0.772
$Pers_{BM25}$	0.388	0.933	0.410	**0.947**	0.359	0.915	0.357	**0.746**	**0.373**	**0.714**	0.336	0.784	0.372	**0.829**
$xQuAD_{BM25}$	0.375	0.924	0.390	0.931	0.357	0.915	0.334	0.688	0.340	0.642	0.327	0.741	0.354	0.789
PIA-Select	0.331	0.861	0.315	0.840	0.350	0.885	0.306	0.652	0.292	0.590	0.323	0.725	0.318	0.742
$PIA\text{-}Select_{BM25}$	0.363	0.892	0.369	0.897	0.356	0.885	0.345	0.670	0.350	0.625	0.340	0.723	0.354	0.766
PxQuAD	0.395	0.930	0.408	0.928	0.379	0.931	0.337	0.714	0.334	0.674	0.341	0.760	0.364	0.807
$PxQuAD_{BM25}$	**0.405**	**0.939**	**0.414**	0.943	**0.394**	**0.933**	**0.361**	0.730	0.354	0.672	**0.371**	**0.800**	**0.382**	0.821

We compute the metrics in two ways: 1) with the assessments given by workers when replying to Q2 (see section 5.1) where they judge how relevant was the result to the evaluated topic; and 2) metrics that make use of the assessments given by workers when replying to Q1, where they judge how relevant is the result to their personal interests. We also provide a breakdown of the obtained results into topics with single keyword queries (K = 1) and topics with two keywords (K = 2). The reason is that we expect, on average, the former to represent more ambiguous topics than the latter.

6.1 Diversity Metrics

Table 1 shows the obtained results regarding diversity metrics from which we can conclude the following. In terms of topic relevance (Table 1 left), most approaches, except PIA-Select, are on par with the commercial search engine baseline. $PxQuAD_{BM25}$ has a significantly better performance than the baseline and plain diversification approaches in terms of ERR-IA and α-nDCG@5. In terms of user relevance (Table 1 right) the improvement of personalized diversity over the baseline and plain diversification is increased, indicating that the personalization factors have a positive effect on the overall view of relevance with respect to the user preference and the diversity of results preferred by the user.

Surprisingly, the plain personalization approach did not show a significant decrease in the diversification values with respect to the baseline, although the obtained values are smaller than the diversification approaches. We hypothesize that this is due to this approach taking into account the original ordering of the search results, given by the baseline, which already performs well in terms of diversity. Thus, this diversification approach resembles the two stage diversification and personalization approach implemented in $xQuAD_{BM25}$ [13] in terms of diversity values.

Regarding query size, the best performing version of IA-Select, $PIA\text{-}Select_{BM25}$, seems to provide better results with topic queries of size K = 2 (more specific queries), whereas for K = 1, the personalized versions of xQuAD –$PxQuAD_{BM25}$ and PxQuAD– seem to perform better regarding topic and user relevance, respectively.

In summary, a general overview of the results indicate that the best performing techniques in terms of diversity metrics are three of our proposed personalized diversification approaches: $PIA\text{-}Select_{BM25}$, PxQuAD, and $PxQuAD_{BM25}$. The latter has a statistically significant difference with respect to the baseline both in terms of topic and user relevance. In subtopic recall, the $PIA\text{-}Select_{BM25}$ approach is the best performing one, as also in terms of topic relevance on more specific topics (K = 2). It is worth noting that all of these approaches outperform significantly the plain personalized approach, $Pers_{BM25}$, both in topic and user relevance. The other personalized diversification approach, PIA-Select, is not able to improve the baseline. This might be due to a negative effect of the probabilistic estimate of the personalized factor on the overall behavior of the PIA-Select algorithm. However, the other variation, $PIA\text{-}Select_{BM25}$, which uses BM25 as personalized estimation factor, appears to perform much better. When compared with the best performing full diversification approach, xQuAD, we can observe that either $PxQuAD_{BM2}$ or $PIA\text{-}Select_{BM25}$ achieve statistically significant results in most of the diversity metrics.

6.2 Accuracy

Table 2 shows the accuracy metrics for the different evaluated approaches, from which we may observe the following. In terms of topic relevance (Table 2 left), values are in general close to the baseline: P@5 values are high for all approaches, whereas the nDCG@5 values, which make use of the graded relevance, are more diverse. We see the expected decrease in performance as measured by nDCG values for the plain diversification approaches (IA-Select, xQuAD). It comes as somewhat unexpected that the best value is obtained by a personalized diversification approach, $PxQuAD_{BM25}$, which is supposed to affect diversity negatively, but that even reaches statistical significance with respect to the baseline. The results obtained by this approach seem to be the highest in almost all metrics regarding topic relevance and queries with size K = 1 and K = 2. These results suggest that there is a further benefit from adding personalized factors to the diversification approach. However, the other personalized diversification approaches do not appear to outperform the baseline, so no definitive conclusion can be drawn from this result.

In terms of user relevance (Table 2 right), there appears to be two clear dominant approaches: $PxQuAD_{BM25}$ and $Pers_{BM25}$. The personalized diversification approach has the best performance in terms of nDCG, outperforming the baseline. It achieves a 9.5% improvement over the baseline, which is statistically significant.

The performance of the plain personalization approach, $Pers_{BM25}$, is on par with the latter, obtaining also a statistical significance with respect to the baseline. It seems that the main difference between these two approaches is that the performance of the $Pers_{BM25}$ approach in more ambiguous queries ($K = 2$) is slightly better than the $PxQuAD_{BM25}$ approach, while on less ambiguous queries ($K = 2$) the latter has better performance.

The last two columns of Table 2 show the F-measure values regarding both the user and topic relevance, in order to assess the combined performance of the approaches taking into account both topic and user relevance. The best values are obtained for the plain personalization approach $Pers_{BM25}$ in terms of P@5 and the diversified personalization approach $PxQuAD_{BM25}$ in terms of nDCG (graded relevance). Both of these approaches achieve statistically significant results with respect to the baseline and the best performing diversity approach, xQuAD.

In summary, while the plain personalization approach, $Pers_{BM25}$, appears to be on par with our best personalized diversification approach $PxQuAD_{BM25}$, in terms of relevance to the user, the former underperforms the baseline in terms of relevance to the topic, while the latter, surprisingly, improves the baseline to this regard, with statistical significance.

6.3 Experiment conclusions

In terms of overall performance, the personalized diversification approach $PxQuAD_{BM25}$ fares best overall: it has the best results in terms of topic-based diversity and is competitive in terms of user-based diversity. Additionally, it has the best accuracy values, together with the plain personalization approach. In most of the obtained values, this approach improves significantly the baseline. Hence, the $PxQuAD_{BM25}$ approach appears to be the best approach to balance both diversification and personalization factors, resulting in competitive values in terms of both diversity and accuracy, with respect to both topic and user relevance.

7. RELATED WORK

To the best of our knowledge, only Radlinski and Dumais [13] have previously investigated a possible combination of personalization and diversification approaches in Web search. They present a two-stage approach: given an initial user query, the search results are first diversified by executing related sub-queries, taken from query reformulations. In a second step, results are personalized using the user profile. Their hypothesis is that by diversifying results there is a higher chance of encountering documents that satisfy the user's interests. In our evaluation, we included a technique based on their two stage approach ($xQuAD_{BM25}$). In our experiments this approach did not achieve significant improvements over plain personalization, and it was outperformed by our personalized diversification approaches. A possible explanation is that the basic personalized approach was applied to the search baseline, which already had competitive diversity performance. Another explanation would be that their approach may further benefit from diversification techniques based on sub-queries, which were not tested in our experiments.

In our work, we exploit social annotation profiles in order to apply personalization factors related to a Web search. The use of a source of information of this type for Web personalization was first suggested by Noll and Meiner [12], who evaluated with real users the use of Delicious user profiles to personalize search, proving the effectiveness of the approach with a simple personalization model. Vallet et al. [18] presented an evaluation framework allowing for offline evaluation of Web search personalization approaches based on social annotations. They tested a number of approaches, including Noll and Meiner's, concluding that the best performing approaches were based either on the BM25 probabilistic model or a simple scalar product. From this work we have adapted the proposed BM25 model as a personalized estimation factor. Our evaluation framework is based on their offline evaluation approach, which has been adapted here to allow for an online evaluation of personalization and diversification methods with real users.

Han et al. [9] presented an improvement over Noll and Meiner [12] and Vallet et al. [18], in which user and document tags are related to the ODP taxonomy. The taxonomy is then exploited to expand tags to related concepts. The authors proposed to use the user's search query to activate concepts within the ODP taxonomy, performing a sort of "contextualization" of user preferences. These preferences are then used to compute the personalization factor to be applied on the search output. Their evaluation, which used the framework presented by Vallet et al., indicated an improvement over the state of the art approaches. In our work, we treated the ODP and tag information as independent sources of information. In future work we will analyze whether relating tags to ODP concepts has a positive impact on personalized diversification approaches.

8. CONCLUSIONS AND FUTURE WORK

In this work we have presented a number of approaches that combine both personalization and diversification components. To this end, we have investigated the introduction of the user as an explicit random variable in two state of the art diversification models: IA-Select [1] and xQuAD [15,16]. Our experiments show that there is a clear benefit in introducing this variable, achieving statistically significant improvements over the baselines that range between 3%-11% in terms accuracy values, and between 3%-8% in terms of diversity values.

These results lead us to conclude that diversification has a case even in the presence of personalization. Personalization in diversity can be seen in two –to some degree equivalent– ways: a) concentrating the extent of diversification in the areas of user interests, and b) seeking relevance to as diverse sides of user preferences as possible. We identify three situations in which our approaches have a potential benefit: 1) user preferences are not necessarily relevant for the query at hand, therefore they may remove uncertainty from it to very variable degrees (from a little to none at all); 2) the system knowledge about user preferences is incomplete (restricted to the observed user activity within the system) and imprecise (evidence of preferences can be implicit, indirect, ambiguous), whereby diversifying the effect of personalization reduces its involved risk; 3) user preferences are themselves diverse, therefore for lack of knowledge about which are most relevant in a given context, it is appropriate to diversify the relevance to the different preference aspects. A similar case as is made for diversification on query intents can thus be made for preference aspects.

We have presented different alternatives to estimate the components that arise from the formal development of our scheme, and we evaluated a number of them. Future work may include further investigation on these alternatives. We have tested two estimation approaches for the user/document and document/category relations, based on social annotations (Crowdflower) and ODP classification (Textwise), respectively. Other ways of approximating these values could be investigated, such as other sources of social profiles (e.g. Facebook, Twitter) or other sources of user profile information (e.g. query and browsing history). Other alternatives for document classification (e.g. clustering, or query reformulations [16]) might be

explored as well. A follow-up evaluation could also take into consideration different values of the combination parameter in xQuAD, which controls the strength of the personalized diversity component on the personalized variation of xQuAD.

The vision of combining diversity and personalization indeed opens a rich array of possibilities. Diversity and personalization can be combined in different ways, some of which we have investigated here, beyond which we see wide room for further research.

9. ACKNOWLEDGMENTS

This work was supported by the national Spanish projects TIN2011-28538-C02-01 and S2009TIC-1542.

10. REFERENCES

[1] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. In *WSDM'09*, pages 5-14, 2009.

[2] Alonso, O. and Baeza-Yates, R. Design and implementation of relevance assessments using crowdsourcing. In *ECIR '11*, pages 153-164, 2011.

[3] Alonso, O. and Mizzaro, S. Can we get rid of TREC assessors? using mechanical turk for relevance assessment. In *SIGIR'09 Workshop on The Future of IR Evaluation*, 2009

[4] Bache, R., Baillie, M., and Crestani, F. Language models, probability of relevance and relevance likelihood. In *CIKM'07*, pages 853-856, 2007.

[5] Carbonell, J. G. and Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR'98*, pages, 335-336, 1998.

[6] Chen, H. and Karger, D. R. Less is More. In *SIGIR'06*, pages 429-436, 2006.

[7] Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. Novelty and diversity in information retrieval evaluation. In *SIGIR'08*, pages 659-666, 2008.

[8] Fernández, M., Vallet, D., and Castells, P. Probabilistic score normalization for rank aggregation. In *ECIR'06*, pages 553-556, 2006.

[9] Han, X., Shen, Z., Miao, C., and Luo, X. Folksonomy-Based ontological user interest profile modeling and its application in personalized search. In *Active Media Technology*, LNCS, vol 6335 pages 34-46.

[10] Lee, J. H.. Analyses of multiple evidence combination. In *SIGIR '97*, pages 267-276, 1997.

[11] Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. Personalized search on the World Wide Web. The Adaptive Web, pages 195-230, 2007.

[12] Noll, M. G., Meinel, C. Web search personalization via social bookmarking and tagging. In *ISWC'07*. pages 367-380, 2007.

[13] Radlinski, F. and Dumais, S. Improving personalized web search using result diversification. In *SIGIR'06*, pages 691-692, 2006.

[14] Rafiei, D., Bharat, K., and Shukla, A. Diversifying web search results. In *WWW'10*, pages 781-790, 2010.

[15] Santos, R., Peng, J., Macdonald, C., and Ounis, I. Explicit search result diversification through sub-queries. In *ECIR '10*, pages 87-99, 2010.

[16] Santos, R. L. T., Macdonald, C., and Ounis, I. Exploiting query reformulations for web search result diversification. In *WWW'10*, pages 881-890, 2010.

[17] Vallet, D. Crowdsourced Evaluation of Personalization and Diversification Techniques in Web Search. In CIR' 11 Workshop (SIGIR 2011), 2011.

[18] Vallet, D., Cantador, I., and Jose, J. Personalizing web search with Folksonomy-Based user and document profiles advances in information retrieval. In *ECIR'10*, pages 420-431, 2010.

[19] Vargas, S. and Castells, P. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. *In RecSys'11*, pages 109-116, 2011.

[20] Vargas, S., Castells, P., and Vallet, D. Intent-Oriented Diversity in Recommender Systems. In *SIGIR'11*, pages 1211-1212, 2011.

[21] Wang, J. and Zhu, J. Portfolio theory of information retrieval. In *SIGIR'09*, pages 115-122, 2009.

[22] Zhai, C., Cohen, W. W., and Lafferty, J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR'03*, pages 10-17, 2003.

[23] Zhang, M. and Hurley, N. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In RecSys'08, pages 123-130, 2008.

[24] Ziegler, C-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In WWW'05, pages 22-32, 2005

Appendix A. PERSONAIZED QUERY ASPECT DISTRIBUTION

In this appendix we offer alternative estimations for the personalized query aspect distribution, $p(c|q,u)$, component:

1. $p(c|q,u) \sim \frac{p(q|c)p(c|u)}{\sum_{c'} p(q|c')p(c'|u)} = \frac{p(c|q)p(c|u)}{\sum_{c'} p(c'|q)p(c'|u)}$

   assuming a uniform aspect prior, and conditional independence between queries and users given an aspect. $p(c|q)$ can be estimated as:

   $$p(c|q) = \sum_d p(c|d,q)p(d|q).$$

2. $p(c|q,u) \sim \frac{p(q|c)p(c|u)}{\sum_d p(q|d)p(d|u)} = \frac{p(c|q)p(c|u)}{p(c)\sum_d p(d|q)p(d|u)/p(d)}$

   assuming conditional independence between queries and users given an aspect. Assuming a small aspect overlap (nr. of aspects per document) the ratio of category and document priors can be calculated as

   $$p(c)/p(d) \sim |\{d \in D | c \text{ is covered by } d\}|$$

   where D could be interpreted as the set of documents being reranked, or the whole collection. For higher aspect overlap, all document-aspect pairs should be taken into account, and normalized by the number of pairs.

3. $p(c|q,u) = \sum_d p(c|d,q,u)p(d|q,u) \sim$

   $\sum_d p(c|d,q)p(d|q,u)$, and $p(c|d,q) \sim \frac{p(c|d)p(c|q)}{\sum_{c'} p(c'|d)p(c'|q)}$,

   assuming conditional independence of aspects and queries given a document, and a uniform aspect prior; or $p(c|d,q) \sim \frac{p(c|d)p(c|q)p(d)}{p(d|q)p(c)}$ without the latter assumption.