Crowdsourced Evaluation of Personalization and Diversification Techniques in Web Search

David Vallet
Universidad Autónoma de Madrid
Cantoblanco, 28049 Madrid, Spain
david.vallet@uam.es

ABSTRACT
Crowdsourcing services have been proven to be a valuable resource for search system evaluation or creation of test collections. However, there are still no clear protocols to perform a user-centered evaluation of approaches that consider user factors, such as personalization or diversification of results. In this work we present two complementary evaluation methodologies to evaluate personalization and/or diversification techniques. Our preliminary results are encouraging, as they are consistent with previous experimental results. This leads us to believe that these evaluation methodologies could allow performing inexpensive user-centered evaluations. As a result of this work, a test collection for personalized and diversified approaches has been made available.

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

General Terms Algorithms, Experimentation.

Keywords Crowdsourcing, personalization, diversification.

1. INTRODUCTION
The evaluation of information retrieval systems with crowdsourcing services is a recent line of research that has been getting increasing attention in recent years. Ultimately, this line of work seeks to optimize the resources necessary to obtain relevance judgments from users, following the Cranfield evaluation paradigm [6]. In TREC related search tasks, this is achieved by substituting experts users with workers from crowdsourcing services as a means of obtaining relevance judgments [2][3][4].

The above studies are applicable to user-independent evaluation tasks. The collected relevance assessments can thus be used to evaluate systems that produce results based on a search topic description and a document collection. However, other types of search systems, such as personalized [7] or session oriented search systems [11] require additional evaluation information, such as user profile information or search interaction information, respectively. While the evaluation of interactive search systems with crowdsourcing services has been already studied [17], there is still little study regarding personalized systems.

In this work, we study crowdsourcing techniques for the evaluation of personalized and diversification methodologies applied to Web search systems. We perform this analysis as we believe there is an inherent relation between both technologies. On the one hand, diversification techniques attempt to deal with ambiguous queries by presenting to the user a list of results that covers all the possible interpretations of the query. Thus, trying to maximize the probability that one of the presented interpretations is relevant to the user [1][9]. On the other hand, the field of personalization attempts a different way of overcoming the Web search problem: rather than covering all possible query interpretations, search results can be tailored to the particular meaning that is relevant to the user’s interests [8]. Both techniques have their shortcomings. Diversification techniques follow a one-size-fits-all approach, which means that it may be that users with particular interests did not find their relevant results at the top of the result set. Personalization techniques make use of user profile representations that are often not accurate enough and thus personalized results are often found to be intrusive by the user. The natural implication of these shortcomings is to study a combination of both models, which has been initially reported in [15].

Two evaluation methodologies for personalized and diversified Web search systems are presented. The first, already reported in [15], is explained in more detail in this work, focusing on the challenges of using workers to perform this type of evaluation. This first methodology is based on a side-by-side comparison of search result lists, which has been previously used in other crowdsourcing evaluations [12], allowing the evaluation of a number of techniques by using pairwise comparisons. The second evaluation methodology follows an evaluation methodology closer to the Cranfield paradigm, as it seeks the collection of relevance judgments for individual results. This information will allow the evaluation of a number of personalization and diversification approaches with standard metrics.

2. RELATED WORK
Most crowdsourcing evaluation studies have focused on how to obtain relevance judgments from workers with a similar quality to those obtained from expert assessors, at a fraction of the cost, both in terms of monetary expenses and time required to collect the judgments [2]. There has been already research on how to measure or improve the quality of this data. Alonso and Baeza [2] propose a methodology to obtain relevance assessment for TREC related topics. Their experiments show that the assessments provided by workers can be comparable to those given originally by expert TREC assessors. Other studies, such as the one propose by Eickhoff and de Vries [1] identify the problem of malicious workers over crowdsourceable tasks and describe possible measures to take in order to make tasks more resistant to such types of fraudulent workers. It is also worth noting that in TREC 2011 a new Crowdsourcing track has been added, which focuses mainly on crowdsourcing strategies for obtaining high quality relevance assessments. In this work we seek to adapt these protocols to allow user-centered evaluations of personalized systems.

Regarding user-centered evaluation methodologies, Zuccon et al. proposed an evaluation approach that makes use of crowdsourcing services to evaluate interactive retrieval systems [17]. The authors proposed a protocol to capture all the interaction information to

Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval.
Copyright is retained by the author(s).

1 Text REtrieval Conference (TREC): http://trec.nist.gov/
perform a proper evaluation of such type of retrieval systems, in addition to acquiring post search information. However, their proposed evaluation framework did not consider user preferences, only search session information. To the best of our knowledge, there is no previous work which analyzes the evaluation of personalized systems using crowdsourcing techniques.

There are a number of studies that have identified crowdsourcing services as a key resource to evaluate diversity techniques in search systems. Agrawal et al. [1] asked workers to assign search results to a fixed small set of ODP topics, in a way that the possible meaning of each result document could be identified. In this work we take a similar approach in our first evaluation methodology. In our second methodology we rather allow users to freely assign a topic to each result, by assigning keyword labels. Tang and Sanderson used a side-by-side evaluation methodology to ask users which two set of results was perceived as more diverse (or better) to evaluate spatial diversification retrieval systems [12]. Our first evaluation methodology is also based on a side-by-side interface, although we present lists of search results while Tang and Sanderson presented results as location points in a map. Side-by-side evaluation was first proposed in a more classic controlled evaluation with on-site subjects [13]. Note that although the evaluation framework proposed in this work is designed to evaluate both personalization and diversification aspects, it can also be used to evaluate only one of these aspects.

3. GENERAL DISCUSSION

In this section, we discuss the main challenges that we faced in order to evaluate personalized systems with crowdsourcing services.

3.1 Evaluating Worker’s Quality

The main characteristic of the evaluation of personalized search systems is that relevance assessments are provided subjectively by each user, who judges the relevancy of a search result with respect to the current search topic and to the user’s own interests. That is, the same result can be judged as irrelevant by one user and as relevant by another while evaluating the same search topic. Hence, in order to evaluate the quality of the collected relevance assessments, it is not sufficient to compare these with those relevance assessments provided by expert users, as it is done in user-independent tasks [2][6]. Thus, additional quality control mechanisms have to be considered.

In our evaluation methodologies we chose to use a controlled environment, similar to the one mentioned in [1], where control questions can be presented to the worker in a way that at least we can ensure that the worker is inspecting the results to some degree. Each unit of work presented to workers contains the topic description and a link to our controlled Web application environment. In our evaluation framework we included special work units that contain a control question in order to ensure that the worker inspects the list of results. In essence, these control units contain results that are artificially deviated from the search topic. Our premise is that workers who are making random relevance assessments will end up marking these results as relevant, at which point we can have a high confidence that the workers was not making a quality assessment. In this case, the environment will return a false confirmation key to the worker that will mark this work unit as failed. An accumulation of failed work units will mark the worker as untruthful, and will eventually ignore their input and prevent the malicious worker from being remunerated.

These control questions are different in each evaluation methodology – they are detailed in Sections 4 and 5.

3.2 Evaluation Topics

In order to evaluate personalization and diversification techniques, the evaluation topics have to meet a main requirement: query topics have to have some degree of ambiguity, as it is only in the case of ambiguous topics where personalization or diversification techniques have a reason to be applied. The procedure of creating evaluation topics is explained following.

Similar to Rafiei et al. [9] and Santos et al.[10], we identified possibly ambiguous queries from Wikipedia, and used the disambiguation pages to extract the possible related subtopics or categories. From each query, we chose no less than 4 and no more than 7 related subtopics. In the case that more subtopics were presented in the disambiguation page, we chose those subtopics which returned a larger number of results. We used the Bing Web search API1 to obtain the results from each topic and subtopic. As Santos et al. [10] suggest, we use the subtopic definition from Wikipedia to obtain results related to each of the sub-topicalities of the query. In this way, we can suppose that if a Web page appears on the results of a subtopic, it is related to this subtopic and to the general topic. This process resulted in 23 evaluation topics. The baseline of our evaluations is the set of results returned by the commercial search engine for the original topic query.

3.3 Obtaining the User Profile

A test collection for personalized systems must include some type of user profile representation, which serves as the input of the evaluated personalization mechanism. In the context of Web search tasks, we propose two ways of obtaining this profile information: an ad-hoc pre-questionnaire which relates the user interests to the different possible meanings of a specific topic, and a more general and realistic user representation based on the social profile of the user. The former is a relatively easy way to obtain accurate user preferences over a specific topic, although it has the additional cost of requiring the user to fill the questionnaire for every topic, which is an unrealistic situation. The latter is a recent trend on personalization systems, based on the analysis of the social behavior of users in order to extract a representation of the user interests. More specifically, we extract a user profile representation from the annotation actions of the user in social tagging services such as Delicious2 [14][16]. The main problem of this methodology is that potential workers are limited to those that participate in this type of social services. Both techniques are explained in the following two sections, where each evaluation framework is detailed.

4. SIDE-BY-SIDE EVALUATION

In the side-by-side evaluation methodology, we present two different lists of results to the worker, each of them related to one of the evaluated techniques. Figure 1 shows an example on how two results lists are presented to the user for evaluation. Previous to this comparison, workers have to fill a small pre-questionnaire in order to indicate their personal preference for the different meanings of the topic that they are going to evaluate. The evaluation procedure is the following: 1) we present the topic associated to the work unit to the subject, and show the possible subtopics related to the topic, as extracted from Wikipedia; 2) in order to

---

1 Bing API: http://www.bing.com/developers/
2 Open Directory Project (ODP): http://www.dmoz.org
3 Delicious bookmarking service: http://www.delicious.com
obtain the specific preferences of the workers, we ask them to indicate a level of interest for each subtopic; 3) the topic and interest information is used as input of two of the evaluated approaches, chosen randomly; 4) the system presents an anonymized side-by-side comparison page with the two resulting search result pages (10 results per list); 5) the worker evaluates each result lists as a whole.

We designed a small questionnaire to collect the subjects’ opinions about the presented results. It consisted of three questions with 5-point liker scales: Q1) This result list is relevant to the topic (Topic); Q2) This result list adjusts to my interests (Interests); Q3) Overall, how would you rate this result list? (Overall). Additionally, users were asked to indicate which of the two result pages they preferred, if any (QP). Note that this interface allows the evaluation of any number of personalized or diversification approaches, as the evaluation framework chooses two techniques at random to evaluate. Questions Q1-Q3 allow the comparison of the Likert scales between systems, while question QP allows a more direct comparison between two specific techniques. Each work unit took ~45 s. on average to be completed. This means that a large amount of evaluations could be collected easily.

In order to prevent the effect of malicious workers, we created a set of special control work units. In these units workers are always presented with a result list relevant to the topic (the output of any of the algorithms being evaluated) and a result list product of a search topic which is somehow related to the evaluated topic, but it is clearly irrelevant once inspected. For instance, for the topic “Sun”, which has a number of different meanings (e.g., Sun newspaper, the technologic company, a music band, the star, etc.) we presented the result list of the search query “tropics”. There is a 50/50 chance that the worker randomly answers the correct question, but eventually the worker will reply incorrectly to the control question, and in this case we can suppose that there is a high chance that the worker is malicious.

5. RESULT ASSESSMENT EVALUATION

We now present an evaluation framework that allows a finer grained evaluation of both personalization and diversification techniques, in which standard evaluation metrics can be applied. Another difference of this evaluation methodology with the side-by-side evaluation is that, when a reasonable amount of relevance assessments are collected from each worker, the collected information can be stored in order to construct a test collection. This collection will contain the necessary information in order to evaluate and compare other personalization and diversification approaches.

In this framework, workers have to perform relevance assessments over individual results. In order to access the topic evaluation, the worker has to have input her/his Delicious profile account, from which the public information will be gathered to use as input of a different number of personalization approaches [14][16]. Evaluated diversification approaches have to incorporate a mechanism to decide how related is each result page to each possible meaning of the query. A possible technique is to classify result documents to ODP categories, as done by Rafiei et al. [9].

The whole evaluation process is the following: 1) the worker enters her/his Delicious profile account name, from which the preferences are extracted; 2) we load the 300 top search results of the topic associated to the work unit, which were obtained previously using the Bing API; 3) we use each evaluated personalization or diversification technique to reorder the original results (we only evaluated reordering techniques, although other types of techniques can be easily incorporated); 4) the top 10 results from each evaluated technique are aggregated to a single pool of results; 5) results are presented to the worker, in a random order; 6) the worker evaluates each result individually.

Figure 2 shows the interface element that asks for the worker’s relevance assessment, by presenting three questions. The first question, QR1, asks how relevant is the result regarding the personal interests of the worker. The second question, QR2, asks how relevant is the result regarding the evaluated topic. The third question, QR3, asks the worker to assign a possible subtopic meaning to each result (see Figure 2 for an example). Note that workers have the possibility to input their own subtopic description, choosing the ‘Other’ option and these descriptions are shared among workers. QR1 and QR2 can be used along with evaluation metrics to evaluate how adequate is a personalization technique. An efficient personalization algorithm should be positioning on top those results that score high values for both QR1 and QR2, indicating that the result is related to the interests of the user, but does not deviate from the specific information need represented.
Figure 2: Example of result relevance assessment for topic “Keyboard”.

by the search topic. Additionally, question QR3 can be used to apply diversity evaluation metrics to the same techniques, such as intent-aware metrics [1] or oNDCG. Each work unit contained on average ~25 results to evaluate when evaluating four different personalization approaches and took ~8 min. on average to be responded by each worker. As we expect to gather less workers for this evaluation (due to the Delicious account restriction), we believe that is better to have a longer evaluation time per worker in order to gather more detailed relevance assessments.

Two different mechanisms are used to avoid malicious workers in this evaluation framework. First, in order to avoid malicious workers to input a Delicious account at random, we ask workers to validate their account by bookmarking a random URL associated to their profile – in this way we can asseverate that they have access to the account. We also force the account to be at least 3 months old and to contain a minimum of 30 bookmarks, so that we avoid the creation of an account specific for the work unit. Second, we use control work units that include, at a random position, a single artificial result which is clearly irrelevant to the topic. The premise is that a user that is not inspecting the whole list in detail will fail to identify the control result as irrelevant. In addition, the worker has to choose the result topic as ‘Other’, since the result does not fall into any of the predefined subtopics. Thus, there is little chance that a malicious worker answering randomly or consistently to the evaluation interface will grade the artificial result correctly, and failing to do so will return an incorrect confirmation key to the worker.

6. EARLY EXPERIMENTS
We performed the user experiments using CrowdFlower5, a crowdsourcing service that allows the creation of golden topics, which were used by our introduced mechanisms of user quality control. Early experiments are reported for both evaluation frameworks, after some consideration on the effectiveness of both systems in detecting malicious workers.

6.1 Malicious Workers Detection
Regarding the detection of malicious workers, both interfaces seemed quite effective on their detection. On average, ~20% of workers were flagged as malicious, and a closer inspection of the gathered results indicated that most of them were indeed inputting random answers, random Delicious accounts, or just inputting confirmation codes at random. In some occasions in the relevance assessments interface a worker failed a golden topic because the random artificial result, which was supposed to be irrelevant to the search topic, happened to be somehow relevant to the topic. But the worker was not flagged as malicious, since the worker had to fail two or more golden topics for this to happen. This problem can be solved by choosing this artificial result more carefully.

Those assessments that were marked as trustful were later inspected to check if any worker had input clearly fake relevance assessments (e.g., all irrelevant, all relevant). We did not find relevance assessments with such characteristics. This is to be expected, as the control units for both systems are designed to detect random, as well as all irrelevant and all relevant assessments. There is still the possibility that a worker making random assessments would have passed by chance the control units. We expect to have only few of these cases (if any) and thus do not have a significant impact on the evaluation results.

6.2 Side-by-side Experiments
In this experiment, we used the side-by-side evaluation framework to evaluate a standard personalization approach (Pers), Agrawal et al.’s pure diversification approach (IA-S) [1], and a Personalized diversification approach (PIA-S) presented in previous work [15], which combines both diversification and personalization factors. The goal of the later approach is to promote those results related to the user’s interests. However, this approach still includes results related to other possible interpretations of the query. In this way, if the profile is not accurate enough, the users will be still able to find relevant results in the top positions.

In this experiment we were able to collect ~850 topic judgments from 190 different workers in approximately one week. We following present a summary of the obtained results, to check the validity of the evaluation methodology.

Table 1 shows the average Likert-scale values for each of the control questions. As this evaluation methodology collects accurate user preferences by means of the pre-questionnaires, the results are as expected. The obtained results indicate that workers had a general preference for those approaches that had some personalization component over the baseline (the original Web search results) and the pure diversification approach (IA-S), as results are more tailored.

5 CrowdFlower: http://www.crowdflower.com
to the workers indicated preferences. Workers preferred the pure personalization approach (Pers) over those that included some diversity component (PIA-S, IA-S). These differences were statistically significant (Mann-Whitney U, p < 0.05). The Mann-Whitney U test was chosen as we collected Likert scales values, which were non-parametric. The side-by-side comparison also indicated a statistically significant preference of users for the personalized approaches (Wilcoxon signed rank test, p < 0.05). We did not find any statistical evidence that users had a preference over the baseline and the IA-S approach, which suggest that the baseline was effective at diversifying results.

In order to test the hypothesis that it was the unrealistic accuracy of the user profile what made workers prefer the personalization approaches over the diversification approaches, we modified the evaluation system to manipulate the user profiles in order to include a noise level of I, which indicates the number of category preferences per user that are assigned a random value. We carried out two additional user evaluations with a level of noise of I=1 and I=2 for the pure personalization algorithm and the combination of diversification and personalization.

Table 2 shows the results of this evaluation. The lambda differences with respect to the original Likert values indicate that, as expected, the personalization approach is penalized when using less accurate user profiles, and is in this situation where the PIA-S approach, which also incorporates diversification factors, is less affected by inaccurate user profiles. These differences were statistically significant (Mann-Whitney U, p < 0.05). These results are encouraging, as the degradation of the pure personalization approach as more noise is introduced in the user profile is to be expected. This is an indication that the collected relevance assessments are reliable.

Table 1: Average values for Likert scales

| Question Baseline Pers PIA-S IA-S |
|-------------------------------|-----------------|-----------------|-----------------|
| Q1 (Topic)                   | 3.805           | 4.304           | 4.209           | 3.813           |
| Q2 (Interests)               | 3.117           | 4.216           | 3.973           | 3.219           |
| Q3 (Overall)                 | 3.305           | 4.272           | 3.899           | 3.284           |

Table 2: Q3 (Overall) Likert values with noisy user profiles

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Q3</th>
<th>Q3 (I = 1)</th>
<th>Q3 (I = 2)</th>
<th>Q3 (Pers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pers</td>
<td>4.27</td>
<td>3.92</td>
<td>4.306</td>
<td>-6.5%</td>
</tr>
<tr>
<td>PIA-S</td>
<td>3.90</td>
<td>3.81</td>
<td>3.71</td>
<td>-2.3%</td>
</tr>
<tr>
<td>IA-S</td>
<td>3.81</td>
<td>3.81</td>
<td>3.71</td>
<td>-4.7%</td>
</tr>
</tbody>
</table>

6.3 Relevance Assessment Experiments

In this experiment we used the relevance assessment interface to evaluate different personalization methodologies based on the Delicious social tagging profiles. These personalization approaches make use of the Delicious profile information in order to compute a personalization score for each result that is used to reorder the search results according to the user’s interests and the result’s profile in Delicious. The reader can consult our previous work [14] in order to get a description of each personalization algorithm. Three different personalization algorithms were evaluated: a personalization algorithm based on previous work by Xu et al. [16] (Cosine), a personalization algorithm based on a scalar product between the user social and the document annotations (Scalar), and our own approach, based on the BM25 retrieval model (BM25).

We run the experiment during one month, with which we collected ~4K individual relevance judgments. This evaluation information is made freely available for the evaluation of personalization and diversification techniques.

Table 3: NCDG@10 values for personalization approaches

<table>
<thead>
<tr>
<th>Question</th>
<th>Baseline</th>
<th>Cosine</th>
<th>Scalar</th>
<th>BM25</th>
</tr>
</thead>
<tbody>
<tr>
<td>QR1 (Topic)</td>
<td>0.672</td>
<td>0.568</td>
<td>0.567</td>
<td>0.565</td>
</tr>
<tr>
<td>QR2 (Preferences)</td>
<td>0.323</td>
<td>0.302</td>
<td>0.352</td>
<td>0.342</td>
</tr>
</tbody>
</table>

We now show some preliminary results with these algorithms and compare the obtained results with those obtained previously by using an automatic topic evaluation methodology, presented in [14]. As an evaluation metric, we use the NDCG@10 measure, where Likert scales of questions QR1 and QR2 are mapped to relevance weights assigned to each result. Table 3 shows the obtained results. We observe similarities between these results and those obtained previously [14]. First, regarding how the results adjust to the user preferences (QR2), we observe that, as our previous work pointed out, the state of the art approach based on the cosine similarity (Cosine) underperforms the baseline, whereas the other two personalization approaches (Scalar and BM25) perform similarly and outperform the baseline. The results obtained in QR1 are to be expected, as some results, that maybe have less relation with the original query topic, are pushed into the top positions due to them being matched with the user preferences. These similarities with our previous experiments are encouraging, and validate to some degree the presented evaluation methodology, although further experimentation is needed to obtain a definite conclusion.

7. CONCLUSIONS AND FUTURE WORK

We have presented two evaluation methodologies applicable to both personalization and diversification Web search systems. We believe that both methodologies are complementary. On the one hand, the side-by-side comparison interface allows a fast and reasonable way to compare different personalization and diversification approaches. On the other hand, the relevance assessment interface allows for a more realistic and fine grain evaluation of the different approaches. Additionally, it allows the creation of test evaluation collections for personalized and diversification approaches, by including the subjective relevance assessments of each worker; their subjective mapping of each result to a subtopic; and their social user profile representation gathered from Delicious.

The experiments presented for the second evaluation framework are still in an early stage, and so far we have only evaluated pure personalization approaches. In future work we will implement and evaluate a number of diversification approaches, which will allow further evaluating our experimental frameworks and our test collections.

8. REFERENCES


6 http://ir.ii.uam.es/~david/webdivers/


