On Diversifying and Personalizing Web Search

David Vallet, Pablo Castells
Universidad Autónoma de Madrid
Cantoblanco, 28049 Madrid, Spain
{david.vallet, pablo.castells}@uam.es

ABSTRACT
Diversification and personalization methods are common approaches to deal with the one-size-fits-all paradigm of Web search engines. We performed a user study with 190 subjects where we analyzed the effects of diversification and personalization methods in a Web search engine. The obtained results suggest that our proposed combination of diversification and personalization factors may be a way to overcome the notion of intrusiveness in personalized approaches.

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

General Terms: Algorithms, Experimentation.

Keywords: Diversity, personalization, Web search

1. INTRODUCTION
With the ever increasing content made available on the Web, search systems struggle with finding the information required by users. The problem of selecting which results are relevant to a user is aggravated by two related characteristic of Web search systems: they follow a one-size-fits-all paradigm and represent the user's information need as a set of keywords. The latter characteristic results often in ambiguous or too broad search topics, which makes difficult to adopt the one-size-fits-all paradigm, as many different personal views of the topic have to be covered. Two different ways of overcoming these problems have been proposed in recent years: personalization and diversification approaches.

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, they try to maximize the probability that one of the presented interpretations is relevant to the user [1][4]. However, diversification models still follow a one-size-fits-all approach, which means that it may be that users with particular interests do not find their relevant results at the top of the result set. On the other hand, the field of personalization attempts a different way of overcoming the Web search problem: rather than adjusting to the one-size-fits-all approach, and trying to cover the different meanings of an ambiguous query, search results can be tailored to the particular meaning that is relevant to the user's interests [2]. However, the goal of personalization has yet to be met: user profiles are often not accurate enough and thus personalized results are often found to be intrusive by the user. This could be because personalization techniques take too much risk by only showing results related to the user profile representation.

In this paper we propose to meet halfway the above two approaches. We perform a user study in order to analyze if personalization and diversification techniques can be combined to overcome the one-size-fits-all paradigm and the intrusiveness felt from diversification and personalization techniques, respectively. Radlinski and Dumais [3] study the application of diversification techniques as a previous step to personalization. Their goal is to maximize the probability that a relevant result to the user is found in the diversified list. Still, their approach relies on an accurate representation of the user's interests. We suggest that this problem of inaccurate personalization approaches can be overcome by adding personalization components to a diversification approach.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

2. EVALUATED METHODS
In this section, we describe the different personalization and diversification approaches to be evaluated in our user study.

Personalization (Pers). As a pure personalization approach, we define a typical approach which represents interests of the user as a set of preferred topics and ranks documents with respect to their relation to each topic. This approach can be defined as a scoring function of a document related to a query and a user:

$$Pers(d|q,u) = \sum_c V(d|q,c)P(c|u,q)$$

where $V(d|q,c)$ is the quality (relevance) of the document given a query and a topic category and $P(c|u,q)$ is the relevance of the category to the user for the current query. The effect of this approach is that if the user has an interest towards only a single category related to a query, only documents related to this category will be shown to the user. If the user has multiple interests related to a query, the degree of interest to each topic will be used to decide the ranking of each document. Note that in our experiments this approach is applied over a diversified result list, and thus it can be considered an adaptation of the state of the art [3].

Intent Aware Select (IA-S). As a pure diversification algorithm, we implement the diversification approach by Agrawal et al [1], which finds a set S of documents of size k that maximizes:

$$P(S|q) = \sum_c P(c|q)\left(1 - \prod_{d\in S}(1 - V(d|q,c))\right)$$

where $P(c|q)$ is the distribution probability of a category belonging to the query, which can be supposed to be uniform if no other information is available.

Personalized Intent Aware Select (PIA-S). We propose to incorporate a personalized factor into the intent aware approach proposed by Agrawal et al. [1]. In our personalized intent aware approach, the probability distribution $P(c|q)$ in the objective function is substituted by the probability that the category is relevant to the user given a query, i.e. $P(c|u,q)$:

$$(S|q,u) = \sum_c P(c|u,q)\left(1 - \prod_{d\in S}\big(1 - V(d|q,c)\big)\right)$$

The goal here is to promote those results that belong to a category related to the user's interests. However, the objective function still encourages including results from other categories with lower values. In this way, if the profile is not accurate enough, the users will be still able to find relevant results in the top positions.

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

3. EXPERIMENTS

In order to test our hypothesis, we performed a user centered study. Similar to Rafiei et al. [4] and Santos et al. [5], we identified possibly ambiguous queries from Wikipedia, and used the disambiguation pages to extract the possible related subtopics or categories. From each subtopic, 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 a well-known commercial search engine to obtain the results from each topic and subtopic. As Santos et al. [5] 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 related to the general topic. As we did not have access to the result scores, the relevance score of a document regarding a subtopic was computed using a rank-based normalization of the subtopic results. This score was computed as $V(d|q,c) = 1/n$, where n is the position of the document in the result set. This process resulted in 23 evaluation topics. The baseline of our evaluation is the set of results returned by the commercial search engine for the original topic query.

In order to evaluate the different approaches, we built a Web-based evaluation interface to perform a user-centered evaluation, with the following steps: 1) present a random topic to the subject, and show the possible subtopics related to the topic, as extracted from Wikipedia; 2) to obtain the interests of the subjects, we ask them to indicate a level of interest for each topic, which was used to estimate $P(c|u,q)$; 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).

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. We performed the user experiment using a crowdsourcing service (http://crowdflower.com), collecting 351 side-by-side evaluations from 60 users.

3.1 Experiment Results

Table 1 shows the average Likert-scale values for each of the control questions. The obtained results indicate that users had a general preference for those approaches that had some personalization component over the baseline and the pure diversification approach (IA-S). Within the personalized approaches, users 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 side-by-side comparison also indicated a statistically significant preference of users for the personalized approaches (Wilcoxon signed ranked test, $p < 0.05$). We did not find any statistical evidence that users had a preference either over the baseline or over the IA-S approach, which suggests that the commercial search engine was effective at diversifying results – this was later confirmed with a visual inspection.

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

Algorithm	Q3	Q3 (I = 1)	ΔQ3	Q3 (I = 2)	ΔQ3
Pers	4.27	3.92	-8.3%*	3.61	-15.5%*
PIA-S	3.90	3.81	-2.3%	3.71	-4.7%

So far, the obtained results indicate a better performance of classic personalization approaches (Pers) over diversified or a combination of both approaches. However, the preference feedback in the previous study was obtained explicitly form the user for each topic. This assumption is unrealistic in common scenarios, as this kind of explicit feedback requires an extra effort from users, which is usually not accepted. In other words, this aspect of the experimental setup introduces an artificial advantage in the Pers system. Personalization approaches usually rely on automatic (more imprecise) preference learning methodologies based on implicit user feedback, therefore resulting in less accurate representations of the user interests and lower performance. Hence, in order to complement our study, we simulate a situation in which we do not have such accurate user preference information: 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 $P(c|u,q)$ 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. We collected over 500 additional topic judgments from 130 distinct users.

Table 2 shows the results of this evaluation. The lambda differences with respect to the original Likert values indicate that, as expected, the pure personalization approach is penalized when using less accurate user profiles. These differences were statistically significant (Mann-Whitney U, $p < 0.05$). However, the PIA-S approach, which incorporates diversification factors, does not have such a negative impact due to noisy preferences. Moreover, at a noise level of $I = 2$, users preferred the PIA-S approach when compared to the Personalization approach (Wilcoxon, $p < 0.05$). These results suggest that our proposed approach, which combines personalization and diversification factors, is more robust to less accurate user preference representations.

To conclude, we performed a user study in order to inspect the preference of real Web search users towards personalization and diversification approaches. Our results suggest that in the idealistic case in which the user interests are highly accurate, a pure personalization approach is the best performing approach. However, in a more realistic scenario, where e.g. the users would not manually build their profiles, our proposed approach, PIA-S, which combines personalization and diversification factors, performed better.

Acknowledgments. This work is supported by the Spanish Government (TIN2008-06566-C04-02), and the Government of Madrid (S2009TIC-1542).

4. REFERENCES

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

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

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

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

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