Side-by-side evaluation

An alternative offline evaluation method, previously used for example by [33], consists of presenting users with two alternative rankings side-by-side and ask which they consider best. The advantage of this method is that an actual judgement is made of which ranking is the best one, although users evaluate the entire presented ranking whilst in real life situations they might only look at the first couple of results, potentially biasing the results. Judging two rankings next to each other is considerably faster than judging k documents per query, but it still requires a long offline evaluation exercise. Additionally, an evaluator has to provide a new assessment for each distinct ordering of documents that is investigated. This makes it hard to use such judgments to tune reranking parameters.

Clickthrough-based evaluation

One online evaluation approach involves looking at the query logs and click-through data from a search engine on large scale (e.g. used by [8]). The logs record which search result was clicked for a given query, allowing the evaluator to check if the clicked result would be positioned higher in a personalized ranking. This allows for testing many parameter configurations and also does not require any user effort as their actions are recorded as they go, reflecting their natural environment and browsing behavior. However, the method can have difficulties in assessing whether a search personalization strategy actually works. First, users are more likely to click a search result presented at a high rank, although these are not necessarily most or more relevant [14]. It is also unsuccessful in assessing whether the results that have been brought up from a page lower down would have been relevant as they would rarely have been clicked if originally shown at low rank. On top of that, we also have no access to such large scale usage and user profile data for this experiment.

Alternatively, both personalized and non-personalized rankings can be shown online to users, with metrics such as mean clickthrough rates and positions being computed. However, [21] showed that such an approach is not sensitive enough to detect small differences in relevance with thousands of query impressions as we could obtain in an online experiment.

Interleaved evaluation

The final online evaluation option we consider, which to our knowledge has not been used before for the evaluation of personalized search, is interleaved evaluation [13, 21]. Interleaved evaluation combines the results of two search rankings by alternating between results from the two rankings while omitting duplicates, and the user is presented with this interleaved ranking. The ranking that contributed the most clicks over many queries and users is considered better. Radlinski et al. [21] showed that this approach is much more sensitive to changes in ranking quality than other click-based metrics. It has also shown to correlate highly with offline evaluations with large numbers of queries [20]. On top of that, this method does not require any additional effort from the user, providing an evaluation users naturally use search engines. However, one evaluation only provides an assessment for one particular ranking, and thus an evaluation is required for each parameter configuration being investigated. It is also the hardest evaluation technique for showing improvements as, opposed to other metrics, bringing a slightly relevant page up from rank 8 to rank 5 will not help if the most relevant page is at rank 1.

4.1 Evaluation Plan

This last approach, interleaved evaluation, best reflects real user experience, and would be preferred for evaluating a personalized search system. However, the user profile generation and re-ranking steps both have a large number of possible parameters and it is infeasible to perform an online evaluation for all of them. Hence, we start with an offline NDCG based evaluation to pick the optimal parameter configurations that we then evaluate with the more realistic online interleaved evaluation.

5. OFFLINE EVALUATION

This section first describes how we collected relevance judgments for offline evaluation. Next, we describe how this was used to identify the most promising personalization strategies.

5.1 Relevance Judgements

To obtain personalized relevance judgments, six participants who had installed the AlterEgo plugin recording their browsing behavior were recruited. At that point, two months of browsing history had been recorded and stored for each of them.

The process and format of the evaluation sessions was kept very similar to the relevance judgment session conducted by Teevan et al [31]. Each participant was asked to judge the relevance of the top 50 web pages returned by Google for 12 queries according to the criteria presented in Table 3. The documents were presented in a random order. Full web pages were evaluated instead of snippets because the user is only interested in the actual web being relevant for his information need.

Every participant was asked to provide 600 judgments, 50 relevance judgements for each of 12 queries. The first query participants were asked to judge was their own name (Firstname Lastname) as a warm-up exercise. Next, each participant was presented with a list of 25 general search queries in a random order, consisting of 16 queries taken from the TREC 2009 Web Search track queries and 9 other UK focussed queries such as “pub”, “football” and “cambridge”. All users were asked to judge 6 of these queries. Examples of some of the queries chosen by multiple people can be seen in Table 4. Finally, each participant was presented with their most recent 40 search queries (from their browsing history) and were asked to judge 5 for which they remembered the returned results could have been better.

On average, it took each participant about 2.5 hours to complete this exercise. Particularly interestingly, all users mentioned that during to the exercise they came across really interesting websites which they did not know existed before, indicating that there is a potential for search personalization to enrich the set of returned results.

5.2 Results and Discussion

The following parameters were investigated for profile generation, representing the different steps shown in Figure 1:

- All combinations of the six different input data sources with three possible values for α (namely 0, 1 and normalized)