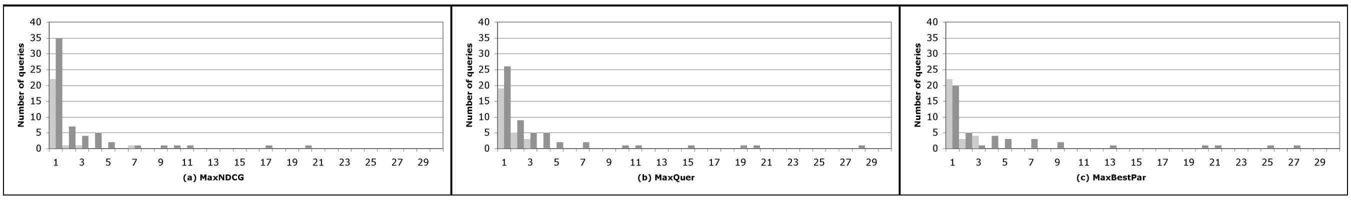

Figure 6: Rank differences for deteriorated (light grey) and improved queries (dark grey) for `MaxNDCG`, `Max-Quer`and `MaxBestPar`

provements. However, `MaxQuer`still improves more queries than the ones it makes worse, indicating that it is still a successful approach. `MaxBestPar`generally makes about the same number of queries worse, however its improvement percentage is 3.5% higher than for `MaxQuer`and it also leaves fewer search queries unchanged. `MaxNDCG`again clearly seems to be the most successful personalization approach, having the highest change and improvement rate and the lowest harming rate, improving 2.7 times more queries than it harms.

Figure 6 shows, for each of the personalization strategies, the distribution of rank changes for all queries that were improved or became worse. It can be seen that for large majority of the deteriorated queries, especially for `MaxNDCG`, the clicked result only loses 1 rank compared to the original ranking. The majority of clicked results that improved a query gain 1 rank as well, however there are quite a few relevant results that are bumped up 2, 3, 4 or even 5 ranks. For each of the personalization strategies, the average rank deterioration is about 1.38 and the average rank improvement is around 3.5, indicating that the gains achieved by personalization are on average more than twice as high as the losses experienced on a query that was originally better. In other words, if personalization was unsuccessful for a query, a user could expect to find the result he was looking for at a rank which is on average only 1 lower than were it originally was. However, when personalization is successful, the result the user is looking for will on average be 3 or 4 ranks higher than original.

Figure 7 shows how much personalization is achieved through the suggested approaches at each rank. It can be seen that most re-ranking is done after rank 10, having little or no influence on user's search experience. Less actual re-ranking is done in the first 5 ranks due to the rank normalization going on, which has already proven to be a worthwhile approach. `MaxQuer` does the least re-ranking in general, whilst `MaxBestPar` does most re-ranking after rank 10.

In summary, it can be seen that the personalization strategy that uses Title, Metadata keywords and Extracted noun phrases as input data using relative weighting, no filtering and TF-IDF as the term weighting scheme, is the most effective according to both the voting method and the number of improved/deteriorated queries, yielding significant improvements over default web search.

7. CONCLUSION & FUTURE WORK

In this paper, we have investigated personalized web search in which we first try to learn a user's long-term interests and then attempt to re-rank the first 50 search results returned by a search engine in a user profile and click history based approach using full browsing history as the input data. We propose a set of personalization techniques that significantly outperform both default Google ranking and the best previous personalization methodologies, which are also compared to each other for the first time in both small scale offline and large scale online experiments. Our methods also seem to be the first profile based approaches, applied to all queries that manage to successfully personalize queries that have a potential for personalization and does not harm queries that are non-personalizable. This is also the first large scale personalized search and online evaluation work for general web search that was not carried out at a search company.

We present a method in which user data and personalization performance results can be collected on large scale in a straightforward way using a browser plug-in based approach.

We discover that the key to using web pages to model a user is to not treat a web page as a normal document, but to treat is as a structured document out of which several types of data can be extracted. We also find that applying advanced NLP techniques like term extraction and parsing can be beneficial for search personalization. The suggested methods can be implemented straightforwardly and are feasible at large scale. One of the outcomes of this paper is a personalized search Firefox add-on that can be publicly downloaded[9] and used without altering the user's browsing experience. The source code is also available for download for the research community[10].

There are a number of directions that can still be investigated. A first option would be to check whether the suggested methods and set of parameters can still be expanded and improved and whether they can benefit from having more data available. Using other field, such as headings in HTML, were not explored. On top of that, a Firefox add-on has access to other behavioral information, such as time spent on a page, amount of scrolling, text selection and mouse activity, that we do not explore. Similarly to the Teevan approach, we could also make use of more personal data like files on their hard drive and e-mails.

In all of the experiments described, the baseline system used the first 50 search results return by the default Google ranker. However, Google also offers personalized search using geo-location and click-through based information. As the details of these algorithms are not publicly known and because it allowed for a more straightforward implementation, we decided not to compare to the personalized version. In future experiments, a more thorough comparison could be done against better baselines that were beyond the scope of our research.

Finally, one could also look into using the extracted profiles for purposes other than personalized search. After the

[9] http://alterego.caret.cam.ac.uk

[10] http://github.com/nicolaasmatthijs/AlterEgo