Table 5: Investigated profiles for interleaved evaluation

Name	Method	Avg NDCG	Improved
`MaxNDCG`	TF-IDF, RTitle, RMKeyw, RCCParse, NoFilt - LM, Look At Rank, Visited	0.573	48/72
`MaxQuer`	TF, RTerms, RCCParse, NoFilt - LM, Look At Rank, Visited	0.567	52/72
`MaxNoRank`	TF, RMKeyw, NoFilt - LM, Look At Rank, Visited	0.520	13/72
`MaxBestPar`	BM25, RTitle, RMKeyw, RTerms, NoFilt - LM, Look At Rank, Visited	0.566	45/72

ranking that outperforms web ranking based on user data has to our knowledge not yet been achieved. While this may be a result of overfitting the parameters given our small offline dataset, we observed this effect often that we do not believe this to be the case.

A different metric that can be used for comparison of personalization methods is to look at the number of queries for which the NDCG score improved, as shown in Figure 3. PClick improves fewest, 13 out of 72 queries, compared to the 44 improved queries for the Teevan approach, despite obtaining a higher NDCG score. This is because the PClick method only works on repeated queries by bringing up pages on which the user previously clicked. While this will only occasionally cause personalization, it makes bigger improvements when it reranks results. Also, the Teevan approach has a negative effect on some of the queries as well. `MaxNDCG`improves performance on 48 queries, `MaxQuer`improves 52.

5.2.2 Parameter Configuration

Given all the investigated user profile and re-ranking combinations, we look at how all of these parameters alter the overall strategy performance. We summarize the one-way effects of all individual parameters and explore their influence on the average NDCG score. In Figure , we consider a certain parameter λ (e.g. filtering, term weighting, etc.). For every other parameter setting, we counted how often the different possible values of λ were most helpful. We repeated this experiment for every parameter group. These approaches do not examine interaction effects, so at the end of this section we give an overview of which parameter combinations achieved the best performance. Note that most parameters only make a small difference in NDCG, even if they are preferred in most cases.

In terms of re-ranking strategies, it is clear that there is one approach that significantly outperforms all other ones, indicating that this approach should be used for all generated profiles. Using the Matching approach for snippet weight calculation is significantly worse ($p < 0.01$) than using Unique Matching. Using a Language Model based on the user profile performs significantly better ($p < 0.01$) than both Matching and Unique Matching and is in 67% of the investigated profiles the best choice. Unique Matching only occasionally performs better than a Language Model when the user profile is very much keyword based and does not use Extracted Nouns, Title, etc. as input data. Except for a single case, multiplying the snippet weight by 10 for URLs that have previously been visited performs significantly better ($p < 0.01$) than not keeping this into account. This proves that a combination of a user profile based and click based personalization strategy can be used successfully. Not normalizing the obtained snippet weights by their original rank always performs worse than when the Google rank is taken into account.

When looking at the input data used to generate a user's profile, it can be seen that all data sources are helpful in improving personalization performance, except when the full web page text is used. Not using the full text performs significantly better ($p < 0.01$) than when this is included. This indicates that treating web pages like a normal document or a bag of words does not work, presumably due to its noisy nature. Using metadata keywords, extracted terms and the page title all yield significant ($p < 0.01$) improvements over not including them. Metadata description and extracted noun phrases also give a significant ($p < 0.05$), but less strong, improvement. The different input data sources can be ranked in terms of helpfulness for personalization: metadata keywords, extracted terms, title, metadata description and extracted noun phrases. However, a combination of the most helpful data sources does not necessarily achieve the best performance, as strong parameter interactions exists between the different sources. It can also be seen that using relative weighting performs consistently better than giving every term of every data source a weight of 1. This can be explained by the fact that the most helpful data sources are the ones that contain the lowest number of terms, becoming more numerous as the data sources become less helpful.

The charts show that in general no term filtering performs significantly better ($p < 0.01$) than using Google N-Gram based filtering, which in its turns performs significantly better ($p < 0.01$) than WordNet noun filtering. WordNet filtering seems to filter out too many actual relevant terms, which is reflected in its lower average NDCG score. Even though no filtering performs significantly better than Google N-Gram filtering, the actual difference in NDCG score is very small (0.001), which can be explained by the fact that N-Gram filtering only filters out non-sensical terms, which are rare in snippets, and not the words that harm in personalization.

When looking at term weighting methods, it seems that Term Frequency performs significantly worse ($p < 0.01$) than TF-IDF and BM25. BM25 performs on average significantly better ($p < 0.05$) than TF-IDF. However, when looking at the number of times a given parameter value is the best option for a parameter configuration, we can see that TF-IDF and BM25 are the best option for roughly the same number of approaches. BM25 seems to work better in general when the input data is richer and thus noisier, as it normalizes the term weights. TF-IDF is in general more successful when the selected set of input sources are more keyword focussed.

5.2.3 Relevance Distribution

Of the 3,600 relevance judgements collected in the evaluation session, 9% were Very Relevant, 32% Relevant and 58% as Non-Relevant. The relevance judgements distribution for the default Google rank and the `MaxNDCG` re-ranking approach shown in Figure 5 indicates that web rank already manages to place the biggest part of the Very Relevant results in the top 5 results. However, the relevance distribu-