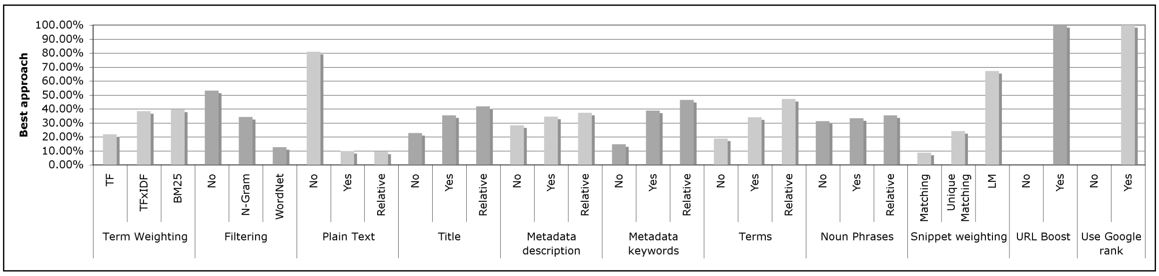

Figure 4: Percentage-wise number of times an investigated parameter performs best for a given parameter configuration

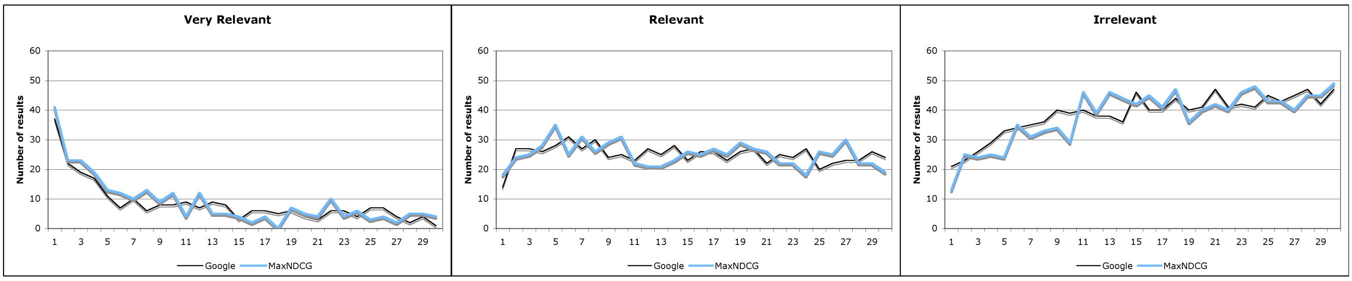

Figure 5: Distribution of relevance at rank for the Google and `MaxNDCG` rankings

tion for the `MaxNDCG` strategy shows that personalized search manages to add more Very Relevant results into the top 5 of the ranking, but it mainly succeeds in adding Very Relevant results in between rank 5 and 10. This is expected as the personalization strategy keeps the Google rank into account and is thus less aggressive at very high ranks.

The Relevant results are evenly spread over all ranks, with slightly more weight towards the first half of the ranking. The personalization strategy manages to add a few more Relevant results to the top 5 of the different rankings. Irrelevant results are present at all ranks, but become more dominant after rank 10. Personalized search manages however to remove quite a large number of Irrelevant results out of the top 10 results and pushes them down into the lower regions of the rankings.

5.2.4 Selected Profiles

Given the discussed results, we selected 3 promising profiles to carry on with in the interleaved evaluation experiment. The chosen profiles are `MaxNDCG`, as it yielded the highest normalized NDCG, `MaxQuer`, as it improved on most queries and `MaxBestPar` as it uses the best parameters according to our greedy learning approach. All of these selected profiles use Language Model snippet weighting, take into account the Google rank, and multiply the weight of previously visited URLs by 10 and the number of visits.

6. ONLINE EVALUATION

The second stage in the evaluation process is a large scale online interleaved evaluation. It is very important that this evaluation is run over a long enough period of time to get a sufficient number of queries per user, and that a reasonable amount of users participate in order to get quantitative and reliable results. It is crucial that participants are able to do the evaluation during their day-to-day lives as realistic information needs come up. This attempts to assess whether any of the 3 personalization strategies selected in section 5 yield an actual improvement in a user's search experience. It is also critical that in doing this evaluation exercise, users' search experience is not altered and they no noticeable differences can be detected in the user interface.

Algorithm 1 Team-Draft Interleaving

```
Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset$;
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ do
    if $(|TeamA| < |TeamB|) \vee$
        $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ then
        $k \leftarrow \min_i \{i : A[i] \notin I\}$ ... top result in A not yet in I
        $I \leftarrow I + A[k]$; ........................ append it to I
        $TeamA \leftarrow TeamA \cup \{A[k]\}$ ..... clicks credited to A
    else
        $k \leftarrow \min_i \{i : B[i] \notin I\}$ ... top result in B not yet in I
        $I \leftarrow I + B[k]$ ........................ append it to I
        $TeamB \leftarrow TeamB \cup \{B[k]\}$ .... clicks credited to B
    end if
end while
Output: Interleaved ranking $I$, $TeamA$, $TeamB$
```

In this section, we first describe the details of the large scale online interleaved evaluation. Next, we discuss the results obtained and find out what the best personalization strategy is and whether it improves over Google. ADD: Given our limited offline relevance data, if we didn't do this, we might overfit.

6.1 Interleaving Implementation

Therefore, an updated version of the AlterEgo Firefox add-on was developed and all of the volunteers who downloaded and installed the initial version were asked to upgrade to this second version. The add-on detects when a