Table 6: Search Queries and Team votes per personalization strategy

Method	Queries	Google Vote	Re-ranked Vote
MaxNDCG	590	220 (37.3%)	370 (62.7%)
MaxQuer	621	295 (47.5%)	326 (52.5%)
MaxBestPar	636	290 (45.7%)	346 (54.3%)

Table 7: Rank quality implications for each Personalization Strategy

Method	Unchanged	Improved	Deteriorated
MaxNDCG	429 (67.5%)	152 (23.9%)	55 (8.7%)
MaxQuer	444 (71.6%)	112 (18.1%)	64 (10.3%)
MaxBestPar	403 (68.3%)	128 (21.6%)	59 (10.1%)

user submits a web search via Google and sends the search query, the unique user identifier and the current page number over to the server. Next, the server requests the first 50 search results from Google for the given query and picks one of the three personalization strategies at random. The selected strategy is used to calculate the snippets weights and re-ranking is done accordingly.

In our experiments, we use the Team-Draft interleaving algorithm, as described in Algorithm 1, to interleave the original Google ranking and the re-ranked personalized version. This algorithm is motivated by how sports teams are often assigned in friendly games. Given a pool of available players (ranking A and B), two captains (one for TeamA and one for TeamB) take turns picking their next favorite player in the set of remaining players, subject to a coin toss every turn deciding which team captain gets to pick first. More details about this approach can be found in [21]. The combined ranking is presented to the user and clicks on any of the results are recorded. If results from the personalized ranking end up being clicked more often, it is a strong indicator that the personalization is successful.

The Team-Draft interleaving algorithm, as described in Algorithm 1, is then utilized to interleave the original Google ranking with the new personalized re-ranked version. In doing this, 50% of the search results are assigned to Team A, which represents the default Google ranking, and 50% is assigned to Team B, which represents the personalized ranking. A click that is recorded for either of these teams is considered a vote for the ranking they represent. The interleaved results are sent back to the user's browser, where they replace the original list of results, but keeping the surrounding page elements and thus not altering the search environment.

To avoid presenting slightly different rankings every time a search page is refreshed, both the random personalization strategy selector and the random bit inside the interleaving algorithm were seeded with a combination of unique identifier, query and the current hour.

An additional user experience consideration that had to be taken into account was re-ranking performance. A user profile had an average size of about 3.4 Megabyte, and all of them were kept in memory for fast access. Also, the user profiles were automatically updated on an hourly basis.

6.2 Results and Discussion

In this section, we present the results from the interleaved evaluation experiment. We expect it to be harder to show

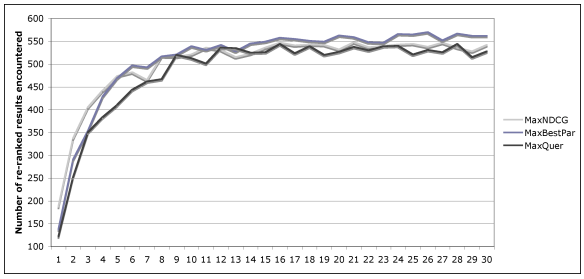

Figure 7: Degree of personalization per rank

actual improvements for this evaluation since bringing a relevant result up to a rank beneath the most relevant result will likely not make a difference in a user's behavior. This experiment should however give a clearer idea of whether our suggested personalization methods yield an improvement in real life. The previous evaluation phase also had a strong emphasis on ambiguous queries, whilst the share of ambiguous queries should be lower in this evaluation, making it harder to successfully deploy search personalization.

This interleaved evaluation experiment was run over three weeks, for the 41 users who updated the plugin. A total of 2,715 queries were submitted, of which 1,847 received an actual click on a search result. 76% of these search queries were unique and 24% were repeated queries.

Each click on a search result represented a vote for either the original Google ranking or the re-ranked personalized ranking and the user profile that was used to do the actual re-ranking. The total number of votes for each and each strategy can be seen in Table 6, showing that all of the investigated personalization approaches yield an actual improvement over the default Google ranking. `MaxNDCG`significantly outperforms ($p < 0.01$) web ranking by more than both `MaxQuer`and `MaxBestPar`. `MaxQuer`and `MaxBestPar`both outperform web ranking as well, although their improvements are not statistically significant. The collected votes suggest that `MaxNDCG`is the most convincing and best performing personalization strategy, which is in line with the findings from the first evaluation experiment. Although `MaxBestPar`performs better than `MaxQuer`in the current experiment, which is in contrast with the NDCG scores obtained in the relevance judgements session, the difference between these algorithms is not statitically significant.

The effect personalization has on the set of search queries for the different strategies can be seen in Table 7. The Unchanged column indicates the number of queries for which personalization did not make a difference, i.e. the clicked result was at the same rank for both the default ranking and the personalized ranking. The *Improved* column shows the number of occasions in which the clicked result was brought up due to search personalization, whilst the *Deteriorated* column shows the amount of queries for which the clicked result was lower in the ranking due to personalization. The trends that can be seen from the voting results are consistent with these numbers, however the actual differences between the different approaches seem smaller. On average, about 70% of the search queries are untouched by search personalization, 20% of the queries are improved and 10% becomes worse. `MaxQuer`seems like the least effective approach, having the highest percentage of unchanged queries, the highest percentage of deterioration and the lowest percentage of im-