**Table 3: Relevance judgements guidelines**

| |
|---|
| (a) Select *Irrelevant* if the document is not useful and not interesting to you. |
| (b) Select *Relevant* if the document is interesting to you, but is not directly about what you were hoping to find or if the document is somewhat useful to you, meaning that it touches on what you were hoping to find (maximum 1 paragraph), but not very extensively. |
| (c) Select *Very Relevant* if the document is useful or very interesting to you, i.e. it is what you were hoping to find. |

**Table 4: Common queries used by different participants**

| Query | Users |
|---|---|
| Cambridge | 6 |
| GPS | 4 |
| Website design hosting | 3 |
| Volvo | 3 |

- Three filtering methods: No Filtering, WordNet noun filtering and Google N-Gram filtering using a threshold of 1,000
- The three term weighting methods: TF, TF-IDF and BM25

In terms of re-ranking the search results, the following set of parameters were investigated:

- The four snippet weighting methods: Matching, Unique Matching, Language Model and PClick
- Whether or not to consider the original Google rank
- Whether or not to give extra weight to previously visited URLs (using $v = 10$)

For every profile and ranker combination, the mean personalized NDCG was measured as follows:

$$NDCG@p = \frac{1}{Z}\sum_{i=1}^{p}\frac{2^{rel_i}-1}{\log_2(1+i)} \quad (11)$$

where $rel_i$ is determined from the relevance judgements (non-relevant = 0, relevant = 1 and very relevant = 2) and $Z$ is such that the maximum NDCG for each query is 1. In all of these following results, we compare NDCG scores that have been averaged across all queries and all users.

### 5.2.1 Baseline Comparisons

To assess how well our different personalization strategies performed, we compared the outcome of the experiment with several baselines as shown in Figures 2 and 3. The reported scores are normalized DCG scores for all 72 queries, and all statistical analyses were performed using a paired T-Test.

The baseline methods we compare to are the default Google ranking, the Teevan method and the PClick method. The results obtained appear to agree with those reported in [31] and [8]. We find the Teevan approach is significantly ($p < 0.05$) better than Google and PClick is significantly ($p < 0.01$) better than Google. Note that in our implementation of the Teevan algorithm we only make use of the browsing history as input data, as we do not have access to the user's files or e-mails, which may disadvantage it.

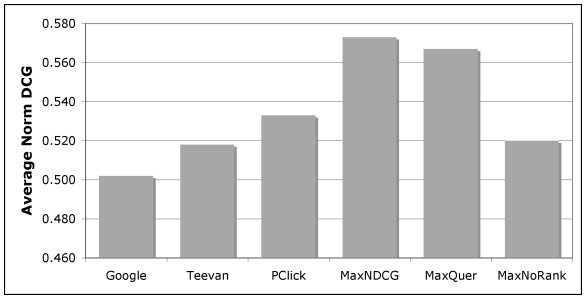


**Figure 2: Average Normalized DCG scores comparing baseline system performance to best personalization strategies**

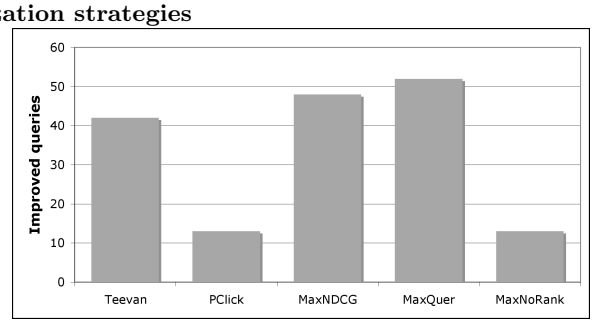


**Figure 3: Comparing baseline system performance to best personalization strategies in terms of number of improved queries**

We selected four user profile and re-ranking parameter settings, summarized in Table **??**, to further evaluate. The first, `MaxNDCG`, yielded the highest average NDCG score on the offline dataset. It involves using Title, Metadata Keywords and Extracted noun phrases as input data using relative weighting, no filtering, and TF-IDF as the weighting method. The second, `MaxQuer`had the highest number of improved queries. It consisted of using Extracted Terms and Extracted noun phrases as input data using relative weighting, no filtering and TF as the weighting method. The third, `MaxNoRank`, was the best method that does not take the original Google ranking into account. It uses Metadata keywords as input data, no filtering and TF as the weighting method. Finally, `MaxBestPar`was obtained by a greedy selection strategy in each parameter was selected one at a time and then locked. It is very similar to `MaxNDCG`, consisting of using Title, Metadata Keywords and Extracted Terms as input data using relative weighting, no filtering and BM25 for term weight calculation. All of these selected profiles used Language Model snippet weighting, the Google rank was taken into account and previously visited URLs received addition weight per Equation 9 with $v = 10$.

`MaxNDCG`and `MaxQuer`are both significantly ($p < 0.01$) better than default Google, Teevan and PClick. `MaxNDCG`, with an average NDCG of 0.573, yields a 14.1% improvement over Google, and `MaxQuer`, with an average NDCG of 0.567, yields a 12.9% improvement over Google.

Interestingly, despite `MaxNoRank`ignoring the Google rank, it obtaining an NDCG score of 0.520 that is significantly ($p < 0.05$) better than Google and better than Teevan. A