Baseline Method

- Linear combination of 77 features
- 4 fields x 17 textual features, e.g. TF, LM, BM25, ...
- 9 numeric features, e.g. # of answers/PVs, date, ...
- Weights are optimized by Coordinate Ascent (CA).

Extended BM25F as Ranking Features

- The BM25F is a non-linear function so that adding it as features may improve the model. (cf. neural net.)
- Moreover we use numeric fields as well as TFs.

\[
\begin{align*}
    k_1: & \text{Non-linear conversion} \\
    \alpha(D) &= \sum_{f \in D_N} v(f, D) \cdot \log \frac{N - df(t) + 0.5}{df(t) + 0.5} \\
    w(t, D) &= \sum_{f \in D} \frac{t(f, t, D) \cdot \log \frac{N - df(t) + 0.5}{df(t) + 0.5}}{\sum_{f \in D} [1 - \beta_t] + \beta_t \cdot \frac{df(f, D)}{\text{avgLen}(f)}}
\end{align*}
\]

- We tried 3 settings of BM25F. (Naive: All fields; SERP: Fields on SERPs, SERP+: Fields prominent on SERPs)
- Adding the 3 settings as features more or less improved the offline score. (On nDCG@10, ~+10%)

Cross Validation

- Five-fold cross validation improved the offline score. (.380 → .412 on nDCG@10, +8.4%)

nDCG@10 as Objective Function

- Initially we used MAP as the objective function of CA.
- Because quality of lower-ranked documents may be important in the greedy optimization process.
- Finally directly using nDCG@10 improved its score. (.396 → .419 on nDCG@10, +5.7%)

Evaluations

Feature Importance

- Setting 0.0 to the weight of each feature, we re-calculated nDCG@10 scores of the resulting rankings.
- The lower the score is, the more important the feature is.

Offline Test Results

- *: Statistically significant (p < 0.05) difference from YJRS based on Student’s paired t-test.
- Our run achieved the 2nd-best nDCG@10, 3rd-best ERR@10, and best Q-measure scores.
- The differences on Q were statistically significant.

Online Test Results

- *: Statistically significant (p < 0.05) difference from YJRS based on the t-test/Pearson’s chi-square test
- Our method achieved the 2nd-largest total credit.
- Difference from the 1st was not stat. significant whereas one from the 3rd was.
- Our run consistently achieved the win-loss ratios better than 0.5 against all the other runs.
- In stat. significantly larger number of PVs, our run won.

Conclusions

- Our method performed well due to its robustness.
- The BM25F is useful as learning-to-rank features.
- Well-known classical techniques, namely Coordinate Ascent and cross validation, are still useful.