Evaluations

- In this task, all runs were evaluated online with the Pairwise Preference Multileaving (PPM) method.
- When a user submits a query, PPM mixes multiple rankings for the query into a list and show it to the user. Clicks on some questions indicate the user’s preference in questions. The PPM gives credits to rankings which accord with the preference.
- The entire evaluation process spanned the first round (61 runs, 164k PVs) and the second round (30 runs, 149k PVs).
  - Note: We list only our runs (with *) and other team’s best runs due to space limitation.

First Round of Multileaved Comparisons

- CA + BM25F (ID: 92), CA + Translated (ID: 93) > ListNet (IDs: 113, 100)
  - The performance of our linear combination runs were not stable.
  - Performance of our CA + Basic runs were quite different although they share the same feature set.
  - Performance of our CA + All run was quite worse than ones of our CA + BM25F and our CA + Translated runs which use only subsets of features.
  - Our LightGBM runs did not work. This may be because of small training data.

Second Round of Multileaved Comparisons

- ListNet (IDs: 113, 100), CA + BM25F (ID: 92) > CA + Translated (ID: 93) > CA + Basic (ID: 95)
  - The ListNet runs occupied better positions than in the first round.
  - The other tendencies were same as the ones in the first round.

Conclusions

- Among our runs, some combinations of a relatively simple learning-to-rank method and a reasonable feature set performed well.
- Namely, our ListNet (+ Basic) and CA + BM25F methods performed well.
- Our linear combination runs resulted in unstable performance whereas ListNet runs, which used only basic features, were quite promising.

References

- [5] https://github.com/microsoft/LightGBM/ntcir14, June 10-13, 2019, NII, Tokyo, Japan