THUIR at the NTCIR-17 FairWeb-1 Task: An Initial Exploration of the Relationship Between Relevance and Fairness

Yiteng Tu[2], Haitao Li[1], Zhumin Chu[1], Qingyao Ai[1], Yiqun Liu[1]

1 Department of Computer Science and Technology, Tsinghua University, Zhongguancun Laboratory, Beijing 100084, China

2 Renmin University of China, Beijing 100872, China

❖ Introduction

➢ We participated in the NTCIR-17 Fairweb-1 Task.

- ➢ We utilize several different methods in all 5 submitted runs including reranking, learning-to-rank, and search result diversification algorithms to deal with the group fairness problem in web search.
- ➢ The official results indicate that our approaches achieve the best results on all relevance and fairness metrics.

❖ Our Methods

➢ Run 1: Sparse Retrieval

- ➢ We choose two classic sparse retrieval algorithms, BM25 and QLD.
- ➢ We respectively conduct the retrieval using the two algorithms for *Q-queries*, *D-queries*, and *QD-queries*.
- ➢ We add RM3 pseudo-relevance feedback to both algorithms and repeat the retrieval process.
- ➢ We choose RRF method to integrate ranking lists generated by different types of queries and different retrieval algorithms.

➢ Run 2: LightGBM

- ➢ We employ MonoBERT and MonoT5 models to rerank all retrieved documents of the three types of queries in Run 1.
- ➢ For each query, we use the 12 sparse retrieval scores from Run 1 and the 6 neural reranker scores as the features to conduct learning-to-rank through a lightweight learning-to-rank model, LightGBM.

➢ Run 3: Query Augmentation

- ➢ We incorporate fairness information into the semantics of queries by simply adding the entity attribute information to the query text.
- ➢ For example, for movie topics that need to consider regional fairness we simply add a suffix *", and these movies are from Africa/America/Antarctica/..."* to the query.
- ➢ We generate a ranking list via the MonoT5 reranker for each value of an attribute. We combine ranking lists of different values of the same attribute by random sampling and then utilize RRF to merge the results of different attributes.

➢ Run 4 & Run 5: PM2 & xQuAD

- ➢ We attempt two different ways of estimating attribute scores of each candidate document.
  - ➢ One is to extract possible entities and obtain attribute information about them through web crawlers. Scores are calculated from the ratio of the attribute values.
  - ➢ The other is simply approximating the document's attribute distribution through the relative proportions of related term appeared in the document.
- ➢ Then we try two search result diversification algorithms, PM2 and xQuAD, to balance both relevance and fairness factors.

❖ Results and Analysis

Table 1: The official relevance evaluation of our runs.

	R topics		M topics		Y topics		All topics	
Run	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU
THUIR-QD-RG-1	0.1918	0.6013	0.1608	0.4400	0.1099	0.4026	0.1542	0.4813
THUIR-QD-RG-2	**0.2638**	**0.6560**	0.2280	0.6923	0.1144	0.3749	0.2021	**0.5744**
THUIR-QD-RR-3	0.2276	0.5804	**0.2653**	**0.7230**	0.1293	0.3919	0.2074	0.5651
THUIR-QD-RR-4	0.2460	0.5957	0.2518	0.6859	**0.1438**	**0.4404**	**0.2139**	0.5740
THUIR-D-RR-5	0.1421	0.5351	0.1223	0.5316	0.1009	0.3649	0.1218	0.4772

Table 2: The official fairness evaluation over the R topics of our runs.

Run	Mean GF^{JSD} (GENDER)	Mean GF^{NMD} (HINDEX)	Mean GF^{RNOD} (HINDEX)	Mean GFR
THUIR-QD-RG-1	0.5823	0.5569	0.5257	0.5698
THUIR-QD-RG-2	**0.5831**	**0.5841**	**0.5352**	**0.5914**
THUIR-QD-RR-3	0.4987	0.5247	0.4875	0.5222
THUIR-QD-RR-4	0.5086	0.5164	0.4720	0.5254
THUIR-D-RR-5	0.5351	0.5080	0.4841	0.5181

Table 3: The official fairness evaluation over the M topics of our runs.

Run	Mean GF^{JSD} (ORIGIN)	Mean GF^{NMD} (RATINGS)	Mean GF^{RNOD} (RATINGS)	Mean GFR
THUIR-QD-RG-1	0.3395	0.4025	0.3684	0.3827
THUIR-QD-RG-2	**0.5684**	0.6330	**0.5788**	**0.6132**
THUIR-QD-RR-3	0.5391	**0.6433**	0.5683	0.6101
THUIR-QD-RR-4	0.5332	0.6118	0.5435	0.5875
THUIR-D-RR-5	0.4900	0.5307	0.4983	0.5066

Table 4: The official fairness evaluation over the Y topics of our runs.

Run	Mean GF^{NMD} (SUBSCS)	Mean GF^{RNOD} (SUBSCS)	Mean GFR
THUIR-QD-RG-1	0.3830	0.3638	0.3832
THUIR-QD-RG-2	0.3423	0.3141	0.3445
THUIR-QD-RR-3	0.3601	0.3297	0.3608
THUIR-QD-RR-4	**0.4112**	**0.3809**	**0.4107**
THUIR-D-RR-5	0.3550	0.3396	0.3523

➢ Our methods outperform others on all relevance metrics and fairness metrics.

- ➢ Run 2, 3, and 4 are significantly better than Run 1 and Run 5 thanks to the powerful fine-grained reranker even in the zero-shot scenario.

➢ If a method performs well in terms of relevance, it also has strong performance in terms of fairness.

- ➢ Relevance and fairness can be jointly optimized within a certain degree.
  - ➢ Search results with higher relevance contain more relevant entities.
  - ➢ These large amounts of randomly distributed related entities can facilitate further optimization towards fairness.

❖ Conclusions

- ➢ **We participate in the NTCIR-17 FairWeb-1 task and submit 5 runs using various methods.**
- ➢ **We achieve first place in all metrics.**
- ➢ **Our results indicate that relevance and fairness are not in opposition to some degree and it is possible to achieve their joint optimization.**

Email: yitengtu16@gmail.com