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
yitengtu16@gmail.com

Introduction

- How to ensure exposure fairness without hurting the quality of search result relevance has become an important challenge for search engines.

- We participate in the NTCIR-17 FairWeb-1 task and explore the impact of various methods on both relevance and fairness.

Our Methods

Table 1: An overview of THUIR's submissions on the NTCIR-17 FairWeb-1 Task

Run Number	Run Name	Description
Run 1	THUIR-QD-RG-1	Directly aggregate the retrieved results of sparse retrieval by RRF
Run 2	THUIR-QD-RG-2	Learning-to-rank based on sparse and dense relevance features
Run 3	THUIR-QD-RR-3	Add feature information to the query text for reranking
Run 4	THUIR-QD-RR-4	A search result diversification algorithm, PM2
Run 5	THUIR-D-RR-5	A search result diversification algorithm, xQuAD

Our Methods

Run 1: Sparse Retrieval

Data Process

- Document
  - Extract the main text from HTML text with bs4 package
- Query
  - 3 topics: researchers (R topics), movies (M topics), Youtube contents (Y topics),
  - 2 sections: *Query* and *Description*
  - Construct 3 types of queries: *Q-queries*, *D-queries* and *QD-queries* based on the two sections

Document Retrieval

- Sparse Retrieval
  - BM25 & QLD
  - w/ & w/o RM3 pseudo relevance feedback
  - *Q/D/QD-queries*
  - A total of 12 ranking lists
- Fusion
  - Reciprocal Rank Fusion (RRF)
  - Only integrate 8 results of *D/QD-queries*

Our Methods

Run 2: LightGBM

Reranking

- Models
    - MonoBERT (*castorini/ monobert-large-msmarco*)
    - MonoT5 (*castorini/monot5-3b-msmarco-10k*)
- Rerank all retrieved documents of the three types of queries (*Q/D/QD*) in Run 1
- 6 features

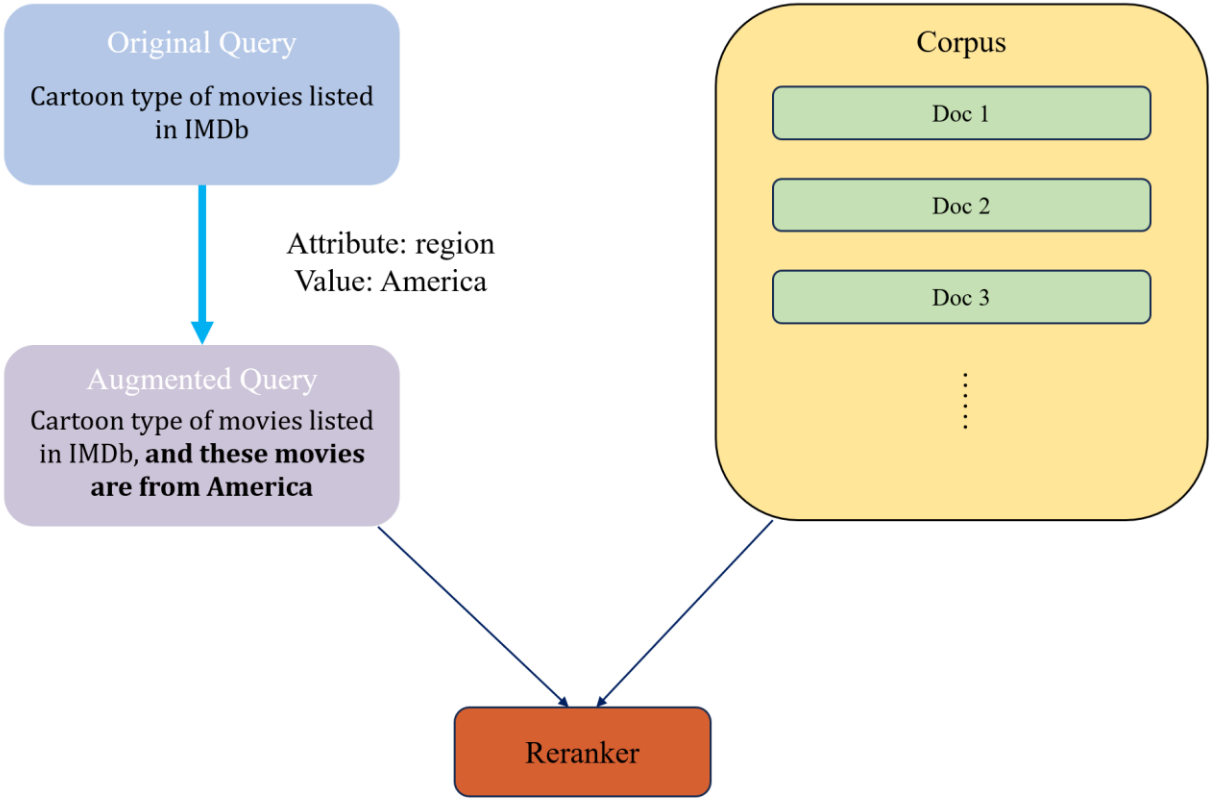
Learning-to-rank

- Features Selection
    - 12 sparse retrieval scores
    - 6 neural reranker scores
- Model: LightGBM
    - lightweight, efficient, easy to use
- Training Data
    - NTCIR WWW2-3

Our Methods

Run 3: Query Augmentation

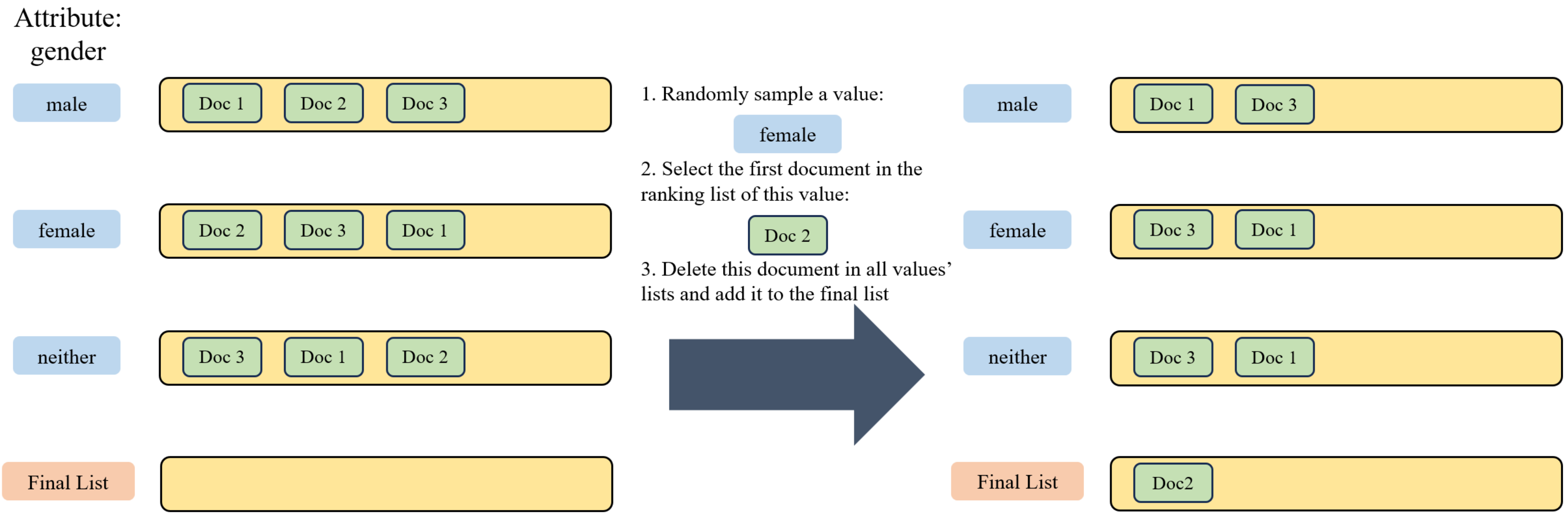
Query Augmentation and Ranking:



Our Methods

Run 3: Query Augmentation

Fusion:



Our Methods

Run 4 & 5: PM2 & xQuAD

Estimate Attribute Scores

- Based on the proportion of the relevant entities appear in the document
- M topics and Y topics
  - Named Entity Extraction: person, organization and location names
  - Web Crawler
- R topics
  - Gender: proportion of gender-related terms (such as he/she/his/her/him……)
  - H-index: There's nothing we can do about it ……

Search Result Diversification

- The target fairness distribution of an attribute can be regarded as the distribution of subtopic importance in search result diversification algorithms
- Algorithms
  - PM2 (relevance score is based on neural reranker scores)
  - xQuAD (relevance score is based on provided baseline retrieval scores)

Experimental Results

- Relevance Metrics

Table 2: The official relevance evaluation over different query topics, where R, M, and Y respectively represent queries about researchers topics, movies topics, and YouTube contents topcis. We present 5 runs by THUIR, 6 baseline runs, as well as the optimal results of other participants. ">" means stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials and $\alpha = 0.05$ [17]) among all 28 runs. For example, in terms of ERR on all topics, THUIR-QD-RR-4 statistically significantly outperforms the runs ranked at 18 through 28. All the results are from [22].

	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(>27-28)	0.4813(>26-28)
THUIR-QD-RG-2	**0.2638**	**0.6560**	0.2280	0.6923	0.1144	0.3749	0.2021(>22-28)	**0.5744(>22-28)**
THUIR-QD-RR-3	0.2276	0.5804	**0.2653**	**0.7230**	0.1293	0.3919	0.2074(>20-28)	0.5651(>23-28)
THUIR-QD-RR-4	0.2460	0.5957	0.2518	0.6859	**0.1438**	**0.4404**	**0.2139(>18-28)**	0.5740(>22-28)
THUIR-D-RR-5	0.1421	0.5351	0.1223	0.5316	0.1009	0.3649	0.1218(>27-28)	0.4772 (>26-28)
Best of Other Participants	0.2131	0.5582	0.2434	0.5819	0.1365	0.3755	0.1847(>26-28)	0.4977(>26-28)
run.bm25-depThre3-Q	0.1989	0.5489	0.1712	0.5035	0.0471	0.2202	0.1390(>27-28)	0.4242(>26-28)
run.bm25-depThre3-D	0.1509	0.4801	0.1564	0.4337	0.0266	0.1735	0.1113(>27-28)	0.3624(>27-28)
run.qld-depThre3-Q	0.1567	0.5518	0.1653	0.4958	0.0459	0.2514	0.1226(>27-28)	0.4330(>26-28)
run.qld-depThre3-D	0.1749	0.5695	0.1187	0.3728	0.0442	0.2194	0.1126(>27-28)	0.3872(>27-28)
run.qljm-depThre3-Q	0.2104	0.4971	0.2114	0.6026	0.0266	0.2010	0.1495(>27-28)	0.4336(>26-28)
run.qljm-depThre3-D	0.1459	0.4361	0.1478	0.4883	0.0520	0.2424	0.1152(>27-28)	0.3889(>27-28)

Experimental Results

- Fairness Metrics over R topics

Table 3: The official fairness evaluation over the R topics. We present 5 runs by THUIR, 6 baseline runs, as well as the optimal results of other participants. ">" means stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials and $\alpha = 0.05$) among all 28 runs.

Run	Mean GF^{JSD} (GENDER)	Mean GF^{NMD} (HINDEX)	Mean GF^{RNOD} (HINDEX)	Mean GFR
THUIR-QD-RG-1	0.5823(>26-28)	0.5569(>26-28)	0.5257(>26-28)	0.5698(>26-28)
THUIR-QD-RG-2	**0.5831(>26-28)**	**0.5841(>26-28)**	**0.5352(>26-28)**	**0.5914(>26-28)**
THUIR-QD-RR-3	0.4987(>26-28)	0.5247(>26-28)	0.4875(>26-28)	0.5222(>26-28)
THUIR-QD-RR-4	0.5086(>26-28)	0.5164(>26-28)	0.4720(>26-28)	0.5254(>26-28)
THUIR-D-RR-5	0.5351(>26-28)	0.5080(>26-28)	0.4841(>26-28)	0.5181(>26-28)
Best of Other Participants	0.5374(>26-28)	0.5195(>26-28)	0.4866(>26-28)	0.5274(>26-28)
run.bm25-depThre3-Q	0.5096(>26-28)	0.4977(>26-28)	0.4605(>26-28)	0.5064(>26-28)
run.bm25-depThre3-D	0.4694(>26-28)	0.4400(>26-28)	0.4155(>26-28)	0.4550(>26-28)
run.qld-depThre3-Q	0.5356(>26-28)	0.5152(>26-28)	0.4807(>26-28)	0.5227(>26-28)
run.qld-depThre3-D	0.5497(>26-28)	0.5306(>26-28)	0.4975(>26-28)	0.5389(>26-28)
run.qljm-depThre3-Q	0.4315(>26-28)	0.4362(>26-28)	0.3999(>26-28)	0.4428(>26-28)
run.qljm-depThre3-D	0.4120(>26-28)	0.4038(>26-28)	0.3824(>26-28)	0.4101(>26-28)

Experimental Results

- Fairness Metrics over M topics

Table 4: The official fairness evaluation over the M topics. We present 5 runs by THUIR, 6 baseline runs, as well as the optimal results of other participants. ">" means stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials and $\alpha = 0.05$) among all 28 runs.

Run	Mean GF^{JSD} (ORIGIN)	Mean GF^{NMD} (RATINGS)	Mean GF^{RNOD} (RATINGS)	Mean GFR
THUIR-QD-RG-1	0.3395(>27-28)	0.4025(>27-28)	0.3684(>27-28)	0.3827(>27-28)
THUIR-QD-RG-2	**0.5684(>27-28)**	0.6330(>27-28)	**0.5788(>27-28)**	**0.6132(>27-28)**
THUIR-QD-RR-3	0.5391(>27-28)	**0.6433(>26-28)**	0.5683(>27-28)	0.6101(>27-28)
THUIR-QD-RR-4	0.5332(>27-28)	0.6118(>27-28)	0.5435(>27-28)	0.5875(>27-28)
THUIR-D-RR-5	0.4900(>27-28)	0.5307(>27-28)	0.4983(>27-28)	0.5066(>27-28)
Best of Other Participants	0.4768(>27-28)	0.5169(>27-28)	0.4758(>27-28)	0.4996(>27-28)
run.bm25-depThre3-Q	0.4135(>27-28)	0.4623(>27-28)	0.4283(>27-28)	0.4484(>27-28)
run.bm25-depThre3-D	0.3401(>27-28)	0.3993(>27-28)	0.3630(>27-28)	0.3789(>27-28)
run.qld-depThre3-Q	0.4275(>27-28)	0.4668(>27-28)	0.4351(>27-28)	0.4528(>27-28)
run.qld-depThre3-D	0.3122(>27-28)	0.3507(>27-28)	0.3208(>27-28)	0.3353(>27-28)
run.qljm-depThre3-Q	0.4716(>27-28)	0.5462(>27-28)	0.4871(>27-28)	0.5205(>27-28)
run.qljm-depThre3-D	0.4273(>27-28)	0.4606(>27-28)	0.4211(>27-28)	0.4456(>27-28)

Experimental Results

- Fairness Metrics over Y topics

Table 5: The official fairness evaluation over the Y topics. We present 5 runs by THUIR, 6 baseline runs, as well as the optimal results of other participants. ">" means stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials and $\alpha = 0.05$) among all 28 runs.

Run	Mean GF^{NMD} (SUBSCS)	Mean GF^{RNOD} (SUBSCS)	Mean GFR
THUIR-QD-RG-1	0.3830(>27-28)	0.3638(>27-28)	0.3832(>27-28)
THUIR-QD-RG-2	0.3423(>27-28)	0.3141(>27-28)	0.3445(>27-28)
THUIR-QD-RR-3	0.3601(>27-28)	0.3297(>27-28)	0.3608(>27-28)
THUIR-QD-RR-4	**0.4112(>27-28)**	**0.3809(>27-28)**	**0.4107(>27-28)**
THUIR-D-RR-5	0.3550(>27-28)	0.3396(>27-28)	0.3523(>27-28)
Best of Other Participants	0.3315(>27-28)	0.3157(>27-28)	0.3428(>27-28)
run.bm25-depThre3-Q	0.2112	0.2039	0.2121
run.bm25-depThre3-D	0.1777	0.1731	0.1733
run.qld-depThre3-Q	0.2451	0.2391	0.2453
run.qld-depThre3-D	0.2155	0.2100	0.2147
run.qljm-depThre3-Q	0.2071	0.2038	0.2024
run.qljm-depThre3-D	0.2425	0.2329	0.2377

Discussion

- Good relevance ranking method also performs well in fairness
  - Search results with higher relevance contain **more relevant entities**
  - The attributes of related entities should exhibit **randomness** without extra factors like popularity and personalization in real search engines
  - **Therefore, we can achieve a win-win situation for both relevance and fairness to some extent**

Conclusion

- We participate in the NTCIR-17 FairWeb-1 task and submit 5 runs with 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.

Thanks!

Q&A